THÈSE DE DOCTORAT DE

L'UNIVERSITÉ PARIS VI – PIERRE ET MARIE CURIE

Discipline : Systématique

École Doctorale Diversité du Vivant

Présentée par M. Blaise LI pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ PARIS VI – PIERRE ET MARIE CURIE

Préparée sous la direction de M. Guillaume LECOINTRE

UMR 7138 CNRS – Équipe Phylogénie

Département Systématique et Évolution, Museum National d'Histoire Naturelle

Sujet de la thèse :

Fiabilité des clades et congruence taxinomique

Application à la phylogénie des téléostéens acanthomorphes

soutenue le 17/09/2008

devant le jury composé de (dans l'ordre alphabétique) :

M. Pierre DARLU	Rapporteur
M. Jean-Yves DUBUISSON	Examinateur
M. François-Joseph LAPOINTE	Rapporteur
M. Guillaume LECOINTRE	Directeur de thèse
Mme Olga OTERO	Examinatrice

Résumé

Si le but de la reconstruction phylogénétique est d'avoir une idée des relations de parenté réelles entre les êtres vivants, il est bon de ne pas se contenter d'un simple arbre obtenu par l'analyse combinée d'un ensemble de données. En effet, même des clades robustes apparaissant dans un tel arbre peuvent ne pas être fiables. La confiance dans une affirmation phylogénétique ne peut émerger qu'après une comparaison de résultats obtenus par des données indépendantes.
Dans un premier temps, la présente thèse propose de mesurer la fiabilité d'un clade à partir d'un indice de répétition prenant en compte le nombre d'occurrences obtenues pour ce clade sur un ensemble d'analyses de données indépendantes, c'est-à-dire peu susceptibles de donner lieu aux mêmes biais de reconstruction. Plus un clade est obtenu un nombre élevé de fois de cette façon, plus il peut être considéré comme fiable. Il est également tenu compte de la présence ou non de clades eux-mêmes répétés et incompatibles avec le clade d'intérêt. Plus un clade est contredit par un clade possédant un grand nombre d'occurrences, moins il doit être considéré comme fiable.
Dans une deuxième partie, l'indice de répétition est calculé à partir d'une série d'analyses mettant en jeu environ 200 taxons et basées sur quatre marqueurs nucléaires : Rhodopsine, MLL4, IRBP et RNF213 (ce dernier étant utilisé ici pour la première fois). Ces marqueurs sont analysés suivant des méthodes probabilistes, séparément et en combinaisons de 2, 3 ou 4, ce qui permet de bénéficier des avantages de l'analyse combinée tout en ayant des séries de résultats indépendants à comparer.
Les résultats de l'analyse de fiabilité sont ensuite synthétisés sous forme d'arbres incluant en priorité les clades les plus fiables, suivant des méthodes gérant de plusieurs façons les différences d'échantillonnages taxinomiques entre les jeux de données.
Les arbres de synthèse obtenus permettent de préciser la structure de la phylogénie des téléostéens acanthomorphes (Actinopterygii : Teleostei). De nouveaux clades fiables sont identifiés à plusieurs niveaux de résolution, et de nouveaux taxons sont placés dans la phylogénie des téléostéens acanthomorphes.

Mots clés

Acanthomorpha, Clades, Congruence taxinomique, Fiabilité, Phylogénie, *Supertree*, Support

Adresse du laboratoire où la thèse a été préparée

UMR 7138 CNRS IRD MNHN UPMC – Systématique, Adaptation, Évolution
Université Paris VI – Pierre et Marie Curie
Bâtiment A, 4ème étage, Case 5
7, quai Saint Bernard
75252 Paris Cedex 05
téléphone : 01 44 27 58 01

English title

Reliability of clades and taxonomical congruence
Application to the phylogeny of acanthomorph teleosts

Abstract

The goal of a phylogeny is usually to get an idea of the real relationships between living organisms. It is thus not a good idea to be satisfied with the tree obtained by a simple combined analysis of the data. Indeed, even robust clades in such a tree may not be reliable. Confidence about a phylogenetic statement can only stem from a comparison between results obtained from independent data.
This thesis begins by proposing a repetition index to measure reliability. This index takes into account the number of occurrences for a clade over a set of trees obtained from independent data, that is, data unlikely to be subject to the same reconstruction biaises. The more a clade occurs this way, the more it can be considered reliable. The index also takes into account clades that are not compatible with the clade under focus and that are also repeated. The highest the number of occurrences of a clade's contradictor, the less reliable it should be considered.
In the second part of the thesis, the repetition index is calculated from a series of analyses involving about two hundred taxa. The analyses are primarily based on four nuclear markers: Rhodopsin, MLL4, IRBP and RNF213 (the latter being used for the first time). These markers are analysed under probabilistic models, separately and in combinations of 2, 3 or 4. This allows to take advantage of combined analysis while still being able to compare sets of independent results.
The results are then summarized into trees including mostly reliable clades. Several ways of dealing with the differences between the taxonomic samplings of the datasets are used.
The synthesis trees allow to refine the structure of the acantomorph (Actinopterygii: Teleostei) phylogeny. New reliable clades are identified, at several resolution levels, and new taxa are placed into the phylogeny of acanthomorph teleosts.

Keywords

Acanthomorpha, Clades, Phylogeny, Reliability, Supertree, Support, Taxonomical congruence

Remerciements (et quelques sigles)

Merci à Agnès DETTAÏ pour son *coaching* à la paillasse.
Merci à Harold LOPPARELLI et au département de physique de l'ENS (École Normale Supérieure) pour m'avoir fait découvrir André ADOUTTE et la phylogénie. Merci à mes vénérés maîtres de biologie, dont les enseignements m'ont guidé jusqu'ici et merci à ma famille pour m'avoir amené jusqu'à eux.
Merci au méta-conseiller qu'est le forum de l'ENS[1] ainsi qu'à la communauté du logiciel libre pour leur aide technique précieuse. Merci à Pedro CORDEIRO qui m'a conseillé les modules OBI (Ouvroir de Bio-Informatique), merci à Joël POTHIER et ses collègues qui les ont fait vivre.
Merci aux divers informatheux que j'ai pu rencontrer ainsi qu'à Mickaël AGOLIN, Nathanaël CAO et Alexandre HASSANIN, *for valuable discussion*, comme ils disent dans les revues.
Merci à l'UPMC (Université Pierre et Marie Curie), à l'EDDV (École Doctorale Diversité du Vivant) et à l'UMR (Unité Mixte de Recherche) 7138, son directeur (Hervé LE GUYADER), son personnel administratif (Catherine LAGRENADE, Philippe LEBALLEUR et Danièle MERKILED) et son antenne du MNHN (Museum National d'Histoire Naturelle) pour m'avoir accueilli en tant que doctorant.
Merci à Philippe BEAREZ, Bruno CHANET, Wei-Jen CHEN, Guy DUHAMEL, Samuel IGLESIAS, Joelle LAI, Sébastien LAVOUÉ, Catherine OZOUF, Leo SMITH, à l'aquarium de la Porte Dorée et à plein d'autres pour les morceaux de viande ayant livré le précieux ADN (Acide Désoxyribo-Nucléique) sans lequel cette thèse aurait manqué de concret.
Merci au SSM (Service de Systématique Moléculaire) et au CNS (Centre National de Séquençage, dit « Génoscope ») sans qui l'ADN sus-cité aurait eu du mal à s'exprimer pleinement.
Merci à Martine DESOUTTER et autres ichthyologues sans qui mes arbres ne voudraient vraiment rien dire.
Merci à Jean-François FLOT, Jérôme MURIENNE et aux *reviewers* de *Systematic Biology* pour certains conseils bibliographiques.
Merci à Timarcha, au club naturaliste de l'ENS et à mon monitorat à l'UVSQ (Université Versailles – St.-Quentin) pour m'avoir fait prendre l'air un peu.
Merci aux membres du jury pour avoir accepté d'en être.
Merci à Guillaume LECOINTRE qui malgré ses multiples activités et fonctions a tout de même réussi à encadrer mon travail.

Les dessins de poissons utilisés pour illustrer ce document proviennent de `Fishbase` (FROESE et PAULY, 2006).

[1] *A.k.a. Normal' Coprocessor.*

Table des matières

Introduction **8**

Contexte scientifique général 8
Mesures de qualité en phylogénie 9
Cohérence et fiabilité 10
Congruence taxinomique et combinaisons partielles 12
Cas d'étude : les acanthomorphes 13

I. Mesurer la fiabilité des clades 16

Définitions et autres conventions **17**

1. Le cahier des charges d'un indice de fiabilité des clades **20**

1.1. Prendre en compte les occurrences indépendantes des clades 20
1.2. Gérer l'incompatibilité entre résultats 20
1.2.1. Intrus et échappés 21
1.2.2. Fiabilité d'hypothèses incompatibles 21
1.3. Gérer la comparabilité des résultats 22

2. Proposition d'indice de fiabilité et application **23**

2.1. Définition de schémas de partitionnement 23
2.2. Confrontation des clades incompatibles 24
2.3. Fiabilité et méta-fiabilité 25
2.4. Article accepté 26

3. Perfectionnements de l'indice **46**

3.1. Prendre en compte le comportement du programme de phylogénie utilisé face à l'absence de signal 46
3.1.1. Qu'est-ce qu'un jeu de données sans signal ? 47
3.1.2. Comment la méthode de reconstruction pourrait jouer sur la distribution des arbres ? 47
3.2. Utiliser les probabilités postérieures des clades 48
3.3. Définir des domaines de validité 48

4. Synthétiser la fiabilité des clades sous forme d'un arbre **51**
4.1. Avec des relations complètes . . . 51
4.2. Avec des relations partielles . . . 52
4.2.1. Synthèse par MRP . . . 52
4.2.2. Synthèse inspirée par le `MinCut` *Supertree* . . . 52

II. Phylogénie des téléostéens acanthomorphes **63**

5. Les acanthomorphes **64**
5.1. Brève présentation . . . 64
5.2. Nouvelle donne moléculaire . . . 64

6. Données utilisées **66**
6.1. Taxons . . . 66
6.2. Marqueurs . . . 73
6.2.1. Les ADN ribosomiques . . . 73
6.2.2. La Rhodopsine . . . 74
6.2.3. MLL4 . . . 74
6.2.4. IRBP . . . 74
6.2.5. C-mos . . . 75
6.2.6. TMO-4c4 . . . 75
6.2.7. RAG1 . . . 75
6.2.8. RNF213 . . . 75
6.3. Travail moléculaire . . . 76
6.3.1. Extraction . . . 76
6.3.2. Amplification par PCR . . . 76
6.3.3. Séquençage . . . 78
6.4. Nettoyage et alignement . . . 79

7. Analyses **80**
7.1. Analyse primaire . . . 80
7.1.1. Jeux de données . . . 80
7.1.2. Méthode d'analyse . . . 81
7.2. Analyse secondaire . . . 82
7.3. Article soumis . . . 83

8. Résultats **152**
8.1. Nouvelles données . . . 152
8.2. Résultat des analyses . . . 157

8.2.1. Analyses primaires 157
8.2.2. Analyses secondaires 169

III. Discussion 181

9. Discussion des relations de parenté entre acanthomorphes 182
9.1. La base de l'arbre 183
9.2. Les sous-clades de γ 185
9.2.1. Le clade δ 185
9.2.2. Le clade Q 188
9.2.3. Le clade F 189
9.2.4. Le clade L 190
9.2.5. Le clade η 192
9.3. Les relations au sein du clade η 193
9.3.1. Les limites du clade N 195
9.3.2. Les relations au sein du clade X 196

10. Nouvelles propositions méthodologiques 199
10.1. Mieux gérer les taxons rares 199
10.2. Utiliser la fiabilité pour contraindre des réanalyses 200
10.3. Utiliser la fiabilité comme critère d'optimalité 200
10.3.1. Quel type d'éléments utiliser ? 201
10.3.2. Comment passer de la fiabilité des triplets à celle de l'arbre ? 202
10.4. Placer des indices sur l'arbre de synthèse 202

11. Discussion générale sur la fiabilité 203

Conclusion 204
Conclusions méthodologiques 204
Conclusions phylogénétiques 205
Conclusions techniques 206

Annexe 216
Représentation informatique des relations phylogénétiques 216

Introduction

Contexte scientifique général

La phylogénie consiste à construire des arbres représentant des relations de parenté putatives entre êtres vivants (actuels ou fossiles) à partir d'observations sur un échantillonnage de taxons terminaux, par exemple des individus représentant leur espèce ou leur genre. Morpho-anatomie, protéines, ADN, comportements sont les principales sources de caractères pouvant servir à comparer les taxons. L'utilisation des séquences d'ADN, de par la relative facilité (PCR[2], séquençage automatisé) avec laquelle on peut obtenir un grand nombre de caractères homologues, est très répandue en phylogénie, quand les taxons étudiés sont actuels ou quand on dispose de spécimens anciens dans état de conservation suffisamment bon pour permettre l'extraction d'ADN. On parle alors de « phylogénie moléculaire ». La phylogénie moléculaire, en synergie avec l'accessibilité croissante des moyens informatiques, a suscité le développement de nouvelles méthodes d'analyses ; des modèles d'évolution des séquences ont permis l'amélioration des méthodes de distance et ont été à l'origine du développement spectaculaire de l'application à la phylogénie des méthodes probabilistes (maximum de vraisemblance et inférence Bayésienne). Les trois grandes familles de méthodes de construction des phylogénies utilisées actuellement sont donc la méthode cladistique (maximum de parcimonie), les méthodes phénétiques (méthodes de distance) et les méthodes probabilistes. C'est dans ce contexte scientifique riche en réflexions méthodologiques, allant de pair avec des questionnements d'ordre épistémologique, que se situe mon travail de thèse ; il comporte un aspect méthodologique motivé par des considérations épistémologiques concernant la façon d'aborder la disponibilité de plusieurs sources de données. Il comporte également un aspect « phylogénie moléculaire » typique, avec obtention de séquences d'ADN, et analyse des ces données sous des modèles d'évolution nécessitant un usage intensif de moyens informatiques. Ce travail s'inscrit dans la continuité des thèses de Wei-Jen CHEN (CHEN, 2001) et Agnès DETTAÏ (DETTAÏ, 2004) toutes deux également dirigées par Guillaume LECOINTRE, et appliquées à la phylogénie du groupe des téléostéens acanthomorphes. Cette continuité est visible à la fois du point de vue méthodologique et du point de vue du sujet d'étude (ma thèse est l'occasion de compléter et d'étendre les jeux de données moléculaires sur les téléostéens acanthomorphes obtenus par les deux précédents doctorants).

[2] *Polymerase Chain Reaction*, procédure d'amplification d'ADN *in vitro*, l'une des principales activités des utilisateurs du SSM.

Mesures de qualité en phylogénie

Le succès de la phylogénie, qui serait l'obtention des relations de parenté réelles[3] entre les taxons étudiés, est difficile à évaluer puisqu'on n'a pas une connaissance directe du passé qui offrirait un point de comparaison[4]. Ce succès est conditionné par la qualité des données, la représentativité de l'échantillonnage, et l'adéquation des méthodes utilisées au champ d'étude et aux données disponibles. La qualité des résultats obtenus est mesurée de diverses manières, dont certaines peuvent se traduire par des indices de *support* pour les clades :

1. la *robustesse* mesure la stabilité des résultats face à la perturbation des données (bootstrap (FELSENSTEIN, 1985), jacknife, voire perturbations de l'échantillonnage taxinomique) ;
2. la *sensibilité* mesure la stabilité des résultats, à données fixées, face à la diversité des méthodes et des modèles utilisables lors de la reconstruction phylogénétique (« tapis navajo ») ;
3. la *fiabilité* mesure le degré de confiance que l'on peut avoir dans l'affirmation que les résultats sont proches de la réalité ; on verra qu'elle peut être approchée par la stabilité des résultats face à la diversité des sources de données.

Par support, ou soutien, on désigne une catégorie générale d'indications concernant les relations de parenté présentes dans un arbre[5].
Les deux premières catégories de support mentionnées ci-dessus, robustesse et sensibilité, permettent de se faire une idée de la qualité des données, des méthodes les plus appropriées au traitement des données ou des modèles les plus justes à utiliser dans le cadre de la méthode choisie. Elles reflètent la force du signal présent dans un jeu de données particulier ; si le signal est fort, on aura beau torturer les données de diverses manières toutes plus raffinées les unes que les autres, elles persisteront dans leur version des faits[6]. On serait tenté de s'incliner devant une telle

[3] Idéalement, il faudrait distinguer plusieurs niveaux auxquels on peut définir des relations de parenté. Les individus étudiés sont porteurs d'une mozaïque de gènes qui n'ont pas forcément tous suivi les mêmes voies généalogiques. Se baser sur des gènes pour inférer des relations de parenté entre individus est donc un moyen détourné qui peut se révéler plus ou moins inapproprié selon les cas. Ces questions sont particulièrement importantes pour la génétique des populations et pour la phylogénie de groupes à fort taux de transferts horizontaux, comme les bactéries.

[4] Notons toutefois l'existence de phylogénies expérimentales réalisées en laboratoire avec des organismes à temps de génération très court ainsi que la possibilité, assez utilisée pour le test de méthodes, de simuler une phylogénie *in silico* qu'on essaie ensuite de reconstruire par la méthode à tester (exemples : HILLIS et BULL, 1993; WIENS et REEDER, 1995; WIENS, 1998).

[5] Indices de Bremer (BREMER, 1988), longueurs de branches, nombres de synapomorphies sont donc également des formes de support, mais qui n'entrent pas forcément très bien dans les catégories particulières dont il est question ici. L'indice de Bremer d'un clade étant le nombre de pas supplémentaires qu'il faut accepter pour trouver une topologie ne contenant pas ce clade, il pourrait être considéré comme une mesure de sensibilité ; la sévérité avec laquelle on applique le principe de parcimonie peut en effet être considérée comme un paramètre de la méthode cladistique.

[6] Au risque d'insister dans le mauvais goût, précisons l'analogie tortionnaire : des données robustes seraient des données qui continueraient à dire la même chose, même après avoir subi les plus atroces mutilations ; des données sensibles seraient des données qui changeraient facilement d'avis selon la manière dont est mené

persistance (si les données sont vraiment porteuses d'un signal clair, pourquoi douter des relations bien soutenues ?), mais il faut garder à l'esprit que certains savent très bien mentir (il peut y avoir de forts biais dans le signal d'un jeu de données), voire sont convaincus de ce qu'ils racontent, mais ce qu'ils savent ou disent ne correspond pas toujours à ce que l'on cherche vraiment à savoir (l'histoire d'un gène peut différer de l'histoire des individus : DOYLE, 1992; MADDISON, 1997). Ceci nous amène au troisième type de mesure, la fiabilité, qui est probablement le plus délicat à obtenir, en raison de l'inaccessibilité de la réalité passée. Une définition pratique et simple de la fiabilité est impossible. En effet, la croyance en la vérité d'un résultat comprend une dimension psychologique. Et même si, dans un souci d'objectivité, on parvient à mettre de côté cet aspect là, on est encore loin de pouvoir proposer une mesure directement applicable à un résultat scientifique. Si un résultat donné fait partie de ce qu'on appelle le langage de la science, s'il peut être énoncé sous une forme comprise par les locuteurs du même langage (c'est-à-dire, les scientifiques de la même discipline), l'appréciation de la correspondance entre ce résultat et la réalité se situe, elle, à un autre niveau, hors de la science elle-même. Ainsi, on dit qu'un énoncé sur la fiabilité d'un résultat fait partie d'un méta-langage ; en l'occurrence, un ensemble d'énoncés portant sur les énoncés de la science. Seuls de hardis épistémologues ou sémanticiens se hasardent à formaliser ce méta-langage sous forme d'équations, et s'ils parviennent a en tirer des formules, telles que celle du « degré de confirmation » (CARNAP, 1950), on peut se demander si elles ont jamais été appliquées à des cas concrets en sciences expérimentales sans graves dénaturations. De telles formules ne peuvent que donner des idées sur les propriétés idéales que devraient avoir des mesures de fiabilité. On devra donc se contenter d'approcher la fiabilité par des mesures utilisant directement le langage de la science. En phylogénie au moins, ce langage est praticable, et j'espère le pratiquer ici de manière compréhensible.

Cohérence et fiabilité

Un point d'approche de la fiabilité est la cohérence qui peut exister au sein de l'ensemble de résultats dans lequel s'inscrit une inférence phylogénétique[7]. Cette cohérence peut être interne, quand les seuls résultats considérés sont ceux produits lors d'analyses, faites dans un même cadre méthodologique, d'un ensemble bien circonscrit de données. Elle fait partie de ce qu'on appelle le *contexte de découverte*. Elle peut aussi être une cohérence « externe », si l'on décide de prendre en compte l'ensemble des résultats déjà produits indépendamment concernant le domaine d'étude. La difficulté qu'il y a à formaliser une mesure de la fiabilité incite à se pencher d'abord sur la cohérence interne à une étude. C'est ce qui sera fait ici. Cependant, une fiabilité dérivant d'une seule étude, même bien menée et validée par une publication dans une revue à

l'interrogatoire. Mais le défaut de cette analogie est qu'il serait en fait plus juste de parler de la robustesse d'un fait énoncé par le torturé, puisqu'on parle de la robustesse d'un clade, et non pas de la robustesse des données elles-mêmes.

[7]Des témoignages concordants incitent à une certaine confiance dans la véracité des faits.

comité de lecture, n'aura jamais le cachet psychologique et social qui résulte de la reproduction des résultats de façon indépendante par plusieurs équipes de recherche. L'aspect social de la fiabilité relève du *contexte de validation*. Ce n'est qu'après avoir atteint une bonne fiabilité « externe » qu'un fait entrera dans le cadre des connaissances objectives reconnues par une communauté scientifique.

La cohérence interne des résultats d'une étude phylogénétique peut être envisagée selon deux approches : analyses simultanées ou congruence taxinomique.

Fiabilité en analyse simultanée : le pouvoir de persuasion du témoin

On peut choisir une méthode qui garantisse un résultat unique et cohérent en lui-même ; c'est la démarche en général adoptée par les tenants du « *total evidence* » dans son acception classique (KLUGE, 1989). Les données utilisées sont mises à contribution lors d'une analyse simultanée avec l'espoir que le signal véritablement historique — partagé en principe par ces données — se distingue du bruit constitué par les biais individuels de chaque jeu de données. On espère en effet que les raisons poussant l'analyse d'une partie des données à produire un résultat non conforme à la réalité diffèrent la plupart du temps d'une partie à l'autre. Ainsi, les biais liés à chacune de ces parties sont différents et se moyennent lors de l'analyse simultanée, contrairement au signal véritablement historique, qui s'additionne, se renforce. Dans le cadre de cette démarche, la fiabilité peut être approchée par la robustesse et la sensibilité des résultats, qui indiquent à quel point un signal dominant (qu'on espère être le bon) s'impose.
Il est d'autant plus légitime de passer ainsi d'une telle forme de support à la fiabilité que toute l'information disponible a été prise en compte lors de l'analyse. Or, l'information disponible ne se réduit pas aux caractères mis dans la matrice ; il faudrait idéalement prendre en compte toutes les autres connaissances ayant trait au problème étudié (*background knowledge*). Ceci peut être difficile à mettre en œuvre (LECOINTRE et DELEPORTE, 2005) et n'est pas toujours bien formalisable, mais on peut cependant dégager certains faits qu'il semble raisonnable de prendre en compte. Par exemple, moins les données disponibles sont abondantes, plus il est probable qu'un biais structurant une partie d'entre elles s'impose face aux autres par ailleurs trop faiblement structurées[8] (PHILIPPE et DOUZERY, 1994; BRINKMANN *et al.*, 2005). Une pure analyse simultanée est donc insuffisante pour donner une évaluation fiable de la fiabilité. Ce constat amène naturellement à se placer dans la deuxième catégorie d'approches, la congruence taxinomique, qui consiste à comparer les résultats obtenus à partir de différentes sources d'information (MICKEVICH, 1978; MIYAMOTO et FITCH, 1995).

[8] Si l'un des témoins a un plus grand talent rhétorique que les autres, sa version des faits risque de l'emporter, quelle qu'en soit la véracité.

Fiabilité en analyses séparées : la concordance des témoignages

Afin de détecter les biais liés à chaque partie des données, il peut s'avérer intéressant de les séparer lors de leur analyse. Les éléments de résultat communs à toutes les parties pourront alors être distingués du reste, et considérés comme fiables (MIYAMOTO et FITCH, 1995; CHEN *et al.*, 2003; DETTAÏ, 2004). C'est donc bien ici aussi par une forme de cohérence que l'on détectera les résultats fiables, une cohérence entre analyses. Cependant, en gardant les données séparées on court le risque que pour certaines des parties analysées, la quantité de données soit insuffisante pour qu'émerge un signal clair, voire que le signal émergeant lors d'une analyse donnée soit un signal non-historique. Dans la pratique on est en effet plus souvent satisfait du résultat d'une analyse combinée, que de ceux des analyses séparées. L'arbre de l'analyse combinée présente souvent un plus grand nombre de nœuds résolus de façon robuste que les arbres des analyses séparées. Si les arbres résultant de chacune des analyses séparées sont peu résolus, il n'y aura pas grand chose à dire à l'issue de l'analyse de congruence taxinomique. Quand on dispose d'un nombre suffisant de jeux de données indépendants, on peut alors adopter une démarche hybride entre analyse combinée et analyses séparées. C'est la méthode des combinaisons partielles proposée par DETTAÏ et LECOINTRE (2004).

Congruence taxinomique et combinaisons partielles

Comparaison de résultats indépendants

La congruence taxinomique repose sur la comparaison d'arbres résultant de l'analyse de données indépendantes. Dans la pratique, des données indépendantes sont des données pour lesquelles on n'a pas de raison particulière de soupçonner une cause partagée pouvant provoquer des biais communs lors des analyses. Ainsi, les résultats communs à plusieurs analyses de données indépendantes devraient avoir pour cause la seule chose commune aux jeux de données ; le fait que les taxons d'où sont tirés ces données ont une histoire évolutive unique (MIYAMOTO et FITCH, 1995), et donc que les transformations des caractères, quels qu'ils soient, sont porteuses d'une part de signal véritablement phylogénétique. Le signal véritablement phylogénétique de chaque jeu de données devrait permettre de reconstituer des éléments de l'unique histoire des taxons. En toute rigueur, ceci n'est vrai que dans la mesure où l'histoire des marqueurs phylogénétique est la même que celle des taxons. On fera l'hypothèse que considérer que les marqueurs phylogénétiques ont la même histoire que les taxons n'est pas une mauvaise approximation dans le cas de la phylogénie à grande échelle des acanthomorphes, le cas d'étude sur lequel porte la présente thèse[9].

[9]Ceci serait beaucoup plus problématique dans une étude portant sur des eubactéries, par exemple, car les échanges d'ADN sont fréquents dans ce groupe.

Combinaisons indépendantes

La méthode des combinaisons partielles consiste à combiner une partie des données seulement, de manière à garder des jeux de données indépendants. Si l'on utilise quatre jeux de données indépendants A, B, C et D, la combinaison des jeux de données A et B est indépendante du jeu de données C et du jeu de données D. Une analyse de congruence taxinomique portant sur les jeux de données $A \cup B$, C et D sera donc valide. En analysant A et B de façon combinée, on espère faire émerger le signal historique présent dans les jeux de données. En considérant cette combinaison partielle, on se donne une chance de mettre en évidence des résultats qui n'apparaissaient ni dans l'analyse de A, ni dans celle de B, et qui, s'ils apparaissent dans l'analyse de C, de D ou mieux encore de chacun de ces deux autres jeux de données, pourront être considérés comme fiables. Une telle démarche, utilisant toutes les combinaisons partielles possibles, a été mise en œuvre dans les travaux liés à la thèse d'Agnès DETTAÏ (DETTAÏ, 2004; DETTAÏ et LECOINTRE, 2004) et a donné lieu à l'établissement manuel de très grands tableaux comparatifs. Plus le nombre de jeux de données est grand, plus la méthode est lourde à appliquer. Le problème de l'automatisation d'une telle méthode est donc posé, ainsi que celui de la synthèse des résultats de l'analyse de congruence taxinomique. C'est cette approche de la fiabilité, sa formalisation sous forme d'indices à attribuer à des clades et son automatisation qui constituent l'objet de la première partie de ma thèse.

Cas d'étude : les acanthomorphes

La deuxième partie concerne la mise en œuvre de cette approche à un cas concret : la phylogénie des téléostéens acanthomorphes. Une mise en pratique est d'autant plus nécessaire que la partie théorique, méthodologique, a été suscitée par un problème pratique à résoudre ; l'automatisation de l'analyse de congruence taxinomique.
Le passage de la théorie à la pratique est l'occasion de prendre conscience de certaines contraintes. La complexité informatique des méthodes proposées est elle raisonnable ? A-t-on les moyens techniques suffisants pour les mettre en œuvre ? Jusqu'à quel point peut-on pratiquer l'approche par combinaisons partielles de façon exhaustive quand le nombre de taxons ou de marqueurs disponibles augmente ? Que faire quand il manque des données ? L'indice de fiabilité obtenu est-il satisfaisant ? Semble-t-il trop conservatif ? Pas assez ? Quel est son comportement face aux cas d'attraction des branches longues ? Ce sont là des questions qui se sont posées au fur et à mesure de la thèse : lors de l'implémentation sous forme de programme informatique, de l'application à des cas-test, de la soumission de la méthode à une revue au comité de lecture consciencieux et enfin de la mise en pratique sur les jeux de données accumulés en fin de thèse.

Les téléostéens acanthomorphes (ROSEN, 1973) — revenons à nos poissons[10] — constituent

[10]Cette petite plaisanterie est l'occasion de rappeler à l'éventuel lecteur non-phylogénéticien que les téléostéens,

un groupe de vertébrés très riche en espèces (un peu plus de 16000 espèces, réparties en 311 familles selon la classification retenue par NELSON, 2006) et au sein duquel les relations phylogénétiques entre familles et entre ordres n'avaient donné lieu qu'à très peu d'hypothèses jusqu'à assez récemment (LAUDER et LIEM, 1983; STIASSNY et MOORE, 1992; JOHNSON et PATTERSON, 1993; MOOI et JOHNSON, 1997; IMAMURA et YABE, 2002). C'est un cas unique chez les vertébrés. Pour les oiseaux ou pour les mammifères, de nombreuses relations avaient été proposées sur la base de la morphologie, et nombre de ces relations inter familiales n'ont que partiellement été remises en cause par les phylogénies moléculaires ; ce sont les relations entre ordres qui sont en train d'être élucidées (HARRISON *et al.*, 2004; SPRINGER *et al.*, 2004; MAYR, 2007). Dans le cas des squamates, des hypothèses avaient été faites, mais les phylogénies moléculaires récentes sont en train de nous donner une vision renouvelée des relations au sein de ce groupe (FRY *et al.*, 2005). Dans le cas des acanthomorphes, les études moléculaires peinent à faire apparaître quelques grands clades au contenu encore imprécis, et l'essentiel des relations entre ces clades est encore inconnu (CHEN *et al.*, 2003; DETTAÏ et LECOINTRE, 2005; MIYA *et al.*, 2005; ORRELL *et al.*, 2006; SMITH et CRAIG, 2007).

Peut-on se permettre d'affirmer que les acanthomorphes sont un groupe idéal pour travailler sur la fiabilité des clades ? Il est en tout cas très probable que l'orientation méthodologique qu'a prise ma thèse est due aux questions phylogénétiques que suscite ce groupe et aux contraintes posées par l'échelle temporelle concernée par la diversification des acanthomorphes.

Le manque de connaissances *a priori* sur la monophylie de nombreuses grandes subdivisions des acanthomorphes impose l'utilisation d'un échantillonnage très varié. En effet, on ne peut pas aisément choisir quelques familles représentatives de la diversité du groupe puisque les regroupements de familles déjà admis sont rares ; une famille n'est dans la plupart des cas représentative que d'elle-même. Par ailleurs, étant donnés les temps de divergence au sein d'un groupe comme les acanthomorphes, l'obtention d'amorces de PCR qui marchent aussi bien pour tout l'échantillonnage et dans des conditions uniques n'est pas garantie, surtout s'il faut en plus que le marqueur amplifié soit porteur d'un signal phylogénétique approprié pour le groupe étudié. En conséquence, l'accumulation d'un jeu de données aussi utile que complet nécessite du temps, une bonne organisation, et/ou beaucoup d'expérience pratique (idéalement, les trois à la fois). En somme, obtenir les données est un travail ingrat et peu valorisable en tant que tel dans une publication. Qui lirait un immense tableau consignant la multitude des PCR essayées[11] avant d'obtenir enfin quelques dizaines de séquences convenables[12] ?

font partie de ce qu'on appelle les poissons en français courant. Ce sont des actinoptérygiens ; des poissons à nageoires rayonnées (exemples : saumon, carpe, morue, esturgeon), groupe-frère des sarcoptérygiens, ceux à nageoires charnues (exemples : cœlacanthe, dipneuste, autruche, babouin, serpent python bicolore de rocher) dont font partie les tétrapodes (ces derniers ont, sauf exception, les nageoires tellement charnues qu'ils n'ont pas vraiment l'allure de poissons, avec leurs quatre pattes).

[11] Et qui prendrait plaisir à mettre en page un tel tableau pour l'annexer à un article ?

[12] Pour comparer avec un autre travail technique, la mise au point d'un programme informatique m'a semblée plus stimulante intellectuellement que la mise au point de conditions de PCR.

Avec un échantillonnage de taille différente et un nombre de marqueurs nettement plus grand, les contraintes techniques auraient été sur certains points assez différentes de celles dans lesquelles nos propositions méthodologiques sur la fiabilité des clades ont été faites. L'approche par combinaisons partielles n'aurait peut-être pas été envisagée, ou alors aurait rapidement été mise de côté s'il avait fallu faire des milliers d'analyses différentes. Avec 10 marqueurs, à moins de n'avoir qu'un très faible nombre de taxons, faire les 1023 analyses nécessaires aurait pris un temps déraisonnable. En ne considérant que les combinaisons à un ou deux marqueurs parmis les 10, il y a déjà 55 analyses à faire : $C_{10}^{1} + C_{10}^{2} = 10 + 45$.

Première partie .

Mesurer la fiabilité des clades

Définitions et autres conventions

Pour limiter les ambiguïtés dans le texte, il vaut mieux poser des définitions au début et essayer de s'y tenir. C'est ce que je me propose de faire ci-dessous. Des définitions plus formelles et moins ambiguës sont possibles mais pas forcément nécessaires ni faciles à manipuler. Certains des termes définis ici seront à nouveau discutés plus loin. D'autres seront à peine évoqués[1].

Taxon terminal : élément de base manipulé lors de la construction d'un arbre phylogénétique. Les taxons terminaux sont les feuilles des arbres phylogénétiques. On parle parfois d'OTU (*Operational Taxonomic Unit*). Par simplification je parlerai de « taxon », tout court[2]. Dans les exemples abstraits, je m'efforcerai de représenter les taxons par des minuscules d'imprimerie (a, b, c, *etc.*)[3].

Domaine de validité : ensemble de taxons sur lequel l'analyse d'un ou de plusieurs jeux de données a été menée afin d'y établir des relations. Ces relations peuvent être décrites par des structures telles que des bipartitions, des clades, des sous-arbres, des arbres (voir ci-dessous). Le domaine de validité d'une relation (quand elle en a un bien défini, ce qui n'est pas nécessairement le cas) est celui de l'analyse ayant produit cette relation. La comparaison des résultats d'analyses portant sur des domaines de validité différents peut poser problème ; je parlerai notamment du problème de non-recouvrement des échantillonnages taxinomiques. Une des manières de traiter ce problème est de définir un domaine de validité pour la comparaison, qui est l'intersection des domaines de validité des relations à comparer. Certains des taxons impliqués dans les relations à comparer pourront être absents de cette intersection. Les domaines de validité seront abordés plus en détail dans la section 3.3 (voir page 48).

Relation : déclaration concernant la distinction de groupes parmi un ensemble de taxons. Des questions importantes rencontrées lors de mon travail sont celles de savoir dans quelle mesure on peut comparer des relations et comment combiner plusieurs relations pour construire un arbre phylogénétique.

[1] Le besoin de préciser certains termes découle parfois de problèmes pratiques rencontrés lors de l'implémentation informatique de la mesure de fiabilité.

[2] Un taxon non-terminal est un regroupement de taxons terminaux qui forment un clade.

[3] Je représenterai parfois un ensemble de taxons en accolant les lettres les représentant : abc, souvent avec des parenthèses : (abc). Dans le cas d'exemples non-abstraits, on sépare les noms des taxons par des virgules : $(Lateolabrax, Howella, Epigonus)$.

Relation de parenté : relation orientée ; c'est-à-dire dans laquelle il existe des groupes de taxons de statut différent, une hiérarchie entre taxons. Les relations de parenté peuvent prendre la forme de « *n-taxon statements* » (voir WILKINSON, 1994). Un arbre phylogénétique comporte un certain nombre de relations de parenté. À une relation non-orientée peuvent correspondre différentes relations orientées selon la position de la racine.

Relation partielle : relation n'étant pas définie sur un domaine de validité particulier ou dans laquelle le statut de certains taxons du domaine de validité n'est pas précisé. Des relations non partielles définies sur des domaines de validité différents ne sont pas directement comparables[4]. On pourra se ramener à un domaine de validité commun : l'intersection des domaines de validité des relations à comparer. Une autre démarche est de ne considérer que des relations partielles. Une relation non partielle est décomposable en relations partielles.

Relation élémentaire : relation (informative) concernant un ensemble minimal de taxons. Une relation élémentaire non-orientée concerne 4 taxons[5] ; il existe 3 relations élémentaires possibles pour les taxons a, b, c et d, qu'on pourra noter $(ab|cd)$, $(ac|bd)$ et $(ad|bc)$[6]. Une relation de parenté élémentaire concerne 3 taxons ; la relation de parenté élémentaire « a et b sont plus proches entre eux qu'ils ne le sont de c » pourra être notée $((ab)c)$ ou $(ab|c)$, c étant implicitement reconnu comme « extérieur ».

Paire incluse : relation orientée partielle particulière dans laquelle il est seulement précisé que deux taxons sont plus proches entre eux, sans préciser par rapport à quel(s) autre(s) taxon(s). À une paire incluse correspondent plusieurs triplets possibles. Cette relation n'a de sens que si l'ensemble des taxons concernés est défini. C'est un cas particulier du *nesting* défini par ADAMS (1986). On notera la paire incluse impliquant les taxons a et b $\{ab\}$.

Bipartition : paire de deux ensembles de taxons d'intersection vide. Correspond à une branche interne d'un graphe non-cyclique connexe, en particulier d'un arbre. Quand on parlera de bipartition sans plus de précisions, c'est en général qu'il s'agira d'une relation non-orientée.

Clade : bipartition orientée ; c'est-à-dire qu'une des deux parties de la bipartition est dite extérieure et l'autre intérieure. Dans le cadre d'un arbre enraciné, la partie intérieure est celle ne comprenant pas la racine ; c'est un groupe monophylétique. Par simplification, je parlerai en général de clade pour désigner l'ensemble de taxons constituant la partie intérieure[7]. On dira qu'une paire incluse est induite par un clade si les taxons la composant sont tous les deux dans la partie intérieure de ce clade.

[4] Que dire des taxons présents dans le domaine de validité d'une relation mais absents de celui de l'autre ?

[5] Sur un graphe connexe, non-orienté et non-cyclique, il existe toujours une branche séparant n'importe quelle feuille de n'importe quel sous-graphe comprenant n'importe quelle paire de deux autres feuilles. Autrement dit, dans l'arbre de la vie, toutes les relations non-orientées à 3 taxons sont vraies ; il faut donc un minimum de 4 taxons pour avoir des relations non-orientées informatives.

[6] L'ordre des taxons dans chaque groupe est indifférent, et le côté de la barre verticale dans lequel se situe un groupe de taxons est également indifférent (du moment que les deux groupes ne sont pas du même côté).

[7] Toutefois, il faut bien noter que si deux clades ont la même partie intérieure, ils ne sont pas identiques s'ils sont définis sur des domaines de validité différents ; leurs parties extérieures diffèreront.

Sous-arbre : ensemble de taxons dont la topologie interne est précisée. On peut voir un sous-arbre comme un clade muni d'un ensemble de clades d'intersection vide, tous inclus dans ce clade, tous définis sur le même domaine de validité, et chacun étant lui-même éventuellement muni d'un tel ensemble de sous-clades, et ce récursivement. Un sous-arbre est donc un clade muni de sous-arbres. Un sous-arbre est une relation de parenté composite.

Arbre enraciné : cas particulier de sous-arbre, dans lequel tous les taxons du domaine de validité sont inclus, sauf éventuellement ceux du groupe extérieur. C'est un graphe non-cyclique orienté et connexe dont les branches externes mênent aux taxons. Quand on déracine un arbre, ses clades deviennent des bipartitions.

Contradiction : fait que des relations ne peuvent pas être présentes ensemble dans un même arbre. Dans un domaine de validité donné, deux clades se contredisent si et seulement si ils ont au moins un taxon en commun et chacun au moins un taxon propre. Quand deux clades ne se contredisent pas, on dit qu'ils sont compatibles. Deux relations élémentaires ne peuvent se contredire que si elles portent sur le même ensemble de taxons. La compatibilité deux à deux dans un ensemble de relations partielles ne garantit pas la compatibilité globale de ces relations ; en particulier, il ne suffit pas de disposer d'un ensemble de relations élémentaires portant chacune sur un ensemble de taxons différent pour pouvoir les assembler en un arbre[8].

Précision : mon travail ayant été mené en plein paradigme de la phylogénie moléculaire, j'aurai peut-être tendance à parler de « gène » là où « marqueur phylogénétique » conviendrait. Une bonne partie des réflexions qui vont suivre devraient pouvoir s'appliquer également à des analyses de données morphologiques, comportementales, ou autres.

[8] Et c'est bien regrettable, car en supposant que pour chaque ensemble de 4 taxons, on ait déterminé la relation élémentaire la plus fiable, on disposerait alors d'une méthode de *supertree* très intéressante.

1. Le cahier des charges d'un indice de fiabilité des clades

1.1. Prendre en compte les occurrences indépendantes des clades

Le « matériau » de base de la fiabilité d'un clade sera le nombre de fois qu'il apparaît lors des analyses des données. Il importe que ce nombre ne soit compté que sur des analyses indépendantes les unes des autres, afin de limiter les risques qu'un même biais soit à l'origine de plusieurs occurrences d'un clade[1].

Pour remplir cette condition, on s'efforcera de ne pas séparer des données pour lesquelles on soupçonne une évolution coordonnée. On pourra par exemple veiller à ce que deux marqueurs moléculaires codant des protéines interagissant assez directement biologiquement soient regroupés dans le même ensemble[2]. Le degré d'intégration des organismes étant parfois très élevé, il est impossible d'affirmer avec certitude que deux gènes évoluent indépendamment et l'on est obligé de se contenter des connaissances actuelles en biologie pour partitionner les données à inclure dans l'analyse.

Après avoir ainsi regroupé les marqueurs en parties indépendantes, on disposera d'un ensemble de jeux de données élémentaires indépendants à partir desquels une analyse de fiabilité pourra être menée. Ces jeux de données élémentaires pourront être assemblés en jeux de données combinés dans le cadre de l'utilisation de la méthode des combinaisons partielles.

1.2. Gérer l'incompatibilité entre résultats

Des relations phylogénétiques sont dites compatibles s'il est possible de concevoir un arbre les contenant toutes. Le problème de la compatibilité des clades se manifeste sous deux aspects s'agissant de la fiabilité. L'obtention de manière indépendante de deux clades incompatibles mais

[1] Il faut éviter que les suspects se soient mis d'accord entre eux avant l'interrogatoire.

[2] Il est intéressant de noter que contrairement aux préoccupations de BULL *et al.* (1993), le problème n'est pas ici de savoir si l'on peut combiner des jeux de données, mais de savoir s'ils peuvent ne pas être combinés (voir MIYAMOTO et FITCH, 1995, p. 68).

proches devrait-elle renforcer la fiabilité de ces clades ? Que penser de deux clades incompatibles mais fiables ?

1.2.1. Intrus et échappés

Dans des cas pratiques concrets, le phylogénéticien reconnaît « à l'œil » un certain nombre de clades répétés plus ou moins strictement. L'analyse de certaines parties des données produit par exemple un clade α et l'analyse d'autres parties des données permet de reconnaître des clades ne différant de α que par quelques taxons en plus (intrus) ou en moins (échappés) (voir CHEN *et al.*, 2003; DETTAÏ et LECOINTRE, 2004). Il est tentant de vouloir tenir compte de ces clades approximativement répétés dans le calcul d'un indice de fiabilité. Une première tentative de formalisation des notions d'intrus et d'échappés a consisté à les définir relativement à un clade de référence α — un clade strictement répété au moins une fois, par exemple. Un intrus étant alors défini comme un taxon présent de manière répétée dans α et un échappé étant un taxon de α présentant une position alternative (en dehors de α) répétée.
Il apparaît cependant que ces définitions sont difficilement applicables dans le cas d'une procédure systématique de traitement des résultats. Définir ce qu'est une position alternative répétée est particulièrement délicat.

1.2.2. Fiabilité d'hypothèses incompatibles

L'autre aspect, celui de la fiabilité à donner à deux hypothèses incompatibles, offre une manière détournée de gérer le problème des taxons intrus ou échappés. Si deux clades sont incompatibles, l'un des deux au moins n'est pas vrai, et ne devrait donc pas être fiable. Le plus fiable des deux devrait être celui qui présente le plus d'occurrences indépendantes. La différence entre le nombre d'occurrences d'un clade et celui de son contradicteur, c'est-à-dire le clade avec lequel il est incompatible, devrait donner une meilleure idée de la fiabilité que le nombre d'occurrences brut. En fait, un clade aura probablement plusieurs contradicteurs. C'est *a priori* le plus fiable de ces contradicteurs qu'il faut prendre en compte pour évaluer sa fiabilité, car c'est celui qui est le plus susceptible d'être vrai. Maintenant, voyons ce que deviennent les intrus et les échappés si l'on adopte cette façon d'évaluer la fiabilité. Si un taxon est régulièrement retrouvé au sein d'un clade donné, on obtient un contradicteur de ce clade, qui aura un nombre d'occurrence d'autant plus grand que le taxon intrus se placera à chaque fois à la même position. Si au contraire, ce taxon s'insère rarement dans le clade et en des positions variables, la contradiction apportée sera faible. De même si un taxon « s'échappe » pour former régulièrement un clade non compatible avec son clade « d'origine », on obtiendra une diminution importante de la fiabilité du clade, alors que si le taxon s'échappe rarement, et sans donner lieu à plusieurs occurrences d'un même clade contradicteur, la fiabilité ne devrait pas être beaucoup diminuée. Un raisonnement en termes de clades semble plus facilement formalisable qu'un raisonnement en termes de position de taxons.

On se préoccuperait alors non plus de localiser des intrus et des échappés, mais de chercher des clades contradictoires avec le clade de référence.

1.3. Gérer la comparabilité des résultats

Il est difficile d'obtenir de grands jeux de données sans qu'il manque des taxons pour certains marqueurs. L'absence d'un taxon empêche l'occurrence de tout clade le contenant. Et sur tout clade obtenu par l'analyse d'une partie présentant des taxons manquants pèse le soupçon suivant : Que serait-il devenu s'il ne manquait pas de taxons ? Ces clades auraient-ils disparu par l'insertion en leur sein d'un des taxons manquants ? Considérons l'exemple suivant :
Soient A et B deux parties des données (jeux de données élémentaires ou combinaisons partielles), indépendantes. Soit t un taxon manquant à la partie B. Soient α un clade produit par l'analyse de A et β un clade produit par l'analyse de B tels que $\beta \cup t = \alpha$.
L'analyse de A à qui on aurait retiré t aurait peut-être produit une nouvelle occurrence de β. Peut-on considérer l'occurrence de α comme une « corroboration faible » de β ? Si t est le seul taxon manquant dans B et qu'il n'y a pas de taxon de B qui ne soit pas également dans A, on doit pouvoir considérer que α corrobore β. À présent, imaginons qu'il existe un taxon s distinct de t, présent dans l'un des jeux de données et absent dans l'autre. Si s est absent de B, il est possible que s'il avait été présent, l'analyse des données l'aurait placé au sein du clade β, provoquant ainsi la formation d'un clade incompatible avec α. Si s est absent de A, il est de même possible que s'il avait été présent, il se serait placé au sein du clade α, provoquant ainsi la formation d'un clade incompatible avec β. Et que dire des taxons absents à la fois de A et de B ? Ils sont potentiellement nombreux à constituer une menace pour nos résultats présents, et ce, même si A et B ont un recouvrement taxinomique parfait ! Jusqu'à quel point faut-il spéculer sur les taxons manquants ? Selon le degré de prudence dont on fait preuve et le degré de complexité auquel on est prêt à s'attaquer, plusieurs attitudes sont possibles. Rejeter toute possibilité de comparer les résultats serait probablement trop prudent. On peut s'en tenir strictement à ce que disent les données disponibles et constater les cas manifestes d'incompatiblité entre relations obtenues[3]. On peut tenter de prendre en compte les cas de « corroboration faible » comme celui décrit plus haut (voir WILKINSON *et al.*, 2005 pour une approche plus détaillée des différents cas possibles). On peut aussi simplifier radicalement le problème, au risque de déroger au principe du *total evidence* en éliminant les taxons qui ne sont pas présents dans tous les jeux de données. Se pose alors la question de savoir à quel stade éliminer les taxons. On peut le faire avant même l'analyse primaire des données ou bien en garder certains, selon la quantité de données manquantes que l'on accepte lors des analyses combinées et enlever ces taxons seulement au moment de la comparaison des résultats.

[3]Par exemple, indépendamment des taxons manquant par ailleurs, les relations $ab|cd$ et $ac|bd$ sont incompatibles.

2. Proposition d'indice de fiabilité et application

Dans un premier temps, j'ai proposé un indice de fiabilité applicable uniquement sur des analyses au recouvrement taxinomique parfait. L'idée générale de cet indice a été présentée en congrès (LI et LECOINTRE, 2005) et lors de séminaires, puis a fait l'objet d'un article finalement accepté par *Zoologica Scripta* (LI et LECOINTRE, in press). Le présent chapitre présente les options retenues ainsi que l'article soumis.

2.1. Définition de schémas de partitionnement

L'indice de fiabilité proposé repose sur un comptage du nombre d'occurrences des clades dans le cadre de la méthode des combinaisons partielles. Ce comptage est effectué sur des « schémas de partitionnements », c'est à dire des ensembles de jeux de données indépendants. Le schéma de partitionnement le plus « naturel » dit « schéma de partitionnement de base », est constitué des parties des données minimales non séparables, qu'on appelle les jeux de données élémentaires. Ce sont les différents marqueurs (ou combinaisons de marqueurs, si nécessaire) dont on a estimé qu'ils pouvaient être analysés séparément sans risques de voir un même biais s'exprimer lors de plusieurs analyses. À partir de ces jeux de données élémentaires, on assemble toutes les combinaisons possibles[1]. Ces différentes combinaisons sont utilisées pour constituer les autres schémas de partitionnement sur lesquels on compte des occurrences indépendantes pour les clades. Si l'on a par exemple trois jeux de données élémentaires, notés A, B et C, on pourra considérer les schémas de partitionnement suivants[2] :

1. (A, B, C) (le schéma de partitionnement de base)
2. $(A \cup B, C)$
3. $(A \cup C, B)$

[1] On analyse toutes les combinaisons de 2, 3, 4… jeux de données élémentaire. Avec n jeux de données élémentaires, on peut obtenir $2^n - 1$ jeux de données différents (en comptant les jeux de données élémentaires séparés ainsi que la combinaison des n ensemble, cette dernière constituant un schéma de partitionnement à elle toute seule).

[2] Un schéma de partitionnement correspond à ce qu'on appelle en mathématiques une partition des jeux de données élémentaires. Cependant, le mot « partition » est parfois utilisé par les phylogénéticiens pour désigner un des jeux de données d'une combinaison.

4. $(A, B \cup C)$
5. $(A \cup B \cup C)$ (cas particulier où il n'y a qu'une seule partie)

Chaque schéma de partitionnement offre des occasions supplémentaires de trouver des clades répétés. L'intérêt de combiner une partie des données est de réduire les effets stochastiques et autres artefacts de reconstruction, ce qui augmente les chances que l'analyse des données ainsi combinées produise des clades « vrais », mais cela réduit le nombre maximal d'occurrences indépendantes qu'il est possible de trouver pour un clade. Pour chaque schéma de partitionnement des données, on compte le nombre d'occurrences pour un clade donné. Le calcul de l'indice de fiabilité se fera à partir du plus grand nombre d'occurrences obtenu sur l'ensemble des schémas de partitionnement examinés. Il est tout à fait envisageable que le meilleur indice pour un clade α ne soit pas obtenu pour le même schéma de partitionnement des données que celui permettant d'obtenir le meilleur indice pour un clade β ; les erreurs stochastiques et autres artefacts qui empêcheraient le clade α d'apparaître peuvent très bien être surmontés par certaines façons de combiner les données qui pourraient ne pas être celles qui permettent de surmonter les obstacles à l'apparition de β.

2.2. Confrontation des clades incompatibles

Une fois chaque clade muni d'un nombre maximal d'occurrences, on établit la liste de ses contradicteurs. Comme expliqué en 1.2.2, on retranche du nombre d'occurrences maximal de chaque clade le plus grand nombre maximal d'occurrences rencontré parmis ses contradicteurs. On obtient ainsi un indice de répétition d'ordre 1. Si α est le clade considéré, on note cet indice $R_1(\alpha)$. À la lumière de cet indice de premier ordre, il peut s'avérer que le contradicteur le plus fiable de α n'est plus celui qui avait le nombre maximal d'occurrences le plus grand. Or, on veut que la fiabilité d'un clade soit affectée par le nombre d'occurrences de son contradicteur le plus fiable. On calcule donc un indice de répétition d'ordre, 2 $R_2(\alpha)$, qui est la différence entre le nombre maximal d'occurrences de α et celui de son meilleur contradicteur d'après R_1 (s'il y a plusieurs meilleurs contradicteurs, on utilise le plus grand nombre maximal d'occurrences trouvé parmis ceux-ci). On détermine ensuite le meilleur contradicteur de α à la lumière de R_2 afin de calculer un indice de répétition d'ordre 3 (R_3). On recommence cette procédure jusqu'à ce que pour tout clade α, la valeur de l'indice de répétition soit stable et on retient comme indice de répétition final R_f le dernier indice de répétition calculé. Il est possible que pour certains clades aucun meilleur contradicteur définitif ne soit déterminé de façon stable. L'ensemble du système des indices de répétition parcourt alors une série d'états successifs de façon périodique[3]. On

[3]Le nombre de clades étant fini, le système, qui peut se résumer en un ensemble de paires (α, β) où β est le meilleur contradicteur (éventuellement provisoire) de α ne peut avoir qu'un nombre fini d'états. Si le système fluctue suffisament longtemps sans atteindre un état stable, il finit par repasser par un état déjà atteint. L'état du système à l'ordre n étant déterminé par l'état à l'ordre $n-1$, quand le système atteint un état déjà atteint, il vient de boucler un cycle et est déterminé à repasser successivement par la même série d'états.

prend dans ce cas comme indice de répétition final la moyenne des indices de répétition d'ordres successifs sur une période de fluctuation du système. Ceci est un moyen arbitraire de ne pas trancher explicitement entre plusieurs meilleurs contradicteurs potentiels[4].

2.3. Fiabilité et méta-fiabilité

L'indice de répétition final a pour valeur maximale possible le nombre de jeux de données élémentaires disponibles. Une conséquence de cette propriété est que la fiabilité peut être d'autant plus grande que le nombre de jeux de données est grand. Ceci est *a priori* souhaitable, c'est pourquoi l'indice de répétition a été retenu sous la forme proposée ici, sans mise à l'échelle. Cependant ce choix a été contesté à plusieurs reprises par les *reviewers* des versions successives de l'article en raison du comportement de l'indice dans certaines circonstances.
En effet on peut se demander s'il est légitime qu'un clade α obtenu deux fois et contredit une fois ait la même fiabilité qu'un clade β obtenu 11 fois et contredit 10 fois. Sans mise à l'échelle, les deux situations donnent une fiabilité de 1. Avec mise à l'échelle du nombre de jeux de données, α aurait une fiabilité de $1/3$, et β une fiabilité de $1/21$. β pourrait être considéré comme moins fiable que α car il a un contradicteur fortement répété, ce qui n'est pas le cas de α. Cette manière de juger la fiabilité est compatible avec un indice ramené à l'échelle. L'autre manière de considérer les choses, allant dans le sens de l'indice proposé, est de dire que α n'est pas plus fiable que β car il est très peu répété (deux occurrences seulement) ; ce qui diffère entre les cas de α et β étant la fiabilité de la fiabilité. Dans le cas de α, on ne peut pas dire grand-chose sinon que ce clade est la moins mauvaise hypothèse jusqu'à preuve du contraire. Dans le cas de β, on peut affirmer avec une certaine confiance qu'il y a un réel conflit entre les données analysées, ce qui explique la faible fiabilité du clade. α est fiable, mais peu significativement, β est significativement peu fiable. Ce sont deux manières d'arriver à un même indice de répétition.
Considérons à présent un clade γ obtenu 5 fois sur 5 jeux de données et un clade δ obtenu 5 fois sur 10 jeux de données, mais jamais contredit. Avec l'indice de répétition proposé, les deux clades auraient une fiabilité de 5. La confiance qu'on peut avoir en γ est tirée du fait qu'il a été obtenu avec tous les jeux de données. La confiance qu'on peut avoir en δ provient du fait que ce clade est la seule hypothèse suggérée par les analyses. Avec un indice mis à l'échelle, γ aurait une fiabilité de 1, et δ une fiabilité de $1/2$. Avec un tel indice, la plus faible fiabilité de δ serait l'expression d'une sorte de « déception » de ne pas voir ce clade confirmé autant qu'il aurait pu l'être.

[4] On pourrait choisir celui qui a la meilleure moyenne, mais ça ferait peut-être repartir le système en boucle et il faudrait recommencer… Après tout, pourquoi pas ; c'est une piste à explorer.

2.4. Article accepté

L'article qui suit est en cours de publication dans *Zoologica Scripta*. La version présentée est celle pré-publiée en ligne.

Formalizing reliability in the taxonomic congruence approach

Blaise Li & Guillaume Lecointre

Submitted: 22 February 2008
Accepted: 25 July 2008
doi:10.1111/j.1463-6409.2008.00361.x

Li, B. & Lecointre, G. (2008). Formalizing reliability in the taxonomic congruence approach. — *Zoologica Scripta*, **, ***–***.

In the 'total evidence' approach to phylogenetics, the reliability of a clade is implicitly measured by its degree of support, often embodied in a robustness index such as a bootstrap proportion. In the taxonomic congruence approach, the measurement of reliability has been implemented by various consensus or supertree methods, but was seldom explicitly discussed as such. We explore a reliability index for clades using their repetition across independent data sets. All possible combinations of the elementary data sets are used to compose the sets of independent data sets, across which the repetitions are counted. The more a clade occurs across such independent combinations, the higher its index. However, if other repeated clades occur that are incompatible with that clade, its index is decreased to take into account the uncertainty resulting from conflicting hypotheses. Results can be summarized through a greedy consensus tree in which clades appear according to their repetition indices. This index is tested on a 73 acanthomorph taxa data set composed of five independent molecular markers and multiple combinations of them. On this particular application, we confirm that reliability as defined here and robustness (estimated by bootstrap proportions obtained from a 'total evidence' approach) should be clearly distinguished.

Corresponding author: *Guillaume Lecointre, Équipe 'Phylogénie', UMR 7138 'Systématique, Adaptation, Évolution', Museum National d'Histoire Naturelle, Département Systématique et Évolution, case postale 26, 57 rue Cuvier, 75231 Paris cedex 05, France. E-mail: lecointr@mnhn.fr*
Blaise Li, Université Paris VI — Pierre et Marie Curie, UMR 7138, 43, rue Cuvier, Paris, France, 75005. E-mail address: blaise.li@normalesup.org

Introduction

With the increasing amount of molecular data available for phylogenetics comes an increasing hope for more extensive and well-resolved phylogenies. Indeed, the more diverse the sources of evidence, the better the expected quality of the results (Hempel 1965; Mahner & Bunge 1997), provided that the evidence is relevant to the problem under focus (Carnap 1950; Lecointre & Deleporte 2005). Quality is usually measured by some support values attached to the nodes of a phylogenetic tree. Support may be robustness (resistance to data perturbation), sensitivity (resistance to variations in the analysis method) or other kinds of measures such as decay indices or, in a supertree context, the measures proposed by Bininda-Edmonds (2003) or Cotton *et al*. (2006) (see also Wilkinson *et al*. 2003). The better the support values, the more the phylogeneticist will consider that the relationships are reliable. However, not all support measures are equally relevant to reliability assessment. The purpose of the present article is to propose a support value, the repetition index, which is designed to provide an appropriate measure of reliability for clades.

Materials and methods

Approach

In line with Carnap's degree of confirmation, reliability is the degree of credit we give to a statement at a given time, ideally taking into account all available data and knowledge relevant to this statement. In phylogenetics, the reliability issue cannot be addressed without considering how multiple data sets are handled: are all available data combined in a single matrix, or not? What are the criteria for considering a given clade reliable in each approach?

In the approach consisting in combining all the data, often called 'total evidence', one tends to trust the clades obtained inasmuch as they are based on the 'coherence' (see Rieppel 2004a,b; Kearney & Rieppel 2006) of all available characters. In the most common 'total evidence' practice, the reliability of a clade is implicitly (or even explicitly; Douady *et al*. 2003) associated with its degree of support, often measured with a Bremer support, a bootstrap proportion or a Bayesian posterior probability. Indeed, as all the available data have been gathered into a single matrix, reliability cannot be obtained otherwise.

No claim to original French government works. Journal compilation © 2008 The Norwegian Academy of Science and Letters • Zoologica Scripta, 2008

In the 'taxonomic congruence' approach, the naturalness of data partitions is justified by positive biological knowledge (Miyamoto & Fitch 1995). The biologist fully recognizes the background knowledge justifying why a given gene can be considered independent of another one.[1] After the separate analyses, the results do not obligatorily end up with a strict consensus tree: actually, there are many ways to summarize the results (Bryant 2003). The issue about how to assess the reliability of a clade in a taxonomic congruence approach has received rather poor explicit interest until recently. In a way, most consensus techniques are implicitly extracting those clades we might have good reasons to trust. However, the term 'reliability' has never been used for that, except in rare cases (Lockhart *et al*. 1995, p. 673; Bryant 2003, p. 5; Brinkmann *et al*. 2005; Lecointre & Deleporte 2005). To have access to reliability, one must take into account other criteria than the pure global 'coherence' among individual characters used in the 'total evidence' approach. Reliable results are results that are supported by congruence among multiple independent relevant sources of information (Rodrigo *et al*. 1993; Rieppel 2004a; Lecointre & Deleporte 2005; see also Grande 1994). One should for instance consider corroboration among trees produced by the analyses of genes that are hypothesized to evolve independently.[2] This approach has been explicitly used by Chen *et al*. (2003) and Dettaï & Lecointre (2004, 2005) but without full formalization. Others (Bininda-Edmonds 2003; Seo *et al*. 2005; Wilkinson *et al*. 2005; Cotton *et al*. 2006; Moore *et al*. 2006) have devised procedures that, under certain assumptions of independence of source trees, could include some reliability evaluation, but without explicitly using this word, using the more general term 'support' instead. We will now examine how the concept of reliability we described here could be formalized so as to be computerized into a repetition index for clades.

Taxonomic congruence from independent data sets

Among the scientific community, credit is given to phylogenetic hypotheses that have been obtained from independent data sets and teams. The more a clade is recovered by the analyses of independent data, the more it is reliable (for a similar idea developed in a supertree perspective, see Pisani & Wilkinson 2002, p. 154). Independent *elementary data sets* have thus to be delineated, the set of which defining what we call the *elementary partitioning scheme* of the available data.

Independence of the data sets is important because there are many reasons why the tree obtained by the analysis of a particular data set might not represent accurately the species interrelationships. Each data set analysis might yield a tree that somewhat differs from the species tree (Maddison 1997). However, the hope is that the trees do not all differ in the same way if they are built from independent data sets. By '*independent*' we mean 'unlikely to be subject to the same causes of incongruence with respect to the true species tree'. The decision whether two data sets can be kept separate or not is based on biological background knowledge, for instance, knowledge pertaining to the functions of the genes used as evolutionary markers, or about strong differences in evolutionary pressures (differences in free mutational space, composition bias, etc., suggesting that resulting tree reconstruction artefacts will not be the same). The analyses of two genes will unlikely yield similar results by pure chance. And since the genes are supposed to be independent, similar results should not be caused by the same artefact, but by the shared feature of the genes: the common ancestry of the taxa bearing them. For some molecular phylogenetic markers, little information may be available. In such cases, when there is *a priori* no reason to suspect that two markers are not independent, the practitioner might want to take the risk to suppose independence. The elementary data sets are the data sets that cannot be further split into independent data sets.

Once the independent data sets have been defined, they should be analysed separately with the appropriate method. Then, the number of occurrences of a clade among the obtained trees is a first indication of its reliability.[3] Starting from this basic indicator, the repetition index may now be refined as follows.

Improving reliability by considering partial data combinations

One of the criticisms against separate analyses is that partitioning data favours stochastic errors. Indeed, trees from smaller data sets are usually more sensitive to stochastic effects of homoplasy than trees from larger ones. As a result, some clades could fail to be repeated because of this 'size effect' (see how Pisani & Wilkinson 2002, p. 153, discuss weak and strong phylogenetic signal). The problem can be partly avoided by examining trees obtained by partial combinations of the elementary independent data sets (Dettaï & Lecointre 2004). If these elementary data sets are A, B and C, and are analysed separately, they constitute the *elementary partitioning scheme*. Their *partial combinations* are $A \cup B$, $A \cup C$ and $B \cup C$, each of which is a data set that can be analysed and compared with the results of the analyses of other independent data sets (C, B and A, respectively). Using these partial combinations, other *partitioning schemes* can be defined, where the

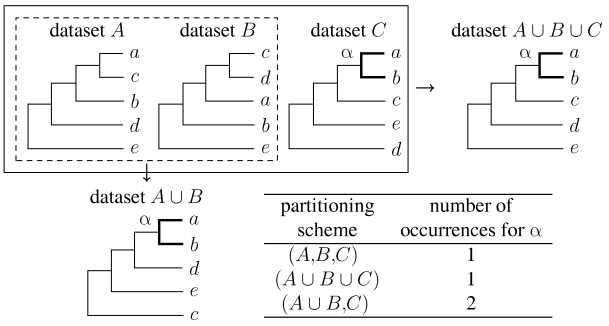
[1] *The use of biological knowledge is not restricted to taxonomic congruence approaches; knowledge in molecular evolution, for instance, is used in sophisticated model-based 'total evidence' practices.*

[2] *This implies the use of some external knowledge: the knowledge justifying the independence of the data partitions.*

[3] *Note that counting the occurrences of a clade is easy when all data sets have the same set of taxa, but otherwise not. This will be discussed later.*

elementary data sets are associated into sets of independent data sets. Here, those partitioning schemes would be ($A \cup B$,C) ($A \cup C$,B) and (A,$B \cup C$). Note that within a partitioning scheme, the results of the analyses can be compared to evaluate reliability because the constituting data sets are assembled so as to be independent: no elementary data set is present twice in a partitioning scheme. ($A \cup B \cup C$) is also a partitioning scheme to take into account, but since it has only one data set, it cannot provide numbers of occurrences higher than 1.

The elementary partitioning scheme, (A,B,C), is the one with the maximum number of independent data sets, but those data sets are the most prone to stochastic effects. The potential interest of the partial combination approach is to limit the stochastic effects usually impairing taxonomic congruence, as illustrated by the following example. Suppose that in the real species tree, there is a clade α that the analysis of data set C recovers, but that the analyses of data sets A and B separately fail to recover. Combining A and B might overcome the biases and stochastic effects that prevented α from being recovered in the separate analyses. In that particular case, using the ($A \cup B$,C) partitioning scheme provides 2 occurrences for clade α, whereas it occurred only once in the elementary partitioning scheme (see Fig. 1). For a given clade, the maximum number of independent occurrences will be obtained for a particular partitioning scheme. This scheme probably achieves the optimal way of combining the elementary data sets regarding the signal supporting the clade under focus. Indeed, a high number of occurrences is something improbable if no signal is supposed. The fact that the clade is repeated is better explained if one supposes that the combinations present in the partitioning scheme allowed common signal to emerge above noise. Therefore, the reliability of this clade should be derived from this partitioning scheme. It involves sufficient combination to overcome stochastic effects, but not too much; this allows to 'test' the clade across independent trees. Indeed, in too big a combination, there are fewer possibilities of independent occurrences, and there is even a risk for the clade to be lost because of a strong bias in one of the elementary data sets. The partial combination approach can be seen as a means of extracting more information from the data when some of the elementary data sets have weak phylogenetic signals. However, this is a computationally intensive procedure when the elementary data sets are numerous (with n elementary data sets, there are 2^n–1 analyses to do, including the total combination). In such a case, the method described in the rest of this article could also be applied using only the elementary partitioning scheme, hoping that a majority of the data sets express their historical signal.



partitioning scheme	number of occurrences for α
(A,B,C)	1
($A \cup B \cup C$)	1
($A \cup B$,C)	2

Fig. 1 Simplified example illustrating the potential usefulness of the partial combination approach. Considering only the elementary partitioning scheme (A,B,C), clade α is only recovered by one data set. With the partial combination approach, all possible partitioning schemes are explored (though we show only three of them here). Among them, there is one in which two independent data sets recover clade α. The combination of data sets A and B has overcome stochastic effects that prevented clade α from being recovered when these data sets were kept separate. The repetition index for α will thus be based on partitioning scheme ($A \cup B$,C).

To summarize, a provisional repetition index can be computed the following way:

1 separate the data into independent elementary data sets;

2 analyse every elementary data set and every possible partial combination of them;

3 for each partitioning scheme (i.e., for each set of independent data sets), count the number of occurrences of each clade that appeared in at least one of the analyses (this number cannot be higher than the number of data sets in the considered partitioning scheme, thus, its maximal possible value is the number of elementary data sets);

4 for each clade recorded in step 3, retain as repetition index the best number of occurrences obtained among all possible partitioning schemes (the partitioning scheme providing the highest repetition for a clade may be different from the one providing the highest repetition for another one).

This provisional repetition index can be expressed by the following formula:

$R(\alpha) = \max_{D \in PSc}(\Sigma_{d \in D}\ \delta_{\alpha,d})$, where PSc is the set of the partitioning schemes, D is the set of the data sets constituting a partitioning scheme in PSc, d is a data set in D, and δ is 1 if α is produced by the analysis of the data set d, 0 otherwise.

It is basically a number of occurrences of a clade in a set of independent analyses, hence the sum over data sets. The more independent data sets there are, the highest the reliability may be. This justifies the use of a sum. What is to be summed however, is subject to discussion: bootstrap proportions, percentages in majority-rule consensuses of equally optimal trees, Bayesian posterior probabilities, raw all-or-nothing occurrences? All these possibilities lead to a repetition index, which has the dimension of a number of occurrences. Here, we simply use occurrences (the 1 or 0 represented by δ), but the methodology presented here could be equally applied using the other options.

No claim to original French government works. Journal compilation © 2008 The Norwegian Academy of Science and Letters

This provisional repetition index can be computed for bipartitions in unrooted trees instead of clades. It can also be used in cases where some taxa are missing from some of the elementary data sets. However, in that case, the values of the repetition indices for the clades containing taxa missing from some data sets could be lower because of the lack of taxonomic overlap; some data sets would be unable to produce those clades. For the sake of simplicity, we will now suppose that there is a total taxonomic overlap between the elementary data sets.

Dealing with contradiction among clades

For some reason (a mistake in the delineation of the elementary data sets because of a lack of biological background knowledge, for instance) two clades can be incompatible but both repeated. In such a case, at least one of the two clades certainly does not reflect the history of the taxa (we neglect here reticulate evolution); it should not be considered reliable. Without any other assumption, one cannot tell which one of the two clades is not 'correct' — and perhaps both are incorrect. Therefore, the reliabilities of both clades should be decreased. We suggest decreasing the value of the repetition index of a clade by subtracting the repetition index of its contradictor. This would lead to an index that is basically a difference between numbers of occurrences. Actually, a clade is likely to have multiple contradictors among all clades occurring at least once through the analyses of all elementary data sets and their partial combinations. It seems meaningful to consider only the most reliable of them, that is, the one with the highest repetition index. Note that this requires that the repetition indices of all the clades contradicting the clade under focus are known, which necessitates successive approximations because the indices of the contradictors depend on the indices of their own contradictors. This can be formalized as follows.

The formerly defined repetition index, $R(\alpha)$, will be called the *first order repetition index* for clade α and noted $R_1(\alpha)$.

A clade β is said to be contradicting α if the three following conditions are true (see Berry & Gascuel 2000, p. 275):

1 $\alpha \cup \beta \neq \alpha$ (β contains at least one taxon that α has not);

2 $\alpha \cup \beta \neq \beta$ (β does not contain α);

3 $\alpha \cap \beta \neq \varnothing$ (β contains at least one taxon that α has).

Contradiction is a reciprocal relationship. It could be reformulated this way: α and β contradict one another if and only if they have at least one shared taxon and at least each a specific taxon. Contradiction is incompatibility. Clades that contradict one another are clades that are not compatible and reciprocally; they cannot be both in the same tree. Such a relationship is straightforward when the clades come from trees having the same leaves, but it is problematic when there are missing taxa (see Bininda-Edmonds 2003). We don't know where the missing taxa would be placed if they were present, and one could imagine cases where the addition of taxa to clades can make them switch from contradiction to compatibility or vice versa [see Wilkinson *et al*. (2005) for a description of the different possible situations]. This explains why the present article is restricted to cases with fully overlapping taxonomic samplings.

Assuming that α has some contradictors, let β_1 be its *'best' contradictor* according to the R_1 index (the one with the highest R_1). A second order repetition index for α can now be defined:

$$R_2(\alpha) = R_1(\alpha) - R_1(\beta_1)$$

However, according to R_2, β_1 might not be the best contradictor of α any more because it is also contradicted. It is possible that there is another contradictor of α, β_2, that is not so much contradicted as β_1 is, so that $R_2(\beta_2) > R_2(\beta_1)$. Thus, the calculation of the repetition index for α should be reconsidered by defining a third order repetition index:

$R_3(\alpha) = R_1(\alpha) - R_1(\beta_2)$, where β_2 is the best contradictor of α according to R_2. If there is more than one best contradictor according to R_2, the highest R_1 value found among these contradictors is used to calculate R_3.

This can be repeated, calculating a fourth order repetition index considering R_3 to find the next best contradictor, and so on. At each step of the process, a contradiction network in which each clade has a provisional best contradictor is implicitly established. In the most simple case, each clade will have a stable unique best contradictor, allowing the calculation of a *final repetition index* $R_f(\alpha) = R_1(\alpha) - R_1(\beta_f)$, where β_f is the best contradictor of α according to R_f (see Table 1).

In the other cases, the process of calculating the next order repetition indices will be periodic: as the number of clades is finite, the contradiction network between them has a finite number of possible configurations, so if none of these networks is stable, the state of the system will change until it comes back to a state already reached before. In such cases, since we cannot tell for each clade which contradictor is the best, we consider that each configuration of the contradiction network in a period ought to be taken into account with the same weight in the determination of the final repetition index. The mean repetition index, $\bar{R}$, over a period will thus be taken as final repetition index. This amounts to decreasing $R_1(\alpha)$ by the mean value of R_1 over the successive best contradictors of α.

Application to acanthomorph phylogeny

Data sets. Five independent elementary data sets with a common taxonomic sampling of 73 taxa have been gathered as a case study (Dettaï 2004). The data sets are the following molecular markers:

1 a mitochondrial data set comprising partial 12S and 16S rDNA for a total length of 828 base pairs (bp). They are kept together because both are elements of the mitochondrial ribosome, thus potentially subject to common evolutionary

Table 1 Example illustrating the computation of the repetition index taking into account contradiction among clades. The taxa involved in this example are designed by the letters a through *j*. It is assumed that there are only five clades and that their first order repetition indices (R_1) have already been calculated. The contradictors of α are β_1 and β_2, the best one being β_1 ($R_1(\beta_1) = 4$). The contradictors of β_1 are α and γ_1. The contradictors of β_2 are α and γ_{12}. After calculating the second order repetition indices (see the procedure in the text), the best contradictor for α has changed: it is now β_2 ($R_2(\beta_2) = 0$). After calculating the third order repetition index, no best contradictor has changed. This index can thus be taken as the final repetition index.

Clades	$\alpha = (a,b,c,d)$	$\beta_1 = (a,b,e)$	$\beta_2 = (c,d,f)$	$\gamma_1 = (e,g,h)$	$\gamma_2 = (f,i,j)$
R_1	3	**4**	3	5	3
Best contradictor	$\boldsymbol{\beta_1}$	γ_1	γ_2	β_1	β_2
R_2	3 – 4 = –1	4 – 5 = –1	3 – 3 = **0**	5 – 4 = 1	3 – 3 = 0
Best contradictor	$\boldsymbol{\beta_2}$	γ_1	γ_2	β_1	β_2
R_3	3 – 3 = 0	4 – 5 = –1	3 – 3 = 0	5 – 4 = 1	3 – 3 = 0
Best contradictor	β_2	γ_1	γ_2	β_1	β_2

constraints, and physically linked, being both on the mitochondrial chromosome;
2 partial sequences of 28S rDNA (C1-C2, D3, D6 and D12 domains). The concatenated length is 801 bp;
3 partial Rhodopsin gene (759 bp);
4 partial Mixed Lineage Leukaemia-like exon 26 (MLL, 552 bp);
5 partial Interphotoreceptor Retinoid Binding Protein module 1 (IRBP, 713 bp).

Bathypterois (Chlorophtalmoidei) was the only outgroup taxon for which we could gather the sequences for the five elementary data sets.

Tree construction method. The alignment was done by hand with SeAl (Rambaut 2002) v2.0a11 carbon. Ambiguous zones in the rDNA data sets were removed. The alignment was submitted to TreeBase (study accession number: S2152, matrix accession number: M4084).

There are 31 possible combinations of the five elementary data sets (including the 'total evidence' combination).

They were successively analysed under maximum parsimony using PAUP* (Swofford 2002) version 4.0b10 for Macintosh (PPC), each with 1000 RAS + TBR rounds, using a nexus batch file (see the nexus file on TreeBase).

The 50% majority-rule consensus trees were used to count the occurrences of the clades. The full combination was bootstrapped (1000 pseudosamples each submitted to 50 replications of RAS + TBR, 'multrees' option turned off) to compare robustness and reliability as defined here. Further data processing was done on a GNU/Linux system with the help of shell scripts and Python 2.3.4 (http://www.python.org/) scripts.

Results

Three clades occurred five times, and obtained the maximal repetition index ($R_f = 5$). These are the clades that occur for each elementary data set. Six clades occurred four times and were only slightly contradicted ($R_f = 3$). 10 other clades reached a repetition index of 2, and a total of 35 clades had a repetition index equal to or higher than 1. All these clades could be provisionally considered reliable (until new data is available) because they occur at least once more than their 'best' contradictor.

The 'total evidence' tree is presented in Fig. 3. The majority-rule consensus of the trees resulting from the separate analyses of the 5 elementary data sets stands for the result of a typical taxonomic congruence study (see Fig. 4). The raw numbers of occurrences of the clades, their repetition indices and their bootstrap supports from the full data combination are written on the trees.

The repetition indices of the clades were plotted against their bootstrap supports. Both the Spearman and Kendall rank correlation coefficients between repetition indices and total evidence bootstrap supports were –0.33.

To synthesize the results concerning acanthomorphs relationships, several methods are possible to build summary trees based on clades reliabilities. One could build a clade-taxon matrix, where each clade recorded from the phylogenetic analyses of the elementary data sets and their partial combinations is weighted according to its repetition index, and each taxon is coded 1 when present in the clade and 0 when not. This matrix could then be analysed under maximum parsimony or compatibility, leading to trees akin to MRP or MRC supertrees. We propose another summary tree, explicitly devised to include reliable clades. It is akin to the greedy consensus method (see Bryant 2003 and Bandelt & Dress 1992, p. 244):
1 group clades having the same repetition index, arrange these groups in descending order of repetition index. Within each group, group clades according to the maximum number of occurrences and order these groups according to this criterion;
2 for each group of clades having equal repetition index and equal maximum number of occurrences, beginning with the

No claim to original French government works. Journal compilation © 2008 The Norwegian Academy of Science and Letters

'best' one (the most reliable), eliminate the clades within the group that are not compatible with the clades already retained. Then, retain the remaining clades of the group if they are mutually compatible and repeat this step with the next group. If the remaining clades in a group are not mutually compatible, they are all discarded. This process should not discard clades with high repetition indices because a clade having contradictors with the same repetition index and same number of occurrences has its best contradictor being at least as reliable as it is. This ensures that it does not have a high repetition index (proof in the case where all clades have a stable best contradictor available as supplementary material); **3** assemble the 'greedy summary tree' by combining the clades that have been retained.

The resulting tree is presented in Fig. 2. In this synthesis tree, one clade was not present in the tree obtained from the 'total evidence' combination, and two clades were absent from all five separate analyses. The elementary partitioning scheme provided the highest number of occurrences for 18 out of the 35 clades present in this tree. The other partitioning schemes contributing the most to the greedy summary tree are ones implying three elementary data sets and one partial combination involving 28S; with 12S and 16S (highest number of occurrences for 12 clades), with IRBP (11 clades) and with MLL (11 clades).

Discussion

A few properties of the repetition index

The partial combination approach leads us to having different sets of independent data sets over which to sum occurrences: the partitioning schemes. The repetition index is based on occurrences within a partitioning scheme. The maximum value potentially can be achieved with the elementary partitioning scheme because it is the one with the highest number of independent data sets. Thus, the maximum value of the repetition index is the number of elementary data sets. For instance, in the present study, the maximum repetition index cannot be 31, but five, as not all possible combinations can be in the same partitioning scheme (MLL + IRBP is not independent from MLL + Rhodopsin). The 'best' clades are those that are recovered by the analyses of every elementary data set.

The minimum possible value of the repetition index is the opposite of the maximum value. That would be the case for a clade that never occurs, and that is contradicted by one of the best clades.

The more elementary data sets are combined within a partitioning scheme, the better the chances to overcome stochastic errors, but the less there are independent data sets in the partitioning scheme. So if a clade fails to appear in the analyses of each elementary data set, the partial combination approach has no chance to give it the maximum value. However, this maximum can increase by the addition of new independent data sets. This raises the question of how to compare that index among different studies. One could think that it is necessary to divide the index by the number of elementary data sets included in the study to allow comparison among studies dealing with different numbers of data sets. The maximum possible value of the repetition index would then be 1, whatever the quantity of data and trees at hand. This seems not appropriate because adding more data should allow an improvement of the maximum reliability. The repetition index we propose reflects the quantity of trees supporting a clade, which is a relevant information. Actually, the indices from different studies can be compared without rescaling. Suppose we have made a study comprising three data sets. If we obtain a particular clade with 2 of the data sets and a contradictor with the other data set, the repetition index of that clade will be 1 (2 occurrences minus 1 contradiction). The reliability of this clade is low, and it cannot be higher as long as we do not analyse new data. But meanwhile, it is still more reliable than any of its contradictors. Now, suppose we add 10 new data sets to the study. In case most new data recover the clade, its reliability should increase, which will be reflected by an increase in its repetition index. However, if five of the new data sets support the clade and the five other support its contradictor, the reliability should not be improved. This will be reflected by the fact that the repetition index of the clade would still be 1 (7 occurrences minus 6 contradictions). The reliability will not be decreased either. A repetition index of 1 indicates the same level of reliability (rather low, but positive nonetheless) whatever the number of data sets used in the study. What changes is that with such a persistent low reliability, we may now hypothesize that there is some conflict between two true historical signals; this could be a sign of reticulate evolution. Monitoring the evolution of repetition indices when the number of data sets grows could be done following a procedure similar to the one devised by Struck *et al*. (2006) in a 'total evidence' context.

Robustness and reliability

The results show that, assuming that reliability and robustness can be assessed by our repetition index and by bootstrap proportions from the full combination, respectively, those two pieces of information about the results are poorly correlated. Indeed, if we consider reliable clades that have a repetition index equal or higher than 1 and robust those with a 70% or higher bootstrap support (value chosen according to Hillis & Bull 1993 and Lecointre *et al*. 1994), 30 reliable clades are not robust and 1 robust clade is not reliable. Thus, although it may not always be easy to justify the independence of the data sets, we are inclined to think that using only a total evidence approach is not suitable, because it dismisses interesting information; it does not allow us to determine which clades

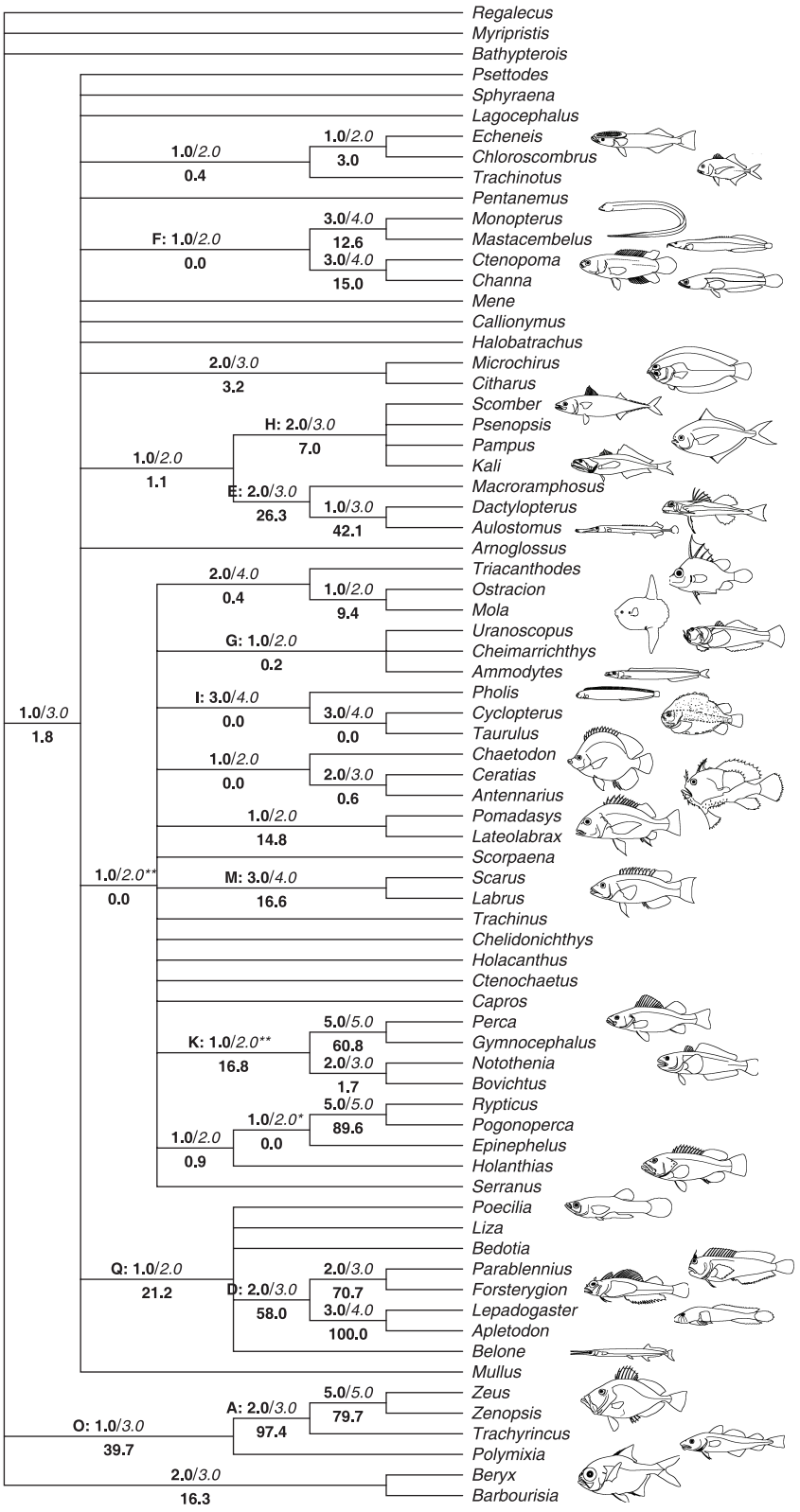


Fig. 2 Tree summarizing the most repeated clades with the greedy summary method described in the text. This tree is composed of the 'best' intercompatible clades according to the repetition index. The bootstrap supports in the 'total evidence' analysis are in bold, below the branches. The repetition indices (in bold) and the maximum number of occurrences of the clades are above the branches. *: the simultaneous analysis of all data sets did not recover this clade. **: these clades could not be recovered without combining at least two elementary data sets. Some reliable clades are lettered according to Dettaï & Lecointre (2005).

No claim to original French government works. Journal compilation © 2008 The Norwegian Academy of Science and Letters

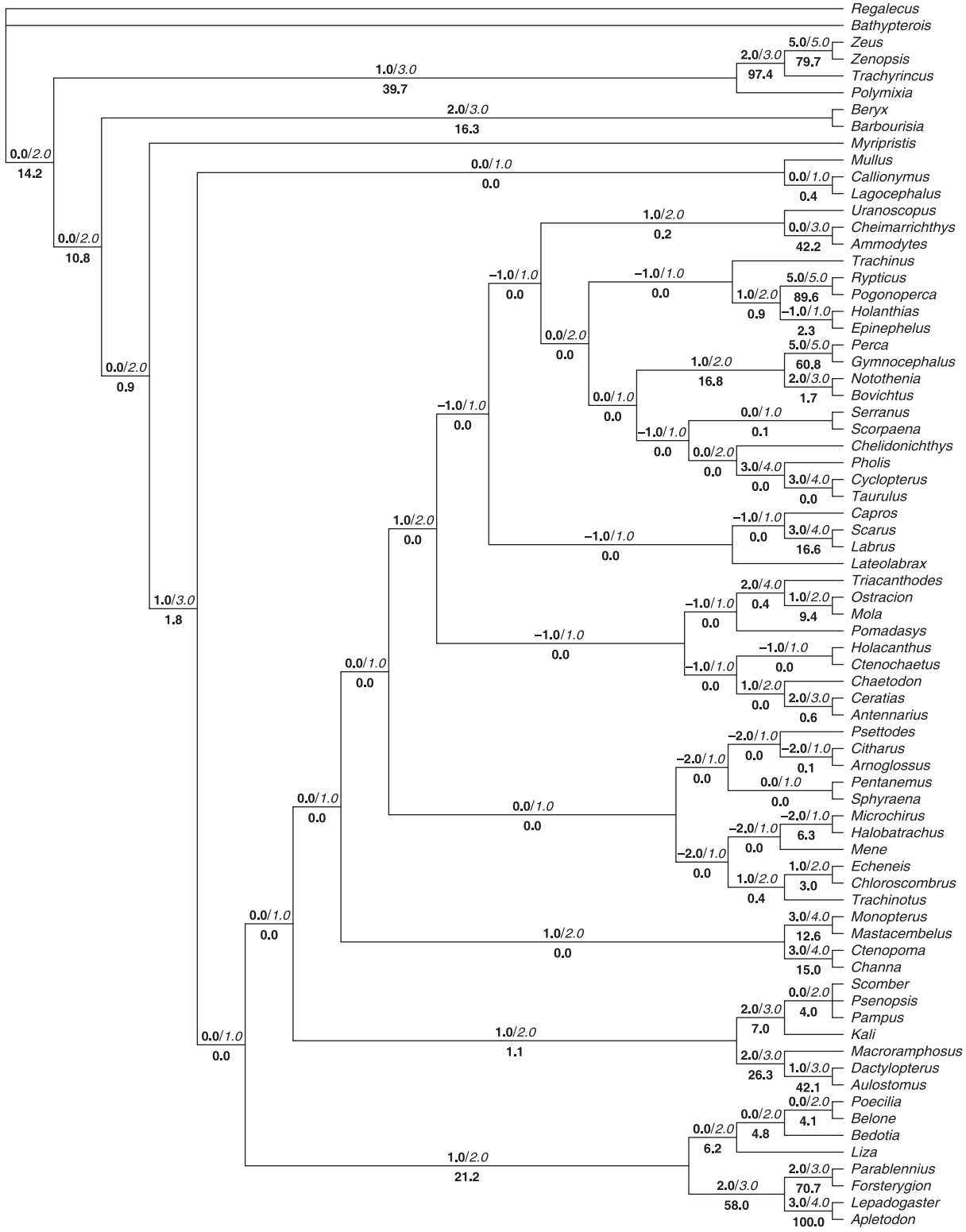


Fig. 3 50% majority-rule consensus of the equally most parsimonious trees obtained by the analysis of the combination of all five elementary data sets. The bootstrap supports are in bold, below the branches. Above the branches are the repetition indices (in bold) and the maximum number of occurrences of the clades.

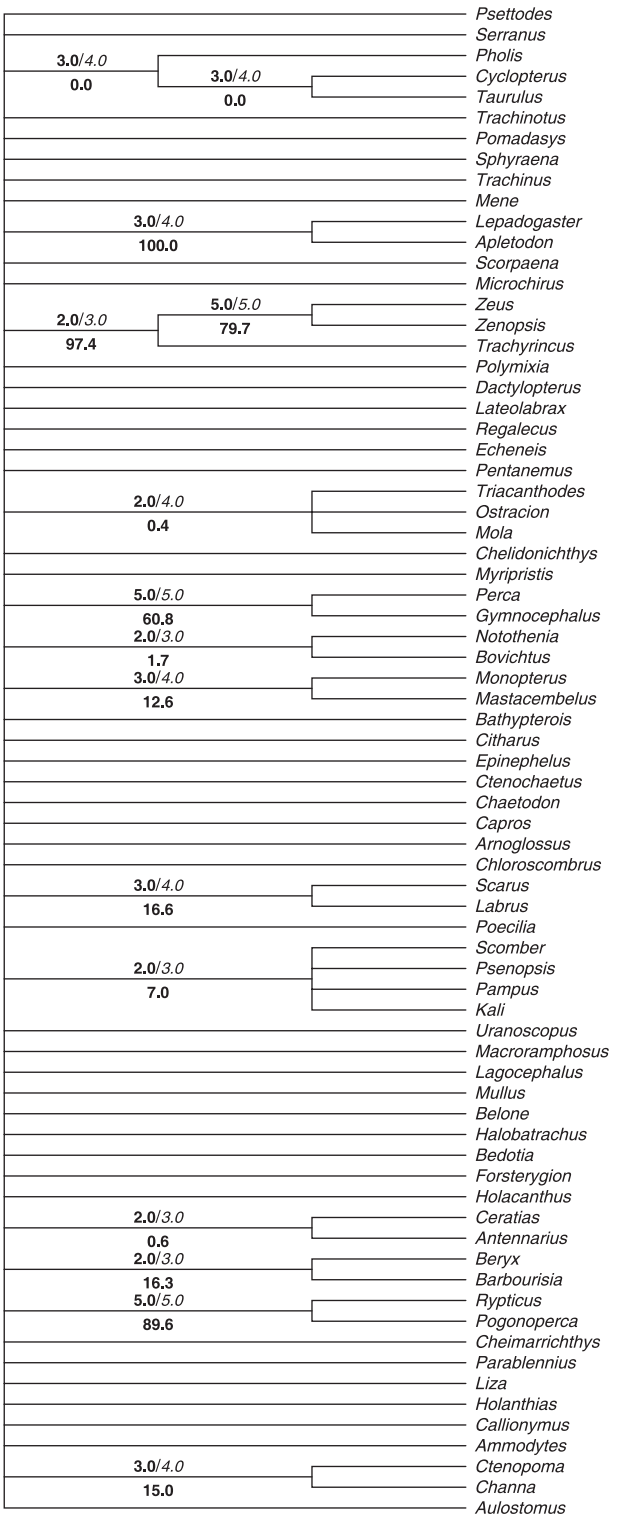


Fig. 4 Majority-rule consensus of the consensus trees obtained by the analyses of the five elementary data sets. The bootstrap supports in the 'total evidence' analysis are in bold, below the branches. The repetition indices (in bold) and the maximum number of occurrences of the clades are above the branches.

may be trusted to represent patterns resulting from species history and which may not. Sometimes, robust clades may be misleading.

Many authors, by using bootstrap proportions obtained on the full data combination, exploit the advantage of positive effects of sequence length on bootstrap proportions and may thus be inclined to trust more clades than they should. Indeed, strong bootstrap proportions are not necessarily linked to species common ancestry, even in the tree based on all available data. For instance, if the genes history is not the same as the species history (horizontal transfer, paralogy ...), or when compositional biases or long-branch attraction artefacts in a single data set are strong enough to impose the wrong topology to the tree based on the full combination (Chen *et al*. 2003; Phillips *et al*. 2004; Brinkmann *et al*. 2005). In that sense, the present method is more conservative than the 'total evidence' approach. Only the clades more repeated than their contradictors are considered reliable.

Relation to previous works

It should be noted that in Dettaï & Lecointre (2004) there was no explicit notion of contradiction among clades as in the present article. Instead, these authors used a concept of 'intruders' and 'escapees'. When a clade is repeated in several data sets, another data set could exhibit the 'same' clade minus one taxon (an 'escapee'), or plus one taxon (an 'intruder'). This allowed them to take into account clades that were 'almost the same' as a repeated clade. However, this approach seemed difficult to formalize in a way that could be implemented into a computer program. A way to circumvent this intruder and escapee issue would be to consider reduced components (also called n-taxon statements by Wilkinson (1994, 1996), or partial splits). In our present approach, clades that are almost the same will be detected by the fact that they are not strongly contradicted, and not by the fact that they are 'almost repeated' (i.e., having 'intruders' or 'escapees') as in Dettaï & Lecointre (2004). When a taxon escapes from a reference clade, the result is a clade that is compatible with the reference clade, but the escapee will be part of a clade that contradicts it. If the same taxon escapes several times from the reference clade and participates in the same contradictory clade [a repeated position in Dettaï & Lecointre (2005)], there will be a 'good' contradictor for the reference clade. Otherwise, the clade will not be really contradicted, it will only be less repeated.

Bininda-Edmonds (2003) and Wilkinson *et al*. (2005) have devised support measures for clades in supertrees that include developed considerations about compatibility and contradiction. Their measures are based on support of supertree relationships from source tree relationships. These measures can thus be interpreted as reliability measures when the source trees of the supertree analysis are based on independent data sets

No claim to original French government works. Journal compilation © 2008 The Norwegian Academy of Science and Letters

(which is often the case since one usually aims for accurate supertrees). They present the interest of being applicable even when some taxa are missing. Their works, however, are focused on the clades already present in the supertree under study whereas in the present article, reliability assessment precedes summary tree building. Another idea that produces some sort of numerical reliability assessment is the bootsrap-derived procedures used by Seo *et al.* (2005) or Moore *et al.* (2006). Their approaches consist in resampling genes or source trees, respectively. This can provide an implicit reliability aspect to bootstrap proportions if the genes or source trees from which the resampling is made can be hypothesized to be independent from one another.

How to synthesize the results?

Synthesizing — in the taxonomic congruence framework — the results concerning the reliability of clades into a tree can be seen as a form of consensus construction. Simply using the majority rule consensus of the results of the separate analyses is, however, too conservative (see how Fig. 4 is poorly resolved) and does not take into account the information added by the partial combination approach. The tree from the 'total evidence' analysis (see Fig. 3) seems to be more useful in our present test case; most reliable clades are recovered. However, the principles underlying its construction do not guarantee this: a strong bias from one data set could mislead the whole reconstruction and prevent a reliable clade from being present. That is one of the reasons for using consensus methods based on the repetition indices. Instead of simply mapping these indices on trees obtained from usual methods, we propose the greedy summary tree method described earlier, because it is designed for selecting the most reliable clades (see Bandelt & Dress 1992, p. 244 for a similar approach, but based on another support value). In addition, two methods based on matrix representation were used, to compare with that greedy summary tree. These methods weight the bipartitions in the matrix according to their repetition indices. One is derived from matrix representation with parsimony (MRP, Baum & Ragan 2004) and the other from matrix representation with compatibility (MRC, Rodrigo 1996; Ross & Rodrigo 2004). The second, also akin to an asymmetric median tree (AMT, Phillips & Warnow 1996) was unfortunately too time-consuming, using the clique program from Felsenstein (2004), to be applied successfully to our 73 taxa data set. Note that for that reason, Bryant (2003, p. 6) recommends to use the greedy consensus method. We resolved the problem the same way, using the greedy summary tree (Fig. 2). The difference between the greedy summary tree and the classical greedy consensus is that the former was constructed from the reliability indices of the clades, not their raw number of occurrences. The greedy summary tree was the same tree as the MRC-derived tree on a test with 16 taxa and six elementary data sets but was dramatically faster to construct. Our MRP-derived method differs from standard MRP mainly by the fact that contradiction among clades is taken into account before writing the clade-taxon matrix. A more MRP-like approach would have consisted in taking every clade occurrence as a column in the matrix and letting the parsimony analysis manage the contradictions. Here, we weight the clades according to their final repetition indices. This also makes the difference between our MRC-derived approach and standard MRC or AMT. The implications of such a difference remain to be studied.

Acanthomorph phylogeny

Some groups identified with letters by Dettaï & Lecointre (2005) are found here (see Fig. 2). The monophyly of the group 'A', comprising gadiforms (cods) and zeioids (dories) is recovered as reliable (R_f = 2.0), also confirming the findings of Chen *et al.* (2000, 2003) and Miya *et al.* (2003). The monophyly of the group 'O', uniting *Polymixia* (beardfish) to the previous groups, also found in the same previous studies, is considered here as reliable (R_f = 1.0). Clade 'F' is found again (R_f = 1.0), uniting channids (snakeheads), anabantoids (climbing gouramies), mastacembeloids (swamp eels) and synbranchiforms (spiny eels). Clade 'E' (R_f = 2.0) contains part of the syngnathiforms (pipefishes and horsefishes) and dactylopteriforms (flying gurnards). Clade 'H' (R_f = 2.0) groups parts of the trachinoids (*Kali*) and parts of the Scombroidei (here, the mackerel), with Stromateoidei (butterfishes). Clades 'E' and 'H' are sister-taxa (R_f = 1.0). Clade 'M' (R_f = 3.0) shows a sister-group relationship between labrids (wrasses) and scarids (parrotfishes). Dettaï & Lecointre (2005) showed that this component of the ex-labroids actually was not related to other labroids like cichlids. Clade 'K' (R_f = 1.0) is showing a sister-group relationship between Antarctic fishes (the Notothenioidei) and percids (perches). Clade 'I' (R_f = 3.0) groups cottoids (sculpins) with zoarcoids (eelpouts). Clade 'G' (R_f = 1.0) groups components of the Trachinoidei (stargazer, sandlances) and Cheimarrichthydae (torrentfishes). Clade 'Q' (R_f = 1.0) groups atherinomorphs (guppies), mugiloids (mullets), blennioids (blennies) and gobiesociforms (clingfishes), the latter two forming clade 'D' (R_f = 2.0). Some of those groups appeal the polyphyly of traditional taxa, most of them poorly defined (Perciformes, Scorpaeniformes, Trachinoidei, Labroidei, Paracanthopterygii).

Some clades from Dettaï & Lecointre (2005) are not recovered. Clade 'N', grouping lophiiforms (anglerfishes), tetraodontiforms (pufferfishes), chaetodontids (butterfly-fishes) and *Capros*, is not recovered, because one of the tetraodontiforms, *Lagocephalus*, is subject to recurrent long branch attraction (LBA). The same can be said of clade 'L' (carangids, menids, flatfishes, echeneids, centropomids, sphyraenids, polynemids), from which *Arnoglossus* — a bothid flatfish — 'escapes' because of its high mutation rate in several genes.

Recurrent LBA tends to influence the summary tree: in Fig. 2, *Arnoglossus*, *Callionymus*, *Mullus* and *Lagocephalus*, four taxa showing recurrent long branch attraction, are placed in an unresolved position within a large clade. Other clades cannot be recovered here probably because the present data set is reduced compared with that of the study of Dettaï & Lecointre (2005). Only the genera present in all data sets have been taken into account here because — for the moment — the method does not deal with missing taxa.

On the other hand, some clades not previously labelled with letters are reliable here, for example, the node splitting 'basal' acanthomorphs (*Regalecus*, clade 'O' and beryciforms) from the rest of the sampling (R_f = 1.0). Nodes of medium depth exhibit poor values (see Fig. 2).

Conclusion

The present index confirms many of the new acanthomorph clades repeatedly found by the recent molecular phylogenies. Work still needs to be done to extend the methodology to cases without perfect taxonomical overlap between data sets. It should also be noted that the repetition index can be misled by recurrent long branch attraction. This can probably be softened by using more elaborate reconstruction methods as the simple maximum parsimony that was used for the present article.

Acknowledgements

Thanks to Olaf Bininda-Edmonds, Nathanael Cao, Dan Faith, Jean-François Flot, François-Joseph Lapointe, Jérôme Murienne, Davide Pisani and Mark Wilkinson for excellent comments. Thanks to Mahendra Mariadassou for help with mathematics. Thanks to Wei-Jen Chen and Agnès Dettaï for most of the data used in this study.

B.L. was supported by a PhD fellowship 'Allocation couplée' (Ministère de l'Éducation Nationale, de la Recherche et de la Technologie).

The figures were produced using TikZ (http://sourceforge.net/project/pgf/) and TreeGraph (http://www.math.uni-bonn.de/people/jmueller/extra/treegraph/, Müller & Müller 2004). The pictures on Fig. 2 are drawn from Fishbase (Froese & Pauly 2006).

References

Bandelt, H.-J. & Dress, A. (1992). Split decomposition: a new useful approach to phylogenetic analysis of distance data. *Molecular Phylogenetics and Evolution*, *3*, 242–252.

Baum, B. & Ragan, M. (2004). The MRP method. In O. Bininda-Edmonds (Ed.) *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life* (pp. 17–34). Dordrecht, the Netherlands: Kluver Academic.

Berry, V. & Gascuel, O. (2000). Inferring evolutionary trees with strong combinatorial evidence. *Theoretical Computer Science*, *240*, 271–298.

Bininda-Edmonds, O. (2003). Novel versus unsupported clades: assessing the qualitative support for clades in MRP supertrees. *Systematic Biology*, *52*(6), 839–848.

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. *Systematic Biology*, *54*(5), 743–757.

Bryant, D. (2003). A classification of consensus methods for phylogenetics. In M. Janowitz, F. Lapointe, F. McMorris, B. Mirkin & F. Roberts (Eds) *Bioconsensus* (pp. 1–21). Piscataway, NJ: American Mathematical Society Publications-DIMACS.

Carnap, R. (1950). *Logical Foundations of Probability*. Chicago: University of Chicago press.

Chen, W.-J., Bonillo, C. & Lecointre, G. (2000). Taxonomic congruence as a tool to discover new clades in the acanthomorph (Teleostei) radiation. In *Program Book and Abstracts, 80th Annual Meeting ASIH, La Paz, México, June 14–20, 2000* (p. 369). American Society of Ichthyologists and Herpetologists, American Society of Ichthyologists and Herpetologists.

Chen, W.-J., Bonillo, C. & Lecointre, G. (2003). Repeatability of clades as a criterion of reliability: a case study for molecular phylogeny of Acanthomorpha (Teleostei) with larger number of taxa. *Molecular Phylogenetics and Evolution*, *26*, 262–288.

Cotton, J., Slater, C. & Wilkinson, M. (2006). Discriminating supported and unsupported relationships in supertrees using triplets. *Systematic Biology*, *55*(2), 345–350.

Dettaï, A. (2004). *La phylogénie des Acanthomorpha (Teleostei) inférée par l'étude de la congruence taxinomique*. PhD Thesis. Université Paris VI Pierre et Marie Curie.

Dettaï, A. & Lecointre, G. (2004). In search of nothothenioid (Teleostei) relatives. *Antarctic Science*, *16*(1), 71–85.

Dettaï, A. & Lecointre, G. (2005). Further support for the clades obtained by multiple molecular phylogenies in the acanthomorph bush. *Comptes Rendus Biologies*, *328*, 674–689.

Douady, C., Delsuc, F., Boucher, Y., Doolittle, F. & Douzery, E. (2003). Comparison of bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. *Molecular Biology and Evolution*, *20*(2), 248–254.

Felsenstein, J. (2004). *PHYLIP. Phylogeny Inference Package*, Version 3.6. Seattle: Department of Genome Sciences and Departement of Biology, University of Washington.

Froese, R. & Pauly, D. (2006). Fishbase. World Wide Web electronic publication. Available Via http://www.fishbase.org.

Grande, L. (1994). Repeating patterns in nature, predictability, and 'impact' in science. In L. Grande & O. Rieppel (Eds) *Interpreting the Hierarchy of Nature*, 1st edn (pp. 61–84). New York: Academic Press.

Hempel, C. (1965). *Aspects of Scientific Explanation, and Other Essays in the Philosophy of Science*. New York: Free press.

Hillis, D. & Bull, J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology*, *42*(2), 182–192.

Kearney, M. & Rieppel, O. (2006). Rejecting 'the given' in systematics. *Cladistics*, *22*, 369–377.

Lecointre, G. & Deleporte, P. (2005). Total evidence requires exclusion of phylogenetically misleading data. *Zoologica Scripta*, *34*(1), 101–117.

Lecointre, G., Philippe, H., Lê, H. L. V. & Le Guyader, H. (1994). How many nucleotides are required to resolve a phylogenetic

No claim to original French government works. Journal compilation © 2008 The Norwegian Academy of Science and Letters • Zoologica Scripta, 2008

problem? The use of a new statistical method applicable to available sequences. *Molecular Phylogenetics and Evolution*, *3*(4), 292–309.

Lockhart, P., Penny, D. & Meyer, A. (1995). Testing the phylogeny of swordtail fishes using split decomposition and spectral analysis. *Journal of Molecular Evolution*, *41*, 666–674.

Maddison, W. (1997). Gene trees in species trees. *Systematic Biology*, *46*(3), 523–536.

Mahner, M. & Bunge, M. (1997). *Foundations of Biophilosophy*. Berlin: Springer.

Miya, M., Takeshima, H., Endo, H., Ishiguro, N., Inoue, J., Mukai, T., Satoh, T., Yamaguchi, M., Kawaguchi, A., Mabuchi, K., Shirai, S. & Nishida, M. (2003). Major patterns of higher teleostean phylogenies: a new perspective based on 100 complete mitochondrial DNA sequences. *Molecular Phylogenetics and Evolution*, *26*, 121–138.

Miyamoto, M. & Fitch, W. (1995). Testing species phylogenies and phylogenetic methods with congruence. *Systematic Biology*, *44*(1), 64–75.

Moore, B., Smith, S. & Donoghue, M. (2006). Increasing data transparency and estimating phylogenetic uncertainty in supertrees: approaches using nonparametric bootstrapping. *Systematic Biology*, *55*(4), 662–676.

Müller, J. & Müller, K. (2004). TreeGraph: automated drawing of complex tree figures using an extensible tree description format. *Molecular Ecology Notes*, *4*, 786–788.

Phillips, M., Delsuc, F. & Penny, D. (2004). Genome-scale phylogeny and the detection of systematic biases. *Molecular Biology and Evolution*, *21*(7), 1455–1458.

Phillips, C. & Warnow, T. (1996). The asymetric median tree: a new model for building consensus trees. *Discrete Applied Mathematics*, *71*, 311–335.

Pisani, D. & Wilkinson, M. (2002). Matrix representation with parsimony, taxonomic congruence, and total evidence. *Systematic Biology*, *51*(1), 151–155.

Rambaut, A. (2002). *Se-Al, Sequence Alignment Editor*, Version, 2.0a11. Oxford, OX1 3PS, UK: Department of Zoology, University of Oxford, South Parks Road.

Rieppel, O. (2004a). The language of systematics, and the philosophy of 'total evidence'. *Systematics and Biodiversity*, *2*(1), 9–19.

Rieppel, O. (2004b). What happens when the language of science threatens to break down in systematics: a popperian perspective. In D. Williams & P. Forey (Eds) *Milestones in Systematics* (pp. 57–100). CRC Press.

Rodrigo, A. (1996). On combining cladograms. *Taxon*, *45*(2), 267–274.

Rodrigo, A., Kelly-Borges, M., Bergquist, P. & Bergquist, P. (1993). A randomisation test of the null hypothesis that two cladograms are sample estimates of a parametric phylogenetic tree. *New Zealand Journal of Botany*, *31*, 257–268.

Ross, H. & Rodrigo, A. (2004). An assessment of matrix representation with compatibility in supertree construction. In O. Bininda-Edmonds (Ed.) *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life* (pp. 35–63). Dordrecht, the Netherlands: Kluver Academic.

Seo, T.-K., Hirohisa, K. & Thorne, J. (2005). Incorporating gene-specific variation when inferring and evaluating optimal evolutionary tree topologies from multilocus sequence data. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(12), 4436–4441.

Struck, T., Purschke, G. & Halanych, K. (2006). Phylogeny of Eunicida (Annelida) and exploring data congruence using a partition addition bootstrap alteration (PABA) approach. *Systematic Biology*, *55*(1), 1–20.

Swofford, D. (2002). *PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods)*, Version 4.0b10. Sunderland, Massachusetts: Sinauer Associates.

Wilkinson, M. (1994). Common cladistic information and its consensus representation: reduced Adams and reduced cladistic consensus trees and profiles. *Systematic Biology*, *43*(3), 343–368.

Wilkinson, M. (1996). Majority-rule reduced consensus trees and their use in bootstrapping. *Molecular Biology and Evolution*, *13*(3), 437–444.

Wilkinson, M., Lapointe, F.-J. & Gower, D. (2003). Branch lengths and support. *Systematic Biology*, *52*(1), 127–130.

Wilkinson, M., Pisani, D., Cotton, J. & Corfe, I. (2005). Measuring support and finding unsupported relationships in supertrees. *Systematic Biology*, *54*(5), 823–831.

Supplementary material

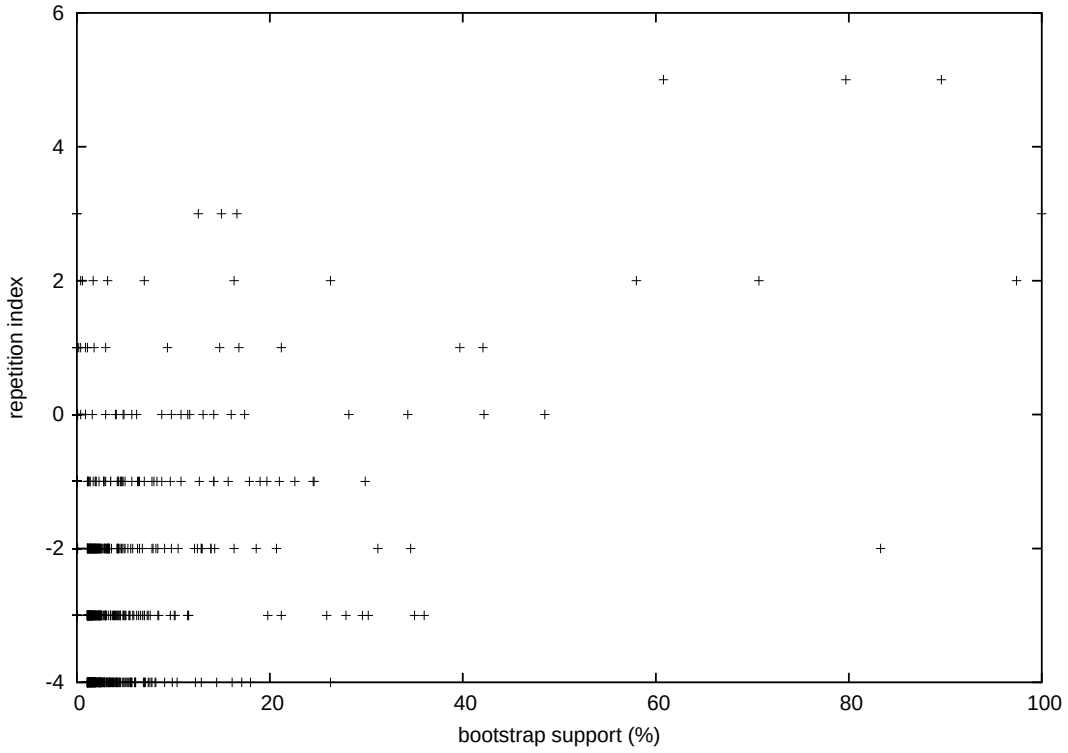


Figure 1: Repetition indices of the clades plotted against their bootstrap supports in the 'total evidence' analysis.

Table 1: Taxonomic sampling (in bold; sequences not present in Dettaï, 2004 or in Dettaï & Lecointre, 2005)

genus	28S	12S	16S	Rhodopsin	MLL	IRBP
Ammodytes	AY141689-92	AY141380	AY141450	AY141306	AY362234	**DQ309871**
Antennarius	AY372752-53	AY368287	AY368304	AY368324	AY362215	DQ168037
Apletodon	AY141557-60	AY141348	AY141418	AY141274	AY362213	DQ168039
Arnoglossus	AY141593-96	AY141358	AY141428	AY141283	AY362228	**DQ309872**
Aulostomus	AY141577-80	AY141353	AY141423	AY141279	AY362226	DQ168040

Table 1: Continued

genus	28S	12S	16S	Rhodopsin	MLL	IRBP
Barbourisia	DQ021385, 93	AY368290	AF221881	AY368333	AY362264	DQ168041
Bathypterois	AY141473-76	AY141326	AY141396	AY141257	AY362219	DQ168042
Bedotia	AY141525-28	AY141339	AY141409	AY141267	AY362271	DQ168043
Belone	AY141529-32	AY141340	AY141410	AY141268	AY362273	DQ168044
Beryx	AY141513-16	AY141336	AY141406	AY141265	AY362238	DQ168045
Bovichtus	AY141661-64	Z32702	**AY520101**	AY141299	AY362283	DQ168046
Callionymus	AY141541-44	AY141344	AY141414	AY141270	AY362225	DQ168047
Capros	AY141501-04	AY141333	AY141403	AY141262	AY362233	DQ168048
Ceratias	AY141505-08	AY141334	AY141404	AY141263	AY362270	DQ168049
Chaetodon	AY372701-04	**AJ748314**	AF055613	AY368312	AY362240	DQ168050
Channa	AY141569-72	AY141351	AY141421	AY141277	AY362241	DQ168051
Cheimarrichthys	AY141693-96	AY141381	AY141451	AY141307	AY362229	DQ168052
Chelidonichthys	AY141609-12	AY141362	AY141432	AY141287	AY362284	DQ168053
Chloroscombrus	AY141717-20	AY141387	AY141457	AY141313	AY362223	DQ168054
Citharus	AY372697-00	AF542220	AY157325	AY141323	AY362232	DQ168055
Ctenochaetus	AY141745-48	AY141394	AY141464	AY141320	AY362242	DQ168057
Ctenopoma	AY141573-76	AY141352	**AY662702**	AY141278	AY362210	DQ168058
Cyclopterus	AY372737-38	AY368284	AY368299	AY368316	AY362218	**DQ309873**
Dactylopterus	AY141589-92	AY141357	AY141427	AY141282	AY362243	DQ168059
Echeneis	AY141725-28	AY141389	AY141459	AY141315	AY362245	DQ168062
Epinephelus	AY141629-32	AY141367	AY141437	AY141291	AY362227	DQ168064
Forsterygion	AY141549-52	AY141346	AY141416	AY141272	AY362276	DQ168065
Gymnocephalus	AY141649-52	AY141373	AY141443	AY141296	AY362278	DQ168068
Halobatrachus	AY372743-44	AY368286	AY368308	AY368323	AY362246	DQ168069

Table 1: Continued

genus	28S	12S	16S	Rhodopsin	MLL	IRBP
Holacanthus	AY141753-56	AF055593	AF055614	AY141322	AY362214	DQ168072
Holanthias	AY141625-28	AY141366	AY141436	AY141290	AY362209	DQ168073
Kali	AY141697-00	AY141382	AY141452	AY141308	AY362224	DQ168074
Labrus	AY141737-40	**AF414201**	AY141462	AY141318	AY362222	DQ168075
Lagocephalus	AY141601-04	AY141360	AY141430	AY141285	AY362221	DQ168076
Lateolabrax	AY141633-36	AY141369	AY141439	AY141293	AY362253	DQ168078
Lepadogaster	AY141553-56	AY141347	AY141417	AY141273	AY362247	DQ168080
Liza	AY141521-24	AY141338	AY141408	AY141266	AY362248	DQ168082
Macroramphosus	AY141581-84	AY141354	AY141424	AY141280	AY362206	DQ168083
Mastacembelus	AY141561-64	AY141349	AY141419	AY141275	AY362249	DQ168084
Mene	AY141729-32	AY141390	AY141460	AY141316	AY362250	DQ168085
Microchirus	AY141597-00	AY141359	AY141429	AY141284	AY362275	DQ168086
Mola	AY141605-08	AY141361	AY141431	AY141286	AY362251	DQ168087
Monopterus	AY141565-68	AY141350	AY141420	AY141276	AY362252	DQ168088
Mullus	AY372719-21	AY368277	AF227680	Y18666	AY362231	DQ168090
Myripristis	AY141517-20	AY141337	AY141407	U57539	AY362265	DQ168091
Notothenia	AY141673-76	Z32712	Z32731	AY141302	AY362282	DQ168093
Ostracion	AY372722-23	AY368281	AF137213	AF137213	AY362207	DQ168095
Pampus	AY141701-04	AY141383	AY141453	AY141309	AY362220	DQ168096
Parablennius	AY141545-48	AY141345	AY141415	AY141271	AY362255	DQ168097
Pentanemus	AY141733-36	AY141391	AY141461	AY141317	AY362272	DQ168098
Perca	AY141645-48	AY141372	AY141442	AY141295	AY362279	DQ168099
Pholis	AY141657-60	AY141375	**AF420459**	AY141298	AY362285	DQ168100
Poecilia	AY141533-36	AY141342	**U80051**	AY141269	AY362203	DQ168102

Table 1: Continued

genus	28S	12S	16S	Rhodopsin	MLL	IRBP
Pogonoperca	AY372711-14	AY141368	**AF297322**	AY141292	AY362256	DQ168103
Polymixia	AY372724-26	AF049730	AF049740	AY368320	AY362208	DQ168104
Pomadasys	DQ021392	AY368293	AY368298	DQ021404	AY363643	AY362230
Psenopsis	AY141705-08	AY141384	AY141454	AY141310	AY362269	DQ168107
Psettodes	AY372717-18	AY368282	AY368302	AY368332	AY362259	DQ168108
Regalecus	AY372729-30	AY368292	AY368296	AY368328	AY362266	DQ168109
Rypticus	DQ021391	AY368295	**AF297327**	AY368329	AY362257	DQ168111
Scarus	AY141741-44	AY141393	AY141463	AY141319	AY362212	DQ168112
Scomber	AY141709-12	AY141385	**AB032521**	AY141311	AY362237	DQ168113
Scorpaena	AY141617-20	AY141364	AY141434	AY141288	AY362236	DQ168114
Serranus	AY141621-24	AY141365	AY141435	AY141289	AY362202	DQ168115
Sphyraena	AY141713-16	AY141386	AY141456	AY141312	AY362254	DQ168118
Taurulus	AH011857	AY141363	AY141433	U97275	AY362217	DQ168121
Trachinotus	AY141721-24	AY141388	AY141458	AY141314	AY362263	DQ168120
Trachinus	AY141681-84	AY141378	AY141448	AY141304	AY362277	DQ168123
Trachyrincus	AY372708-10	AY368280	AY368301	AY368318	AY362289	DQ168124
Triacanthodes	DQ021383, 96	AY368289	AY368311	AY368331	AY362258	DQ168125
Uranoscopus	AY141685-88	AY141379	AY141449	AY141305	AY362239	DQ168126
Zenopsis	AY372748-50	AY368278	AY368300	AY368314	AY362286	DQ168127
Zeus	AY141493-96	AY141331	AY141401	Y14484	AY362287	DQ168128

Why two clades with the same repetition index, same number of occurrences and contradicting one another should not have a high repetition index?

Let α and β be two clades that have the same repetition index R and the same number of occurrences M, and that contradict one another.

Let's also assume that each clade has a stable best contradictor.

If β is the best contradictor of α, then their repetition index is 0.

Is it possible that α and β have a strictly positive repetition index?

Let r be the function that to a clade γ associates its final repetition index $r(\gamma)$.

Let O be the function that to a clade γ associates its maximum number of occurrences (or sum of supports) over independant analyses $O(\gamma)$.

Let C be the function that to a clade γ associates its best contradictor $C(\gamma)$.

$r(\gamma) = O(\gamma) - O(C(\gamma))$

Let C_i be the sequence of the clades such that $C_0 = \alpha$ and for $i \geq 0$: $C_{i+1} = C(C_i)$.

Let O_i be the sequence of the number of ocurrences of the clades C_i. For $i \geq 0$: $O_i = O(C_i)$.

Let r_i be the sequence of the repetition indices of the clades C_i. For $i \geq 0$: $r_i = r(C_i)$.

$C_0 = \alpha$

$O_0 = O(\alpha) = M$

$r_0 = r(\alpha) = R$

Let us suppose (1): $R > 0$.

β is one of the contradictors of α, therefore $r_1 \geq r(\beta) = R$.

If $r_1 = R$, then $O_1 \geq O(\beta) = M$.

Then $R = r_0 = O_0 - O_1 \leq M - M = 0$: impossible because (1).

Therefore (2): $r_1 > R$.

$r_0 = O_0 - O_1$, therefore $O_1 = O_0 - r_0 = M - R$.

α is one of the contradictors of C_1, therefore $r_2 \geq r(\alpha) = R$.

If $r_2 = R$, then $O_2 \geq O(\alpha) = M$.

Now $O_1 = M - R$ and $r_1 = O_1 - O_2$.

Therefore $r_1 = O_1 - O_2 \leq M - R - M = -R$: impossible because (1) and (2).

Therefore $r_2 > R$.

$r_1 = O_1 - O_2$, therefore $O_2 = O_1 - r_1 = M - R - r_1 < M - 2R$.

Property P_i: $r_i > r_{i-2} \geq R$ and $O_i < M - iR$.

Let us try to demonstrate P_i, for $i \geq 2$.

P_2 is true: $r_2 > R = r_0$ and $O_2 < M - 2R$.

Let j be an interger such that $j > 2$.

Let us suppose that P_i is true for each integer i such that $2 \leq i < j$.

C_{j-2} is a contradictor of C_{j-1}, therefore $r_j \geq r_{j-2}$.

If $r_j = r_{j-2}$, then $O_j \geq O_{j-2}$.

Then $r_{j-1} = O_{j-1} - O_j \leq O_{j-1} - O_{j-2}$.

Now $O_{j-1} - O_{j-2} = -r_{j-2}$ and $r_{j-2} > R$ (from (2) if $j = 3$, or from P_{j-2}).

Therefore $r_{j-1} < -r_{j-2} < -R$: impossible because $r_{j-1} > R$ (from P_{j-1}).

Therefore $r_j > r_{j-2} \geq R$.

$r_{j-1} = O_{j-1} - O_j$, therefore $O_j = O_{j-1} - r_{j-1}$.

Now $O_{j-1} < M - (j-1)R$ and $r_{j-1} > R$ (from P_{j-1}).

Therefore $O_j < M - (j-1)R - R = M - jR$: P_j is true.

P_2 is true, therefore P_i is true for every integer i such that $i \geq 2$.

For i an integer such that $i > M/R$, $O_i < 0$ (from P_i).

However, by definition, $O_i \geq 0$ for every value of i.

Therefore (1) cannot be true:

Two clades with the same repetition index and the same number of occurrences that contradict one another cannot have a strictly positive repetition index (in the case where all clades have a stable best contradictor).

If the repetition index is an average, it's more difficult to see if α and β have a low repetition index.

References

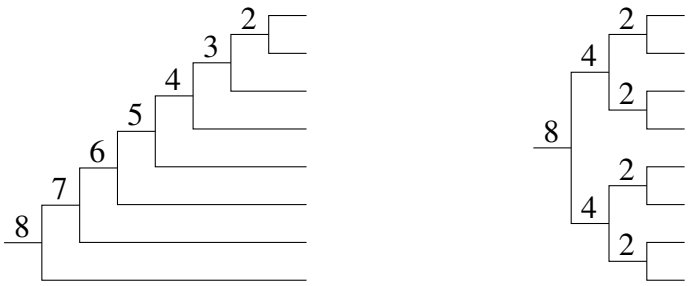
Dettaï, A. (2004), *La phylogénie des Acanthomorpha (Teleostei) inférée par l'étude de la congruence taxinomique*. Ph.D. thesis, Université Paris VI Pierre et Marie Curie.

Dettaï, A. & Lecointre, G. (2005), Further support for the clades obtained by multiple molecular phylogenies in the acanthomorph bush. *Comptes Rendus Biologies*, *328*, 674–689.

3. Perfectionnements de l'indice

3.1. Prendre en compte le comportement du programme de phylogénie utilisé face à l'absence de signal

Moins la présence d'un clade est probable *a priori* (c'est-à-dire sans tenir compte des données), plus son apparition lors d'une analyse est signe de fiabilité. Il serait donc bon de pouvoir disposer de probabilités d'apparition d'un clade *a priori*, dans le cas ou les données ne contiendraient aucun signal. Au premier abord, on pourrait penser que cette probabilité dépend juste de la taille du clade, un clade étant alors un tirage aléatoire de taxons parmi l'échantillonnage, sans remise. En fait, la probabilité d'obtenir un clade de telle ou telle taille dépend des probabilités relatives des types de topologies possibles et les probabilités des différents types d'arbres dépendent de la méthode de reconstruction employée. Si la méthode de reconstruction tend à favoriser des topologies déséquilibrées (type « chenille »), on obtiendra une distribution des tailles de clades plus uniforme que si la méthode tend à favoriser des topologies équilibrées (type « cerises »).



Type 'chenille' Type 'cerises'

Figure 3.1. Les deux types de topologies extrêmes. Les tailles des clades sont indiquées aux nœuds.

Il convient donc de définir ce que serait un jeu de données sans signal et de déterminer la distribution des arbres obtenus par l'application de la méthode de reconstruction choisie à de tels

jeux de données. Ceci pourrait se faire par la répétition un grand nombre de fois de la procédure suivante[1] :

1. tirage d'un jeu de données aléatoire (sans signal) ;
2. application de la méthode de reconstruction à ces données ;
3. enregistrement de la distribution des tailles des clades obtenus.

Il faudrait ensuite calculer la fréquence globale de chaque clade sur tous les arbres obtenus lors des répétitions. Cette fréquence pourrait être retranchée de chaque occurrence lors du calcul de l'indice de fiabilité ; une occurrence d'un clade peu probable est un signe de fiabilité plus significatif qu'une occurrence d'un clade plus probable.

3.1.1. Qu'est-ce qu'un jeu de données sans signal ?

Pour avoir un jeu de données sans signal, il faut éviter que certains groupes de séquences se distinguent systématiquement par un biais compositionnel. Il faut donc tirer la composition de chaque séquence dans la même loi de probabilité ; ne pas reprendre la composition des séquences originales pour créer des séquences aléatoires à partir de ces compositions. Faut-il pour autant prendre à chaque fois une composition uniforme pour chaque séquence ? Un jeu de données aléatoire doit il avoir certains aspects réalistes ?
Pour simplifier, on pourra admettre que tirer n'importe quel caractère dans une loi uniforme sera une façon appropriée d'obtenir un jeu de données sans signal.

3.1.2. Comment la méthode de reconstruction pourrait jouer sur la distribution des arbres ?

Il y a au moins un élément qui influence la fréquence des clades qu'on peut obtenir ; le choix d'un groupe extérieur. On doit donc choisir les mêmes méthodes d'enracinement et les mêmes groupes externes que lors des analyses des vrais jeux de données pour obtenir les justes fréquences d'apparition des clades *a priori*.
Il se pourrait bien que le type d'heuristique employé biaise la distribution des arbres obtenus ; c'est peut-être même la principale raison pour laquelle il faudrait analyser des données aléatoires plutôt que d'essayer de calculer analytiquement les probabilités d'obtenir les clades.

[1] Le problème qui se pose alors est le suivant : si la méthode de reconstruction employée est lente, il n'est pas possible d'explorer de façon satisfaisante l'espace des arbres possibles pour pouvoir admettre que les fréquences mesurées d'apparition des clades sont proches de leur probabilité d'apparaître en l'absence de signal. Peut-on se rabattre sur la seule taille des clades à la place de leur composition exacte afin d'avoir un espace plus réduit à explorer ? Il faut pour cela au moins une condition : que les arbres ne soient pas enracinés, car l'enracinement introduit un biais. Il est peut-être possible d'appliquer une version moins poussée de l'exploration des arbres lors de l'analyse des matrices aléatoires que lors de l'analyse des vraies données.

3.2. Utiliser les probabilités postérieures des clades

La méthode Bayésienne d'inférence phylogénétique, très employée pour les phylogénies moléculaires, fournit une estimation de la probabilité des clades *a posteriori*, dans le cadre du modèle d'évolution des séquences utilisé et d'après les données analysées. Ces probabilités, dites « probabilités postérieures » ont valeur de fiabilité dans le cadre de l'analyse de ces données. Elles pourraient donc servir de base à un indice de fiabilité construit autour de la comparaison d'analyses indépendantes. Au lieu de compter un nombre entier d'occurrences d'un clade pour un schéma de partitionnement, il suffirait de ne compter que des occurrences partielles. L'indice de fiabilité d'ordre 1 serait obtenu en sommant les probabilités postérieures obtenues par le clade pour chacune des analyses constituant le schéma de partitionnement. De cette manière, l'indice de fiabilité aurait plus de finesse : Un clade ne serait pas simplement présent ou absent d'une analyse puisque lors d'une analyse par la méthode Bayésienne, le résultat ne se résume pas à un simple arbre, mais à une liste de clades plus ou moins fréquemment échantillonnés. La même façon de procéder pourrait s'appliquer aux pourcentages de *bootstrap* ou de *jacknife*, mais ces valeurs de support sont probablement moins bien corrélées à la fiabilité que les probabilités postérieures.

3.3. Définir des domaines de validité

Un des principaux problèmes de l'indice de répétition proposé dans LI et LECOINTRE (in press) est de nécessiter que toutes les comparaisons de clades soient effectuées sur le même ensemble de taxons. Si le nombre de taxons qu'il faudrait ainsi mettre de côté est grand, surtout s'il s'agit essentiellement de taxons ne manquant que pour une faible proportion des jeux de données élémentaires, on peut choisir de les inclure dans les analyses et de les retirer des arbres obtenus au moment du calcul de l'indice de répétition[2]. L'indice ainsi calculé n'est valable que sur l'ensemble de taxons qui correspond à l'intersection des ensembles de taxons utilisés dans les analyses. On définit un « domaine de validité » pour la comparaison des analyses, qui est l'intersection de ce qu'on pourrait appeler les domaines de validité des analyses à comparer. Cette procédure peut être étendue afin de calculer des fiabilités valables sur d'autres ensembles de taxons plus larges, mais prenant en compte moins d'analyses. Prenons un exemple simpliste. On note A, B et C les jeux de données élémentaires. On note V_A, V_B et V_C les ensembles de taxons présents dans les jeux de données correspondant. Les taxons sont notés a, b, c, d, e, f, g, h et i. Supposons qu'on ait les compositions suivantes : $V_A = \{a, d, e, g, h, i\}$, $V_B = \{b, d, f, g, h, i\}$, $V_C = \{c, e, f, g, h, i\}$. Le plus petit domaine de validité est $V_A \cap V_B \cap V_C = \{g, h, i\}$. On peut y calculer l'indice de répétition tel qu'il est décrit dans LI et LECOINTRE (in press), en prenant en compte tous les schémas de partitionnement. Si l'on admet des données manquantes dans les analyses combinées,

[2] Cette façon de procéder relève d'un « *total evidence* taxinomique ».

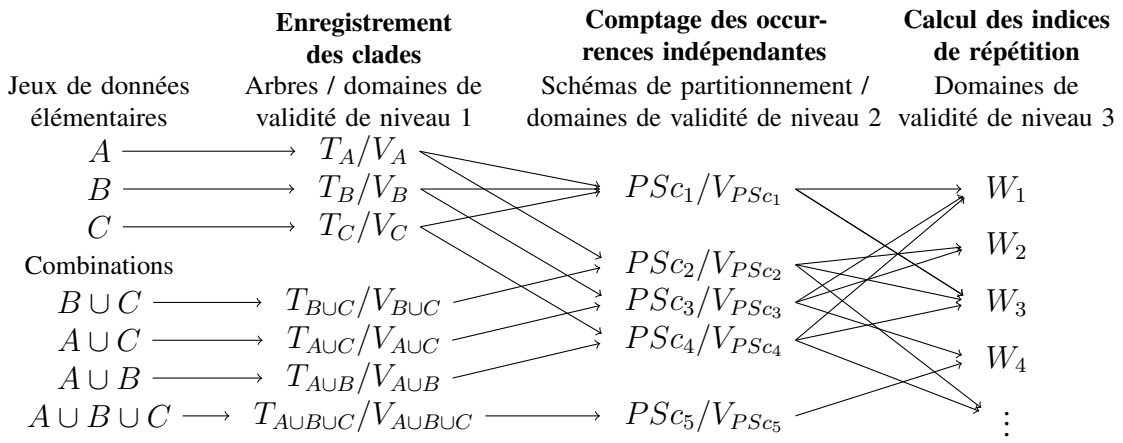


Figure 3.2. Les trois niveaux de domaines de validité. Ceux du premier niveau (notés V_X) sont les ensembles de taxons des arbres (T_X) obtenus par les analyses des jeux de données (X). Dans cet exemple, il y a trois jeux de données élémentaires, soit sept jeux de données en tout, et donc sept arbres et sept domaines de validité de niveau 1. Les domaines de validité de deuxième niveau (V_{PSc_i}) sont les intersections des domaines de validité de premier niveau des jeux de données indépendants composant les schémas de partitionnement (PSc_i). Seuls les schémas de partitionnement complets sont représentés ici. On compte les nombres d'occurrences des clades dans le cadre d'un schéma de partitionnement, en examinant les arbres constituant le schéma de partitionnement, après en avoir ôté les taxons hors du domaine de validité associé au schéma de partitionnement. Les domaines de validité de niveau 3 (W_i) sont les intersections de toutes les combinaisons possibles de domaines de validité de schéma de partitionnement. Les indices de répétition sont associés à ces domaines de validité de niveau 3. Ils sont basés sur les plus grands nombres d'occurrences des clades (une fois les taxons hors du domaine de validité de niveau 3 élagués) trouvés parmis les schémas de partitionnement dont le domaine de validité contient au moins le domaine de validité de niveau 3. Seuls certains domaines de validité de niveau 3 sont représentés ici.

les autres jeux de données analysés ont les compositions suivantes : $V_{A\cup B} = \{a, b, d, e, f, g, h, i\}$, $V_{A\cup C} = \{a, c, d, e, f, g, h, i\}$, $V_{B\cup C} = \{b, c, d, e, f, g, h, i\}$, $V_{A\cup B\cup C} = \{a, b, c, d, e, f, g, h, i\}$. $V_{A\cup B\cup C}$ est le plus grand domaine de validité, mais la seule analyse qui peut être prise en compte pour calculer un indice valable dans ce domaine est l'analyse combinée totale[3]. Si l'on veut pouvoir calculer un indice de répétition prenant en compte le schéma de partitionnement $(A \cup B, C)$, par exemple, le plus grand domaine de validité utilisable est $V_{A\cup B} \cap V_C = \{e, f, g, h, i\}$; tout taxon supplémentaire manquerait dans les résultats de l'analyse de $A\cup B$ ou de C. $(A\cup B, C)$ peut être utilisé pour le calcul d'un indice de répétition associé à tout domaine de validité inclus dans $V_{A\cup B} \cap V_C$.

On peut en fait définir trois niveaux de domaines de validité (voir figure 3.2) :

[3]L'indice de répétition d'un clade se ramènerait alors simplement à 1 si le clade est présent et 0 s'il est absent. Si l'on veut raffiner, on peut prendre une valeur de support, comme une probabilité postérieure ou une proportion dans une analyse de bootstrap ; on se ramène à la situation décrite dans l'introduction, page 11.

1. Le domaine de validité d'une analyse primaire (domaine de validité de premier niveau) est l'ensemble des taxons présents dans un jeu de données (combiné ou non) analysé pour obtenir un arbre. Ces domaines de validité sont ici notés V_A, V_B, V_C, $V_{A\cup B}$, $V_{B\cup C}$, $V_{A\cup C}$, $V_{A\cup B\cup C}$.
2. Le domaine de validité d'une comparaison ou d'un schéma de partitionnement (domaine de validité de niveau 2) est l'intersection des domaines de validité d'un ensemble d'arbres à comparer. Comme une comparaison ne peut se faire qu'entre arbres obtenus à partir de données indépendantes, ces jeux de données sont choisis de manière à former un schéma de partitionnement (voir page 23). Afin d'exploiter le plus d'information possible, on pourra également considérer des schémas de partitionnement partiels, c'est-à-dire dans lesquels certains jeux de données élémentaires ne sont pas représentés, ce qui permettra de prendre en compte les taxons rares. Par exemple, (A, B) est un schéma de partitionnement partiel qui permettra de prendre en compte les taxons présents à la fois dans les jeux de données A et B, mais absents du jeu de données C. Si PSc_1 est le schéma de partitionnement $(A, B\cup C)$, on pourra noter son domaine de validité V_{PSc_1}, et on aura : $V_{PSc_1} = V_A \cap V_{B\cup C}$.
3. Le domaine de validité d'un indice de fiabilité (domaine de validité de niveau 3) est l'intersection des domaines de validité des schémas de partitionnements parmis lesquels on a choisi les meilleurs nombres d'occurrences des clades. Quand tous les jeux de données ont les mêmes taxons, tous les schémas de partitionnement ont le même domaine de validité, et on peut se contenter d'un seul indice de fiabilité, qui prend en compte tous les schémas de partitionnement à la fois. Si le recouvrement taxinomique n'est pas parfait, on peut potentiellement tirer de l'information utile de toute combinaison de schémas de partitionnement. Plus il y aura de schémas de partitionnement pris en compte en même temps, plus l'indice de fiabilité pourra potentiellement être élevé[4], mais plus le domaine de validité risque d'être restreint.

Définir de multiples domaines de validité permet de prendre en compte plus de taxons mais pose un problème de lisibilité des résultats. Entre différents domaines de validité on pourra rencontrer des clades assez semblables, ne différant que par l'absence de certains taxons appartenant à l'un des domaines de validité et pas à l'autre, chacun avec son propre indice de fiabilité. En l'état, une telle situation impose un effort de synthèse important au lecteur ; on est confronté à un problème semblable à celui des tableaux de répétabilité construits jusqu'à présent (CHEN *et al.*, 2003; DETTAÏ et LECOINTRE, 2004, 2005).

[4]En multipliant les schémas de partitionnement, on multiplie les chances de trouver une façon optimale de combiner les données pour un clade donné (voir page 24).

4. Synthétiser la fiabilité des clades sous forme d'un arbre

Si l'ambition d'un indice de fiabilité est de montrer dans quelle mesure on peut penser que tel ou tel clade représente les relations de parenté réelles entre les taxons, il semble naturel de vouloir synthétiser les résultats en combinant les clades les plus fiables en un arbre, qu'on espère le plus proche possible de l'arbre de parenté réel.

4.1. Avec des relations complètes

Ici, on s'intéresse au cas où les relations à combiner en un arbre sont toutes définies sur un même domaine de validité. C'est le cas quand toutes les analyses utilisées pour le calcul des fiabilités ont été faites sur le même ensemble de taxons ou bien quand on veut synthétiser dans un premier temps les résultats des analyse de fiabilité restreintes à chacun des domaines de validité séparément.
Le but est de simplifier la somme d'informations dont on dispose de façon à n'avoir plus que des relations compatibles que l'on puisse combiner en un arbre. On peut tester la compatibilité d'un ensemble de clades simplement en testant la compatibilité deux à deux des clades ; si tous sont compatibles deux à deux, alors tous sont combinables en un seul arbre (BERRY et GASCUEL, 2000). Pour obtenir un arbre qui représente au mieux les résultats, une manière simple de procéder est alors de rassembler les clades en commençant par les plus fiables, et en éliminant au fur et à mesure les clades incompatibles avec les clades déjà sélectionnés. S'il arrive un moment où plusieurs clades de même fiabilité se présentent en tête des clades restant à ajouter, et si tous sont compatibles avec les clades déjà ajoutés, deux options sont possibles. L'une est de départager les clades selon un critère secondaire, comme le nombre brut d'occurrences (solution utilisée dans LI et LECOINTRE, in press), ou leur robustesse dans l'analyse combinée (mais on a vu que la robustesse n'était pas toujours un signe de fiabilité). L'autre possibilité est d'éliminer en bloc tous ces clades d'égale fiabilité. En principe, des clades incompatibles d'égale fiabilité ne devraient pas avoir une très bonne fiabilité puisque leur fiabilité a été diminuée du nombre d'occurrences d'un clade au moins aussi fiable. On peut donc envisager de rejeter de l'arbre de synthèse tous les clades moins fiables que ceux de la première catégorie de clades équi-fiables mais incompatibles.

4.2. Avec des relations partielles

On a vu en 3.3 que quand les échantillonnages taxinomiques n'étaient pas parfaitement recouvrants, on pouvait définir des domaines de validité sur lesquels calculer des indices de fiabilité. Les clades obtenus dans les différents domaines de validité ne peuvent pas être directement combinés en un arbre comprenant tous les taxons étudiés. Par contre, sur chaque domaine de validité, on peut assembler un arbre de synthèse de la façon décrite en 4.1. On dispose alors de plusieurs arbres synthétisant les résutats de l'analyse de fiabilité et on est ramené à un problème de construction de *supertree*[1].

4.2.1. Synthèse par MRP

La méthode de *supertree* la plus employée est la *Matrix Representation with Parsimony* (MRP, voir BAUM et RAGAN, 2004). Elle consiste à coder toutes les bipartitions trouvées dans l'ensemble des arbres-source sous forme de pseudo-caractères dans une matrice qui sera ensuite analysée par la méthode du maximum de parcimonie. Le codage d'une bipartition se fait comme suit : chaque taxon se voit attribuer l'état 0 ou l'état 1 selon qu'il se situe d'un côté ou de l'autre de la bipartition. S'il n'est pas dans l'arbre d'origine de la bipartition (en ce qui nous concerne, s'il n'est pas dans le domaine de validité de l'analyse de fiabilité d'où provient la bipartition), il est codé comme données manquantes, en général par un point d'interrogation. Pour prendre en compte la fiabilité des clades dans l'assemblage d'un *supertree* destiné à synthétiser l'ensemble des résultats obtenus sur les différents domaines de validité, il paraît naturel de pondérer le caractère correspondant à la bipartition par la fiabilité de cette bipartition. D'autres méthodes basées sur une représentation matricielle sont envisageables, mais la méthode MRP est une des plus simples à mettre en œuvre. Il existe également des méthodes de *supertree* non basées sur une représentation des bipartitions sous forme de caractères, et dont on peut s'inspirer pour faire une synthèse des analyses de fiabilité.

4.2.2. Synthèse inspirée par le `MinCut` *Supertree*

Le `MinCut` *Supertree* (SEMPLE et STEEL, 2000) est un algorithme d'assemblage de *supertree* dérivé de l'algorithme `OneTree` (AHO *et al.*, 1981; NG et WORMALD, 1996). L'algorithme `OneTree` prend en entrée un ensemble de triplets[2] et renvoie un arbre si les triplets sont compatibles ou bien signale l'existence d'une incompatibilité. Il procède de façon récursive en subdivisant l'ensemble des taxons en sous-ensembles dans lesquels il s'applique, jusqu'à

[1] Les méthodes de *supertree* sont les méthodes qui consistent à combiner l'information de plusieurs arbres en un seul. Les méthodes de consensus en sont un cas particulier, restreint à des arbres comportant les mêmes taxons.

[2] Les triplets peuvent être ceux composant des arbres-source. C'est une méthode qui ne marche qu'avec des arbres enracinés.

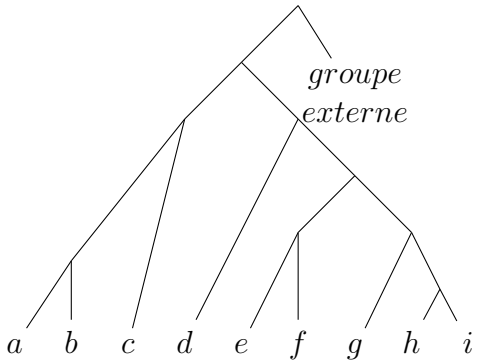


Figure 4.1. Exemple d'arbre pris pour illustrer le déroulement de l'algorithme `OneTree`. L'ensemble des triplets de cet arbre est compatible puisque ces triplets forment un arbre. L'algorithme `OneTree` appliqué à cet ensemble de triplets devrait donc aboutir à la reconstruction de l'arbre.

rencontrer un ensemble dans lequel les triplets révèlent leur incompatibilité ou bien sont trivialement assemblables en un arbre. Pour mieux comprendre le déroulement de l'algorithme, appliquons-le à un ensemble de triplets forcément compatibles ; les triplets de l'arbre représenté sur la figure 4.1.

Pour un ensemble donné de triplets compatibles, un sous-arbre de l'arbre formé par ces triplets sera reconnaissable au fait qu'aucun de ses taxons ne sera dans la partie interne d'un triplet (voir définitions page 17 et suivantes) en compagnie d'un taxon d'un autre sous-arbre de l'arbre. Ainsi, il est possible de séparer les taxons en sous-groupes devant former des sous-arbres. Ceci se fait au moyen d'un graphe dans lequel les sommets représentent les taxons de l'arbre à subdiviser en sous-arbres et où il existe une arête entre deux sommets quand il existe un triplet dont la partie interne est constituée des deux taxons correspondant aux sommets à relier, ce qui revient à mettre une arête entre deux sommets quand il existe une paire incluse[3] constituée des taxons correspondant à ces sommets. Les groupes de taxons destinés à former des sous-arbres sont les composantes connexes[4] du graphe (voir figure 4.2).

L'algorithme `OneTree` est ensuite appliqué aux différents groupes de plus de deux[5] taxons afin d'en déterminer les relations. Lors de l'application de cet algorithme à un groupe de taxons, on construit le graphe de la manière décrite ci-dessus en ne prenant plus en compte que les triplets constitués exclusivement de taxons du groupe. Ainsi, les arêtes du graphe construit à l'étape précédente qui reposaient uniquement sur des triplets dont la partie externe n'est plus dans le groupe considéré, ces arêtes, donc, n'existent plus dans le nouveau graphe. Ceci permet de distinguer de nouveaux sous-ensembles de taxons à assembler en sous-arbres (voir figure 4.3).

[3] Ici aussi, voir les définitions page 17 et suivantes.

[4] Une composante connexe d'un graphe est un ensemble maximal de sommets tels qu'il existe un chemin entre n'importe lesquels de ces sommets, un chemin étant une sucession d'arêtes partageant un sommet.

[5] La topologie des sous-arbres à un ou deux taxons est triviale et n'offre donc pas de nouvelles occasions de découvrir de l'incompatibilité entre triplets.

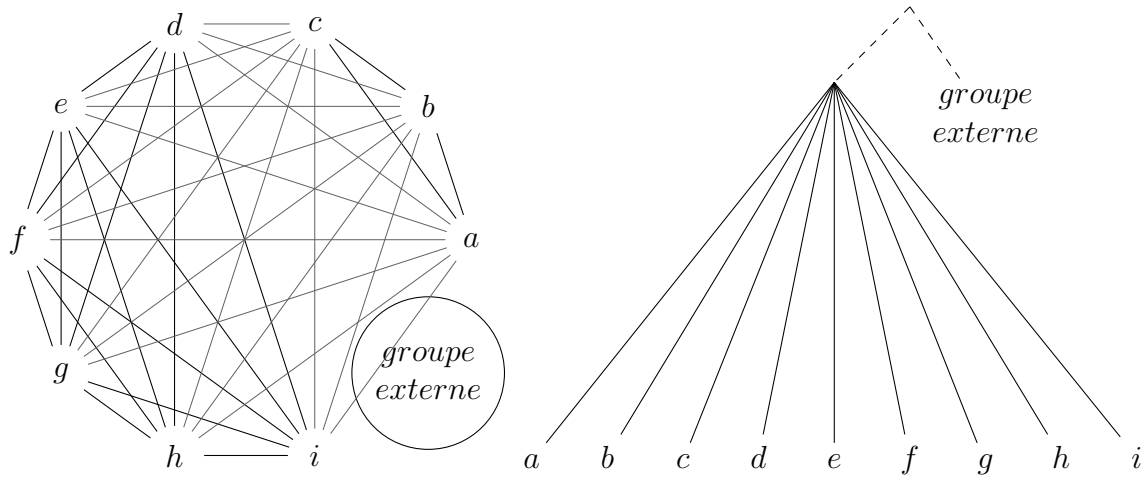


Figure 4.2. Déroulement de l'algorithme `OneTree`. Application aux triplets de l'arbre de la figure 4.1. Étape initiale. Chaque sommet du graphe de gauche représente un taxon. Il y a une arête entre deux sommets quand il existe un triplet dans lequel la partie interne est constituée des deux taxons correspondant, c'est-à-dire quand ces taxons forment une paire incluse. L'arête $a - b$, qui correspond à la paire incluse $\{ab\}$, est soutenue entre autres par les triplets $(ab|c)$ et $(ab|groupe\,externe)$. À ce stade, seul le groupe extérieur n'est dans aucune paire incluse. Les taxons a à i forment une composante connexe du graphe ; ils constituent un sous-arbre de l'arbre en cours de construction (le groupe externe constitue un autre sous-arbre).

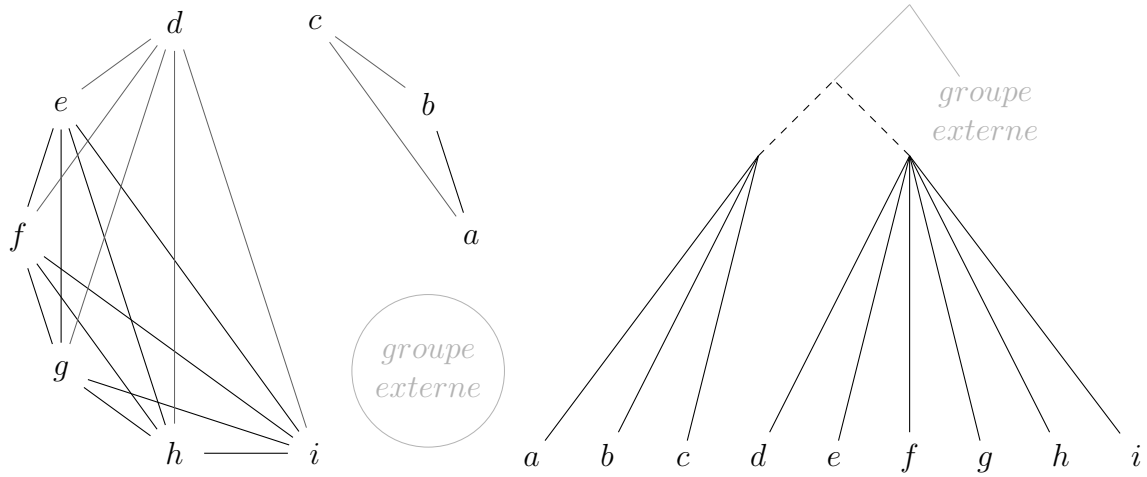


Figure 4.3. Déroulement de l'algorithme `OneTree` (suite). On s'intéresse ici uniquement aux taxons a à i, qui formaient une des composantes connexes du graphe de l'étape précédente. Le groupe extérieur n'est plus pris en compte (d'où la couleur grise choisie ici). Le triplet $(ae|groupe\,externe)$ était le seul à soutenir l'arête $a - e$; cette arête n'est donc plus présente dans le graphe. Les deux composantes connexes obtenues sont constituées des ensembles de taxons a à c et d à i. Le sous-arbre des taxons a à i peut donc être décomposé en deux sous-arbres.

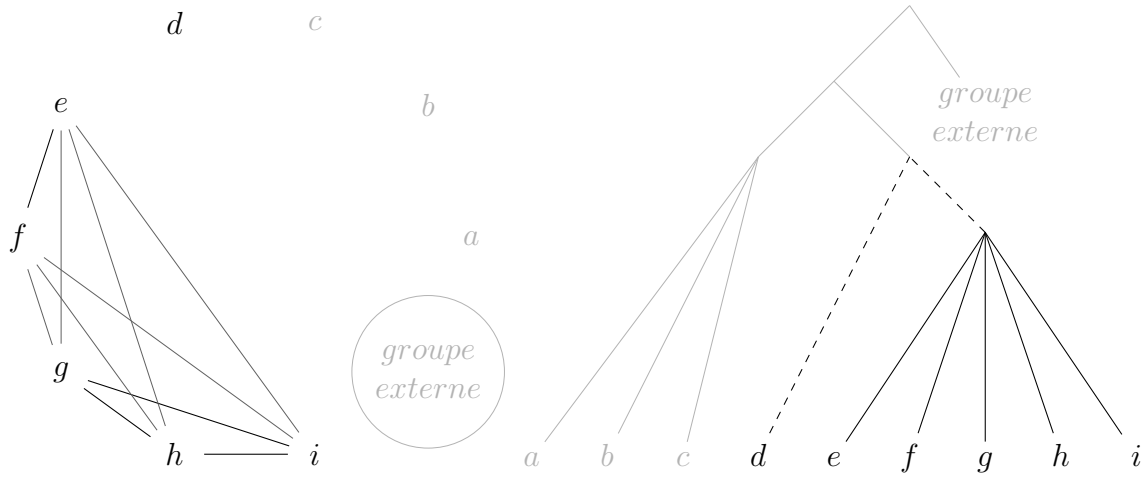


Figure 4.4. Déroulement de l'algorithme `OneTree` (suite). On s'intéresse ici aux taxons d à i. Les triplets comportant d et un des taxons f à i ne sont plus présents puisqu'ils impliquaient le groupe externe ou un des taxons a à c. d est donc isolé du reste du graphe. On peut donc décomposer le sous-arbre des taxons d à i en deux nouveaux sous-arbres (dont l'un n'est en fait constitué que du taxon d).

Si à un moment de l'application de l'algorithme `OneTree`, le graphe construit à partir d'un de ces sous-ensembles est connexe, c'est que les triplets ne sont pas compatibles. L'algorithme `OneTree` échoue alors à construire un arbre. Une variante possible permettant de toujours obtenir un arbre est de laisser les taxons en polytomie en cas d'incompatibilité entre triplets. Dans notre exemple, comme les triplets sont ceux d'un arbre, ils sont compatibles et l'algorithme finit par retrouver la topologie de l'arbre (voir figures 4.4 à 4.7).

Le `MinCut` ***Supertree*** est une variante de l'algorithme `OneTree` dans laquelle, à chaque fois que les triplets sur un ensemble de taxons s'avèrent incompatibles, un ensemble minimal d'arêtes du graphe est coupé afin d'obtenir au moins deux composantes. Pour déterminer cet ensemble minimal d'arêtes à couper (*minimum cut set*), on fusionne préalablement les sommets connectés par des arêtes soutenues par au moins un triplet de chaque arbre-source[6]. Dans le cas qui nous préoccupe, le critère de délétion des arêtes dans le graphe doit être basé sur la fiabilité des relations supportant ces arêtes. Je propose donc une variante que j'appelle `CutMinKeepMax` dans laquelle la déconnexion du graphe se fait par la suppression des arêtes les moins fiables : lors de la construction de l'arbre, quand le graphe des taxons est connexe, on supprime, pour chaque sommet, les arêtes partant de ce sommet les moins bien soutenues en termes de fiabilité

[6]Ceci permet d'obtenir un arbre dans lequel tous les *nestings* au sens de ADAMS (1986) (à ne pas confondre avec des triplets ; un *nesting* peut correspondre à plusieurs triplets, voir WILKINSON, 1994) qui sont partagés par tous les arbres-source sont représentés. PAGE (2002) propose une variante de cet algorithme qui évite également de perdre les *nestings* non contredits. Il signale que, de même que dans le cas du consensus d'ADAMS, les groupes de taxons apparaissant dans un arbre assemblé de cette façon ne sont pas forcément interprétables comme des groupes monophylétiques. Dans le cas $((a,b),(c,d)) + ((a,b),(e,f)) \rightarrow ((a,b),(c,d),(e,f))$, (c,d) et (e,f) ne sont pas contredits en tant que *nestings* mais rien ne permet de dire, par exemple, que e et f sont plus proches entre eux qu'ils ne le sont chacun de d, contrairement à ce que pourrait laisser penser l'arbre obtenu.

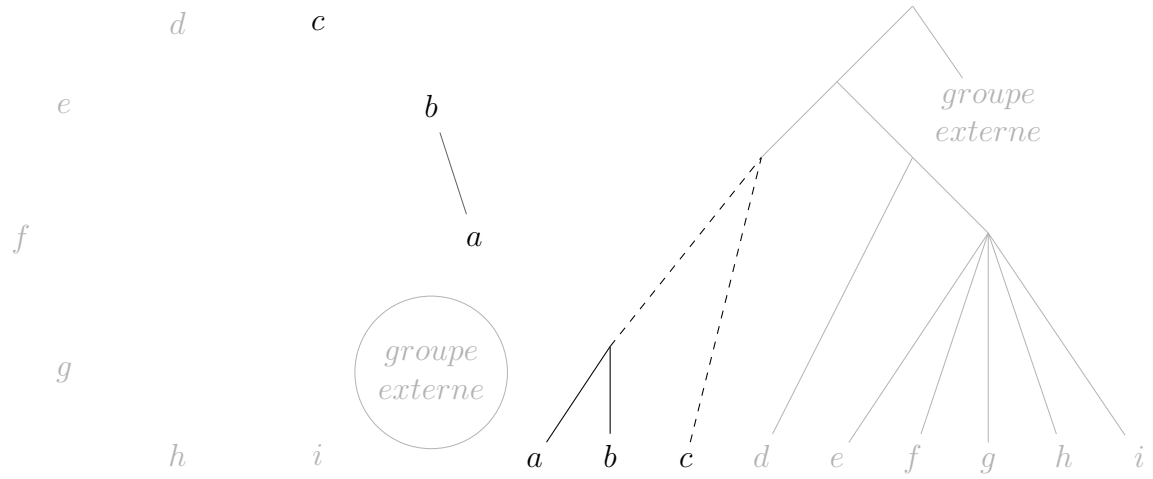


Figure 4.5. Déroulement de l'algorithme `OneTree` (suite). De la même façon que le taxon d s'est trouvé isolé dans l'ensemble des taxons d à i (voir figure 4.4), le taxon c se trouve ici isolé des taxons a et b.

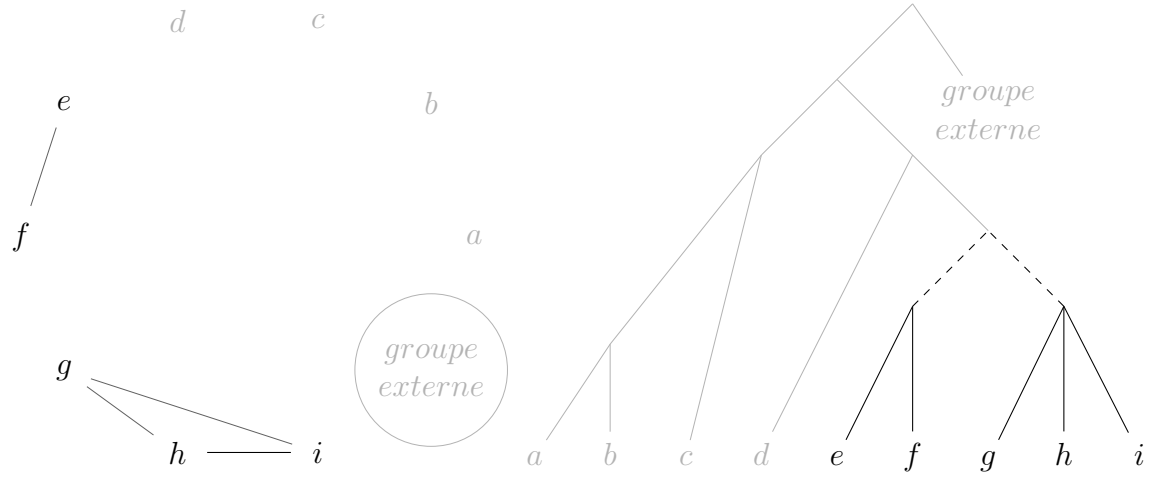


Figure 4.6. Déroulement de l'algorithme `OneTree` (suite). Au sein de l'ensemble des taxons e à i isolé précédemment (figure 4.4), l'absence du taxon d permet à présent de distinguer deux composantes ; la paire de taxons e et f et l'ensemble des taxons g à i. L'élimination des triplets $(eg|d)$, $(eh|d)$, $(ei|d)$, $(fg|d)$, $(fg|d)$ et $(fi|d)$ a en effet fait disparaître les dernières arêtes reliant ces deux composantes.

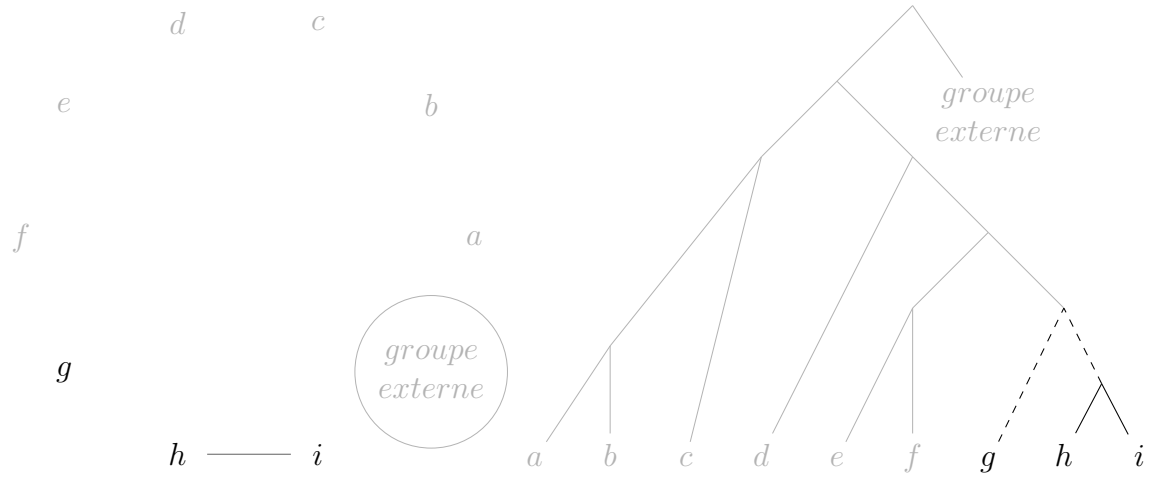


Figure 4.7. Déroulement de l'algorithme `OneTree` (suite et fin). La dernière composante à plus de deux taxons est enfin décomposée en deux sous-arbres. L'arbre d'où ont été tirés les triplets initiaux (figure 4.1) est reconstruit. Supposons qu'en plus de ces triplets, on ait eu le triplet $(gh|i)$, donc la paire incluse $\{gh\}$. L'ensemble des taxons g à i serait resté connexe puisqu'on aurait eu la persistance d'une arête $g - h$. Ceci aurait révélé l'existence d'une incompatibilité dans l'ensemble de triplets. L'algorithme `OneTree` échoue à retourner un arbre quand il est confronté à une telle situation.

(« *cut min* »), sauf si ce sommet n'a plus d'autre arête mieux soutenue (« *keep max* »)[7]. Le soutien d'une arête est l'indice de fiabilité du meilleur clade induisant la paire incluse formée par les deux taxons impliqués dans l'arête (voire figure 4.8).

Pour éviter une trop forte perte de résolution dans l'arbre, les arêtes sont en fait supprimées progressivement (voire figure 4.9), en les décomposant en deux parties : pour une paire de taxons (a, b), on supprime l'arête orientée $a \rightarrow b$ si elle est la moins soutenue des arêtes impliquant a encore présentes, et de même on supprime l'arête orientée $b \rightarrow a$ si elle est la moins soutenue des arêtes encore présentes impliquant b. Une arête n'est considérée comme supprimée du graphe qu'une fois que ses deux composantes ont été retirées. Si le graphe reste connexe après le retrait pour tout taxon de toutes les arêtes orientées (sauf la plus soutenue) partant du sommet représentant ce taxon, alors on reprend le graphe initial et on en retire l'arête (ou les arêtes) restante(s) globalement la (ou les) moins bien soutenue(s) (même si c'est la dernière impliquant ce taxon) puis on recommence les suppressions progressives d'arêtes orientées (figure 4.10).

Les taxons en position la plus basale sont ceux qui ont le moins d'arêtes dans le graphe. La procédure de suppression progressive des arêtes devrait favoriser l'isolement de tels taxons.

[7] Ce « *keep max* » a pour motivation le placement dans l'arbre des taxons présents dans une faible proportion des jeux de données.

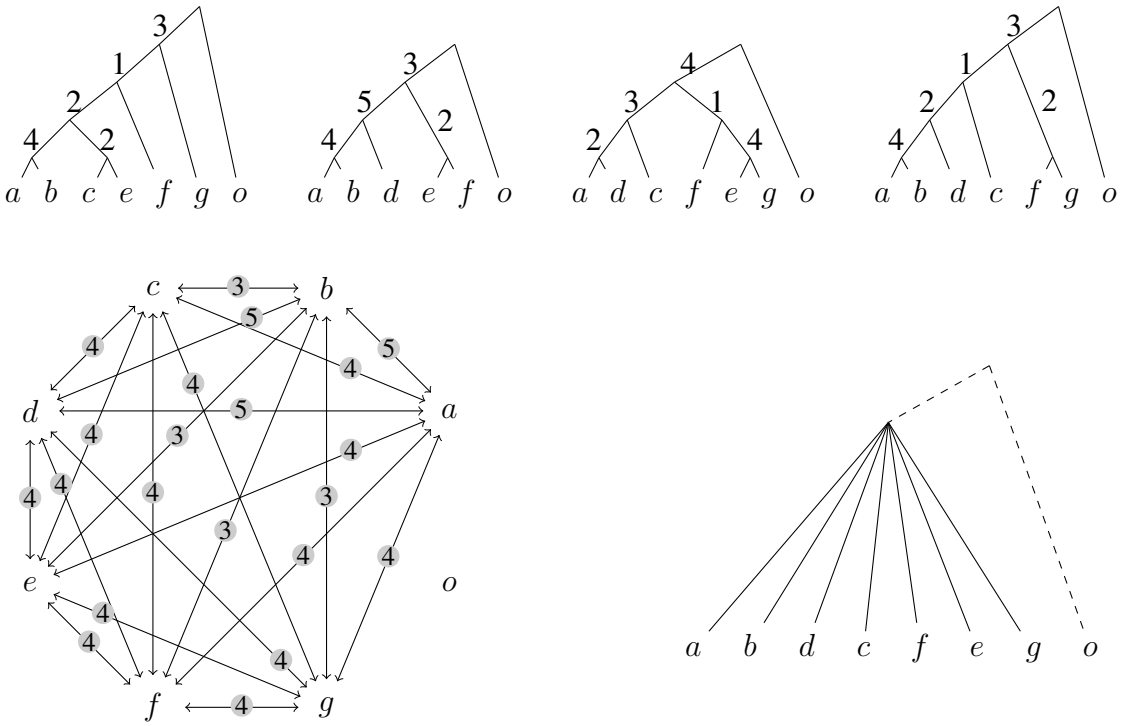


Figure 4.8. Déroulement de l'algorithme `CutMinKeepMax`. Étape initiale. En haut, les 4 arbres-source que l'on veut combiner en un *supertree*. Les supports sont indiqués sur les branches. En bas à gauche, le graphe représentant les paires incluses et leur plus fort support dans les arbres-source. Par exemple, le plus fort support pour la paire incluse $\{ab\}$ est dû au clade $((abd)efo)$ qui a un support de 5 dans le deuxième arbre-source. Le taxon o peut être isolé des autres taxons. En bas à droite, l'arbre en cours de construction.

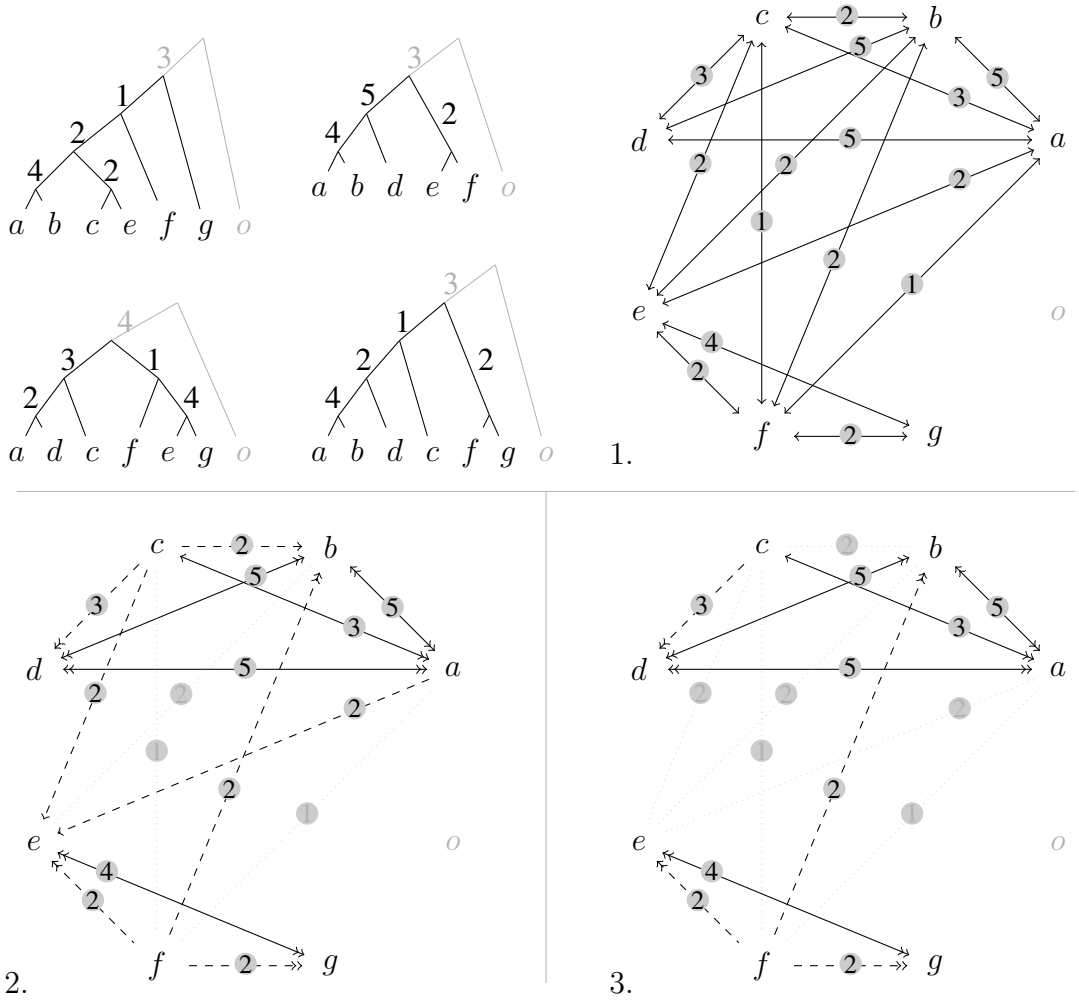


Figure 4.9. Déroulement de l'algorithme `CutMinKeepMax` (suite). L'algorithme est appliqué dans l'ensemble des taxons a à g. En haut à gauche, les arbres-source dans lesquels les parties qui ne sont plus à prendre en compte sont en grisé. 1. Le graphe des paires incluses valables dans l'ensemble des taxons a à g est connexe (o ne compte pas ; il est en grisé). Il faut donc retirer des arêtes pour pouvoir séparer l'arbre en sous-arbres. 2. Pour chaque sommet, les arêtes orientées de plus faible support ont été retirées (exemples : l'arête orientée $a \to f$ avait le plus faible support pour une arête issue de a, pour b, ce sont les arêtes orientées $b \to c$ et $b \to f$). En grisé, les arêtes dont les deux composantes ont été retirées (exemple : $a - f$). En pointillé, celles qui n'ont été que partiellement retirées (exemple : $b \to f$). La composante restante est celle allant dans le sens de la flèche. Une flèche à deux pointes indique une arête qui a le plus fort soutien pour le sommet d'où elle part (exemple : $f \to b$). De telles arêtes ne peuvent dans un premier temps pas être retirées. Le graphe est toujours connexe ; il faut procéder à une nouvelle série de suppressions d'arêtes. 3. Pour chaque sommet, les arêtes orientées « sortantes » de deuxième moins fort support ont été enlevées (exemple : $a \to e$), sauf si elles ont en fait le plus fort support (exemple : $f \to b$). Il ne reste plus que des arêtes de plus fort support pour le sommet d'où elles sont issues, mais le graphe est toujours connexe. Il faut donc recommencer les suppressions à partir du graphe 1. auquel on aura retiré les arêtes les moins soutenues (ici, celles soutenues par une valeur de 1, voir figure 4.10).

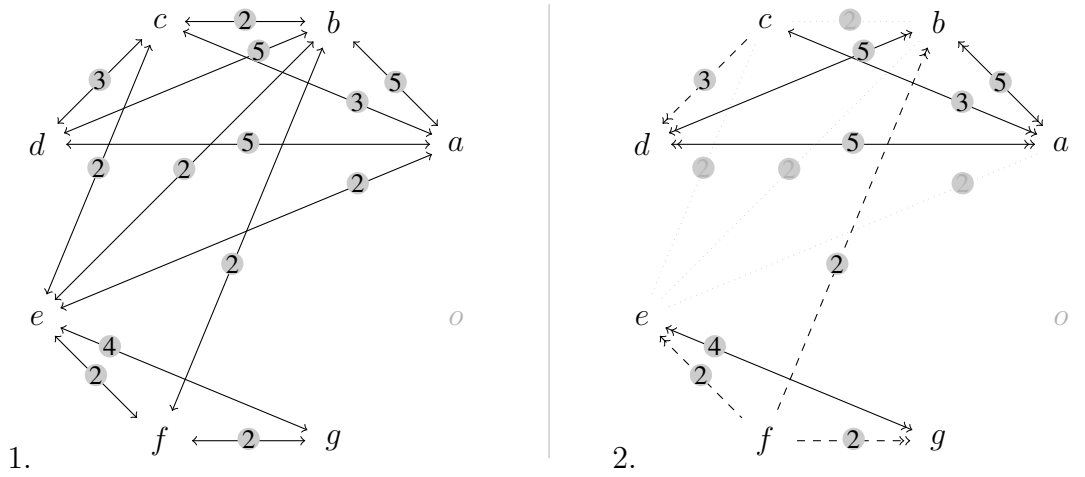


Figure 4.10. Déroulement de l'algorithme `CutMinKeepMax` (suite). 1. Les arêtes du graphe 1. de la figure 4.9 qui n'avaient qu'un support de 1 ont été retirées. Le graphe est toujours connexe ; il faut donc supprimer des arêtes. 2. Pour chaque sommet, les arêtes « sortantes » les plus faibles ont été supprimées, sauf si elles étaient également les arêtes les plus fortes pour le sommet d'où elles sont issues. Il ne reste plus que des arêtes orientées de support maximal pour leur sommet d'origine, mais le graphe est toujours connexe ; il faut donc supprimer les arêtes de support inférieur ou égal à 2 avant de recommencer la suppression progressive des arêtes orientées les moins bien soutenues.

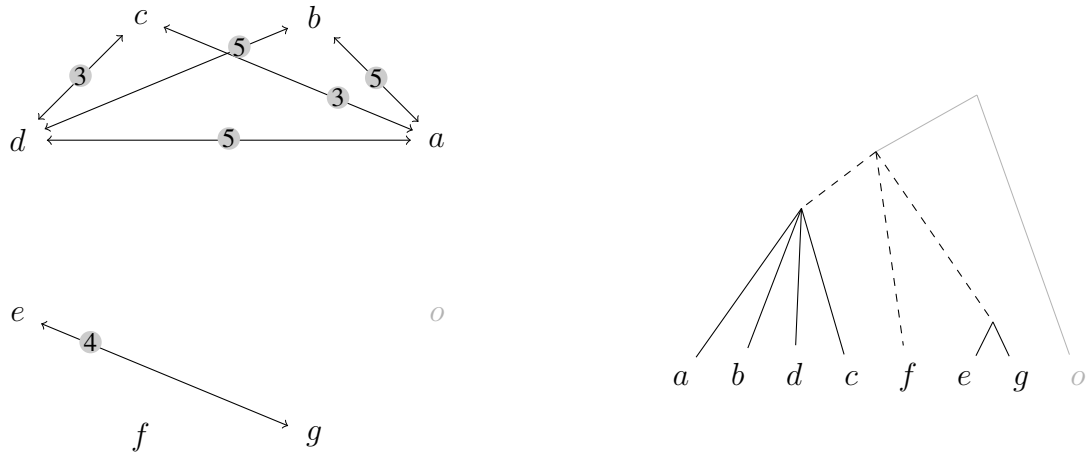


Figure 4.11. Déroulement de l'algorithme `CutMinKeepMax` (suite). Les arêtes de support inférieur ou égal à 2 du graphe 1. de la figure 4.9 ont été supprimées. Cette fois, le graphe n'est plus connexe (à gauche) ; il n'y a donc pas besoin de supprimer d'autres arêtes. L'arbre peut être décomposé en 3 sous-arbres (à droite). Il reste à appliquer l'algorithme dans l'ensemble des taxons a à d (voir figure 4.12), les deux autres composantes formant trivialement des sous-arbres entièrement résolus.

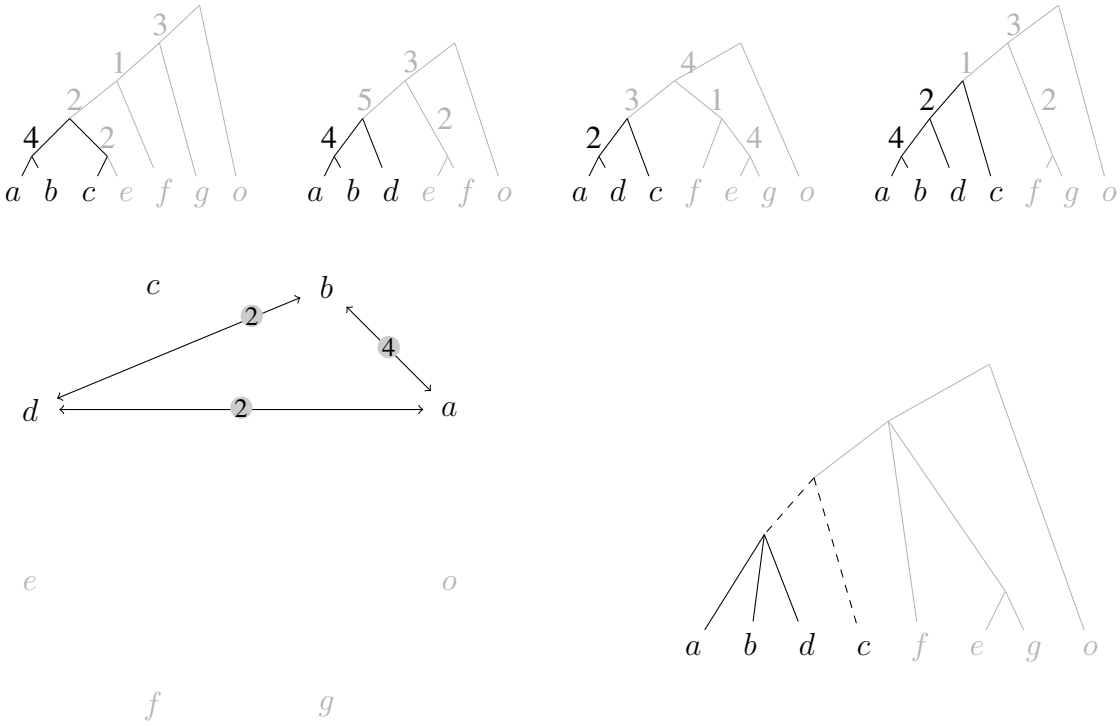


Figure 4.12. Déroulement de l'algorithme `CutMinKeepMax` (suite). Application aux taxons a, b, c et d. En haut, les arbres-source, avec en grisé les parties qu'il ne faut pas prendre en compte. En bas à gauche, le graphe des paires incluses correspondant. Il y a deux composantes dans ce graphe. Le taxon c peut donc être séparé des autres (en bas à droite). L'algorithme sera ensuite appliqué aux taxons a, b et d (voir figure 4.13).

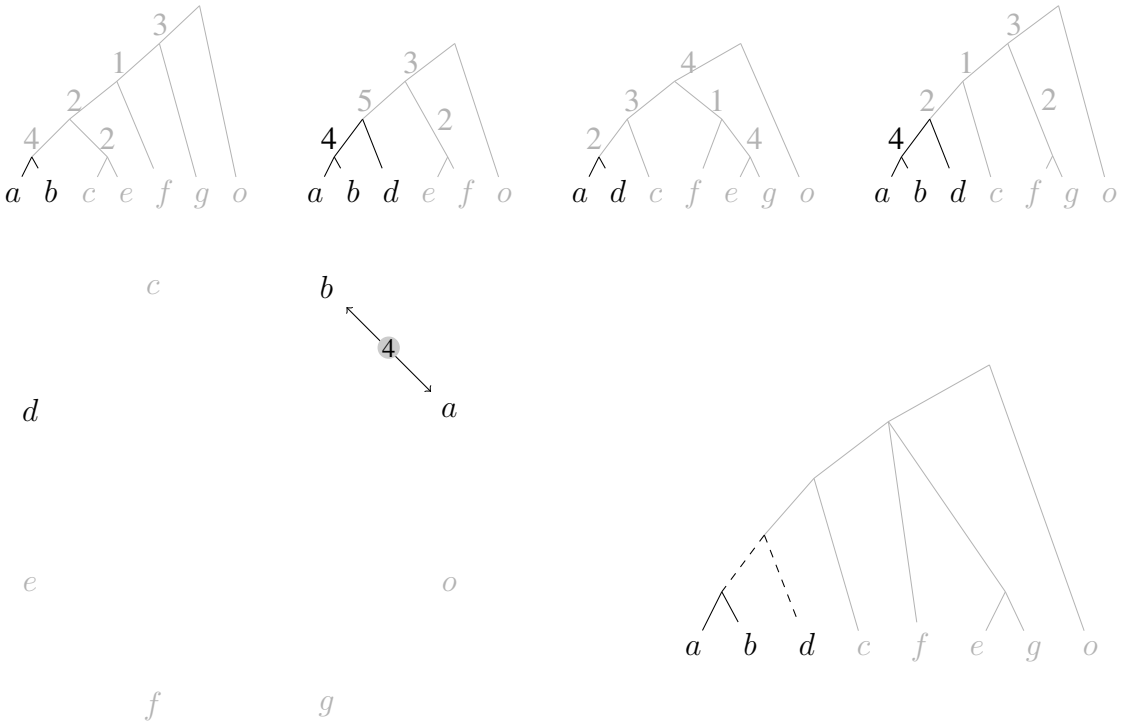


Figure 4.13. Déroulement de l'algorithme `CutMinKeepMax` (suite et fin). En haut, ce qu'il reste des arbres-source quand on se limite aux trois taxons a, b et d. En bas à gauche, le graphe des paires incluses correspondant. La seule paire incluse restante est $\{ab\}$. En bas à droite, la fin de la construction de l'arbre.

Deuxième partie .

Phylogénie des téléostéens acanthomorphes

5. Les acanthomorphes

5.1. Brève présentation

Les téléostéens acanthomorphes sont un clade des actinoptérygiens (les poissons à nageoires rayonnées). Ils doivent leur nom (Acanthomorpha) à ROSEN (1973) qui caractérisa ce groupe par la présence de rayons en épines creuses et non segmentées dans les nageoires. La monophylie de ce groupe a été confirmée morphologiquement (STIASSNY, 1986; JOHNSON et PATTERSON, 1993) et moléculairement (WILEY *et al.*, 2000; CHEN, 2001; MIYA *et al.*, 2001; INOUE *et al.*, 2003).
Les acanthomorphes sont déjà présents au Crétacé, mais leur diversité actuelle résulte d'une radiation qui semble avoir eu lieu au début du tertiaire (voir PATTERSON, 1993).
Ils représentent actuellement plus de 16000 espèces réparties en 311 familles (selon la classification retenue par NELSON, 2006), et leur diversité morphologique est considérable : on trouve dans ce groupe des poissons aussi divers que le guppy (*Poecilia reticulata*), le poisson-lune (*Mola mola*), l'espadon (*Xiphias gladius*), la sole (*Solea solea*), l'hippocampe (*Hippocampus guttulatus*), le bar (*Dicentrarchus labrax*) ou la baudroie (*Lophius piscatorius*).

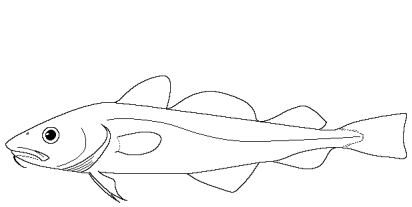
ROSEN (1973) divise les acanthomorphes en deux clades putatifs : les Acanthopterygii et les Paracanthopterygii. Ce deuxième groupe, qui rassemblait Batrachoidiformes, Gadiformes (morues ...), Gobiesociformes (porte-écuelles), Lophiiformes (baudroies ...), Ophidiiformes, Percopsiformes, Polymixiiformes, a été peu à peu démembré (STIASSNY, 1986; STIASSNY et MOORE, 1992; JOHNSON et PATTERSON, 1993). JOHNSON et PATTERSON (1993) distinguent deux clades au sein des Percomorpha : Les Smegmamorpha et les Perciformes. Au sein des Acanthopterygii, les Percomorpha font office de fourre-tout et en leur sein, les Perciformes sont également un groupe dans lequel tout ce qui n'a pas pu être placé ailleurs est mis. Une remise en cause de ces groupes par des données moléculaires était donc prévisible.

5.2. Nouvelle donne moléculaire

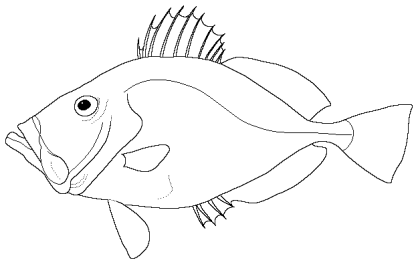
L'arrivée des phylogénies moléculaires a confirmé la polyphylie des Paracanthopterygii (WILEY *et al.*, 2000; CHEN *et al.*, 2003) et celle des Perciformes et des Smegmamorpha (MIYA *et al.*, 2001, 2003; CHEN *et al.*, 2003). De ces nouvelles données se dégagent peu à peu des relations de parenté

inattendues, comme la proximité entre Zeioidei et Gadiformes ou celle entre Tetraodontiformes et Lophiiformes. Un certain nombre de clades considérés comme fiables de par leur corroboration par des données indépendantes ont été désignés par des lettres par Wei-Jen CHEN et Agnès DETTAÏ :

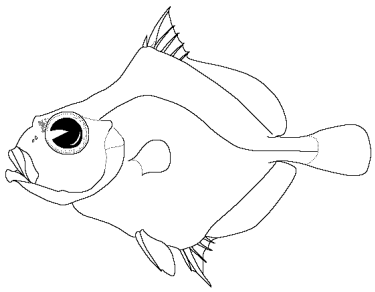
– A : Gadiformes et Zeioidei ;
– D : Blennioidei et Gobiesocidae ;
– E : Aulostomidae, Callionymidae, Dactylopteridae, Centriscidae, Mullidae et Syngnathidae ;
– F : Channidae, Anabantidae et Symbranchiformes ;
– Gu : Ammodytidae, Cheimarrichthyidae et Uranoscopidae ;
– H : Chiasmodontidae, Scombridae et Centrolophidae ;
– I : Cottoidei et Zoarcoidei ;
– Isc : Clade I, Triglidae et Gasterosteidae ;
– K : Notothenioidei et Percidae ;
– L : Carangidae, Echeneidae, Latidae, Menidae, Pleuronectiformes, Polynemidae et Sphyraenidae ;
– N : Caproidae, Chaetodontidae, Acanthuridae, Drepaneidae, Lophiiformes, Pomacanthidae, Siganidae et Tetraodontiformes
– Q : Clade D, Atherinomorpha, Cichlidae et Mugilidae ;
– X : Clade Isc, Clade K, Scorpaenidae, Serranidae et Trachinidae.



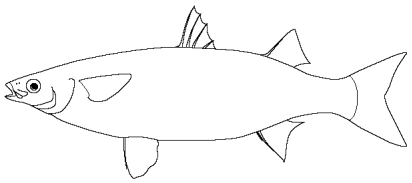
Gadidae



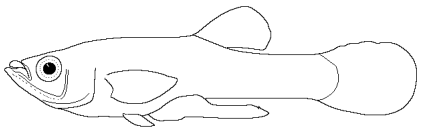
Zeidae



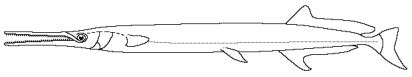
Oreosomatidae



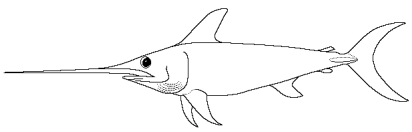
Mugilidae



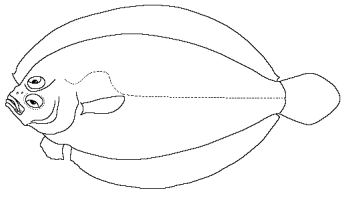
Poeciliidae



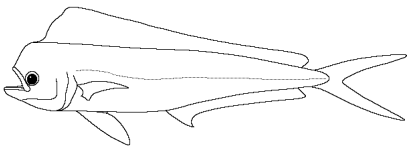
Belonidae



Xiphiidae



Bothidae



Coryphaenidae

6. Données utilisées

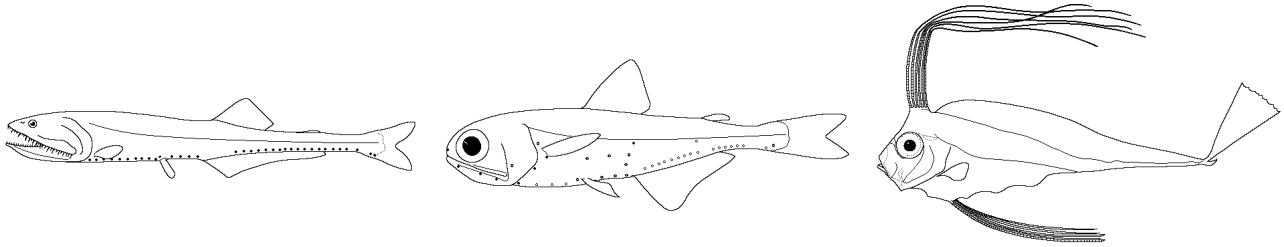
Le jeu de données utilisé dans la présente thèse est l'extension d'une partie des données déjà obtenues par Wei-Jen CHEN et Agnès DETTAÏ. Extension taxinomique (de nouvelles espèces ont été ajoutées) et ajout d'un nouveau marqueur, après quelques tâtonnements.

6.1. Taxons

Par rapport aux données déjà disponibles, l'échantillonnage a été densifié dans certains clades en particulier, mais les occasions d'élargir la diversité des familles représentées ont également été saisies. Certains taxons ont été ajoutés dans le but de « casser les branches longues » ; une espèce présentant des branches longues, en raison d'un taux d'évolution élevé de ses séquences, est plus fiablement placée si des espèces proches mais non sujettes à de telles déformations sont également présentes dans une analyse[1].
L'extension ciblée de l'échantillonnage visait les clade L (recherche du groupe frère des Pleuronectiformes), H (éclatement des Scombroidei entre clades L et H) et N (travail en coopération avec Francesco SANTINI sur la position des Tetraodontiformes) ainsi que les ex-trachinoïdes (éclatement de ce groupe et recherche du groupe frère des Notothenioidei).

Les groupes extérieurs les plus proches des acanthomorphes sont les Ateleopodiformes, les Myctophiformes et les Aulopiformes. Nous n'avons pas d'Ateleopodiformes dans notre échantillonnage. MIYA *et al.* (2005) obtiennent un regroupement des Myctophiformes avec les Lampridiformes (qui sont des acanthomorphes). *Bathypterois* (Aulopiformes) est donc notre plus proche groupe extérieur non ambigu.

L'échantillonnage taxinomique couvert dans la présente thèse est donné dans le tableau 6.1.



Gonostomatidae Myctophidae Trachipteridae

[1] Si les problèmes subsistent, il faut parfois se résoudre à enlever l'espèce qui pose problème et de la remplacer par l'espèce proche.

Tableau 6.1. Échantillonnage taxinomique couvert. Les ordres et les sous-ordres sont ceux retenus par NELSON (2006), et sont présentés dans l'ordre de ce même ouvrage. Les familles sont celles données dans `Fishbase` (FROESE et PAULY, 2006).

Ordre/Sous-ordre	Famille	Genre/Espèce	Nom vernaculaire
Argentiniformes			
Alepocephaloidei			
	Alepocephalidae	*Alepocephalus antipodianus* (Parrott, 1948)	
Stomiiformes			
Gonostomatoidei			
	Gonostomatidae	*Gonostoma bathyphilum* (Vaillant, 1884)	
Aulopiformes			Barracudines et poissons-lézards
Chlorophthalmoidei			
	Ipnopidae	*Bathypterois dubius* Vaillant, 1888	
Myctophiformes			Poissons-lanternes
	Myctophidae	*Electrona antarctica* (Günther, 1878)	
Lampriformes			Opahs et poissons-rubans
	Lampridae	*Lampris immaculatus* Gilchrist, 1904	Opah
	Trachipteridae	*Trachipterus arcticus* (Brünnich, 1788)	
	Regalecidae	*Regalecus glesne* Ascanius, 1772	Roi des harengs
Polymixiiformes			Poissons-chèvres
	Polymixiidae	*Polymixia lowei* Günther, 1859	
	Polymixiidae	*Polymixia nobilis* Lowe, 1838	
	Polymixiidae	*Polymixia sp.*	
Percopsiformes			Omiscos
	Aphredoderidae	*Aphredoderus sayanus* (Gilliams, 1824)	
Gadiformes			Morues, merlans et grenadiers
	Muraenolepididae	*Muraenolepis marmoratus* Günther, 1880	Gadomurène marbrée
	Macrouridae	*Coryphaenoides rupestris* Gunnerus, 1765	Grenadier de roche
	Macrouridae	*Trachyrincus murrayi* Günther, 1887	Grenadier-scie
	Moridae	*Mora moro* (Risso, 1810)	
	Merlucciidae	*Merluccius merluccius* (L., 1758)	Merlu
	Phycidae	*Phycis phycis* (L., 1766)	
	Lotidae	*Enchelyopus cimbrius* (L., 1766)	
	Lotidae	*Gaidropsarus novaezelandiae* (Hector, 1874)	
	Lotidae	*Gaidropsarus sp.*	
	Lotidae	*Gaidropsarus vulgaris* (Cloquet, 1824)	
	Gadidae	*Gadus morhua* L., 1758	Morue
	Gadidae	*Merlangius merlangus* (L., 1758)	Merlan
Ophidiiformes			Abadèches, aurins, brotules et donzelles
Ophidioidei			
	Carapidae	*Carapus boraborensis* (Kaup, 1856)	
	Carapidae	*Echiodon cryomargarites* Markle, Williams & Olney, 1983	
	Ophidiidae	*Lamprogrammus shcherbachevi* Cohen & Rohr, 1993	
Bythitoidei			
	Bythitidae	*Cataetyx laticeps* Koefoed, 1927	
Batrachoidiformes			Poissons-crapauds
	Batrachoididae	*Halobatrachus didactylus* (Bloch & Schneider, 1801)	

Tableau 6.1. (Suite)

Ordre/Sous-ordre	Famille	Genre/Espèce	Nom vernaculaire
Lophiiformes			Baudroies
Lophioidei			
	Lophiidae	*Lophius budegassa* Spinola, 1807	
	Lophiidae	*Lophius piscatorius* L., 1758	Baudroie
Antennarioidei			
	Antennariidae	*Antennarius striatus* (Shaw, 1794)	
Ogcocephaloidei			
	Himantolophidae	*Himantolophus groenlandicus* Reinhardt, 1837	
	Ceratiidae	*Ceratias holboelli* Krøyer, 1845	
Mugiliformes			Mulets
	Mugilidae	*Liza sp.*	
Atheriniformes			Athérines
	Atherinopsidae	*Menidia menidia* (L., 1766)	Capucette
	Bedotiidae	*Bedotia geayi* Pellegrin, 1907	
Beloniformes			Orphies et poissons-volants
	Adrianichthyidae	*Oryzias latipes* (Temminck & Schlegel, 1946)	Medaka
	Exocoetidae	*Cheilopogon heterurus* (Rafinesque, 1810)	Exocet méditerranéen
	Belonidae	*Belone belone* (L., 1761)	Orphie
Cyprinodontiformes			Gambusies, guppys, killis et mollys
	Anablepidae	*Anableps anableps* (L., 1758)	Quatre-yeux à grandes écailles
	Poeciliidae	*Poecilia reticulata* Peters, 1859	Guppy
Stephanoberyciformes			Poromitres et poissons-baleines
	Rondeletiidae	*Rondeletia sp.*	
	Barbourisiidae	*Barbourisia rufa* Parr, 1945	
Beryciformes			Poissons-écureuils et poissons-soldats
Trachichthyoidei			
	Anomalopidae	*Photoblepharon palpebratum* (Boddaert, 1781)	
	Diretmidae	*Diretmoides sp.*	
	Trachichthyidae	*Hoplostethus atlanticus* Collett, 1889	
	Trachichthyidae	*Hoplostethus mediterraneus* Cuvier, 1829	
Berycoidei			
	Berycidae	*Beryx splendens* Lowe, 1834	
Holocentroidei			
	Holocentridae	*Myripristis botche* Cuvier, 1929	
	Holocentridae	*Myripristis sp.*	
	Holocentridae	*Myripristis violacea* Bleeker, 1851	
Zeiformes			
	Oreosomatidae	*Neocyttus helgae* (Holt & Byrne, 1908)	
	Grammicolepididae	*Grammicolepis brachiusculus* Poey, 1873	
	Zeidae	*Zenopsis conchifera* (Lowe, 1852)	
	Zeidae	*Zeus faber* L., 1758	Saint-Pierre
Gasterosteiformes			Épinoches et Hippocampes
Gasterosteoidei			
	Gasterosteidae	*Gasterosteus aculeatus* L., 1758	Épinoche à trois épines
	Gasterosteidae	*Spinachia spinachia* (L., 1758)	Épinoche de mer
	Indostomidae	*Indostomus paradoxus* Prashad & Mukerji, 1929	

Tableau 6.1. (Suite)

Ordre/Sous-ordre	Famille	Genre/Espèce	Nom vernaculaire
Syngnathoidei			
	Syngnathidae	*Hippocampus guttulatus* Cuvier, 1829	Hippocampe moucheté
	Syngnathidae	*Nerophis lumbriciformis* (Jenyns, 1835)	
	Syngnathidae	*Nerophis ophidion* (L., 1758)	
	Syngnathidae	*Nerophis sp.*	
	Syngnathidae	*Syngnathus typhle* L., 1758	
	Fistulariidae	*Fistularia petimba* Lacépède, 1803	Cornette rouge
	Aulostomidae	*Aulostomus chinensis* (L., 1766)	Poisson trompette
	Centriscidae	*Aeoliscus strigatus* (Günther, 1861)	Canif
	Centriscidae	*Macroramphosus scolopax* (L., 1758)	Bécasse de mer
Symbranchiformes			
Symbranchoidei			
	Symbranchidae	*Monopterus albus* (Zuiew, 1793)	
Mastacembeloidei			
	Mastacembelidae	*Mastacembelus erythrotaenia* Bleeker, 1850	
Scorpaeniformes			Rascasses, grondins, limaces et terpugas
Dactylopteroidei			
	Dactylopteridae	*Dactylopterus volitans* (L., 1758)	Grondin volant
Scorpaenoidei			
	Sebastidae	*Sebastes sp.*	
	Scorpaenidae	*Pontinus longispinis* Goode & Bean, 1896	
	Scorpaenidae	*Scorpaena onaria* Jordan & Snyder, 1900	
	Synanceiidae	*Synanceia verrucosa* Bloch & Schneider, 1801	Poisson pierre
	Congiopodidae	*Zanclorhynchus spinifer* Günther, 1880	
Platycephaloidei			
	Triglidae	*Chelidonichthys lucernus* (L., 1758)	Grondin perlon
Cottoidei			
	Cottidae	*Taurulus bubalis* (Euphrasen, 1786)	Chabot de mer
	Agonidae	*Agonopsis chiloensis* (Jenyns, 1840)	
	Agonidae	*Xeneretmus latifrons* (Gilbert, 1890)	
	Psychrolutidae	*Cottunculus thomsonii* (Günther, 1882)	Cotte blème
	Cyclopteridae	*Cyclopterus lumpus* L., 1758	Lompe
	Liparidae	*Liparis fabricii* Krøyer, 1847	Limace gélatineuse
Perciformes			
Percoidei			
	Centropomidae	*Centropomus undecimalis* (Bloch, 1792)	Crossie blanc
	Lateolabracidae	*Lateolabrax japonicus* (Cuvier, 1828)	
	Lateolabracidae	*Lateolabrax sp.*	
	Latidae	*Lates calcarifer* (Bloch, 1970)	Perche barramundi
	Latidae	*Lates niloticus* (L., 1758)	Perche du nil
	Moronidae	*Dicentrarchus labrax* (L., 1758)	Bar
	Moronidae	*Morone saxatilis* (Walbaum, 1792)	Bar d'Amérique
	Percichthyidae	*Howella brodiei* Ogilby, 1899	
	Serranidae	*Acanthistius brasilianus* (Cuvier, 1828)	Serran argentin
	Serranidae	*Cephalopholis urodeta* (Forster, 1801)	Mérou tacheté à queue noire
	Serranidae	*Dermatolepis dermatolepis* (Boulenger, 1895)	Mérou cuir

Tableau 6.1. (Suite)

Ordre/Sous-ordre	Famille	Genre/Espèce	Nom vernaculaire
	Serranidae	*Epinephelus aeneus* (Geoffroy Saint-Hilaire, 1817)	Mérou blanc
	Serranidae	*Holanthias chrysostictus* (Günther, 1872)	
	Serranidae	*Liopropoma fasciatum* Bussing, 1980	
	Serranidae	*Niphon spinosus* Cuvier, 1828	Ara
	Serranidae	*Plectropomus leopardus* (Lacépède, 1802)	Saumonée léopard
	Serranidae	*Pogonoperca punctata* (Valenciennes, 1830)	Savon barbu
	Serranidae	*Pseudanthias squamipinnis* (Peters, 1855)	Barbier rouge
	Serranidae	*Rypticus saponaceus* (Bloch & Schneider, 1801)	Grand savon
	Serranidae	*Serranus accraensis* (Norman, 1931)	Serran ganéen
	Callanthiidae	*Callanthias ruber* (Rafinesque, 1810)	Barbier perroquet
	Plesiopidae	*Assessor flavissimus* Allen & Kuiter, 1976	
	Centrarchidae	*Lepomis gibbosus* (L., 1758)	Perche-soleil
	Percidae	*Gymnocephalus cernuus* (L., 1758)	Grémille
	Percidae	*Perca fluviatilis* L., 1758	Perche
	Priacanthidae	*Priacanthus arenatus* Cuvier, 1829	Beauclaire soleil
	Epigonidae	*Epigonus telescopus* (Risso, 1810)	Sonneur commun
	Apogonidae	*Apogon quadrifasciatus* Cuvier, 1828	
	Apogonidae	*Sphaeramia nematoptera* (Bleeker, 1856)	
	Malacanthidae	*Lopholatilus chamaeleonticeps* Goode & Bean, 1879	Tile chameau
	Sillaginidae	*Sillago sihama* (Forsskål, 1775)	Pêche-madame argenté
	Coryphaenidae	*Coryphaena equiselis* L., 1758	Coryphène dauphin
	Coryphaenidae	*Coryphaena hippurus* L., 1758	
	Echeneidae	*Echeneis naucrates* L., 1758	Rémora
	Carangidae	*Chloroscombrus chrysurus* (L., 1766)	Sapater
	Carangidae	*Gnathanodon speciosus* (Forsskål, 1755)	Carangue jaune
	Carangidae	*Selene dorsalis* (Gill, 1863)	Musso africain
	Carangidae	*Trachinotus ovatus* (L., 1758)	Palomine
	Carangidae	*Trachurus trachurus* (L., 1758)	Chinchard
	Menidae	*Mene maculata* (Bloch & Schneider, 1801)	Luneur
	Leiognathidae	*Leiognathus fasciatus* (Lacépède, 1803)	Sap-sap rayé
	Bramidae	*Pterycombus brama* Fries, 1837	
	Lutjanidae	*Apsilus fuscus* Valenciennes, 1830	Vivaneau fourche
	Lutjanidae	*Lutjanus sebae* (Cuvier, 1816)	Vivaneau bourgeois
	Caesionidae	*Pterocaesio digramma* (Bleeker, 1864)	Fusilier à deux bandes jaunes
	Datnioididae	*Datnioides polota* (Hamilton, 1822)	
	Haemulidae	*Pomadasys perotaei* (Cuvier, 1830)	Grondeur perroquet
	Sparidae	*Spondyliosoma cantharus* (L., 1758)	Dorade grise
	Sciaenidae	*Argyrosomus regius* (Asso, 1801)	Maigre
	Sciaenidae	*Johnius sp.*	
	Sciaenidae	*Micropogonias manni* (Moreno, 1970)	
	Sciaenidae	*Micropogonias sp.*	
	Sciaenidae	*Sciaena sp.*	Courbine
	Polynemidae	*Pentanemus quinquarius* (L., 1758)	Capitaine royal
	Mullidae	*Mullus surmuletus* L., 1758	Rouget de roche
	Toxotidae	*Toxotes sp.*	Poisson archer
	Monodactylidae	*Monodactylus sp.*	

Tableau 6.1. (Suite)

Ordre/Sous-ordre	Famille	Genre/Espèce	Nom vernaculaire
	Kyphosidae	*Microcanthus strigatus* (Cuvier, 1831)	
	Chaetodontidae	*Chaetodon semilarvatus* Cuvier, 1831	
	Drepaneidae	*Drepane africana* Osório, 1892	Forgeron ailé
	Pomacanthidae	*Holacanthus ciliaris* (L., 1758)	Demoiselle royale
	Pomacanthidae	*Pomacanthus maculosus* (Forsskål, 1775)	
	Terapontidae	*Pelates quadrilineatus* (Bloch, 1790)	Violon crépuscule
	Cheilodactylidae	*Nemadactylus monodactylus* (Carmichael, 1819)	Castanette de Saint Paul
	Aplodactylidae	*Aplodactylus punctatus* Valenciennes, 1832	
	Cepolidae	*Cepola macrophthalma* (L., 1758)	
Elassomatoidei			
	Elassomatidae	*Elassoma zonatum* Jordan, 1877	
Labroidei			
	Cichlidae	*Haplochromis nubilus* (Boulenger, 1906)	
	Cichlidae	*Haplochromis sp.*	
	Pomacentridae	*Dascyllus trimaculatus* (Rüppel, 1829)	
	Labridae	*Labrus bergylta* Ascanius, 1767	Vieille
	Labridae	*Xyrichtys novacula* (L., 1758)	
	Scaridae	*Scarus hoefleri* (Steindachner, 1881)	Perroquet de Guinée
Zoarcoidei			
	Zoarcidae	*Austrolycus depressiceps* Regan, 1913	
	Zoarcidae	*Lycodapus antarcticus* Tomo, 1982	
	Pholidae	*Pholis gunnellus* (L., 1758)	Gonelle
	Anarhichadidae	*Anarhichas lupus* L., 1758	Loup atlantique
Notothenioidei			
	Nototheniidae	*Notothenia coriiceps* Richardson, 1844	Bocasse jaune
	Nototheniidae	*Trematomus bernachii* Boulenger, 1902	Bocasson émeraude
	Bovichtidae	*Bovichtus variegatus* Richardson, 1846	
	Bovichtidae	*Cottoperca gobio* (Günther, 1861)	
	Bovichtidae	*Pseudaphritis urvillii* (Valenciennes, 1832)	
	Eleginopsidae	*Eleginops maclovinus* (Cuvier, 1830)	Guite de patagonie
	Channichthyidae	*Chionodraco hamatus* (Lönnberg, 1905)	
	Channichthyidae	*Neopagetopsis ionah* Nybelin, 1947	
	Channichthyidae	*Pagetopsis macropterus* (Boulenger, 1907)	
Trachinoidei			
	Chiasmodontidae	*Kali macrura* (Parr, 1933)	
	Champsodontidae	*Champsodon snyderi* Franz, 1910	
	Pinguipedidae	*Parapercis clathrata* Ogilby, 1910	
	Pinguipedidae	*Pinguipes chilensis* Valenciennes, 1833	
	Cheimarrichthyidae	*Cheimarrichthys fosteri* Haast, 1874	
	Trachinidae	*Echiichthys vipera* (Cuvier, 1829)	Petite vive
	Trachinidae	*Trachinus draco* L., 1758	Grande vive
	Ammodytidae	*Ammodytes tobianus* L., 1758	Lançon équille
	Uranoscopidae	*Uranoscopus albesca* Regan, 1915	
Blennioidei			
	Tripterygiidae	*Forsterygion lapillum* Hardy, 1989	
	Tripterygiidae	*Tripterygion delaisi* Cadenat & Blache, 1970	

Tableau 6.1. (Suite)

Ordre/Sous-ordre	Famille	Genre/Espèce	Nom vernaculaire
	Blenniidae	*Parablennius gattorugine* (L., 1758)	Blennie paon
	Blenniidae	*Salaria pavo* (Risso, 1810)	
Gobiesocoidei			Porte-écuelles
	Gobiesocidae	*Apletodon dentatus* (Facciolà, 1887)	
	Gobiesocidae	*Aspasma minima* (Döderlein, 1887)	
	Gobiesocidae	*Lepadogaster lepadogaster* (Bonnaterre, 1788)	
Callionymoidei			
	Callionymidae	*Callionymus lyra* L., 1758	Dragonnet lyre
	Callionymidae	*Callionymus schaapii* Bleeker, 1852	
Gobioidei			
	Eleotridae	*Ophiocara porocephala* (Valenciennes, 1837)	
	Gobiidae	*Favonigobius reichei* (Bleeker, 1853)	
	Gobiidae	*Periophthalmus barbarus* (L., 1766)	Sauteur de vase
	Gobiidae	*Pomatoschistus minutus* (Pallas, 1770)	Gobie des sables
	Gobiidae	*Pomatoschistus sp.*	
	Gobiidae	*Valenciennea strigata* (Broussonet, 1782)	
	Microdesmidae	*Ptereleotris zebra* (Fowler, 1938)	
Acanthuroidei			
	Scatophagidae	*Selenotoca multifasciata* (Richardson, 1846)	
	Siganidae	*Siganus vulpinus* (Schlegel & Müller, 1845)	Picot renard
	Luvaridae	*Luvarus imperialis* Rafinesque, 1810	
	Acanthuridae	*Ctenochaetus sp.*	
	Acanthuridae	*Ctenochaetus striatus* (Quoy & Gaimard, 1825)	Chirurgien strié
	Acanthuridae	*Naso lituratus* (Forster, 1801)	
Scombroidei			
	Sphyraenidae	*Sphyraena sphyraena* (L., 1758)	Bécune
	Trichiuridae	*Aphanopus carbo* Lowe, 1839	Sabre noir
	Scombridae	*Scomber japonicus* Houttuyn, 1782	Maquereau espagnol
	Xiphiidae	*Xiphias gladius* L., 1758	Espadon
Stromateoidei			
	Centrolophidae	*Psenopsis anomala* (Temminck & Schlegel, 1844)	
	Centrolophidae	*Schedophilus medusophagus* (Cocco, 1839)	Rouffe des méduses
	Nomeidae	*Cubiceps gracilis* (Lowe, 1843)	Dérivant
	Stromateidae	*Pampus argenteus* (Euphrasen, 1788)	Aileron argenté
Anabantoidei			
	Anabantidae	*Ctenopoma sp.*	
Channoidei			
	Channidae	*Channa sp.*	
	Channidae	*Channa striata* (Bloch, 1793)	
Caproidei			
	Caproidae	*Antigonia capros* Lowe, 1843	
	Caproidae	*Capros aper* (L., 1758)	Sanglier
Pleuronectiformes			Poissons-plats
Psettodoidei			
	Psettodidae	*Psettodes belcheri* Bennett, 1831	
	Psettodidae	*Psettodes erumei* (Bloch & Schneider, 1801)	Turbot épineux indien

Tableau 6.1. (Suite)

Ordre/Sous-ordre	Famille	Genre/Espèce	Nom vernaculaire
	Psettodidae	*Psettodes sp.*	
Pleuronectoidei			
	Citharidae	*Citharus linguatula* (L., 1758)	Feuille
	Paralichthyidae	*Syacium micrurum* Ranzani, 1842	Rombou de canal
	Scophthalmidae	*Scophthalmus rhombus* (L., 1758)	Barbue
	Scophthalmidae	*Zeugopterus punctatus* (Bloch, 1787)	Targeur
	Bothidae	*Arnoglossus imperialis* (Rafinesque, 1810)	
	Bothidae	*Bothus podas* (Delaroche, 1809)	Rombou podas
	Achiridae	*Trinectes maculatus* (Bloch & Schneider, 1801)	Sole bavoche
	Soleidae	*Microchirus frechkopi* Chabanaud, 1952	Sole de Frechkop
	Soleidae	*Microchirus variegatus* (Donovan, 1808)	Sole-perdrix
	Soleidae	*Solea solea* (L., 1758)	Sole
Tetraodontiformes			Balistes, poissons-globes et poissons-coffres
Triacanthodoidei			
	Triacanthodidae	*Triacanthodes anomalus* (Temminck & Schlegel, 1850)	
	Triacanthodidae	*Triacanthodes sp.*	
Balistoidei			
	Balistidae	*Balistes sp.*	
	Ostraciidae	*Ostracion cubicus* L., 1758	Poisson-coffre
	Ostraciidae	*Ostracion sp.*	
Tetraodontoidei			
	Tetraodontidae	*Lagocephalus laevigatus* (L., 1766)	Compère lisse
	Tetraodontidae	*Lagocephalus lagocephalus* (L., 1758)	Compère lièvre
	Tetraodontidae	*Takifugu rubripes* (Temminck & Schlegel, 1850)	Fugu
	Tetraodontidae	*Tetraodon nigroviridis* Marion de Procé, 1822	
	Molidae	*Mola mola* (L., 1758)	Poisson-lune

6.2. Marqueurs

Un certain nombre de marqueurs ont été envisagés, mais tous n'ont pas été retenus. Certains n'ont pas été utilisés pour la phylogénie générale des acanthomorphes mais sur une partie de l'échantillonnage seulement[2].

6.2.1. Les ADN ribosomiques

L'ADN ribosomique 28S est un marqueur nucléaire présent en copies répétées, ce qui permettait son séquençage avant l'invention de la PCR. C'est donc naturellement que les premières études sur la phylogénie des acanthomorphes l'ont utilisé (WILEY *et al.*, 2000; CHEN *et al.*, 2003). Les ADN ribosomiques mitochondriaux 12S et 16S étaient déjà séquencés pour un certain

[2]Dans le cadre d'une étude centrée sur la position phylogénétique des Tetraodontiformes en collaboration avec Francesco SANTINI, Cmos, TMO et RAG1 ont été utilisés (article en préparation).

nombre d'espèces quand Wei-Jen CHEN a commencé sa thèse. Le jeu de données a donc été enrichi au cours des thèses de Wei-Jen CHEN et d'Agnès DETTAÏ.
Cependant, les ARN ribosomiques ne sont pas codants. On ne peut donc pas se baser sur un cadre de lecture pour les aligner[3]. Par ailleurs SMITH et WHEELER (2004, 2006), MIYA *et al.* (2005), HOLCROFT (2005), CHEN *et al.* (2007), SMITH et CRAIG (2007) et YAMANOUE *et al.* (2007) utilisent ces marqueurs, ce qui rend les comparaisons de résultats entre équipes moins parlantes[4]. La thèse d'Agnès DETTAÏ ayant montré le faible intérêt de ces marqueurs à l'échelle des acanthomorphes, ils n'ont pas été repris pour l'extension de l'échantillonnage taxinomique.

6.2.2. La Rhodopsine

C'est en fait un rétrogène du gène de la rhodopsine (protéine constituant de certains photorécepteurs). Il n'y a donc pas d'introns. Les sondes (CHEN *et al.*, 2003) sont relativement bonnes et le jeu de données déjà constitué assez conséquent, d'autant plus qu'une portion de ce marqueur est utilisée par le projet `Fishtrace` (`http://www.fishtrace.org/gb/main.htm`).
CHEN *et al.* (2003) ont montré un biais dans l'usage des différentes bases selon les taxons pour ce gène, ce qui perturbe la reconstruction phylogénétique. Ce marqueur a néanmoins été gardé en raison du large échantillonnage taxinomique déjà séquencé, de la relative facilité à obtenir de nouvelles séquences, et de la résolution globalement satisfaisante qu'il apporte pour certains clades.

6.2.3. MLL4

Des amorces pour ce fragment de l'exon 26 du gène *Mixed-Lineage Leukemia-Like 4*[5] ont été mises au point par Agnès DETTAÏ à partir des séquences de *Danio rerio*, *Takifugu rubripes* et *Tetraodon nigroviridis*. Ce marqueur est efficace : avec environ 500 nucléotides, il permet d'obtenir des résultats en assez bon accord avec ceux obtenus par MIYA *et al.* (2003, 2005) à partir de génomes mitochondriaux complets. Il a donc été gardé dans la présente thèse.

6.2.4. IRBP

Ce marqueur, fragment du module 1 du gène 2 de l'*Inter-Retinoid Binding Protein* (une protéine transportant le rétinal et le rétinol dans la rétine), était déjà utilisé en phylogénie des mammifères

[3] Outre les problèmes directement liés à l'alignement, l'absence de cadre de lecture rend par ailleurs la détection d'erreurs de séquençage ou d'interprétation du chromatogramme plus difficile. Certains préfèrent cependant ne pas se fier à un cadre de lecture pour « nettoyer » leurs séquences pour ne pas risquer de sur-interpréter le chromatogramme.

[4] L'équipe de MIYA *et al.*, bien que séquençant des génomes mitochondriaux complets excluait ensuite les ADN ribosomiques 12S et 16S de leurs analyses en raison des difficultés d'alignement. Depuis leur publication de 2005, ils ont réintégré ces deux marqueurs.

[5] Ce gène était initialement nommé MLL, tout court, avant la découverte des autres gènes MLL.

(DEBRY et SAGEL, 2001). Il a été adapté par DETTAÏ et LECOINTRE (2008) aux acanthomorphes, pour lesquels il semble bien résoudre les relations de parenté. Il donc a été retenu ici, malgré quelques difficultés en PCR ; un nombre relativement élevé d'amorces ayant dû être défini.

6.2.5. C-mos

Ce gène, impliqué dans la régulation du cycle cellulaire, a été utilisé avec succès en phylogénie des amphibiens (GRAYBEAL, 1994), des squamates (SAINT *et al.*, 1998) ou des oiseaux (COOPER et PENNY, 1997). Son adaptation aux acanthomorphes a été tentée par Agnès DETTAÏ puis par Francesco SANTINI et moi-même. C-mos s'est avéré trop variable, à toutes les positions de codon, ce qui rend l'obtention d'amorces efficaces assez difficile. Les séquences obtenues avec difficulté ont donné lieu à des résultats préliminaires décevants. C-mos n'a donc pas été retenu pour la phylogénie des acanthomorphes.

6.2.6. TMO-4c4

Ce petit fragment (environ 500 nucléotides), facile à amplifier en PCR (STREELMAN et KARL, 1997), est beaucoup utilisé à moyenne échelle chez les acanthomorphes, voire à grande échelle en combinaison avec d'autres gènes (SMITH et WHEELER, 2004). Toutefois, dans ses papiers postérieurs à 2004, Leo SMITH n'utilise plus TMO-4c4 (SMITH et WHEELER, 2006; SMITH et CRAIG, 2007). Des problèmes de contamination récurrents ont jeté le doute sur un certain nombre de séquences. En outre, les résultats préliminaires n'étaient pas très encourageants. TMO-4c4 n'a donc pas été utilisé ici.

6.2.7. RAG1

Ce marqueur est un fragment d'un gène impliqué dans la production des immunoglobulines (*Recombination-Activating Gene 1*). Il est très utilisé en phylogénie des acanthomorphes (HOLCROFT, 2004, 2005; CHEN *et al.*, 2007; HOLCROFT et WILEY, 2008). Il n'a pas été utilisé dans la présente thèse car un important jeu de données pour ce marqueur est en cours de constitution (Wei-Jen CHEN, communication personnelle) ; il est plus efficace de ne pas faire deux fois le même travail. À l'avenir, il pourrait s'avérer utile de l'intégrer dans une analyse de fiabilité.

6.2.8. RNF213

Le marqueur RNF213 (*Ring Finger Protein 213*[6]) fait partie d'un ensemble de marqueurs sélectionnés par Agnès DETTAÏ à partir de l'examen des génomes complets disponibles sur

[6]RNF213 était initialement nommé C17 orf 27 en raison de sa localisation sur le génome humain.

`Ensembl` (`http://www.ensembl.org/biomart/martview`). Il s'agit d'une portion d'un gène codant une protéine à doigt de zinc. Ce marqueur a été testé dans le cadre de cette thèse.

6.3. Travail moléculaire

Le travail moléculaire repose sur des échantillons de tissus accumulés au fil des ans lors de campagnes de collecte ou donnés par des collègues. Les échantillons sont la plupart stockés dans l'alcool à 70°, souvent au frais, parfois congelés. L'ADN génomique est extrait des tissus, puis utilisé pour l'amplification par PCR (MULLIS et FALOONA, 1987) des marqueurs d'intéret. Les amplifiats jugés convenables sont ensuite séquencés. Les séquences sont enfin traitées informatiquement pour produire les matrices de caractères à analyser.

6.3.1. Extraction

Une bonne partie des extraits d'ADN étaient déjà disponibles, suite aux travaux de Wei-Jen CHEN et d'Agnès DETTAÏ. Quelques uns sont issus d'échanges avec Leo SMITH ou Wei-Jen CHEN. Les nouveaux extraits ont été obtenus selon deux méthodes ; extraction classique au CTAB ou à la machine (ABI PRISM 6100). Dans les deux cas, le principe général est le même et commence toujours par une fragmentation au scalpel d'un petit morceau de tissu. Le tissu est ensuite digéré à $60^{\circ}C$. La suite a pour but de séparer l'ADN des autres constituants du tissu et de le récupérer dans une solution ne nuisant pas au bon déroulement de l'amplification par PCR. L'ADN génomique est éventuellement visualisé aux ultra-Violets après migration par électrophorèse sur gel d'agarose en présence de BET[7]. Les extraits sont conservés à $-20^{\circ}C$.

6.3.2. Amplification par PCR

Cette opération consiste à obtenir un très grand nombre de copies d'un fragment d'ADN compris entre les sites d'hybridation de deux oligonucléotides bien choisis ; les amorces (aussi appelées sondes, voir tableau 6.2).

Ces amorces, une fois hybridées sur l'ADN, permettent l'initialisation de la réplication de l'ADN par une polymérase en présence de désoxyribonucléotides tri-phosphate et de divers additifs, et chacun des réplicats ainsi obtenus peut à son tour être répliqué puisqu'il comprend les sites d'hybridation des amorces. Il faut que les fragments nouvellement synthétisés se séparent de

[7]Le BET, ou bromure d'éthidium est un agent intercalant ; c'est-à-dire qu'il s'intercale entre les bases azotées de l'ADN, ce qui rend ce dernier fluorescent. Une conséquence de cette propriété intercalante est que le BET est également un agent mutagène ; sa manipulation se fait sous haute protection (blouse, décapsuleur pour le tube, gants en nitrile, lunettes de protection, poubelle spéciale).

Tableau 6.2. Sondes utilisées.

Nom	Gène	Séquence	Source
C17F3111	RNF213	GCTGACTGGATTYAAAACCTT	nouveau
C17F3128	RNF213	CCTTTGTGGTGGAYTTYATGAT	nouveau
C17F3150	RNF213	WCTGATGGCNAARGACTTTGC	nouveau
C17R4036	RNF213	GGRATRGANCCNAGCTTTTCAT	nouveau
C17R4096	RNF213	CCANACCAGAGGGATCATRCT	nouveau
C17R4111	RNF213	AACTGTCCAAARTCCCACAC	nouveau
IRBPL1338	IRBP	GTGRAAGGAGAYTTTGATCAGCTC	DETTAÏ et LECOINTRE (2008)
IRBPL922	IRBP	TGATNNCRGTKGCRAGGGCATC	DETTAÏ et LECOINTRE (2008)
IRBPL936	IRBP	CACGGAGGYTGAYNATCTTGAT	DETTAÏ et LECOINTRE (2008)
IRBPL953	IRBP	CNGGAAYYTGARCACGGAGG	DETTAÏ et LECOINTRE (2008)
IRBPU104	IRBP	ATAGTYNTGGACAANTACTGCTC	DETTAÏ et LECOINTRE (2008)
IRBPU110	IRBP	TGGACAAYTACTGCTCRCCAGA	DETTAÏ et LECOINTRE (2008)
MllL2080	MLL4	GTGAACTCMAYCAGTCCTCC	nouveau
MllL2105	MLL4	ACCYTGCGTTGGGARGTGG	nouveau
MllL2158	MLL4	ARAGTAGTGGGATCYAGRTACAT	DETTAÏ et LECOINTRE (2005)
MllU1477	MLL4	AGYCCAGCRGTCATCAAACC	DETTAÏ et LECOINTRE (2005)
MllU1499	MLL4	GTCAATCAGCAGTTCCAGC	DETTAÏ et LECOINTRE (2005)
MllU1570	MLL4	CCCYCAAAAKATCARTGCCAC	nouveau
MllU1590	MLL4	CRGGRGTGATNGACACCAGC	nouveau
Rh1039r	Rhodopsine	TGCTTGTTCATGCAGATGTAGA	CHEN *et al.* (2003)
Rh1073r	Rhodopsine	CCRCAGCACARCGTGGTGATCATG	CHEN *et al.* (2003)
Rh193	Rhodopsine	CNTATGAATAYCCTCAGTACTACC	CHEN *et al.* (2003)
Rh667r	Rhodopsine	AYGAGCACTGCATGCCCT	CHEN *et al.* (2003)

leur matrice pour qu'une nouvelle phase d'hybridation des amorces et de réplication de l'ADN puisse avoir lieu. La PCR est donc une succession de cycles (entre 25 et 65) à trois étapes :

1. dénaturation de l'ADN par chauffage (aux alentours de $94^{\circ}C$), afin que les brins d'ADN se séparent pour laisser les sites d'hybridation des amorces accessibles ;
2. hybridation des amorces, à plus basse température (entre $40^{\circ}C$ et $65^{\circ}C$) ;
3. élongation par la polymérase, à une température comprise entre la température d'hybridation et celle de dénaturation (en général vers $72^{\circ}C$, mais avec certaines polymérases, pour de longues PCR, cette étape peut se faire à $68^{\circ}C$, ce qui évite une trop importante dégradation de l'enzyme).

Ces cycles sont précédés par une étape de dénaturation initiale plus longue, afin de bien séparer les brins de l'ADN génomique, et sont suivis d'une étape d'élongation finale, qui a pour but d'éviter que certains fragments amplifiés soient incomplets, ce qui perturberait le séquençage.

La qualité de l'amplifiat est jugée par migration sur gel (voir page 76) accompagnée d'un marqueur de taille. On suppose la réaction de PCR réussie quand le fragment d'ADN obtenu a la taille attendue d'après la position des sites d'hybridation des amorces[8] et n'est pas accompagné d'amplifiats non désirés qui pourraient perturber le séquençage.
La qualité de l'amplification dépend d'un certain nombre de facteurs, comme la qualité de l'extrait d'ADN utilisé, l'adéquation des amorces à la séquence de leurs sites d'hybridation chez l'individu séquencé, la température d'hybridation, les proportions des réactifs, la cinétique du thermocycleur[9] … Ceci peut nécessiter de longs tâtonnements avant l'obtention d'un bel amplifiat pour chaque espèce[10].

6.3.3. Séquençage

Le séquençage selon la méthode de SANGER *et al.* (1977) repose sur le même principe que la PCR, en utilisant comme matrice de départ l'amplifiat purifié, et en ajoutant au milieu de réaction des didésoxyribonucléotides. Ces derniers n'ont pas de terminaison 3'-OH, ce qui bloque l'élongation du brin d'ADN en train d'être synthétisé. Il en résulte un mélange de brins d'ADN de diverses longueurs qui seront séparés par électrophorèse. Dans le cas du séquençage par séquenceur automatique, les quatre types des didésoxyribonucléotides sont liés chacun à un fluorochrome différent. Les amplifiats de la réaction de séquençage sont soumis à une électrophorèse. Les brins de différentes tailles passent alors successivement devant un détecteur

[8] Sauf cas particuliers, la taille des marqueurs est sensiblement la même entre taxons.
[9] L'appareil qui chauffe et refroidit les tubes où se déroule la PCR est plus ou moins rapide ou lent à porter le mélange réactionnel à la température voulue au moment voulu, ce qui influe par exemple sur la qualité de l'hybridation des sondes et fait qu'un programme de PCR ne donne pas forcément les mêmes résultats d'un thermocycleur à l'autre. Même l'épaisseur des tubes semble avoir une influence.
[10] Un certain nombre d'échecs à la PCR ont été une motivation supplémentaire pour tenter de rendre l'indice de répétition tolérant aux taxons manquants

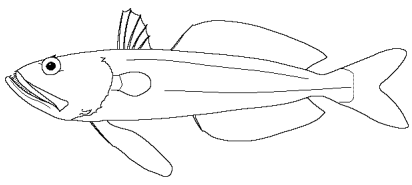
qui enregistre l'intensité et la couleur du signal correspondant au nucléotide terminant le brin. Ainsi, on obtient un chromatogramme qui, après traitement informatique, permet la lecture de la succession des bases de l'ADN séquencé.

Les premières séquences de la thèse ont été obtenues sur le séquenceur du SSM, qui était relativement vieux et peu précis ; les séquences étaient souvent de mauvaise qualité. Au cours de la thèse, un partenariat avec le Génoscope a permis de confier le séquençage à cet organisme, qui dispose de matériel performant.

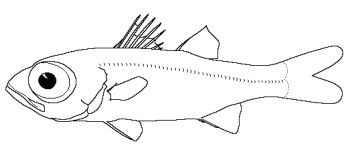
6.4. Nettoyage et alignement

Le « nettoyage » des séquences est la vérification visuelle de la bonne interpétation automatique des chromatogrammes. Les éventuelles erreurs d'interprétation ont été repérées et corrigées à l'aide du logiciel `Sequencher`. Tous les gènes séquencés étant codants, le cadre de lecture a été pris en compte lors de cette opération, en vérifiant qu'aucun codon « stop » ne se trouvait en milieu de gène. Une telle démarche est critiquable car elle suppose *a priori* que le gène est effectivement codant chez tous les individus séquencés. Elle permet toutefois de détecter certaines erreurs réelles dans la séquence, probablement beaucoup plus souvent qu'elle ne conduit à des erreurs supplémentaires d'interprétation des chromatogrammes. En général, les codons « stop » prématurés s'avèrent correspondre à des pics ambigus dans les chromatogrammes. Les parties trop peu fiables ainsi que les zones correspondant aux amorces ont été codées en caractères manquants.

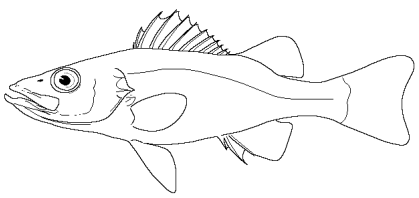
Au fur et à mesure que les séquences étaient considérées comme « propres », elles étaient alignées avec celles obtenues sur `Genbank`[11] ou `Ensembl`[12] sous `Se-Al` (RAMBAUT, 2002), en prenant soin de toujours les nommer selon un format `Genre_espece_option` ou « `_option` » est une extension optionelle permettant de distinguer deux séquences d'individus distincts d'une même espèce, par exemple un numéro d'extraction. Hormis quelques sites du marqueur MLL4, l'alignement est trivial pour les quatre marqueurs. Les parties manquantes des séquences incomplètes ont été codées par des points d'interrogation et l'alignement final a été exporté au format `Nexus`.



Champsodontidae



Epigonidae



Percichthyidae

[11] `http://www.ncbi.nlm.nih.gov/`
[12] `http://www.ensembl.org/index.html`

7. Analyses

L'analyse des données se décompose en deux phases. L'analyse primaire est l'obtention directe d'arbres à partir des jeux de données. L'analyse secondaire est celle qui, à partir de la comparaison des arbres obtenus lors de l'analyse primaire, vise à identifier les clades fiables.

7.1. Analyse primaire

7.1.1. Jeux de données

Les jeux de données analysés sont les quatre marqueurs indépendants (Rhodopsine, MLL4, IRBP et RNF213) et leurs combinaisons, ce qui fait 15 jeux de données en tout.
Les jeux de données ont été assemblés dans un esprit de *total evidence* taxinomique : les analyses sont faites avec plus de taxons que ce qui sera gardé au moment du calcul de l'indice, car l'information apportée par les taxons qui ne seront pas pris en compte est utile au bon placement des autres taxons[1]. Cependant, une situation potentiellement problématique est à éviter : il ne vaut mieux pas que dans une même matrice trop de séquences n'aient aucun site séquencé en commun, sinon leur placement relatif dans l'arbre risque de ne pas reposer sur beaucoup d'information. Seuls les taxons présents dans plus de la moitié des jeux de données élémentaires ont été gardés dans la matrice d'une combinaison donnée. Les combinaisons à deux marqueurs ne comprennent donc que des taxons dont une séquence a été obtenue pour chacun des deux marqueurs. Les combinaisons de trois marqueurs comprennent les taxons séquencés pour au moins deux des trois marqueurs et la combinaison totale comprend les taxons présents pour au moins trois des quatre marqueurs.

Un script python, `concatnexus.py`, a été écrit pour semi-automatiser l'assemblage des matrices. Ce script lit les matrices des jeux de données élémentaires au format `Nexus` et produit une matrice concaténée en demandant l'avis de l'utilisateur quand plusieurs taxons sont disponibles pour représenter un genre[2]. Pour certains genres, des séquences chimères ont été ainsi assemblées quand le taxon ne pouvait pas être représenté par les séquences d'une seule

[1] Dans cette logique, une séquence obtenue pour un seul gène, à partir d'une espèce non identifiée aurait pleinement sa place dans la matrice analysée si l'homologie avec les autres séquences n'est pas douteuse.

[2] `concatnexus.py` ne reconnaît le genre et l'espèce des taxons que si la séquence a été nommée selon la convention donnée page précédente.

espèce du genre. Le script écrit la matrice concaténée au format `Nexus`, les séquences étant désignées par le nom du genre[3], ainsi qu'un fichier de commandes `Nexus` pour chaque jeu de données, qui sélectionne l'ensemble de taxons et de caractères à prendre en compte et lance l'analyse.

7.1.2. Méthode d'analyse

Les analyses primaires ont été faites selon la méthode Bayésienne, telle qu'implémentée dans le logiciel `MrBayes`[4] (`http://sourceforge.net/projects/mrbayes/`). Ce choix découle d'une approche plus pragmatique que philosophique. La comparaison menée entre la méthode Bayésienne et le maximum de parcimonie par Agnès DETTAÏ dans sa thèse met en évidence une bien meilleure cohérence des analyses séparées dans le cas de la méthode Bayésienne. Ceci s'explique par le fait que la méthode Bayésienne est une méthode probabiliste qui met en jeu des modèles d'évolution des séquences, ce qui permet de mieux tenir compte des connaissances biologique, surtout à grande échelle évolutive, échelle à laquelle l'homoplasie est grande. `MrBayes` permet en outre l'utilisation de modèles d'évolution partitionnés par codon, ce qui améliore le réalisme des modèles[5]. Il semble normal qu'au moins les troisièmes positions de codons évoluent selon des modalités différentes des autres car la redondance du code génétique autorise une plus grande variabilité de ces position. Quel que soit le jeu de données et quelle que soit la partition, le modèle $GTR + I + \Gamma$ a été choisi[6]. Il apparaît en effet que c'est un modèle assez souvent proposé par les logiciels destinés à aider au choix du modèle pour des jeux de données de grande échelle évolutive. Choisir le même modèle pour toutes les partitions simplifiait grandement l'automatisation des analyses. De plus, la taille des jeux de données est probablement suffisante pour permettre une estimation convenable des nombreux paramètres du modèle $GTR + I + \Gamma$ (probabilités relatives de substitutions entre nucléotides, proportion de sites invariables, paramètre α de la loi Γ, et ce pour chacune des partitions).

Les analyses ont pu être lancées sur les processeurs de la grappe de calcul du Département

[3] Dans les cas de *Callionymus lyra* et *Callionymus schaapii*, *Coryphaena equiselis* et *Coryphaena hippurus*, *Lates calcarifer* et *Lates niloticus* et *Nerophis lumbriciformis* et *Nerophis ophidion*, la formation de séquences chimères ne présentait pas d'intérêt, car l'une des deux espèces n'était séquencée que pour une sous-partie des marqueurs pour lesquels des séquences de l'autre espèce étaient déjà disponibles. Dans ces cas, les espèces ont été artificiellement élevées au rang de genre pour tromper `concatnexus.py` et permettre leur présence simultanée dans la matrice concaténée. Les noms de pseudo-genre correspondant sont `Callionymuslyra`, `Callionymusschaapii`, `Coryphaenaequiselis`, `Coryphaenahippurus`, `Latescalcarifer`, `Latesniloticus`, `Nerophislumbriciformis` et `Nerophisophidion`.

[4] `MrBayes` utilise le format `Nexus`, c'est pourquoi la matrice concaténée et les commandes générées par `concatnexus.py` sont dans ce format.

[5] Une comparaison grossière avec ce que permet `Phyml` (à savoir, pas de partitionnement du modèle), semble montrer une moins faible sensibilité des analyses faites avec `MrBayes` aux attractions de branches longues et aux regroupements des taxons selon les biais de composition en bases de leur ADN. L'utilisation de `MrBayes` se paie en revanche par un temps de calcul très significativement plus long (dix à cent fois plus long).

[6] Ce n'est pas contradictoire avec le choix d'un modèle partitionné. L'intérêt des modèles partitionnés est également de permettre des paramètres d'évolution du modèle différents selon les partitions.

Systématique et Évolution du Museum National d'Histoire Naturelle, mais malheureusement sans bénéficier du parallélisme réel que devrait permettre `MrBayes`[7]. Le seul parallélisme possible était donc manuel : il s'agissait du lancement simultané des analyses des différentes combinaisons de marqueurs, chacune sur un processeur dédié. Monopoliser 15 processeurs en même temps n'est pas toujours possible, étant donnée la fréquentation fluctuante de la grappe de calcul, mais ce n'était pas le facteur limitant. Le facteur limitant était plutôt la lenteur de certaines analyses impliquant la Rhodopsine, qui auraient bien mérité une parallélisation intra-analyse. Faire évoluer les différentes chaînes de Markov d'une analyse donnée sur autant de processeurs qu'il y a de chaînes n'est pas possible manuellement car les chaînes de Markov sont en interaction[8].

Pour chaque jeu de données, le consensus majoritaire des arbres échantillonnés après convergence (estimée d'après l'allure du graphique produit par la commande `sump` de `MrBayes` représentant l'évolution de la vraisemblance des chaînes froides) a été retenu. Trois des jeux de données ont été trop longs à analyser avec `MrBayes` ; la Rhodopsine seule, combinée avec RNF213 et IRBP et combinée avec RNF213 et MLL4. Pour ces jeux de données, les consensus des 100 arbres obtenus lors d'un bootstrap avec `Phyml` (GUINDON et GASCUEL, 2003) sous un modèle $GTR + I + \Gamma$ non partitionné ont été utilisés à la place.

7.2. Analyse secondaire

Les fichiers contenant les arbres obtenus lors des analyses des jeux de données ainsi que leurs probabilités postérieures (ou pourcentage de bootstrap, pour les trois arbres non obtenus par la méthode Bayésienne) sont rassemblés dans un dossier et lus par le script `rely.py`. Ce script détermine les différents domaines de validité de niveau 3 (voir figure 3.2 page 49) et calcule l'indice de répétition des clades dans chacun de ces domaines. Optionellement, ce script génère une matrice au format nexus représentant les clades, pondérés par leur indice de répétition et/ou, construit un arbre de synthèse selon la méthode exposée page 55. Ces deux options ont été choisies pour obtenir des arbres de synthèse intégrant tous les taxons. La représentation sous forme de matrice a été analysée en maximum de parcimonie par `PAUP*`. L'arbre retenu dans ce cas est le consensus majoritaire des arbres les plus parcimonieux obtenus lors de 20000 recherches avec addition aléatoire des taxons suivies pour la moitié d'exploration des arbres par SPR (*Subtree Pruning and Regrafting*) et pour l'autre moitié d'exploration des arbres par TBR

[7] Pour une raison obscure, quand on demande à utiliser 8 processeurs pour une analyse, ce que la version parallélisée de `MrBayes` devrait pouvoir faire quand deux analyses à 4 chaînes chacune sont lancées, le système de gestion des processeurs de la grappe de calcul mobilise bien 8 processeurs, mais un seul est effectivement occupé à faire évoluer les chaînes de Markov.

[8] L'interaction entre chaînes correspond au deuxième C de l'acronyme MCMCMC (*Metropolis-Coupled Markov Chain Monte Carlo*[9]). Des chaînes de Markov évoluant à des « températures » différentes tentent régulièrement d'échanger leurs états ; elles sont « couplées ».

[9] Aujourd'hui, le Chef vous propose son Montécarle de chaînes de Markov, couplées façon Métropolis.[10]

[10] Ceci était un essai de notes de bas de pages imbriquées.

(*Tree Bisection and Reconnexion*).

7.3. Article soumis

L'article qui suit est soumis à *Molecular Phylogenetics and Evolution*. Les analyses présentées dans cet article sont celles effectuées avec `Phyml` uniquement.

RNF213, a New Nuclear Marker for Acanthomorph Phylogneny

Blaise Li [a] Agnès Dettaï [a] Corinne Cruaud [b] Arnaud Couloux [b]

Martine Desoutter-Meniger [a] Guillaume Lecointre [a,*]

[a] *Équipe 'Phylogénie', UMR 7138 'Systématique, Adaptation, Évolution', Département Systématique et Évolution, Muséum National d'Histoire Naturelle. 57, rue Cuvier, CP26, 75231 Paris cedex 05, France.*

[b] *Genoscope. Centre National de Séquençage. 2, rue Gaston Crémieux, CP5706, 91057 Évry cedex, France.*

Abstract

We show that RNF213 is a nuclear gene suitable for investigating large scale acanthomorph teleosteans interrelationships. The gene recovers many clades already found by several independent studies of acanthomorph molecular phylogenetics and considered as reliable. Moreover, we performed phylogenetic analyses of three other independent nuclear markers, first separately and then of all possible combinations (Dettaï and Lecointre, 2004) of the four genes. This was coupled with an assessment of the reliability of clades using the repetition index of Li and Lecointre (in press). This index was improved here to handle the incomplete taxonomic overlap among datasets. The results lead to the identification of new reliable clades within the 'acanthomorph bush'. Within a clade containing the Atherinomorpha, the Mugiloidei,

Preprint submitted to Elsevier *August 13, 2008*

the Plesiopidae, the Blennioidei, the Gobiesocoidei, the Cichlidae and the Pomacentridae, the Plesiopidae is the sister-group of the Mugiloidei. The Apogonidae

are closely related to the Gobioidei. A clade named 'H' grouping a number of families close to stromateids and scombrids (Stromateidae, Scombridae, Trichiuridae, Chiasmodontidae, Nomeidae, Bramidae, Centrolophidae) is related to another clade named 'E' (Aulostomidae, Macrorhamphosidae, Dactylopteridae). The Sciaenidae is closely related to the Haemulidae. Within clade 'X' (Dettaï and Lecointre, 2004), the Cottoidei, the Zoarcoidei, the Gasterosteidae, the Triglidae, the Scorpaenidae, the Sebastidae, the Synanceiidae, and the Congiopodidae form a clade. Within clade 'L' (Chen et al., 2003; Dettaï and Lecointre, 2008) grouping carangoids with flatfishes and other families (Centropomidae, Menidae, Sphyraenidae, Polynemidae, Echeneidae, Toxotidae, Xiphiidae), carangids are the stem-group of echeneids and coryphaenids and sphyraenids are the sister-group to the Carangoidei. The Howellidae, the Epigonidae and the Lateolabracidae are closely related. We propose names for most of the clades repeatedly found in acanthomorph phylogenetic studies of various teams of the past decade.

Key words: acanthomorpha, RNF213, nuclear marker, phylogeny, reliability

1 Introduction

Introduction

Acanthomorphs are a large group of more than 16,000 teleostean fish species. This monophyletic group is composed of some well supported clades but also of some

* Corresponding author. Tel: (33) 1 40 79 37 51. Fax: (33) 1 40 79 38 44.
Email addresses: blaise.li@normalesup.org (Blaise Li), adettai@mnhn.fr (Agnès Dettaï), lecointr@mnhn.fr (Guillaume Lecointre).

poorly defined large assemblages like percomorphs, perciforms, scorpaeniforms. Most of these have been suspected not to be monophyletic for a long time (Stiassny et al., 2004). Morphology and comparative anatomy are difficult to use for phylogenetic purposes at such a large scale (see Stiassny and Moore, 1992; Johnson and Patterson, 1993). Even after the efforts to clarify acanthomorph interrelationships, synthesized by Johnson and Patterson (1993), many of the large clades retained later appeared strongly contradicted by molecular phylogenies (Chen et al., 2000, 2003, 2007; Dettaï and Lecointre, 2004, 2005, 2008, submitted; Miya et al., 2001, 2003, 2005; Mabuchi et al., 2007; Kawahara et al., 2008; Smith and Wheeler, 2004, 2006; Smith and Craig, 2007). Paracanthopterygii, Acanthopterygii, Euacanthopterygii and Smegmamorpha, for instance, are all in this case. These large divisions had to be broken up because new contradicting groups were supported from independent molecular studies. Percoids also are widely distributed among acanthomorphs, so at least the traditionally defined percomorphs, perciforms, scorpaeniforms and percoids can now be considered polyphyletic (see Dettaï and Lecointre, 2005, 2008; Smith and Craig, 2007).

Moreover, new groups emerged as a result of the analyses in a single matrix of a significant number of taxonomic components, often compared directly for the first time. The picture of large scale acanthomorph fish interrelationships changes rapidly, just like the mammalian tree changed as soon as enough genes were sequenced for all eutherian orders. However the acanthomorph revolution is far from being over. Large scale relationships are still poorly known: areas of irresolution remain and all 311 families (Nelson, 2006) have not yet been sampled. In spite of the recent spectacular advances, 'the bush at the top' (Nelson, 1989) persists.

4

More acanthomorph families must be sampled, and in parallel, a higher number of phylogenetically efficient nuclear markers must be available (Li et al., 2007; Dettaï and Lecointre, 2008).

Reliability of phylogenetic findings is generally considered to be reached when several teams have found the same results independently from independent markers. This can be applied to the work of a single research team, by performing separate phylogenetic analyses of different independent molecular markers and checking for clade repetition across trees. Of course, it is necessary to base the study not only on mitochondrial markers, but also on carefully chosen, functionally and positionally independent nuclear markers. Studies using the partial rhodopsin retrogene (Chen et al., 2003), MLL (Dettaï and Lecointre, 2004, 2005) and IRBP (Dettaï and Lecointre, 2008) showed that these nuclear markers bear information relevant to the phylogeny of acanthomorphs. Other markers must be used more cautiously on this group and at this scale, like for instance 28S rDNA sequences (Chen et al., 2003) and TMO4C4 gene sequences (Smith and Wheeler, 2004). Therefore, there is still a need for additional, high phylogenetic quality markers. The availability of several teleost genomes opens new opportunities for the research of new markers, as also demonstrated by the study of Li et al. (2007). In the present study, we used the Ensembl Biomart mining tool and selected a few candidates markers. A promising locus, RNF213, was amplified for a representative sampling of teleost acanthomorphs, and compared to other large nuclear and mitochondrial datasets. Additionaly, we performed combined and separate phylogenetic analyses with retro-rhodopsin, MLL and IRBP, and assessed the reliability of clades using the repetition index of Li and Lecointre (in press).

5

2 Materials and methods

2.1 Selection of the molecular markers

In previous studies, we used protein-coding nuclear genes as well as mitochondrial (12S and 16S, Chen et al., 2003; Dettaï and Lecointre, 2004, 2005) and nuclear ((28S, Chen et al., 2003) rDNA sequences. Alignment difficulties and low phylogenetic content of these rDNA genes with respect to the acanthomorph issue led us to abandon these markers and focus on carefully chosen nuclear protein-coding genes.

The Biomart mining tool of the Ensembl Portal (Hubbard et al., 2005) release 40 was used to get a list of protein-coding genes shared by *Tetraodon nigroviridis*, *Takifugu rubripes*, and *Danio rerio*, using *T. nigroviridis* as a query. Genes having unique best hits were retained, and checked for divergence and exon length through the Ensembl Portal on all the available teleost genomes. The sequence coding for RNF213 was again blasted (Altschul et al., 1997) on all available teleost genomes to check that it was a single copy marker in all. Last, it was blasted in the CoreNucleotide database of Genbank and all available sequences for acanthomorph species were recovered and used for primer design after alignment with Bioedit (Hall, 2001).

The previously published datasets for the retro-rhodopsin, MLL and IRBP genes were completed with additional taxa (Table 2). To these, we added sequence data from the gene RNF213.

2.2 PCR and sequencing

DNA was extracted mostly from muscle samples stored in 70% ethanol, following the protocol of Winnepenninckx et al. (1993). The primers published in Chen et al. (2003) for Rhodopsin, in Dettaï and Lecointre (2005) for MLL and in Dettaï and Lecointre (2008) for IRBP were used, but 4 new primers were designed for MLL in order to obtain more gadiform sequences, and 6 new primers were used to amplify the new RNF213 marker (Table 1). Most RNF213 sequences could be obtained with primers C17 F3111 and C17 R4111. Various PCR conditions were used, depending on the primers and the DNA sample. Three different polymerases were used for the PCRs: Taq Appligen, QbioTaq and Taq Qiagen. PCRs began with a denaturation phase at 94°C for 2 to 5 minutes and ended with a final elongation phase at 72°C (or 68°C for the longest PCRs using Taq Qiagen) for 4 to 7 minutes. Cycles began with a denaturation phase at 94°C for 20s to 40s, followed by an annealing phase at temperatures ranging from 47°C to 60°C and during from 25s to 45s. The annealing phase was followed by an elongation phase at 72°C (or 68°C for the longest PCRs using Taq Qiagen) for 35s to 2 minutes. The number of cycles ranged from 35 to 60. Purification and sequencing of the PCRs were performed at the Genoscope (`http://www.genoscope.cns.fr/`). The same primers were used for PCR and sequencing. Sequences were checked individually using Sequencher (Gene Codes Corporation) and aligned by hand using Se-Al (Rambaut, 2002). Indels were grouped by 3 so as to fit the coding frame, and adjusted according to the translation in amino-acid sequences. Preliminary distance trees were done using PAUP* (Swofford, 2002) to check for contaminations. Accession numbers are

given in table 2.

2.3 Analysis strategy

When a clade contradicts the previously supported phylogenetic hypotheses, it is necessary to check whether this is due to an artifact. Those can be detected by using different taxonomic samplings, tree reconstruction methods, or, more reliably, by comparing the topology to the one inferred from an independent dataset. This is the primary reason to perform separate analyses in molecular phylogenetics. If selective pressures characterizing mutational space at each position are relatively homogeneous within genes but heterogeneous among genes, the fact that a given clade is recurrently recovered from independent markers is a strong indication of its reliability (Nelson, 1979; Chen et al., 2003; Dettaï and Lecointre, 2004; Lecointre and Deleporte, 2005). Indeed, finding a clade twice independently just by chance is very improbable (Page and Holmes, 1998), and the probability of obtaining exactly the same tree reconstruction artifact from independent genes is also low, although sometimes the notorious long branch attraction artifact can occur with several markers when higher mutation rates affect large parts of the genome of several species of the taxonomic sample. Separate analyses tend to be more subject to stochastic errors because of the shorter length of the analyzed sequences, prone to marker-specific biases or sometimes even reflect different histories. The recovery of a clade in separate analyses of several independent markers in spite of these problems is therefore a strong indication of the reliability of the clade. Moreover, repetition across trees based on independent data is a better indicator than

bootstrap proportions extracted from a crude 'total evidence' (for a review of the origins of that term, see Rieppel, 2004; Lecointre and Deleporte, 2005), because tree reconstruction artifacts can lead to clades with high robustness (Philippe and Douzery, 1994). Additionally, a positively misleading signal from a single gene can impose the topology of some parts of the tree inferred from the combined data (Grande, 1994; Chen et al., 2000, 2003; Chen, 2001). Separate analyses of independent partitions are an efficient way to assess the reliability of clades and to identify marker-specific reconstruction artifacts. The addition of new nuclear markers is still important: the changes between the topologies and reliabilities between Chen et al. (2003) and Dettaï and Lecointre (2004, 2005, 2008) show that the inclusion of new markers with this approach increases the number of repeated clades even if the number of taxa remains almost unchanged (Dettaï and Lecointre, 2005, 2008, submitted).

However, keeping partitions separate has its own risks (see review in Miyamoto and Fitch, 1995; Lecointre and Deleporte, 2005), like stochastic errors. To circumvent these problems, it is interesting to perform both separate and simultaneous analyses (Mickevich, 1978; Bull et al., 1993; Miyamoto and Fitch, 1995). Dettaï and Lecointre (2004, figure 2 therein), Dettaï and Lecointre (2005) and Li and Lecointre (in press) also proposed partial combinations as a way to explore marker-specific topologies and to assess the reliability of clades. In their approach, each elementary dataset is analyzed separately, and every possible combination of the datasets is produced and analyzed too. The occurrences of a clade are recorded across 'partitioning schemes': sets of independent datasets (including combinations with no data overlap). In the present study, we used the same approach, analyzing every

possible combination of the datasets and recording repeated clades from combinations having no marker in common. This approach takes into account both the strengths and the weaknesses of separate and simultaneous analyses. When independent datasets provide the same signal, chances are that such a signal comes from species history. However, stochastic effects affect more strongly small datasets, and separate datasets are smaller; partial combinations limit this problem by allowing analyses of longer sequences. The four nuclear markers were thus assembled in 15 combinations of one to four datasets.

2.4 Datasets design

All genera for which we could obtain sequence data for at least one of the four nuclear markers were used in this study, in order to cover a broad taxonomic area, and because it has been shown that even a single sequence can still convey relevant information for the phylogenetic analyses (Wiens and Reeder, 1995). Missing data can have an effect on phylogenetic reconstruction in some cases, therefore, a rule about the minimum amount of sequences that need to be present for a taxon be included in a dataset was decided. The taxa for which half the markers (or more) were missing for a given combination were not used in that combination. This means that in a combination, a terminal taxon must have sequences for at least two of the markers in a combination of 2 or 3 markers, and sequences for 3 or 4 markers for the combination of the 4 markers. For a few taxa, the sequences for different genes were obtained from different species of the same genus. Using such chimeric sequences at the species level for a study at the interfamilial level should

not be problematic. For *Callionymus*, *Coryphaena* and *Nerophis*, species were not fused in a chimeric sequence because this would not have led to a better taxonomic overlap between the different markers: *Callionymus lyra* was present for the four markers while *C. schaapii* was present for Rhodopsin only, *Coryphaena equiselis* was present for the four markers while *C. hippurus* was present for IRBP only, and *Nerophis lumbriciformis* was present for RNF213, IRBP and Rhodopsin while *N. ophiodon* was present for IRBP only.

2.5 Primary analyses

Considering the taxonomic scale of our study, sequence data were analyzed under probabilistic sequence evolution models. PhyML (Guindon and Gascuel, 2003) was used for its speed, with a GTR + I + Γ model.

To offer an assessment of the role played by the RNF213 sequence data in the resolution of the 'acanthomorph bush', four trees were compared: the tree based on the new RNF213 sequence data, the tree based on the combination of the three other nuclear markers, the tree based on the combination of the four datasets (the 'total evidence' tree) and the MRP supertree displaying/summarizing the reliable clades calculated from the repetition indices of Li and Lecointre (in press), based on partial combinations and validity domains for the four datasets studied here.

2.6 Validity domains: adapting Li and Lecointre's method to datasets with different sets of taxa

To evaluate the reliability of clades, Li and Lecointre (in press) proposed repetition indices based on the partial combinations strategy (Dettaï and Lecointre, 2004). However, the proposed indices are only valid when all trees to be compared have the same set of taxa. Here, restricting all analyses to genera present in all four elementary datasets would have resulted in discarding nearly half the taxa, a considerable loss of information. To avoid this problem, we adopted here a 'prune-to-compare' strategy based on three levels of what we call 'validity domains' (Figure 1).

Analyses of the various dataset combinations were done on different sets of taxa, depending on the proportion of available sequences for a given taxon in the combined datasets. These sets of taxa are the first-level validity domains: the validity domains of the primary analyses. Here our strategy is to take into account the taxa present in at least half the elementary datasets of a given combination. With three elementary datasets A, B and C, noting V_A, V_B and V_C their validity domains, the validity domain of $A \cup B \cup C$ would be:

$V_{A \cup B \cup C} = (V_A \cap V_B) \cup (V_A \cap V_C) \cup (V_B \cap V_C)$ (that is, the taxa that are either in A and B, in A and C or in B and C).

These validity domains are the sets of leaves of the trees resulting from the primary analyses. The number of occurrences of the clades are counted over sets of such trees. In order to take robustness into account, the repetition index can also be

based on sums of bootstrap proportions instead of simple sums of occurrence.

However, counting occurrences makes sense only when the trees taken into account are based on independent data. Relevant partitioning schemes, containing independent non overlapping datasets are therefore defined. For example, among all possible combinations of datasets, A and $B \cup C$ form a valid partitioning scheme because the two parts do not contain elementary datasets in common. A number of clade occurrences may be counted over the two trees obtained from these datasets.

A clade can be recognised and counted only when the trees are defined on the same set of taxa. To obtain a common taxonomic sampling for all trees within a partitioning scheme, a second-level validity domain has been defined and associated to the partitioning scheme through a process we call 'prune-to-count'. Taxa that are not in all datasets in the partitioning scheme are pruned from the trees. If we note PSc_2 partitioning scheme $(A, B \cup C)$, the corresponding second-level validity domain (Figure 1) would be:
$V_{PSc_2} = V_A \cap V_{B \cup C}$

Moreover, in addition to the previous description of a partitioning scheme made by Li and Lecointre (in press), we also took into account all the partial partitioning schemes; those that do not contain all elementary datasets but have a larger second-level validity domain than full partitioning schemes. This avoids loosing information about potentially reliable relationships.

The repetition index for a clade, proposed in Li and Lecointre (in press), is based on a comparison between the best number of occurrences of the clade across all

possible partitioning schemes and the best number of occurrences of its contradictors.

Here, different partitioning schemes (PSc_1, PSc_2, PSc_3, etc.) can have different taxonomic samplings. The repetition index could be computed in the set of taxa common to all partitioning schemes, but this would entail an important loss of terminals. Therefore the repetition index was also computed using smaller sets of partitioning schemes. To compute a repetition index taking into account a given set of partitioning schemes, one must restrict this part of the reliability analysis to the intersection of the validity domains of the partitioning schemes. This second pruning step is performed in order to be able to compare clades from the different partitioning schemes of the set. We call it a 'prune-to-compare' process: If PSc_1, PSc_3 and PSc_4 are the partitioning schemes from where the clades to be compared are drawn, the comparison is made in a third-level validity domain (that we can note W_1, figure 1):

$W_1 = V_{PSc_1} \cap V_{PSc_3} \cap V_{PSc_4}$

This is done for all combinations of partitioning schemes and leads to several third-level validity domains (W_1, W_2, W_3, etc.).

Thus, within each W, each clade has a first order repetition index, which is the maximum number of occurrences found among the involved partitioning schemes, counted on all clades before the pruning process corresponding to the clade under focus. This index is then decreased using the first order repetition index of contradicting clades defined in the same W, following the procedure described in Li and Lecointre (in press). The whole procedure results in clades associated with

repetition indices, each association being defined in a particular W.

The procedure can be summarized as follows:

(1) Analyze each dataset (elementary datasets and all possible combinations thereof).

(2) Arrange the data into partitioning schemes (sets of independent datasets) and determine the validity domain of each of them. The validity domain of a partitioning scheme (second-level validity domain, noted V_{PSc}) is the intersection of the validity domains of its constituting datasets (first-level validity domains, noted V).

(3) For each partitioning scheme, prune the taxa not in the validity domain of the partitioning scheme from the trees and record the clades in the pruned trees with their number of occurrences (whenever a number of clade occurrences is used, a sum of support values could be used instead).

(4) Arrange the partitioning schemes into combinations and determine the validity domain of each of the possible combinations (third-level validity domains, noted W). The validity domain of a combination of partitioning schemes is the intersection of the validity domains of its constituting partitioning schemes. Group together combinations that have the same validity domain.

(5) For each of the preceding validity domains (W), prune the clades found in the associated combinations of partitioning schemes from the taxa that are not in the validity domain. For each distinct resulting clade, keep as first order repetition index the best number of occurrences found among the clades that, once pruned, become the clade under focus.

(6) Within each third-level validity domain, proceed as in Li and Lecointre (in press) to obtain the final repetition indices.

In the present reliability analyses, to accommodate for the uncertainty entailed by the use of heavy heuristics, bipartition occurrences were weighted by their bootstrap supports across 100 resamplings. But pruning taxa from a tree causes the fusion of several internal branches. The boostrap value used to weight the bipartition delimited by such fused branches was the highest support among the fused branches. We consider this choice justified because in order to collapse a clade, one must break its branch. The clade resulting from taxon pruning in a restricted validity domain is the 'heir' of one or more pre-pruning clade(s), as they differ only with respect to the terminals that have been pruned. This new clade may thus be supported by several successive branches in the original tree, each of which has to be broken. This clade can therefore be considered to be as strong as the strongest of its 'ancestors' in the original trees.

2.7 Displaying reliable clades

Some clades repeated in the separate analyses can be absent from the tree based on all available data. This has been shown theoretically (see the clade BCD in Barrett et al., 1991, figure 1) as well as empirically (Dettaï and Lecointre, 2004, figures 4 and 5 therein). This is one of the grounds to conduct both separate and simultaneous analyses (Nixon and Carpenter, 1996). A tree summarizing the clades considered reliable has to be constructed to allow the visualization of the possible discrepancies. To synthesize the results of the reliability analysis, we used

a supertree approach: the bipartitions with positive repetition indices from the various validity domains were gathered in a matrix representation and weighted by their repetition indices. This matrix was analyzed under maximum parsimony.

3 Results

3.1 Effect of RNF213 on support

Figure 2 is the tree based on RNF213 sequence data only. The marker allows to recover clades already recorded in previous molecular phylogenies (Dettaï and Lecointre, 2005, 2008, table 3), namely A, D, E', F, G, H, E'+H, L, M, Q, X, Is, Isc, P (Zeioidei, Gadiformes, Polymixiiformes, Percopsiformes and Lampridiformes, as in Dettaï and Lecointre, 2008, submitted).

Clade N (Dettaï and Lecointre, 2005, 2008; Yamanoue et al., 2007; Kawahara et al., 2008; Holcroft and Wiley, 2008) fails to appear. Lophiiforms and *Siganus* are out of it, though without resolution. Resolution inside N is also very poor. Clade B is the monophyly of Beryciformes *sensu lato* (Chen et al., 2003; Miya et al., 2005), however that clade is not repeated throughout studies. Clades C and O are not recovered because of incomplete taxonomic samplings, while clades I, J, K are contradicted.

Comparing figure 3 to figure 4, it is not clear whether RNF213 sequence data are able to improve resolution. More investigation is needed to establish whether clade N should also include (as in figure 3) Lobotidae, Monodactylidae and Lutjanidae

(they should according to Yamanoue et al., 2007; Holcroft and Wiley, 2008), Leiognathidae, Cepolidae, Labridae, Scaridae and Moronidae (the last three are closely related in Dettaï and Lecointre, submitted), Centrarchidae, Elassomatidae (as suggested in Dettaï and Lecointre, submitted), Callanthiidae, Priacanthiidae, Caesionidae, Malacanthidae, Datnioididae, and Scatophagidae, Sciaenidae and Haemulidae (as in Chen et al., 2007). The previously described clade N plus the families listed above constitute a working hypothesis that we call here 'extended N'. That clade is rendered paraphyletic in figures 4 and 5. If those topologies are to be trusted, it could be extended again to contain Kyphosidae, Aplodactylidae, Cheilodactylidae, Sparidae, Champsodontidae and clades X, G and R.

In figures 3 and 4, clade H is the sister-group of clade E (forming together clade S) while clade H is the sister group of E' in figure 2 (RNF213 data only). This E'+H clade might be due to the lack of taxa of clade E in the RNF213 dataset.

3.2 New results from RNF213 sequence data

The new RNF213 sequence data adds some interesting results, while others come from the addition of new taxa to the previously published samplings.

The ability of RNF213 sequence data to get more clades can be assessed through clades absent from figure 3 and present in figures 2 and 4:

- Clade Q (Dettaï and Lecointre, 2005, 2008) is recovered (Gobiesociformes, Blennioidei, Atherinomorpha, Mugiloidei, Pomacentridae) with a new member, the Plesiopidae;

- Family Plesiopidae is in clade C, probably as sister-group to the Mugiloidei;
- Clade E': Mullidae, Callionymidae, Syngnathidae. That clade is not new, already proposed in Dettaï and Lecointre (submitted) and not contradicted by Kawahara et al. (2008) because of poor support within their clade 'D' and absence of any mullid or callionymid;
- Clade R: Howellidae, Lateolabracidae and Epigonidae. It was already present in Smith and Craig (2007);
- Clade T: Notothenioidei, Trachinidae due to the addition of *Echiichthys*;
- Clade Z: Cottoidei, Zoarcoidei, Gasterosteidae, Triglidae, Scorpaenidae, Sebastidae, Synanceiidae, Congiopodidae. That clade is a beginning of structuration within clade X (Dettaï and Lecointre, 2004).
- Clade S (E+H) because members of E are present in figures 3 and 4;
- Indostomidae is a member of clade F because *Indostomus* is added (also found by several studies, Miya et al., 2003; Kawahara et al., 2008);
- Clade U: *Pampus* (Stromateidae) sister-group of all other members of H in figures 2, 3 and 4;
- Clade V: L+F in figures 3 and 4;
- Clade L': Coryphaenidae and Echeneidae nested within the Carangidae.

Figure 5 is the supertree summarizing the reliable clades. When this tree is compared to figure 2, it appears that some of the new clades listed above were recovered by RNF213 only.

- Clade R: Epigonidae, Lateolabracidae and Howellidae form a clade (as in Smith and Craig, 2007);

- Indostomidae is a member of clade F (as in Miya et al., 2003; Kawahara et al., 2008);
- Clade Q including the Plesiopidae;
- Clade T (Notothenioidei, Trachinidae, contradicting the clade K of Dettaï and Lecointre (2004, 2005).

These clades are an indication that the new marker has the potential to hel to the emergence of some not yet identified reliable clades, while it also has the potential to recover clades previously recorded as reliable, like A, D, Q, H, f1, f2, F, G, Is, Isc, X, L, M.

The comparison of figure 5 with figure 4 is interesting to evaluate whether there are reliable clades that do not appear in the tree based on the simultaneous analysis of all datasets ('total evidence'). It is indeed the case for the following clades:

- Clade W: Apogonidae and Gobioidei;
- Clade M': Sciaenidae and Haemulidae;
- Clade M": Centrarchidae, Moronidae and Elassomatidae;

This discrepancy has already been described (Dettaï and Lecointre, 2004), and is not entirely surprising as the tree based on the whole data can be perturbed by the usual pitfalls of phylogenetic reconstruction like any other tree.

Symmetrically, some clades from the total evidence approach are not found in figure 5:

- Clade E' (Figures 2 and 4);

- Clade U (Figures 2, 3 and 4);
- Clade V (Figures 2, 3 and 4);

These clades are intuitively more problematic because they are found from different trees based on independent data. Clades U and V are striking examples. They are repeated from independent data (Figure 2: RNF213 data only, and figure 3: all other data) while they are not recovered using the multiple combinations protocol.

4 Discussion

4.1 New names for new reliable clades

A number of clades had already been recovered in the previous studies and will not be further discussed (Chen et al., 2000, 2003; Dettaï and Lecointre, 2004, 2005, 2008, submitted): A, C, D, E, F, G, H, J, K, L, N, M, P'. We will focus here on the new results. In both cases, new names are proposed for some of these clades (Table 3)

- Clade Y: Plesiopids (roundheads) are the sister-group of the Mugiloidei (grey mullets). This has not been proposed before, and was not found by Smith and Craig (2007), as they had no mugiloid included.
- The large clade Q contains the Mugiloidei (grey mullets), the Plesiopidae (roundheads), the Atherinomorpha (guppies, pupfishes, silversides, needlefishes), the Pomacentridae (damselfishes), the Blennioidei (blennies) and the Gobiesociformes (clingfishes). The group was present with a more reduced sampling in

Chen et al. (2003) and Miya et al. (2005), as there were no cichlid, no pomacentrid and no plesiopid; Dettaï and Lecointre (2005) added a cichlid but no pomacentrid and no plesiopid. Different studies showed that the Cichlidae were close to a group containing the Atherinomorpha (Chen et al., 2003, 2007; Dettaï and Lecointre, 2005; Mabuchi et al., 2007) and not to the other members of the six-family labroidei: Labridae and Scaridae (Dettaï and Lecointre, 2005; Mabuchi et al., 2007; Chen et al., 2007). Among these studies, the best taxonomic sampling is reached by Mabuchi et al. (2007) showing that the Pomacentridae, the Embiotocidae (surfperches) and the Cichlidae are close to each other (their clade 'B') and to members of what we call here clade Q, while the Odacidae (cales), the Scaridae (parrotfishes) and the Labridae (wrasses) form a clade (their clade 'A') and are members of what we call here the 'extended clade N'. These results are clearly corroborated by Chen et al. (2007) from independent nuclear and mitochondrial sequence data. The Labroidei are therefore diphyletic and the specialized 'labroid' pharyngeal jaw apparatus evolved twice. Here the cichlids (*Haplochromis*) have an undetermined position within clade Q, while the Blennioidei are the sister-group of the Gobiesociformes (clade D, Chen et al., 2003). It must be noticed that the family Pseudochromidae (dottybacks) may also be a member of clade Q according to the position of *Labracinus* in the tree of Mabuchi et al. (2007). This is corroborated by the tree of Smith and Craig (2007) where another pseudochromid, *Pseudochromis*, lies within a clade corresponding to the present clade Q (Cichlidae, Blennioidei, Atherinomorpha). But gobiesociforms and mugiloids are absent in their dataset and some families missing from ours integrate their equivalent of clade Q: Opisthognathidae (jaw-

fishes), Grammatidae and Pholidichthyidae. An interesting feature emerges from the comparison of all these studies. Clade Q may contain the Atherinomorpha, the Mugiloidei, the Plesiopidae, the Pomacentridae, the Cichlidae, the Embiotocidae, the Blennioidei, the Gobiesociformes, the Pseudochromidae, the Opisthognathidae, the Grammatidae and the Pholidichthyidae. A closer look for possible morphological characters uniting all these taxa yielded one candidate. Mooi (1990) records adhesive chorionic filaments arranged around the micropyle of the demersal egg in grammatids, opisthognathids, pomacentrids, plesiopids and apogonids (the latter is not in Q here but in clade W, see on the following page). Such eggs are found in pseudochromids (Mooi, 1993) and in plesiopids *sensu lato* (*i.e.* including acanthoclinids (Mooi, 1993) and notograptids (Gill and Mooi, 1993). Gill and Mooi (1993) also record such demersal eggs with filaments in blennioids (here clade Q) and gobioids (here in clade W), as did Breder and Rosen (1966). Parenti (1993) records these eggs as a synapomorphy of the Atherinomorpha (viviparity in that group being a derived condition). Additionally Smith and Wheeler (2004) mention these eggs in the Cichlidae, and Breder and Rosen (1966) in the Gobiesocidae and the Kurtidae (probably in clade W). These eggs are recorded in 9 of the 12 potential components of clade Q. The Mugiloidei and the Embiotocidae do not have these eggs (Breder and Rosen, 1966), but these conditions may be reversals, as mugiloids are the sister group of plesiopids and embiotocids are grouped with cichlids and pomacentrids. Demersal eggs with filaments may therefore be a synapomorphy of the clade. The presence of this character state also in gobioids, kurtids and apogonids remains to be explained. It could have been gained by convergence, but this needs an

in-depth comparison to explore the homology in and between these two groups. Alternatively, demersal eggs with chorionic adhesive filaments could also be a synapomorphy of a clade W+Q, however more resolution is needed to test that hypothesis.

- Clade W: Apogonidae (cardinalfishes) are closely related to the Gobioidei in our figure 5 (as in Smith and Wheeler, 2006), while *Apogon* is grouped with *Kurtus* in Smith and Craig (2007) in the absence of gobioids. *Kurtus* is found to be the sister-group of *Apogon* and gobies in Smith and Wheeler (2006). Interestingly, horizontal and vertical rows of sensory papillae on the head and body are exclusively shared by Apogonidae, Kurtidae (nurseryfishes), Gobioidei (gobies) and Champsodontidae (crocodile toothfishes) (Johnson, 1993, p. 18). However, here *Champsodon* fails to cluster with the apogonid and gobioidei representatives. On the basis of anatomical data, Prokofiev (2006) stresses a close relationship between the Apogonidae and the Kurtidae (without mentioning the Gobioidei). Comparison of different studies and Smith and Wheeler (2006) suggest that a clade 'W' at least composed of Apogonidae, Kurtidae and Gobioidei is worth being investigated further. Like clade Q, this clade would also be supported by the presence of eggs with chorionic adhesive filaments (see on the previous page). Apogonids and kurtids were not among the potential sister-groups of the Gobioidei identified by Winterbottom (1993): trachinoids, gobiesocoids, hoplichthyids and other scorpaeniforms. However, these were not included in his study.
- Clade T (trachinids more closely related to notothenioids than to percids) contradicts the clade K of Chen et al. (2003) and Dettaï and Lecointre (2004, 2005)

where the perches (Percidae) are the most closely related group to Antarctic fishes (Notothenioidei). Nonetheless, in Chen et al. (2003) as well as in Dettaï and Lecointre (2004, 2005), *Trachinus* (Weeverfish) was always placed very close to clade K. Interestingly, in Smith and Craig (2007) *Bembrops* (Percophidae), *Acanthistius* (Serranidae, Anthiinae) and *Niphon* (Serranidae, Epinephelinae) are inserted between percids and notothenioids and *Trachinus* is branched off far away, among serranids. On the contrary, in Smith and Wheeler (2006) trachinids are more closely related to notothenioids than percids, while *Bembrops* is still the closest to notothenioids. In our summary tree, *Niphon* is among serranids and *Acanthistius* has an undertermined position. Sequence data for the genes analyzed here would be much needed for *Bembrops*, to test their effects on the relative positions of trachinids and percids and get a clearer idea about the sister-group of notothenioids.

- Clade Z is providing more precision within clade X: Synanceiidae (stonefishes), Scorpaenidae (scorpionfishes and rockfishes), Congiopodidae (pigfishes) and Sebastidae (thornyheads) constitute the stem-group of clade Isc. Clade Z cannot be identified in most other studies because of insufficient taxonomic sample overlap. In Smith and Craig (2007) the clade is very well represented by 33 terminals however it is not recovered because the Synanceiidae branches outside it and a clade made of the Bembridae (deepwater flatheads), Plectrogeniidae, and their 'clade E' (notothenioids, percids, percophids, anthines) is included in it. It is important to note that we have identified clade X and clade Z without taking into account single unstable taxa 'escaping' with no determined position. In fact, the problematic four taxa (*Acanthistius*, *Pseudaphritis*, *Liopropoma*, *Plec-*

tropomus) have sequences for one marker only and have question marks for all the other genes. More data are needed to stabilize their position. Moreover, the pinguipedid *Parapercis* (grubfishes and sandperches) is nested among serranids in figure 5. In Smith and Craig (2007) it is close to other trachinoid families like Ammodytidae (sandlances) and Cheimarrichthyidae (torrentfish), like in our figure 2. The position of *Parapercis* in figure 5 must be taken with caution. Indeed, clade G includes *Pinguipes* in figure 5 (along with three formerly 'trachinoid' families Ammodytidae, Uranoscopidae (stargazers), Cheimarrichthyidae) and in figure 2 the two pinguipedids *Pinguipes* and *Parapercis* are both placed in clade G. More sequence data is needed for *Parapercis* before a conclusion can be drawn on the mono- or polyphyly of the Pinguipedidae.

- Clade S (H and E) is recovered figures 3, 4 and 5; it does not appear in other studies because of the lack of overlap between the taxonomic samplings. In Chen et al. (2007), there is a 'backbone' of clade E'+H with *Mullus* (E') associated to *Scomberomorus* and *Psenopsis* (H). The study of Kawahara et al. (2008) dealing with the polyphyly of the Gasterosteiformes from independent sequence datasets using a complete sampling of that order at the family level includes no member of our 'clade H' in . Moreover the position of their 'clade C' (containing *Macrorhamphosus*, *Aulostomus* and dactylopterids) is poorly supported, leaving the question open.
- The Gasterosteiformes are polyphyletic, with indostomids (armored sticklebacks) within clade F (with synbranchiformes) and gasterosteids (sticklebacks) closely related to the Zoarcoidei (eelpouts), a result fully confirmed by independent data in Miya et al. (2003) and in Kawahara et al. (2008) with a much larger

sampling for gasterosteiform. In Kawahara et al. (2008), the Syngnathoidei are paraphyletic, including dactylopterids (flying gunards). Gasterosteoids are closer to the Zoarcoidei and indostomids closer to symbranchiforms. Interestingly here part of the Syngnathoidei (*Macrorhamphosus* and *Aulostomus*) are close to the Dactylopteridae (clade E) however other syngnathoids (*Aeoliscus*, *Syngnathus*, *Hippocampus* and *Nerophis*) never group with them, probably because they have long branches. Though they concluded from comparative anatomy and bone development that indostomids were gasterosteoid gasterosteiforms, Britz and Johnson (2002) mentioned a feature that is shared by indostomids and mastacembelids (though also by most other gasterosteoids): the lack of distal radials in all pterygiophores supporting fin spines at all developmental stages.

- Clade M' (Sciaenidae (croakers) and Haemulidae (grunts)) has not been found by molecular studies because of lack of representatives included for these families. However, Smith and Craig (2007) did sample those two families but they don't appear related to each other in their tree. Also, from partially independent sequence data in Chen et al. (2007), haemulids appear close to lutjanids and sparids while sciaenids are closer to drepanids and chaetodontids.
- The same applies for clade M" grouping the Centrarchidae (sunfishes), the Moronidae (temperate basses) and the Elassomatidae (pygmy sunfishes). Moreover, that clade contradicts the association of the Moronidae in Dettaï and Lecointre (submitted) with some members of the labroids (*i.e.* labrids and scarids) and some members of the polyphyletic trachinoids. In Chen et al. (2007), *Elassoma* is not related to moronids and this family is closer to labrids and scarids. Clade M" should be evaluated again with more taxa.

- Clades L' and L" are structuring the inside of clade L. Carangids (jacks and pompanos) are placed as the stem group of the Echeneoidea (*sensu* Johnson, 1993)), represented here by Echeneidae (remoras) and Coryphaenidae (dolphinfishes). The Sphyraenidae (barracudas) are the sister-group of carangoids (carangids plus Echeneoidea). Johnson (1984, 1993) defined the Carangoidei as the Carangidae, Echeneidae, Rachycentridae, Nematistiidae and Coryphaenidae. Those clades L' and L" have not been found by previous molecular studies because of the lack of representation of these gropups. The exception is Smith and Wheeler (2006), who confirm these two clades. Smith and Craig (2007) did find an equivalent of L but did not find any clade compatible with clade L" as the Sphyraenidae are branched well within L.
- Clade R: *Epigonus*, *Howella*, and *Lateolabrax* form a clade (already found by Smith and Craig, 2007) suggesting close relationships of *Howella*, Lateolabracidae (Asian seaperches) and Epigonidae (deepwater cardinalfishes). Interestingly, the study of Smith and Craig (2007) includes other percichthyid genera (namely *Bostockia*, *Gadopsis*, *Macquaria*, *Nannoperca*), but *Howella* is not grouped with them but within their equivalent of clade R, suggesting the polyphyly of the Percichthyidae. This would not be surprising as the family is known to be poorly defined (Nelson, 1994). Prokofiev (2007) even erected a new family with three genera, (Howellidae) on the basis of several osteological features. Other families like Polyprionidae (wreckfishes), Dinolestidae (long-finned pike), Pentacerotidae (armorheads), Acropomatidae (lanternbellies) appear in Smith and Craig (2007) as more closely related to the clade grouping *Howella*, *Epigonus* and *Lateolabrax* than *Howella* is to the other Percichthyidae. Sequence data for more markers

from all those key taxa would be of interest to confirm the polyphyly of the Percichthyidae.

- Clade P is interesting. As a large clade located at the base of the acanthomorph tree, it has been difficult to find because of long-branch attractions in molecular studies of acanthomorph phylogeny. Long-branch attractions tend to attract the longest branches towards the outgroups (which have long branches by definition, see for example Dettaï and Lecointre, 2005) and create comb-like tree shapes at the most basal parts of the trees. Clade P groups Polymixiiformes (beardfishes), Percopsiformes (trout-perches), Lampridiformes (oarfishes and opahs), Zeioidei (dories), and Gadiformes (cods). The clade appears in Dettaï and Lecointre (2008) and is only partial in Dettaï and Lecointre (2005). It is contradicted by studies using complete mitochondrial sequence data (Miya et al., 2001, 2003, 2005) on a single point: the Lampridiformes are attracted to a non-acanthomorph group (either Myctophiformes or Ateleopodiformes). We propose to name that large-scale clade the Paracanthomorpha. It contains two orders of the former paracanthopterygians (Percopsiformes and Gadiformes) and during the last ten years the polyphyly of the Paracanthopterygii has been demonstrated several times by independent teams and data: the Lophiiformes are members of clade N, the Gobiesociformes members of clade D and the Batrachoidiformes the sister-group of what we call here clade F (Miya et al., 2005).
- Basal Acanthomorpha: present results corroborate that clade P is the most basal among acanthomorphs sampled here and that ophidiifoms (cusk-eels) are the sister-group of non-P and non-beryciform acanthomorphs (as in Miya et al., 2003, 2005).

A number of groups are found in several trees however they are absent from figure 5 and not considered to be reliable. They can be used as working hypotheses.

- Clade U: *Pampus* (Stromateidae) sister-group of all other members of H in figures 2, 3 and 4;
- Clade V: Carangimorpha (L) + Anabantiformes (F) found in figures 2, 3 and 4;
- Extended N: clade N including Lobotidae (tripletails), Monodactylidae (fingerfishes), Lutjanidae (snappers), Leiognathidae (ponyfishes), Cepolidae (bandfishes), Labridae (wrasses), Scaridae (parrotfishes) and Moronidae, Centrarchidae, Elassomatidae, Callanthiidae (groppos), Priacanthiidae (bigeyes), Caesionidae (fusiliers), Scatophagidae (scats), Malacanthidae (tilefishes), Datnioididae (tigerperches), Kyphosidae (sea chubs), Aplodactylidae (marblefishes), Cheilodactylidae (morwongs), Sparidae (porgies), Champsodontidae (crocodile toothfishes), clades X, G, M' and R .

As discussed above, a clade can be considered reliable when it has been recovered on several independent datasets, and whenever possible by several teams independently. New clades are hypotheses of interrelationships that need to be tested through various sources of data by other teams before being accepted by the community. This is why new clades temporarily received letters (as in Chen et al., 2003; Kawahara et al., 2008; Dettaï and Lecointre, 2005, 2008) suggesting that they were working hypotheses. However, as letters differ from one study to another for the same clades, a need for stabilization emerges. Once the new clades are sufficiently corroborated, it becomes necessary to give them names, for convenience' sake. In the case of acanthomorphs, the new names were proposed by Johnson and Pat-

terson (1993). Almost none of the molecular studies of large-scale acanthomorph interrelationships (Chen et al., 2000; Wiley et al., 2000; Miya et al., 2001, 2003; Chen et al., 2003; Dettaï and Lecointre, 2004, 2005, 2008, submitted; Smith and Wheeler, 2004, 2006) proposed new names. It is striking that, while many clades were recovered several times from independent genes and teams, they still remain unnamed. Table 3 proposes names for the clades that have been repeatedly recovered in the molecular phylogenies of acanthomorphs.

All the 309 (Nelson, 2006) acanthomorph families have not yet been included in a single study. However this is not an obstacle and recommendations for names can be made progressively as new families are included. Smith and Craig (2007) added new evidence and summarized results from different studies. They then proposed new and necessary delimitations for serranids, percoids, trachinoids, resurrected epinephelids and niphonids, and created the Moronoidei. We have a single minor point of disagreement with their propositions: they proposed to incorporate the Notothenioidei and the Percophidae into the new Notothenioidea. 'Notothenioidei' has the suborder termination; while 'Notothenioidea' has the super family termination: the second cannot contain the first. Therefore, the name Notothenioidea should be replaced by the name Notothenioidi. The status of our Notothenioidiformes with regard to Smith and Craig's Percoidei will be clarified once sequences of *Niphon*, *Acanthistius* and *Bembrops* will be accessible for the present four molecular markers.

4.2 Supertrees and reliability

In Li and Lecointre (in press), since all trees were built on the same set of taxa, the repetition indices could easily be mapped on the summary tree. Here, the MRP supertree plays the role of a summary tree and is obtained from several partially-overlapping validity domains. Since the repetition indices that were used to weight the clades are only valid in their restricted validity domains, no repetition index is displayed on the supertree. MRP is known to have biases. Even if the bipartitions were weighted according to their repetition indices, the reliability of clades shown in the summary tree holds if the supertree method used is itself accurate. Moreover, our conclusions could be affected by some taxa with undetermined position because the method (supertree based on clades weighted according to Li and Lecointre's repetition index) does not seem to handle well taxa present in only one elementary dataset (e.g. as a result *Acanthistius*, *Plectropomus* and *Liopropoma* do not join serranids, *Aeoliscus* does not group with macrorhamphosids, *Centropomus* is not close to *Lates*, *Balistes* fails to join the Tetraodontiformes, *Pseudaphritis* fails to join notothenioids, *Carapus* is not close to *Echiodon*). The interpretation given to clades present in the 'total evidence' tree (Figure 4) and not in Figure 5 — as a possible bias in the tree from the 'total evidence' — must therefore be taken with caution.

Acknowledgements

This work was supported by the 'Consortium National de Recherche en Génomique', and the 'Service de Systématique Moléculaire' of the Muséum National d'Histoire Naturelle (IFR 101). It is part of the agreement number 2005/67 between the Genoscope and the Muséum National d'Histoire Naturelle on the project 'Macrophylogeny of life' directed by Guillaume Lecointre. Thanks to Damien Hinsinger for some of the new MLL sequences. Thanks to Bruno Chanet for help and advice.

References

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., Lipman, D. J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25 (17), 3389–3402.

Barrett, M., Donoghue, M., Sober, E., 1991. Against consensus. Syst. Zool. 40 (4), 486–493.

Breder, C. M., Rosen, D. E., 1966. Modes of reproduction in fishes. The Natural History Press, New York.

Britz, R., Johnson, D., 2002. "paradox lost": skeletal ontogeny of *Indostomus paradoxus*, and its significance for the phylogenetic relationships of Indostomidae (Teleostei, Gasterosteiformes). Am. Mus. Novit. (3383), 1–43.
URL `http://hdl.handle.net/2246/2872`

Bull, J., Huelsenbeck, J., Cunningham, C., Swofford, D., Waddell, P., 1993. Partitioning and combining data in phylogenetic analysis. Syst. Biol. 42 (3), 384–397.

Chen, W.-J., 2001. La répétitivité des clades comme critère de fiabilité: application à la phylogénie des Acanthomorpha (Teleostei) et des Notothenioidei (acanthomorphes antarctiques). Ph.D. thesis, Université Paris VI Pierre et Marie Curie.

Chen, W.-J., Bonillo, C., Lecointre, G., 2000. Taxonomic congruence as a tool to discover new clades in the acanthomorph (Teleostei) radiation. In: Program Book and Abstracts, 80th Annual Meeting ASIH, La Paz, México, June 14-20, 2000. American Society of Ichthyologists and Herpetologists, American Society of Ichthyologists and Herpetologists, p. 369.

Chen, W.-J., Bonillo, C., Lecointre, G., 2003. Repeatability of clades as a criterion of reliability: a case study for molecular phylogeny of Acanthomorpha (Teleostei)

with larger number of taxa. Mol. Phylogenet. Evol. 26, 262–288.

Chen, W.-J., Ruiz-Carus, R., Ortí, G., 2007. Relationships among four genera of mojarras (Teleostei: Perciformes: Gerreidae) from the western Atlantic and their tentative placement among percomorph fishes. J. Fish Biol. 70, 202–218.
URL http://dx.doi.org/10.1111/j.1095-8649.2007.01395.x

Dettaï, A., Lecointre, G., 2004. In search of nothothenioid (Teleostei) relatives. Antarct. Sci. 16 (1), 71–85.
URL http://dx.doi.org/10.1017/S0954102004

Dettaï, A., Lecointre, G., 2005. Further support for the clades obtained by multiple molecular phylogenies in the acanthomorph bush. C.R. Biol. 328, 674–689.

Dettaï, A., Lecointre, G., 2008. New insights into the organization and evolution of vertebrate IRBP genes and utility of IRBP gene sequences for the phylogenetic study of the Acanthomorpha (Actinopterygii: Teleostei). Mol. Phylogenet. Evol. 48 (1), 258–269.
URL http://dx.doi.org/10.1016/j.ympev.2008.04.003

Dettaï, A., Lecointre, G., submitted. Clade reliability in spiny teleosts (Acanthomorpha) through multiple combinations of independent datasets. Submitted.

Froese, R., Pauly, D., 2006. Fishbase. World Wide Web electronic publication. www.fishbase.org, version (06/2006).
URL www.fishbase.org

Gill, A. C., Mooi, R. D., 1993. Monophyly of the Grammatidae and of the Notograptidae, with evidence for their phylogenetic positions among perciforms. B. Mar. Sci. 52 (1), 327–350.

Grande, L., 1994. Repeating patterns in nature, predictability, and "impact" in

science. In: Grande, L., Rieppel, O. (Eds.), Interpreting the hierarchy of nature, 1st Edition. Academic Press, New York, pp. 61–84.

Guindon, S., Gascuel, O., 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52 (5), 696–704.
URL http://dx.doi.org/10.1080/10635150390235520

Hall, T., 2001. Bioedit version 5.0.6. North Carolina State University, Department of Microbiology.

Holcroft, N. I., Wiley, E. O., 2008. Acanthuroid relationships revisited: a new nuclear gene-based analysis that incorporates tetraodontiform representatives. Ichthyol. Res. Online, 1–10.
URL http://dx.doi.org/10.1007/s10228-007-0026-x

Hubbard, T., Andrews, D., Caccamo, M., Cameron, G., Chen, Y., Clamp, M., Clarke, L., Coates, G., Cox, T., Cunningham, F., Curwen, V., Cutts, T., Down, T., Durbin, R., Fernandez-Suarez, X. M., Gilbert, J., Hammond, M., Herrero, J., Hotz, H., Howe, K., Iyer, V., Jekosch, K., Kahari, A., Kasprzyk, A., Keefe, D., Keenan, S., Kokocinsci, F., London, D., Longden, I., McVicker, G., Melsopp, C., Meidl, P., Potter, S., Proctor, G., Rae, M., Rios, D., Schuster, M., Searle, S., Severin, J., Slater, G., Smedley, D., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Storey, R., Trevanion, S., Ureta-Vidal, A., Vogel, J., White, S., Woodwark, C., Birney, E., 2005. Ensembl 2005. Nucleic Acids Res. 33, 447–453.
URL http://dx.doi.org/10.1093/nar/gki138

Johnson, D., 1984. Percoidei: development and relationships. In: G., M. H., Richards, W. J., Cohen, D. M., Fahay, M. P., Kendall Jr, A. W., Richardson, S. L. (Eds.), Ontogeny and systematics of fishes (based on an international

symposium dedicated to the memory of Elbert Halvor Ahlstrom). Allen Pres, Lawrence, Kansas, pp. 464–498.

Johnson, D., 1993. Percomorph phylogeny: progress and problems. B. Mar. Sci. 52 (1), 3–28.

Johnson, D., Patterson, C., 1993. Percomorph phylogeny: a survey of acanthomorphs and a new proposal. B. Mar. Sci. 52 (1), 554–626.

Kawahara, R., Miya, Masaki an Mabuchi, K., Lavoué, S., Inoue, J., Satoh, T., Kawaguchi, A., Nishida, M., 2008. Interrelationships of the 11 gasterosteiform families (sticklebacks, pipefishes, and their relatives): A new perspective based on whole mitogenomes sequences from 75 higher teleosts. Mol. Phylogenet. Evol. 46, 224–236.

URL http://dx.doi.org/10.1016/j.ympev.2007.07.009

Lecointre, G., Deleporte, P., 2005. Total evidence requires exclusion of phylogenetically misleading data. Zool. Scr. 34 (1), 101–117.

Li, B., Lecointre, G., in press. Formalizing reliability in the taxonomic congruence approach. Article accepted by *Zoologica Scripta*.

Li, C., Orti, G., Zhang, G., Lu, G., 2007. A practical approach to phylogenomics: the phylogeny of ray-finned fish (Actinopterygii) as a case study. BMC Evolutionary Biology 7 (44), 1–11.

URL http://dx.doi.org/10.1186/1471-2148-7-44

Mabuchi, K., Miya, M., Azuma, Y., Nishida, M., 2007. Independent evolution of the specialized phartngeal jaw apparatus in cichlid and labrid fishes. BMC Evolutionary Biology 7 (10), 1–12.

URL http://dx.doi.org/10.1186/1471-2148-7-10

Mickevich, M., 1978. Taxonomic congruence. Syst. Zool. 27, 143–158.

Miya, M., Kawaguchi, A., Nishida, M., 2001. Mitogenomic exploration of higher teleostean phylogenies: A case study for moderate-scale evolutionary genomics with 38 newly determined complete mitochondrial dna sequences. Mol. Biol. Evol. 18 (11), 1993–2009.

Miya, M., Satoh, T., Nishida, M., 2005. The phylogenetic position of toadfishes (order Batrachoidiformes) in the higher ray-finned fish as inferred from partitioned Bayesian analysis of 102 whole mitochondrial genome sequences. Biol. J. Linn. Soc. 85, 289–306.

Miya, M., Takeshima, H., Endo, H., Ishiguro, N., Inoue, J., Mukai, T., Satoh, T., Yamaguchi, M., Kawaguchi, A., Mabuchi, K., Shirai, S., Nishida, M., 2003. Major patterns of higher teleostean phylogenies: a new perspective based on 100 complete mitochondrial DNA sequences. Mol. Phylogenet. Evol. 26, 121–138.

Miyamoto, M., Fitch, W., 1995. Testing species phylogenies and phylogenetic methods with congruence. Syst. Biol. 44 (1), 64–75.

Mooi, R. D., 1990. Egg surface morphology of pseudochromoids (Perciformes: Percoidei), with comments on its phylogenetic implications. Copeia (2), 455–475.

Mooi, R. D., 1993. Phylogeny of the Plesiopidae (Pisces: Perciformes), with evidence for the inclusion of the Acanthoclinidae. B. Mar. Sci. 52 (1), 284–326.

Nelson, G. J., 1979. Cladistic analysis and synthesis: principles and definitions, with a historical note on Adanson's Familles des plantes (1763-1764). Syst. Zool. 28, 1–21.

Nelson, G. J., 1989. Phylogeny of major fish groups. In: Jörnvall, H., Fernholm, B., Bremer, K. (Eds.), The hierarchy of life: molecules and morphology in phylog-

nentic analysis. Nobel Foundation, Excerpta Medica, Amsterdam, pp. 325–336.

Nelson, J., 1994. Fishes of the World, 3rd Edition. John Wiley and Sons, Inc., Hoboken, New Jersey.

Nelson, J., 2006. Fishes of the World, 4th Edition. John Wiley and Sons, Inc., Hoboken, New Jersey.

Nixon, K. C., Carpenter, J. M., 1996. On simultaneous analysis. Cladistics 12 (3), 221–241.
URL http://dx.doi.org/10.1111/j.1096-0031.1996.tb00010.x

Page, R., Holmes, E., 1998. Molecular Evolution. Blackwell Science, Oxford.

Parenti, L. R., 1993. Relationships of the atherinomorph fishes (Teleostei). B. Mar. Sci. 52 (1), 170–196.

Philippe, H., Douzery, E., 1994. The pitfalls of molecular phylogeny based on four species, as illustrated by the Cetacea/Artiodactyla relationships. J. Mammal. Evol. 2 (2), 133–152.

Prokofiev, A. M., 2006. A new genus of cardinalfishes (Perciformes: Apogonidae) from the south China sea, xith a discussion of the relationships between the families Apogonidae and Kurtidae. Journal of Ichthyology 46 (4), 279–291.
URL http://dx.doi.org/10.1134/S0032945206040011

Prokofiev, A. M., 2007. The osteology of *Bathysphyraenops symplex* and the diagnosis of the Howellidae (Perciformes: Percoidei) family. Journal of Ichthyology 47 (8), 566–578.
URL http://dx.doi.org/10.1134/S0032945207080036

Rambaut, A., 2002. Se-Al, Sequence Alignment Editor, version 2.0a11. Department of Zoology, University of Oxford, South Parks Road, Oxford, OX1 3PS, UK.

Rieppel, O., 2004. The language of systematics, and the philosophy of 'total evidence'. Syst. Biodivers. 2 (1), 9–19.

Smith, L., Craig, M., 2007. Casting the percomorph net widely: the importance of broad taxonomic sampling in the search for the placement of serranid and percid fishes. Copeia (1), 35–55.

Smith, L., Wheeler, W., 2004. Polyphyly of the mail-cheeked fishes (Teleostei: Scorpaeniformes): evidence from mitochondrial and nuclear sequence data. Mol. Phylogenet. Evol. 32, 627–646.

Smith, L., Wheeler, W., 2006. Venom evolution widespread in fishes: a phylogenetic road map for the bioprospecting of piscine venoms. J. Hered. 97 (3), 206–217.
URL http://dx.doi.org/10.1093/jhered/esj034

Stiassny, M., Moore, J., 1992. A review of the pelvic girdle of acanthomorph fishes, with comments on hypotheses of acanthomorph intrarelationships. Zool. J. Linn. Soc.-Lond. 104, 209–242.

Stiassny, M. L. J., Wiley, E. O., Johnson, G. D., de Carvalho, M. R., 2004. Gnathostome fishes. In: Cracraft, J., Donoghue, M. J. (Eds.), Assembling the tree of life. Oxford University Press, New-York.

Swofford, D., 2002. PAUP*. Phylogenetic analysis using parsimony (* and other methods), version 4.0b10. Sinauer Associates, Sunderland, Massachusetts.

Wiens, J., Reeder, T., 1995. Combining data sets with different numbers of taxa for phylogenetic analysis. Syst. Biol. 44 (4), 548–558.
URL http://dx.doi.org/10.2307/2413660

Wiley, E. O., Johnson, G. D., Dimmick, W. W., 2000. The interrelationships of acanthomorph fishes: a total evidence approach using molecular and morpho-

logical data. Biochem. Syst. Ecol. 28, 319–350.

Winnepenninckx, B., Backeljau, T., Dewachter, R., 1993. Extraction of high-molecular-weight DNA from mollusks. Trends Genet. 9 (12), 407.

Winterbottom, R., 1993. Search for the gobioid sister group (Actinopterygii: Percomorpha). B. Mar. Sci. 52 (1), 395–414.

Yamanoue, Y., Miya, M., Matsuura, K., Yagishita, N., Mabuchi, K., Sakai, H., Katoh, M., Nishida, M., 2007. Phylogenetic position of tetraodontiform fishes within the higher teleosts: bayesian inferences based on 44 whole mitochondrial genome sequences. Mol. Phylogenet. Evol. 45, 89–101.
URL http://dx.doi.org/10.1016/j.ympev.2007.03.008

Table 1: Primers used in this study.

Table 2: Sequences used in this study. Taxonomy follows Nelson (2006), except for families, that are those present in Fishbase (Froese and Pauly, 2006). Sequences in boldface font are new sequences.

Figure 1: The 3 levels of validity domains. The first-level validity domains (V_X) are the sets of leaves of the trees (T_X) obtained by the analyses of the datasets (X). In this example, 3 elementary datasets are used, which leads to 7 datasets, and thus to 7 trees and 7 first-level validity domains. The second-level validity domains (V_{PSc_i}) are the intersection of the validity domains of the independant datasets involved in the partitioning schemes (PSc_i). Here, only the full partitioning schemes are shown. The occurrences of the clades are counted within a partitioning scheme, across its constituting datasets, after pruning the corresponding trees from the taxa outside the relevant second-level validity domain. The third-level validity domains (W_i) are the intersections of all possible combinations of second-level validity domains. The repetition indices are attached to such third-level validity domains. They are based on the maximum number of occurrences (for the clades once pruned from the taxa outside the third-level validity domain) found among the partitioning schemes whose validity domains span at least the entire third-level validity domain. Only some of the possible third-level validity domains are shown here.

Figure 2 (page 61): Maximum likelihood trees obtained by the analysis of the new RNF213 sequence data matrix under a GTR + I + Γ model with phyml (Guindon and Gascuel, 2003).

Figure 3 (page 62): Maximum likelihood trees obtained by the analysis of the combined matrix Rhodopsin+MLL+IRBP under a GTR + I + Γ model with phyml (Guindon and Gascuel, 2003).

Figure 4 (page 63): Maximum likelihood trees obtained by the analysis of the combined matrix of the four nuclear genes ('total evidence') of the new RNF213 sequence data matrix under a GTR + I + Γ model with phyml (Guindon and Gascuel, 2003).

Figure 5 (page 64): Supertree exhibiting the clades having the highest repetition indices (Li and Lecointre, in press) in the partial combination and validity domains approach. The tree is the majority-rule consensus of the most parsimonious trees obtained by 10,000 RAS + TBR parsimony analyses by PAUP* (Swofford, 2002) of the matrix representing the clades with positive repetition index, weighted by their repetition indices. The values above the branches are the percentage of equally parsimonious trees that include the clade. The illustrations come from Fishbase (Froese and Pauly, 2006).

Table 3: Clades of interest extracted from the separate and multiple combined analyses and repetition index. New names are proposed for some of those reliable clades of the 'acanthomorph bush'. First column lists the contents of the clade; the name given to the clade refers to the last common ancestor to the taxa indicated. Second column indicates by a cross the presence of the clade in figure 2. A blank means that the clade is not recovered in the tree figure 2 and a question mark indicates that the presence of the clade cannot be assessed either because of an incomplete taxonomic sampling or because of irresolution. Third column indicates

presence of clades in figure 3. Fourth column indicates presence of clades in figure 4. The mention 'X-1' indicates that a single taxon escapes from the clade (generally a long branch). Fifth column indicates presence of the clade in the summary tree of figure 5. Sixth column gives the letter associated to the clade in Chen et al. (2003) and Dettaï and Lecointre (2005, 2008) and proposes letters for new clades (P, P', R, S, T, U, V, W, Y, Z, L', L", M', M"). Seventh column refers to the presence of the clade in the study of Dettaï and Lecointre (2008) based on the IRBP gene. Eighth column records the clade in the studies of Miya et al. (2003, 2005) based on independent mitochondrial sequence data; ninth column in the study of Mabuchi et al. (2007) using the same genes as in Miya et al. (2005) and tenth column in the study of Kawahara et al. (2008) also based on the same markers. Eleventh column records the clades in the study of Smith and Craig (2007) based on mitochondrial and nuclear data independent from both Dettaï and Lecointre (2005) and Miya et al. (2005). In the last five columns, question marks are given either when taxonomic sampling is insufficient or the interrelationships unresolved. The last column proposes a name to some of the clades that have been repeatedly found in several previous studies, or reliable clades newly identified by the present study.

Table 1: Primers used in this study.

Primer	Sequence (5'→3')	Forward/Reverse	Source
Rh193	CNTATGAATAYCCTCAGTACTACC	Forward	Chen et al. (2003)
Rh667r	AYGAGCACTGCATGCCCT	Reverse	Chen et al. (2003)
Rh1039r	TGCTTGTTCATGCAGATGTAGA	Reverse	Chen et al. (2003)
Rh1073r	CCRCAGCACARCGTGGTGATCATG	Reverse	Chen et al. (2003)
MLL U1477	AGYCCAGCRGTCATCAAACC	Forward	Dettaï and Lecointre (2005)
MLL U1499	GTCAATCAGCAGTTCCAGC	Forward	Dettaï and Lecointre (2005)
MLL U1570	CCCYCAAAAKATCARTGCCAC	Forward	This study
MLL U1590	CRGGRGTGATNGACACCAGC	Forward	This study
MLL L2080	GTGAACTCMAYCAGTCCTCC	Reverse	This study
MLL 2105	ACCYTGCGTTGGGARGTGG	Reverse	This study
MLL L2158	ARAGTAGTGGGATCYAGRTACAT	Reverse	Dettaï and Lecointre (2005)
IRBP U104	ATAGTYNTGGACAANTACTGCTC	Forward	Dettaï and Lecointre (2008)
IRBP U110	TGGACAAYTACTGCTCRCCAGA	Forward	Dettaï and Lecointre (2008)
IRBP L936	CACGGAGGYTGAYNATCTTGAT	Reverse	Dettaï and Lecointre (2008)
IRBP L953	CNGGAAYYTGARCACGGAGG	Reverse	Dettaï and Lecointre (2008)
IRBP L1338	GTGRAAGGAGAYTTTGATCAGCTC	Reverse	Dettaï and Lecointre (2008)
C17 F3111	GCTGACTGGATTYAAAACCTT	Forward	This study
C17 F3128	CCTTTGTGGTGGAYTTYATGAT	Forward	This study
C17 F3150	WCTGATGGCNAARGACTTTGC	Forward	This study
C17 R4036	GGRATRGANCCNAGCTTTTCAT	Reverse	This study
C17 R4096	CCANACCAGAGGGATCATRCT	Reverse	This study
C17 R4111	AACTGTCCAAARTCCCACAC	Reverse	This study

Table 2:

Order/Suborder	Familly	Genus/Species	Rhodopsine	MLL	IRBP	RNF213
Argentiniformes						
Alepocephaloidei						
	Alepocephalidae	*Alepocephalus antipodianus* (Parrott, 1948)	**EU637933**	-	-	-
Stomiiformes						
Gonostomatoidei						
	Gonostomatidae	*Gonostoma bathyphilum* (Vaillant, 1884)	AY141256	-	-	-
Aulopiformes						
Chlorophthalmoidei						
	Ipnopidae	*Bathypterois dubius* Vaillant, 1888	**AY141257**	AY362219	DQ168042	-
Myctophiformes						
	Myctophidae	*Electrona antarctica* (Günther, 1878)	AY141258	AY36220	-	-
Lampriformes						
	Lampridae	*Lampris immaculatus* Gilchrist, 1904	AY141259	-	DQ168077	-
	Trachipteridae	*Trachipterus arcticus* (Brünnich, 1788)	-	-	**EU638158**	-
	Regalecidae	*Regalecus glesne* Ascanius, 1772	**AY368328**	AY362266	DQ168109	**EU638252**
Polymixiiformes						
	Polymixiidae	*Polymixia nobilis* Lowe, 1838	AY368320	AY362208	DQ168104	-
Percopsiformes						
	Aphredoderidae	*Aphredoderus sayanus* (Gilliams, 1824)	-	-	DQ168038	-
Gadiformes						
	Muraenolepididae	*Muraenolepis marmorata* Günther, 1880	-	**EU638073**	-	-
	Macrouridae	*Coryphaenoides rupestris* Gunnerus, 1765	AY368319	**EU638041**	-	-

Table 2: (continued)

Order/Suborder	Familly	Genus/Species	Rhodopsine	MLL	IRBP	RNF213
	Macrouridae	*Trachyrincus murrayi* Günther, 1887	AY368318	**AY362289**	DQ168124	**EU638270**
	Moridae	*Mora moro* (Risso, 1810)	AY368322	**EU638071**	DQ168089	**EU638227**
	Merlucciidae	*Merluccius merluccius* (L., 1758)	-	**EU638068**	-	-
	Phycidae	*Phycis phycis* (L., 1766)	**EU637994**	-	-	-
	Lotidae	*Enchelyopus cimbrius* (L., 1766)	**EU637958**	-	-	-
	Lotidae	*Gaidropsarus novaezelandiae* (Hector, 1874)	-	**EU638051**	-	-
	Lotidae	*Gaidropsarus sp.*	**EU637961**	-	-	-
	Lotidae	*Gaidropsarus vulgaris* (Cloquet, 1824)	-	-	DQ168067	-
	Gadidae	*Gadus morhua* L., 1758	AF137211	**EU638050**	DQ168066	-
	Gadidae	*Merlangius merlangus* (L., 1758)	AY141260	-	-	-
Ophidiiformes						
Ophidioidei						
	Carapidae	*Encheliophis boraborensis* (Kaup, 1856)	-	-	-	**EU638179**
	Carapidae	*Echiodon cryomargarites* Markle, Williams & Olney, 1983	**EU637956**	-	-	-
	Ophidiidae	*Lamprogrammus shcherbachevi* Cohen & Rohr, 1993	**EU637969**	**EU638058**	**EU638130**	-
Bythitoidei						
	Bythitidae	*Cataetyx laticeps* Koefoed, 1927	**EU637947**	**EU638035**	-	-
Batrachoidiformes						
	Batrachoididae	*Halobatrachus didactylus* (Bloch & Schneider, 1801)	AY368323	AY362246	DQ168069	**EU638205**
Lophiiformes						
Lophioidei						
	Lophiidae	*Lophius budegassa* Spinola, 1807	-	-	-	**EU638217**

Table 2: (continued)

Order/Suborder	Familly	Genus/Species	Rhodopsine	MLL	IRBP	RNF213
	Lophiidae	*Lophius piscatorius* L., 1758	AY368325	AY362274	-	-
Antennarioidei						
	Antennariidae	*Antennarius striatus* (Shaw, 1794)	AY368324	AY362215	DQ168037	-
Ogcocephaloidei						
	Himantolophidae	*Himantolophus groenlandicus* Reinhardt, 1837	**EU637965**	**EU638055**	**EU638125**	-
	Ceratiidae	*Ceratias holboelli* Krøyer, 1845	AY141263	AY362270	DQ168049	**EU638181**
Mugiliformes						
	Mugilidae	*Liza sp.*	AY141266	AY362248	DQ168082	-
Atheriniformes						
	Atherinopsidae	*Menidia menidia* (L., 1766)	**EU637977**	**EU638067**	**EU638137**	-
	Bedotiidae	*Bedotia geayi* Pellegrin, 1907	AY141267	AY362271	DQ168043	-
Beloniformes						
	Adrianichthyidae	*Oryzias latipes* (Temminck & Schlegel, 1946)	-	-	DQ168094	Ensembl
	Exocoetidae	*Cheilopogon heterurus* (Rafinesque, 1810)	**EU637950**	**EU638039**	**EU638113**	**EU638184**
	Belonidae	*Belone belone* (L., 1761)	AY141268	AY362273	DQ168044	-
Cyprinodontiformes						
	Anablepidae	*Anableps anableps* (L., 1758)	**EU637935**	-	-	-
	Poeciliidae	*Poecilia reticulata* Peters, 1859	Y11147	AY362203	DQ168102	**EU638243**
Stephanoberyciformes						
	Rondeletiidae	*Rondeletia sp.*	AY368327	**EU638087**	DQ168110	-
	Barbourisiidae	*Barbourisia rufa* Parr, 1945	AY368333	AY362264	DQ168041	-

Table 2: (continued)

Order/Suborder	Familly	Genus/Species	Rhodopsine	MLL	IRBP	RNF213
Beryciformes						
Trachichthyoidei						
	Anomalopidae	*Photoblepharon palpebratum* (Boddaert, 1781)	**EU637993**	AY362268	DQ168101	**EU638242**
	Diretmidae	*Diretmoides sp.*	-	AY362205	DQ168060	-
	Trachichthyidae	*Hoplostethus atlanticus* Collett, 1889	-	-	**EU638127**	**EU638207**
	Trachichthyidae	*Hoplostethus mediterraneus* Cuvier, 1829	AY141264	AY362267	-	-
Berycoidei						
	Berycidae	*Beryx splendens* Lowe, 1834	AY141265	AY362238	DQ168045	**EU638174**
Holocentroidei						
	Holocentridae	*Myripristis botche* Cuvier, 1929	-	AY362265	DQ168091	-
	Holocentridae	*Myripristis sp.*	**EU637983**	-	-	**EU638230**
Zeiformes						
	Oreosomatidae	*Neocyttus helgae* (Holt & Byrne, 1908)	AY141261	**AY362288**	-	-
	Grammicolepididae	*Grammicolepis brachiusculus* Poey, 1873	**EU637964**	**EU638054**	**EU638124**	-
	Zeidae	*Zenopsis conchifera* (Lowe, 1852)	AY368314	AY362286	DQ168127	**EU638279**
	Zeidae	*Zeus faber* L., 1758	**EU638023**	**AY362287**	DQ168128	-
Gasterosteiformes						
Gasterosteoidei						
	Gasterosteidae	*Gasterosteus aculeatus* L., 1758	**EU637962**	**EU638052**	Ensembl	Ensembl
	Gasterosteidae	*Spinachia spinachia* (L., 1758)	AY141281	AY362261	-	**EU638264**
	Indostomidae	*Indostomus paradoxus* Prashad & Mukerji, 1929	**EU637967**	**EU638057**	-	**EU638209**

Table 2: (continued)

Order/Suborder	Familly	Genus/Species	Rhodopsine	MLL	IRBP	RNF213
Syngnathoidei						
	Syngnathidae	*Hippocampus guttulatus* Cuvier, 1829	AY368330	AY362216	**EU638126**	-
	Syngnathidae	*Nerophis lumbriciformis* (Jenyns, 1835)	**EU637987**	-	**EU638143**	**EU638232**
	Syngnathidae	*Nerophis ophidion* (L., 1758)	-	-	DQ168071	-
	Syngnathidae	*Syngnathus typhle* L., 1758	AY368326	AY362211	DQ168120	-
	Fistulariidae	*Fistularia petimba* Lacépède, 1803	AY141324	-	-	**EU638202**
	Aulostomidae	*Aulostomus chinensis* (L., 1766)	AY141279	AY362226	DQ168040	-
	Centriscidae	*Aeoliscus strigatus* (Günther, 1861)	**EU637931**	-	**EU638100**	-
	Centriscidae	*Macroramphosus scolopax* (L., 1758)	**AY141280**	AY362206	DQ168083	-
Symbranchiformes						
Symbranchoidei						
	Symbranchidae	*Monopterus albus* (Zuiew, 1793)	AY141276	AY362252	DQ168088	**EU638226**
Mastacembeloidei						
	Mastacembelidae	*Mastacembelus erythrotaenia* Bleeker, 1850	**AY141275**	**AY362249**	**DQ168084**	-
Scorpaeniformes						
Dactylopteroidei						
	Dactylopteridae	*Dactylopterus volitans* (L., 1758)	AY141282	AY362243	DQ168059	-
Scorpaenoidei						
	Sebastidae	*Sebastes sp.*	-	-	-	**EU638258**
	Scorpaenidae	*Pontinus longispinis* Goode & Bean, 1896	**EU637996**	**EU638081**	**EU638146**	**EU638247**
	Scorpaenidae	*Scorpaena onaria* Jordan & Snyder, 1900	AY141288	AY362236	DQ168114	**EU638257**
	Synanceiidae	*Synanceia verrucosa* Bloch & Schneider, 1801	**EU638011**	**EU638093**	**EU638156**	**EU638267**

Table 2: (continued)

Order/Suborder	Familly	Genus/Species	Rhodopsine	MLL	IRBP	RNF213
	Congiopodidae	*Zanclorhynchus spinifer* Günther, 1880	**EU638021**	-	**EU638165**	**EU638278**
Platycephaloidei						
	Triglidae	*Chelidonichthys lucernus* (L., 1758)	AY141287	AY362284	DQ168053	**EU638186**
Cottoidei						
	Cottidae	*Taurulus bubalis* (Euphrasen, 1786)	U97275	AY362217	DQ168121	-
	Agonidae	*Agonopsis chiloensis* (Jenyns, 1840)	**EU637932**	**EU638025**	**EU638101**	**EU638167**
	Agonidae	*Xeneretmus latifrons* (Gilbert, 1890)	**EU638018**	**EU638097**	**EU638162**	-
	Psychrolutidae	*Cottunculus thomsonii* (Günther, 1882)	AY368315	AY362260	-	-
	Cyclopteridae	*Cyclopterus lumpus* L., 1758	AY368316	**AY362218**	**EU638116**	-
	Liparidae	*Liparis fabricii* Krøyer, 1847	AY368317	AY362235	DQ168081	-
Perciformes						
Percoidei						
	Centropomidae	*Centropomus undecimalis* (Bloch, 1792)	-	-	-	**EU638180**
	Lateolabracidae	*Lateolabrax japonicus* (Cuvier, 1828)	AY141293	AY362253	DQ168078	**EU638213**
	Latidae	*Lates calcarifer* (Bloch, 1970)	**EU637970**	**EU638059**	DQ168075	**EU638214**
	Latidae	*Lates niloticus* (L., 1758)	**EU637971**	-	-	-
	Moronidae	*Dicentrarchus labrax* (L., 1758)	-	-	**EU638119**	**EU638195**
	Moronidae	*Morone saxatilis* (Walbaum, 1792)	**EU637981**	**EU638072**	**EU638140**	**EU638228**
	Percichthyidae	*Howella brodiei* Ogilby, 1899	**EU637966**	**EU638056**	**EU638128**	**EU638208**
	Serranidae	*Acanthistius brasilianus* (Cuvier, 1828)	-	**EU638024**	-	-
	Serranidae	*Cephalopholis urodeta* (Forster, 1801)	-	**EU638036**	-	-
	Serranidae	*Dermatolepis dermatolepis* (Boulenger, 1895)	-	**EU638045**	-	-

Table 2: (continued)

Order/Suborder	Familly	Genus/Species	Rhodopsine	MLL	IRBP	RNF213
	Serranidae	*Epinephelus aeneus* (Geoffroy Saint-Hilaire, 1817)	AY141291	**EU638049**	AY362227	**EU638201**
	Serranidae	*Odontanthias chrysostictus* (Günther, 1872)	**AY141290**	**AY362209**	**DQ168073**	**EU638206**
	Serranidae	*Liopropoma fasciatum* Bussing, 1980	-	**EU638062**	-	-
	Serranidae	*Niphon spinosus* Cuvier, 1828	**EU637934**	-	-	-
	Serranidae	*Plectropomus leopardus* (Lacépède, 1802)	-	**EU638078**	-	-
	Serranidae	*Pogonoperca punctata* (Valenciennes, 1830)	AY141292	AY362256	DQ168103	**EU638244**
	Serranidae	*Pseudanthias squamipinnis* (Peters, 1855)	-	**EU638083**	-	-
	Serranidae	*Rypticus saponaceus* (Bloch & Schneider, 1801)	AY368329	**AY362257**	**DQ168111**	**EU638253**
	Serranidae	*Serranus accraensis* (Norman, 1931)	AY141289	AY362202	DQ168115	**EU638260**
	Callanthiidae	*Callanthias ruber* (Rafinesque, 1810)	**EU637945**	**EU638034**	**EU638110**	-
	Plesiopidae	*Assessor flavissimus* Allen & Kuiter, 1976	**EU637944**	**EU638032**	**EU638109**	**EU638173**
	Centrarchidae	*Lepomis gibbosus* (L., 1758)	AY742571	**EU638061**	**EU638132**	**EU638216**
	Percidae	*Gymnocephalus cernuus* (L., 1758)	AY141296	AY362278	DQ168068	-
	Percidae	*Perca fluviatilis* L., 1758	AY141295	AY362279	DQ168099	**EU638240**
	Priacanthidae	*Priacanthus arenatus* Cuvier, 1829	**EU637997**	**EU638082**	**EU638147**	-
	Epigonidae	*Epigonus telescopus* (Risso, 1810)	**EU637959**	**EU638048**	**EU638122**	**EU638200**
	Apogonidae	*Apogon fasciatus* (White, 1970)	**EU637940**	-	-	**EU638171**
	Apogonidae	*Sphaeramia nematoptera* (Bleeker, 1856)	**EU638010**	**EU638091**	**EU638154**	-
	Malacanthidae	*Lopholatilus chamaeleonticeps* Goode & Bean, 1879	**EU637973**	**EU638063**	**EU638133**	**EU638218**
	Sillaginidae	*Sillago sihama* (Forsskål, 1775)	**EU638008**	-	-	**EU638262**
	Coryphaenidae	*Coryphaena equiselis* L., 1758	**EU637951**	**EU638040**	**EU638114**	**EU638189**
	Coryphaenidae	*Coryphaena hippurus* L., 1758	-	-	DQ168056	-

Table 2: (continued)

Order/Suborder	Familly	Genus/Species	Rhodopsine	MLL	IRBP	RNF213
	Echeneidae	*Echeneis naucrates* L., 1758	AY141315	AY362245	DQ168062	**EU638197**
	Carangidae	*Chloroscombrus chrysurus* (L., 1766)	AY141313	AY362223	DQ168054	**EU638187**
	Carangidae	*Gnathanodon speciosus* (Forsskål, 1755)	**EU637963**	**EU638053**	**EU638123**	**EU638204**
	Carangidae	*Selene dorsalis* (Gill, 1863)	**EU638006**	**EU638089**	**EU638153**	**EU638259**
	Carangidae	*Trachinotus ovatus* (L., 1758)	AY141314	AY362263	DQ168120	-
	Carangidae	*Trachurus trachurus* (L., 1758)	**EU638013**	-	**EU638159**	**EU638269**
	Menidae	*Mene maculata* (Bloch & Schneider, 1801)	**AY141316**	**AY362250**	DQ168085	**EU638221**
	Leiognathidae	*Leiognathus fasciatus* (Lacépède, 1803)	**EU637972**	**EU638060**	**EU638131**	-
	Bramidae	*Pterycombus brama* Fries, 1837	**EU638001**	**EU638086**	**EU638149**	**EU638251**
	Lutjanidae	*Apsilus fuscus* Valenciennes, 1830				
	Lutjanidae	*Lutjanus sebae* (Cuvier, 1816)	**EU637974**	**EU638064**	**EU638134**	**EU638219**
	Caesionidae	*Pterocaesio digramma* (Bleeker, 1864)	**EU638000**	**EU638085**	**EU638148**	**EU638250**
	Datnioididae	*Datnioides polota* (Hamilton, 1822)	**EU637954**	**EU638044**	**EU638118**	**EU638194**
	Haemulidae	*Pomadasys perotaei* (Cuvier, 1830)	-	AY362230	DQ168105	**EU638246**
	Sparidae	*Spondyliosoma cantharus* (L., 1758)	-	**EU638092**	**EU638155**	**EU638265**
	Sciaenidae	*Argyrosomus regius* (Asso, 1801)	**EU637942**	**EU638030**	**EU638107**	**EU638172**
	Sciaenidae	*Johnius sp.* Bloch, 1793	-	-	**EU638129**	-
	Sciaenidae	*Micropogonias furnieri* (Desmarest, 1823)	**EU637979**	-	-	-
	Sciaenidae	*Micropogonias sp.*	-	-	-	**EU638224**
	Sciaenidae	*Sciaena sp.*	**EU638004**	-	-	-
	Polynemidae	*Pentanemus quinquarius* (L., 1758)	AY141317	AY362272	DQ168098	**EU638239**
	Mullidae	*Mullus surmuletus* L., 1758	**EU637982**	AY362231	DQ168090	**EU638229**

Table 2: (continued)

Order/Suborder	Familly	Genus/Species	Rhodopsine	MLL	IRBP	RNF213
	Toxotidae	*Toxotes sp.*	**EU638012**	**EU638094**	**EU638157**	-
	Monodactylidae	*Monodactylus sp.* Lacépède, 1801	**EU637980**	**EU638070**	**EU638139**	-
	Kyphosidae	*Microcanthus strigatus* (Cuvier, 1831)	**EU637978**	**EU638069**	**EU638138**	**EU638222**
	Chaetodontidae	*Chaetodon semilarvatus* Cuvier, 1831	AY368312	AY362240	DQ168050	-
	Drepaneidae	*Drepane africana* Osório, 1892	AY141321	AY362244	DQ168061	**EU638196**
	Pomacanthidae	*Holacanthus ciliaris* (L., 1758)	AY141322	AY362214	DQ168072	-
	Pomacanthidae	*Pomacanthus maculosus* (Forsskål, 1775)	**EU637995**	**EU638079**	**EU638145**	**EU638245**
	Terapontidae	*Pelates quadrilineatus* (Bloch, 1790)	**EU637991**	-	-	-
	Cheilodactylidae	*Nemadactylus monodactylus* (Carmichael, 1819)	**EU637985**	**EU638075**	**EU638142**	**EU638231**
	Aplodactylidae	*Aplodactylus punctatus* Valenciennes, 1832	**EU637939**	-	-	-
	Cepolidae	*Cepola macrophthalma* (L., 1758)	**EU637948**	**EU638037**	**EU638111**	-
Elassomatoidei						
	Elassomatidae	*Elassoma zonatum* Jordan, 1877	**EU637957**	-	DQ168063	-
Labroidei						
	Cichlidae	*Haplochromis nubilus* (Boulenger, 1906)	-	-	DQ168070	-
	Cichlidae	*Haplochromis sp.*	AB084933	-	-	-
	Pomacentridae	*Dascyllus trimaculatus* (Rüppel, 1829)	**EU637953**	**EU638043**	**EU638117**	**EU638193**
	Labridae	*Labrus bergylta* Ascanius, 1767	AY141318	AY362222	DQ168075	**EU638211**
	Labridae	*Xyrichtys novacula* (L., 1758)	**EU638020**	-	**EU638164**	**EU638277**
	Scaridae	*Scarus hoefleri* (Steindachner, 1881)	AY141319	AY362212	DQ168112	**EU638254**
Zoarcoidei						
	Zoarcidae	*Austrolycus depressiceps* Regan, 1913	AY141297	-	-	-

Table 2: (continued)

Order/Suborder	Familly	Genus/Species	Rhodopsine	MLL	IRBP	RNF213
	Zoarcidae	*Lycodapus antarcticus* Tomo, 1982	**EU637976**	**EU638066**	**EU638136**	-
	Pholidae	*Pholis gunnellus* (L., 1758)	AY141298	AY362285	DQ168100	**EU638241**
	Anarhichadidae	*Anarhichas lupus* L., 1758	**EU637936**	**EU638026**	**EU638103**	**EU638169**
Notothenioidei						
	Nototheniidae	*Notothenia coriiceps* Richardson, 1844	AY141302	AY362282	DQ168093	-
	Nototheniidae	*Trematomus bernachii* Boulenger, 1902	**EU638014**	-	**EU638160**	**EU638271**
	Bovichtidae	*Bovichtus variegatus* Richardson, 1846	**AY141299**	AY362283	DQ168046	**EU638176**
	Bovichtidae	*Cottoperca trigliodes* (Forster, 1801)	AY141300	-	-	-
	Bovichtidae	*Pseudaphritis urvillii* (Valenciennes, 1832)	AY141301	-	-	-
	Eleginopsidae	*Eleginops maclovinus* (Cuvier, 1830)	AY141303	**EU638047**	**EU638121**	**EU638199**
	Channichthyidae	*Chionodraco hamatus* (Lönnberg, 1905)	-	AY362280	-	-
	Channichthyidae	*Neopagetopsis ionah* Nybelin, 1947	**EU637986**	AY362281	DQ16802	-
	Channichthyidae	*Pagetopsis macropterus* (Boulenger, 1907)	**EU637990**	**EU638076**	**EU638144**	**EU638235**
Trachinoidei						
	Chiasmodontidae	*Kali macrura* (Parr, 1933)	AY141308	AY362224	DQ168074	**EU638210**
	Champsodontidae	*Champsodon snyderi* Franz, 1910	**EU637949**	**EU638038**	-	**EU638182**
	Pinguipedidae	*Parapercis clathrata* Ogilby, 1910	-	**EU638077**	-	**EU638238**
	Pinguipedidae	*Pinguipes chilensis* Valenciennes, 1833	**EU637989**	-	-	**EU638234**
	Cheimarrichthyidae	*Cheimarrichthys fosteri* Haast, 1874	**AY141307**	AY362229	DQ168052	**EU638185**
	Trachinidae	*Echiichthys vipera* (Cuvier, 1829)	**EU637955**	**EU638046**	**EU638120**	**EU638198**
	Trachinidae	*Trachinus draco* L., 1758	AY141304	AY362277	DQ168123	**EU638268**
	Ammodytidae	*Ammodytes tobianus* L., 1758	AY141306	AY362234	**EU638102**	**EU638168**

Table 2: (continued)

Order/Suborder	Familly	Genus/Species	Rhodopsine	MLL	IRBP	RNF213
	Uranoscopidae	*Uranoscopus albesca* Regan, 1915	AY141305	AY362239	**DQ168126**	**EU638275**
Blennioidei						
	Tripterygiidae	*Forsterygion lapillum* Hardy, 1989	AY141272	AY362276	DQ168065	**EU638203**
	Tripterygiidae	*Tripterygion delaisi* Cadenat & Blache, 1970	**EU638016**	-	-	**EU638274**
	Blenniidae	*Parablennius gattorugine* (L., 1758)	AY141271	AY362255	DQ168097	**EU638237**
	Blenniidae	*Salaria pavo* (Risso, 1810)	Y18674	-	-	-
Gobiesocoidei						
	Gobiesocidae	*Apletodon dentatus* (Facciolà, 1887)	AY141274	AY362213	DQ168039	-
	Gobiesocidae	*Aspasma minima* (Döderlein, 1887)	**EU637943**	**EU638031**	**EU638108**	-
	Gobiesocidae	*Lepadogaster lepadogaster* (Bonnaterre, 1788)	AY141273	AY362247	DQ168080	**EU638215**
Callionymoidei						
	Callionymidae	*Callionymus lyra* L., 1758	AY141270	AY362225	DQ168047	**EU638177**
	Callionymidae	*Callionymus schaapii* Bleeker, 1852	**EU637946**	-	-	-
Gobioidei						
	Eleotridae	*Ophiocara porocephala* (Valenciennes, 1837)	**EU637988**	-	-	-
	Gobiidae	*Favonigobius reichei* (Bleeker, 1853)	**EU637960**	-	-	-
	Gobiidae	*Periophthalmus barbarus* (L., 1766)	**EU637992**	-	-	-
	Gobiidae	*Pomatoschistus minutus* (Pallas, 1770)	X62405	-	-	-
	Gobiidae	*Pomatoschistus sp.* Gill, 1863	-	**EU638080**	DQ168106	-
	Gobiidae	*Valenciennea strigata* (Broussonet, 1782)	**EU638017**	-	-	-
	Microdesmidae	*Ptereleotris zebra* (Fowler, 1938)	**EU637999**	**EU638084**	-	-

Table 2: (continued)

Order/Suborder	Familly	Genus/Species	Rhodopsine	MLL	IRBP	RNF213
Acanthuroidei						
	Scatophagidae	*Selenotoca multifasciata* (Richardson, 1846)	**EU638002**	**EU638088**	**EU638150**	-
	Siganidae	*Siganus vulpinus* (Schlegel & Müller, 1845)	**EU638007**	**EU638090**	DQ168116	**EU638261**
	Luvaridae	*Luvarus imperialis* Rafinesque, 1810	**EU637975**	**EU638065**	**EU638135**	**EU638220**
	Acanthuridae	*Ctenochaetus sp.*	-	-	-	**EU638190**
	Acanthuridae	*Ctenochaetus striatus* (Quoy & Gaimard, 1825)	AY141320	AY362242	DQ168057	-
	Acanthuridae	*Naso lituratus* (Forster, 1801)	**EU637984**	**EU638074**	**EU638141**	-
Scombroidei						
	Sphyraenidae	*Sphyraena sphyraena* (L., 1758)	AY141312	AY362254	DQ168118	**EU638263**
	Trichiuridae	*Aphanopus carbo* Lowe, 1839	**EU637938**	**EU638028**	**EU638105**	**EU638170**
	Scombridae	*Scomber japonicus* Houttuyn, 1782	AY141311	AY362237	DQ168113	-
	Xiphiidae	*Xiphias gladius* L., 1758	**EU638019**	**EU638098**	**EU638163**	**EU638276**
Stromateoidei						
	Centrolophidae	*Psenopsis anomala* (Temminck & Schlegel, 1844)	AY141310	AY362269	DQ168107	**EU638248**
	Centrolophidae	*Schedophilus medusophagus* (Cocco, 1839)	**EU638003**	**EU660040**	**EU638151**	**EU638255**
	Nomeidae	*Cubiceps gracilis* (Lowe, 1843)	**EU637952**	**EU638042**	**EU638115**	**EU638192**
	Stromateidae	*Pampus argenteus* (Euphrasen, 1788)	AY141309	AY362220	DQ168096	**EU638236**
Anabantoidei						
	Anabantidae	*Ctenopoma sp.*	AY141278	AY362210	DQ168058	**EU638191**
Channoidei						
	Channidae	*Channa sp.*	-	-	-	**EU638183**
	Channidae	*Channa striata* (Bloch, 1793)	AY141277	AY362241	DQ168051	-

Table 2: (continued)

Order/Suborder	Familly	Genus/Species	Rhodopsine	MLL	IRBP	RNF213
Caproidei						
	Caproidae	*Antigonia capros* Lowe, 1843	**EU637937**	**EU638027**	**EU638104**	-
	Caproidae	*Capros aper* (L., 1758)	AY141262	AY362233	DQ168048	**EU638178**
Pleuronectiformes						
Psettodoidei						
	Psettodidae	*Psettodes belcheri* Bennett, 1831	EU637998	AY362259	DQ168108	**EU638249**
Pleuronectoidei						
	Citharidae	*Citharus linguatula* (L., 1758)	AY141323	AY362232	DQ168055	**EU638188**
	Paralichthyidae	*Syacium micrurum* Ranzani, 1842	AY368334	AY362262	DQ168119	**EU638266**
	Scophthalmidae	*Scophthalmus rhombus* (L., 1758)	**EU638005**	-	**EU638152**	**EU638256**
	Scophthalmidae	*Zeugopterus punctatus* (Bloch, 1787)	**EU638022**	**EU638099**	**EU638166**	**EU638280**
	Bothidae	*Arnoglossus imperialis* (Rafinesque, 1810)	AY141283	AY362228	-	-
	Bothidae	*Bothus podas* (Delaroche, 1809)	AY368313	**EU638033**	-	**EU638175**
	Achiridae	*Trinectes maculatus* (Bloch & Schneider, 1801)	**EU638015**	**EU638096**	**EU638161**	**EU638273**
	Soleidae	*Microchirus frechkopi* Chabanaud, 1952	-	-	-	**EU638223**
	Soleidae	*Microchirus variegatus* (Donovan, 1808)	AY141284	AY362275	DQ168086	-
	Soleidae	*Solea solea* (L., 1758)	**EU638009**	-	**DQ168117**	-
Tetraodontiformes						
Triacanthodoidei						
	Triacanthodidae	*Triacanthodes anomalus* (Temminck & Schlegel, 1850)	-	**EU638095**	-	**EU638272**
	Triacanthodidae	*Triacanthodes sp.*	AY368331	-	DQ168125	-
Balistoidei						

Table 2: (continued)

Order/Suborder	Familly	Genus/Species	Rhodopsine	MLL	IRBP	RNF213
	Balistidae	*Balistes sp.*	AF137212	-	-	-
	Ostraciidae	*Ostracion cubicus* L., 1758	-	-	-	**EU638233**
	Ostraciidae	*Ostracion sp.*	AF137213	AY362207	DQ168095	-
Tetraodontoidei						
	Tetraodontidae	*Lagocephalus laevigatus* (L., 1766)	-	AY362221	DQ168076	-
	Tetraodontidae	*Lagocephalus lagocephalus* (L., 1758)	**EU637968**	-	-	**EU638212**
	Tetraodontidae	*Takifugu rubripes* (Temminck & Schlegel, 1850)	-	Ensembl	Ensembl	Ensembl
	Tetraodontidae	*Tetraodon nigroviridis* Marion de Procé, 1822	Ensembl	Ensembl	Ensembl	Ensembl
	Molidae	*Mola mola* (L., 1758)	AF137215	AY362251	DQ168087	**EU638225**

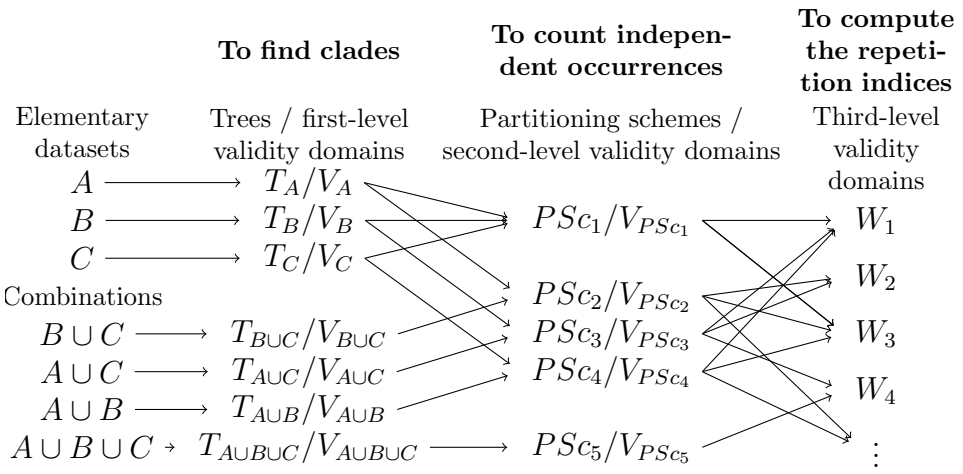

To find clades
To count independent occurrences
To compute the repetition indices
Elementary datasets
Trees / first-level validity domains
Partitioning schemes / second-level validity domains
Third-level validity domains
A
B
C
Combinations
$B \cup C$
$A \cup C$
$A \cup B$
$A \cup B \cup C$
T_A/V_A
T_B/V_B
T_C/V_C
$T_{B\cup C}/V_{B\cup C}$
$T_{A\cup C}/V_{A\cup C}$
$T_{A\cup B}/V_{A\cup B}$
$T_{A\cup B\cup C}/V_{A\cup B\cup C}$
PSc_1/V_{PSc_1}
PSc_2/V_{PSc_2}
PSc_3/V_{PSc_3}
PSc_4/V_{PSc_4}
PSc_5/V_{PSc_5}
W_1
W_2
W_3
W_4

Figure 1.

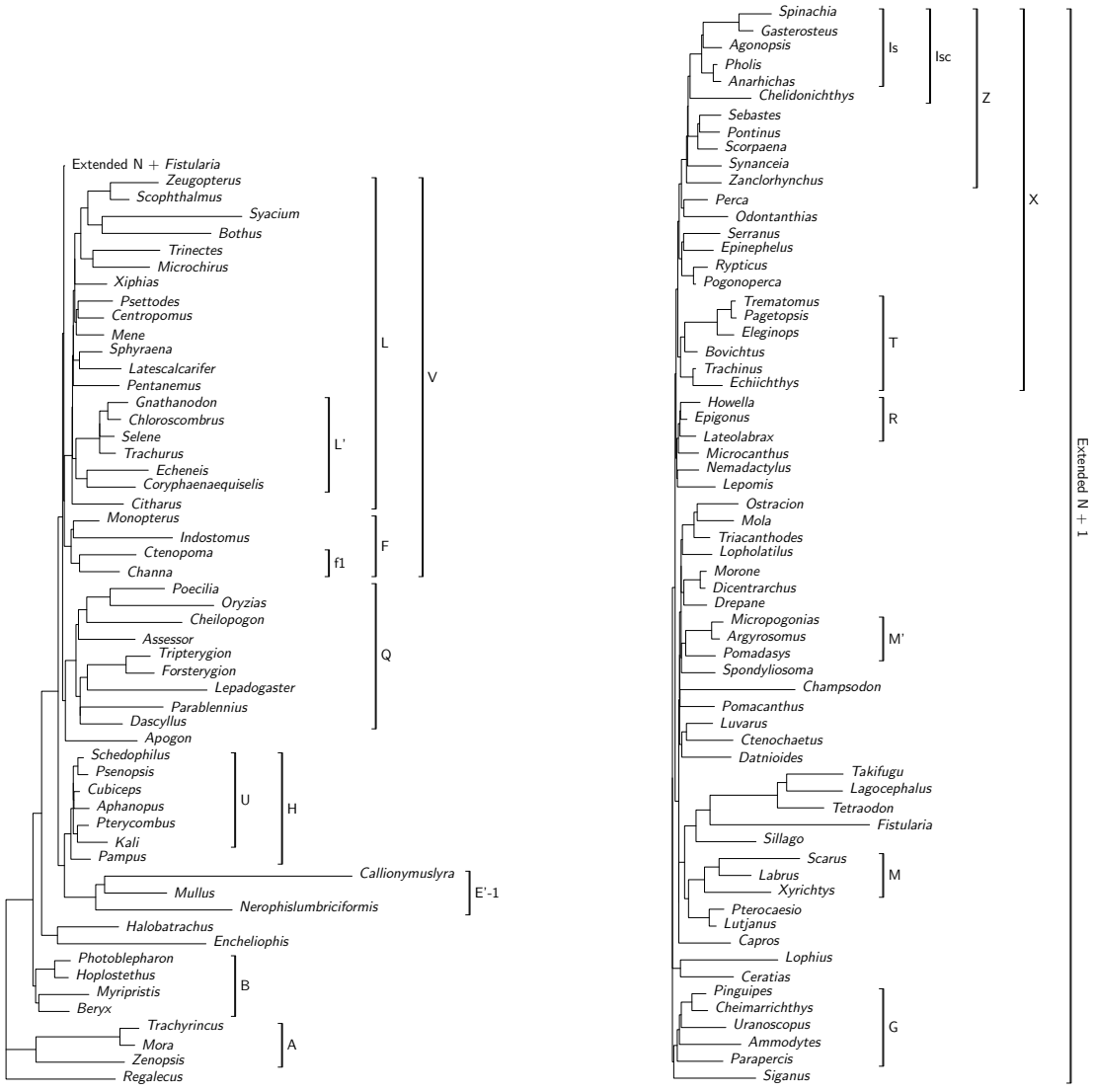

Extended N + Fistularia
Zeugopterus
Scophthalmus
Syacium
Bothus
Trinectes
Microchirus
Xiphias
Psettodes
Centropomus
Mene
Sphyraena
Latescalcarifer
Pentanemus
Gnathanodon
Chloroscombrus
Selene
Trachurus
Echeneis
Coryphaenaequiselis
Citharus
Monopterus
Indostomus
Ctenopoma
Channa
Poecilia
Oryzias
Cheilopogon
Assessor
Tripterygion
Forsterygion
Lepadogaster
Parablennius
Dascyllus
Apogon
Schedophilus
Psenopsis
Cubiceps
Aphanopus
Pterycombus
Kali
Pampus
Callionymuslyra
Mullus
Nerophislumbriciformis
Halobatrachus
Encheliophis
Photoblepharon
Hoplostethus
Myripristis
Beryx
Trachyrincus
Mora
Zenopsis
Regalecus
L
V
L'
F
f1
Q
U
H
E'-1
B
A
Spinachia
Gasterosteus
Agonopsis
Pholis
Anarhichas
Chelidonichthys
Sebastes
Pontinus
Scorpaena
Synanceia
Zanclorhynchus
Perca
Odontanthias
Serranus
Epinephelus
Rypticus
Pogonoperca
Trematomus
Pagetopsis
Eleginops
Bovichtus
Trachinus
Echiichthys
Howella
Epigonus
Lateolabrax
Microcanthus
Nemadactylus
Lepomis
Ostracion
Mola
Triacanthodes
Lopholatilus
Morone
Dicentrarchus
Drepane
Micropogonias
Argyrosomus
Pomadasys
Spondyliosoma
Champsodon
Pomacanthus
Luvarus
Ctenochaetus
Datnioides
Takifugu
Lagocephalus
Tetraodon
Fistularia
Sillago
Scarus
Labrus
Xyrichtys
Pterocaesio
Lutjanus
Capros
Lophius
Ceratias
Pinguipes
Cheimarrichthys
Uranoscopus
Ammodytes
Parapercis
Siganus
Is
Isc
Z
X
T
R
M'
M
G
Extended N + 1

Figure 2.

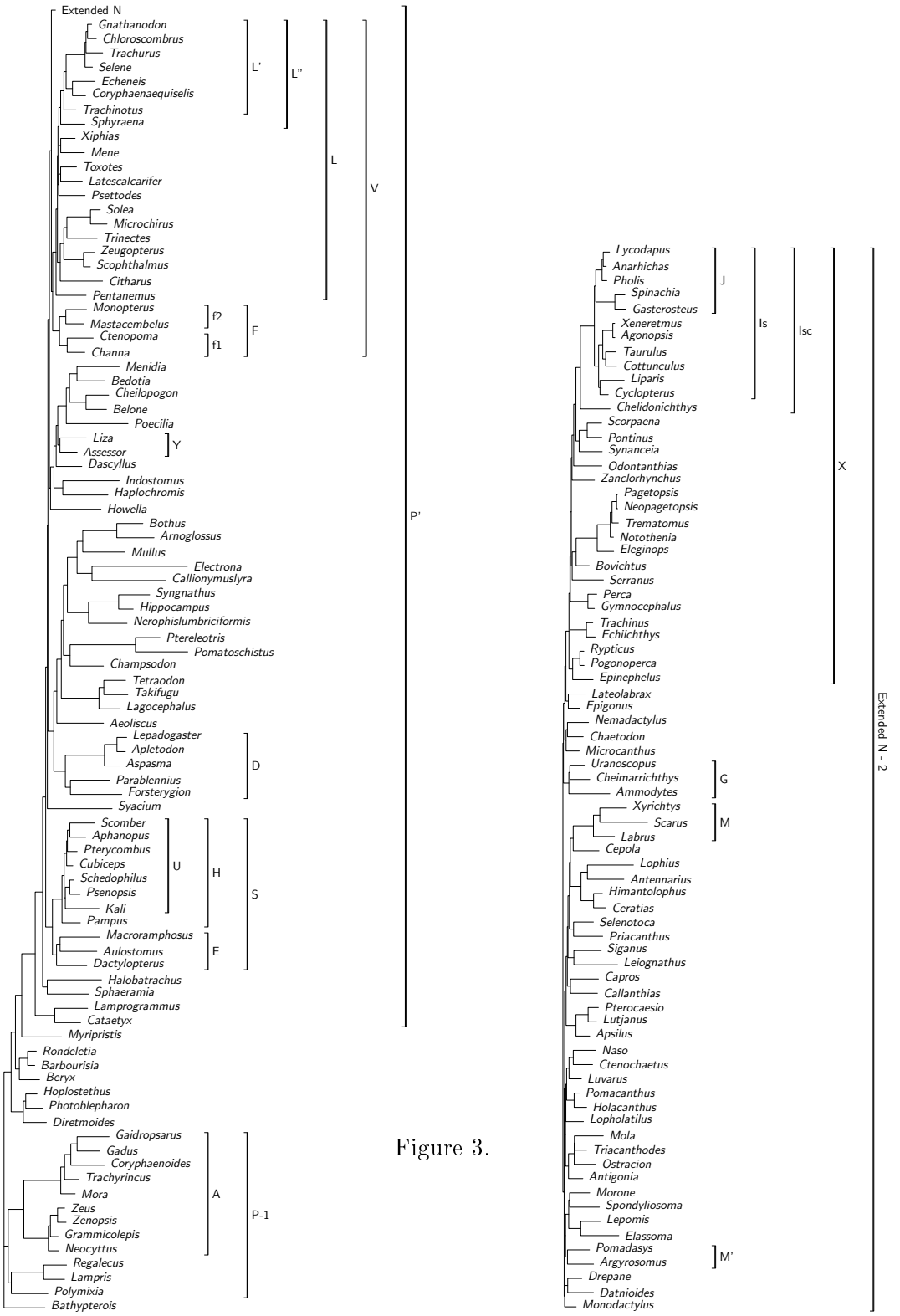
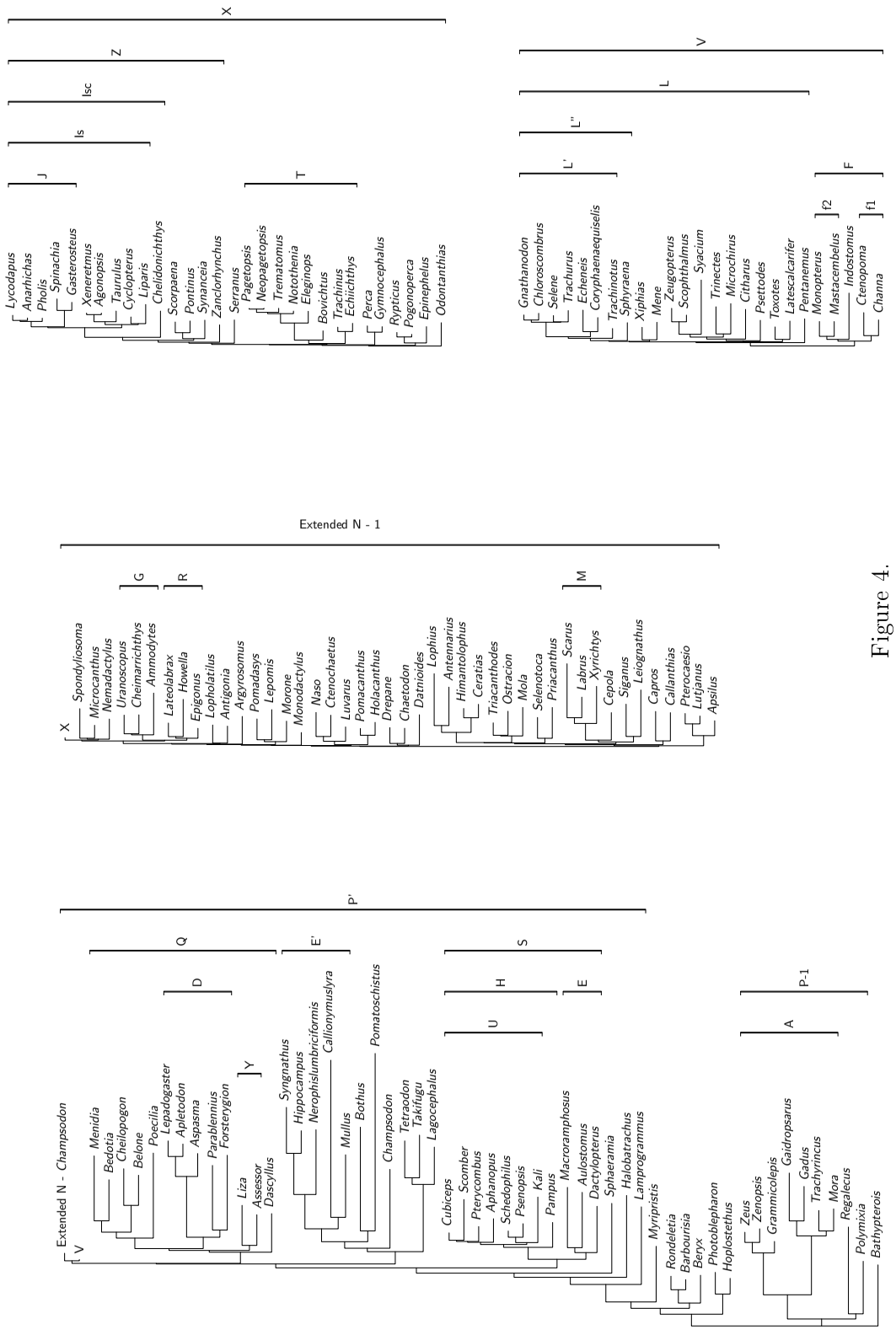

Extended N
Gnathanodon
Chloroscombrus
Trachurus
Selene
Echeneis
Coryphaenaequiselis
Trachinotus
Sphyraena
Xiphias
Mene
Toxotes
Latescalcarifer
Psettodes
Solea
Microchirus
Trinectes
Zeugopterus
Scophthalmus
Citharus
Pentanemus
Monopterus
Mastacembelus
Ctenopoma
Channa
Menidia
Bedotia
Cheilopogon
Belone
Poecilia
Liza
Assessor
Dascyllus
Indostomus
Haplochromis
Howella
Bothus
Arnoglossus
Mullus
Electrona
Callionymuslyra
Syngnathus
Hippocampus
Nerophislumbriciformis
Ptereleotris
Pomatoschistus
Champsodon
Tetraodon
Takifugu
Lagocephalus
Aeoliscus
Lepadogaster
Apletodon
Aspasma
Parablennius
Forsterygion
Syacium
Scomber
Aphanopus
Pterycombus
Cubiceps
Schedophilus
Psenopsis
Kali
Pampus
Macroramphosus
Aulostomus
Dactylopterus
Halobatrachus
Sphaeramia
Lamprogrammus
Cataetyx
Myripristis
Rondeletia
Barbourisia
Beryx
Hoplostethus
Photoblepharon
Diretmoides
Gaidropsarus
Gadus
Coryphaenoides
Trachyrincus
Mora
Zeus
Zenopsis
Grammicolepis
Neocyttus
Regalecus
Lampris
Polymixia
Bathypterois
L'
L''
L
V
P'
f2
f1
F
Y
D
U
H
S
E
A
P-1
Lycodapus
Anarhichas
Pholis
Spinachia
Gasterosteus
Xeneretmus
Agonopsis
Taurulus
Cottunculus
Liparis
Cyclopterus
Chelidonichthys
Scorpaena
Pontinus
Synanceia
Odontanthias
Zanclorhynchus
Pagetopsis
Neopagetopsis
Trematomus
Notothenia
Eleginops
Bovichtus
Serranus
Perca
Gymnocephalus
Trachinus
Echiichthys
Rypticus
Pogonoperca
Epinephelus
Lateolabrax
Epigonus
Nemadactylus
Chaetodon
Microcanthus
Uranoscopus
Cheimarrichthys
Ammodytes
Xyrichtys
Scarus
Labrus
Cepola
Lophius
Antennarius
Himantolophus
Ceratias
Selenotoca
Priacanthus
Siganus
Leiognathus
Capros
Callanthias
Pterocaesio
Lutjanus
Apsilus
Naso
Ctenochaetus
Luvarus
Pomacanthus
Holacanthus
Lopholatilus
Mola
Triacanthodes
Ostracion
Antigonia
Morone
Spondyliosoma
Lepomis
Elassoma
Pomadasys
Argyrosomus
Drepane
Datnioides
Monodactylus
J
Is
Isc
X
G
M
M'
Extended N - 2

Figure 3.


Lycodapus
Anarhichas
Pholis
Spinachia
Gasterosteus
Xeneretmus
Agonopsis
Taurulus
Cyclopterus
Liparis
Chelidonichthys
Scorpaena
Pontinus
Synanceia
Zanclorhynchus
Serranus
Pagetopsis
Neopagetopsis
Trematomus
Notothenia
Eleginops
Bovichtus
Trachinus
Echiichthys
Perca
Gymnocephalus
Rypticus
Pogonoperca
Epinephelus
Odontanthias
Gnathanodon
Chloroscombrus
Selene
Trachurus
Echeneis
Coryphaenaequiselis
Trachinotus
Sphyraena
Xiphias
Mene
Zeugopterus
Scophthalmus
Syacium
Trinectes
Microchirus
Citharus
Psettodes
Toxotes
Latescalcarifer
Pentanemus
Monopterus
Mastacembelus
Indostomus
Ctenopoma
Channa
J
Is
Isc
Z
X
T
L'
L''
L
V
f2
f1
F

Extended N - 1
X
Spondyliosoma
Microcanthus
Nemadactylus
Uranoscopus
Cheimarrichthys
Ammodytes
Lateolabrax
Howella
Epigonus
Lopholatilus
Antigonia
Argyrosomus
Pomadasys
Lepomis
Morone
Monodactylus
Naso
Ctenochaetus
Luvarus
Pomacanthus
Holacanthus
Drepane
Chaetodon
Datnioides
Lophius
Antennarius
Himantolophus
Ceratias
Triacanthodes
Ostracion
Mola
Selenotoca
Priacanthus
Scarus
Labrus
Xyrichtys
Cepola
Siganus
Leiognathus
Capros
Callanthias
Pterocaesio
Lutjanus
Apsilus
G
R
M

Extended N - Champsodon
V
Menidia
Bedotia
Cheilopogon
Belone
Poecilia
Lepadogaster
Apletodon
Aspasma
Parablennius
Forsterygion
Liza
Assessor
Dascyllus
Y
Syngnathus
Hippocampus
Nerophislumbriciformis
Callionymuslyra
Mullus
Bothus
Champsodon
Pomatoschistus
Tetraodon
Takifugu
Lagocephalus
Cubiceps
Scomber
Pterycombus
Aphanopus
Schedophilus
Psenopsis
Kali
Pampus
Macroramphosus
Aulostomus
Dactylopterus
Sphaeramia
Halobatrachus
Lamprogrammus
Myripristis
Rondeletia
Barbourisia
Beryx
Photoblepharon
Hoplostethus
Zeus
Zenopsis
Grammicolepis
Gaidropsarus
Gadus
Trachyrincus
Mora
Regaleucus
Polymixia
Bathypterois
Q
D
E'
P'
U
H
S
E
A
P-1

Figure 4.

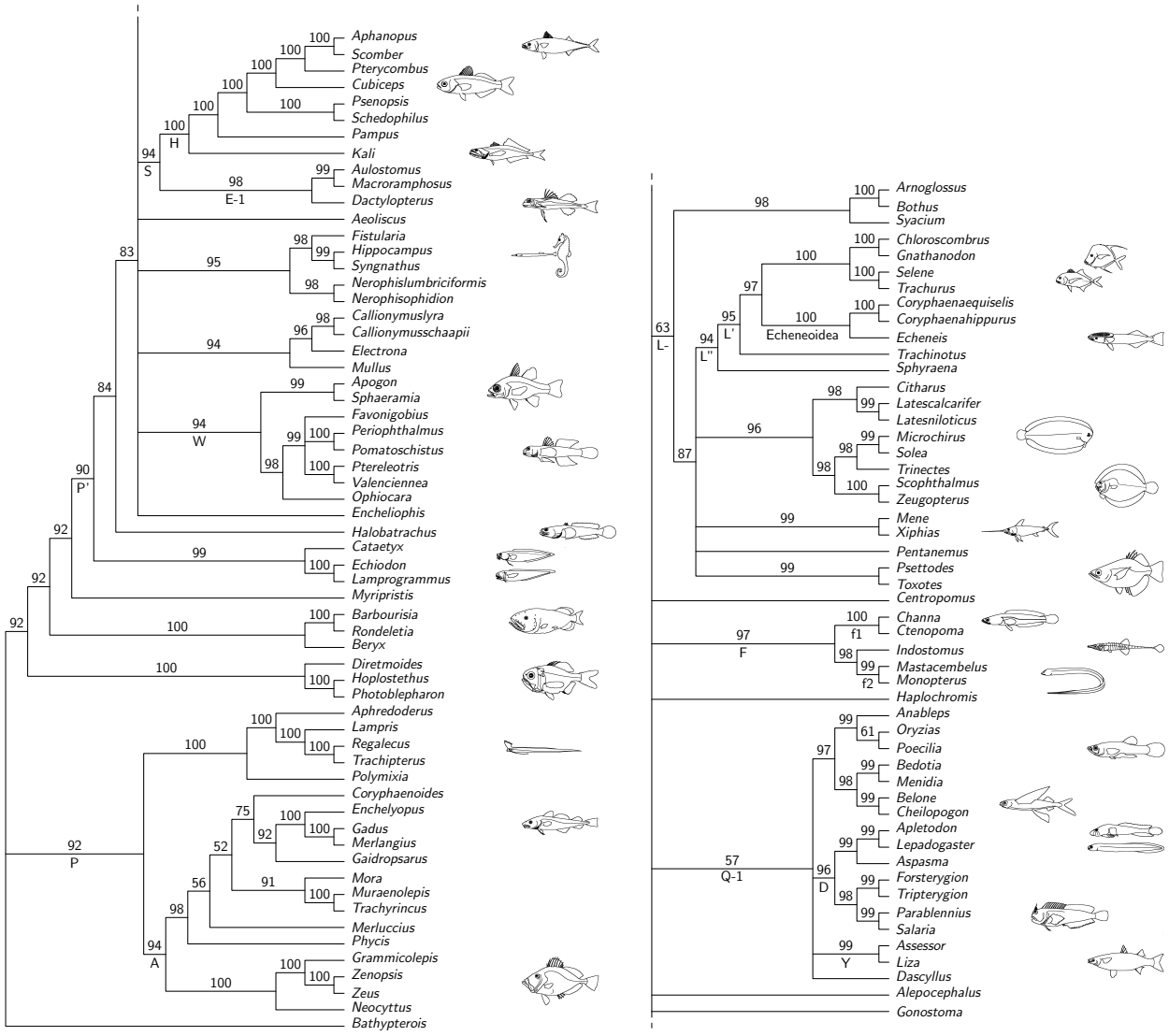

Aphanopus
Scomber
Pterycombus
Cubiceps
Psenopsis
Schedophilus
Pampus
Kali
Aulostomus
Macroramphosus
Dactylopterus
Aeoliscus
Fistularia
Hippocampus
Syngnathus
Nerophislumbriciformis
Nerophisophidion
Callionymuslyra
Callionymusschaapii
Electrona
Mullus
Apogon
Sphaeramia
Favonigobius
Periophthalmus
Pomatoschistus
Ptereleotris
Valenciennea
Ophiocara
Encheliophis
Halobatrachus
Cataetyx
Echiodon
Lamprogrammus
Myripristis
Barbourisia
Rondeletia
Beryx
Diretmoides
Hoplostethus
Photoblepharon
Aphredoderus
Lampris
Regalecus
Trachipterus
Polymixia
Coryphaenoides
Enchelyopus
Gadus
Merlangius
Gaidropsarus
Mora
Muraenolepis
Trachyrincus
Merluccius
Phycis
Grammicolepis
Zenopsis
Zeus
Neocyttus
Bathypterois
S
H
E-1
W
P'
P
A
Arnoglossus
Bothus
Syacium
Chloroscombrus
Gnathanodon
Selene
Trachurus
Coryphaenaequiselis
Coryphaenahippurus
Echeneis
Trachinotus
Sphyraena
Echeneoidea
L'
L''
L-
Citharus
Latescalcarifer
Latesniloticus
Microchirus
Solea
Trinectes
Scophthalmus
Zeugopterus
Mene
Xiphias
Pentanemus
Psettodes
Toxotes
Centropomus
Channa
Ctenopoma
Indostomus
Mastacembelus
Monopterus
Haplochromis
f1
f2
F
Anableps
Oryzias
Poecilia
Bedotia
Menidia
Belone
Cheilopogon
Apletodon
Lepadogaster
Aspasma
Forsterygion
Tripterygion
Parablennius
Salaria
Assessor
Liza
Dascyllus
Alepocephalus
Gonostoma
D
Y
Q-1

Figure 5.

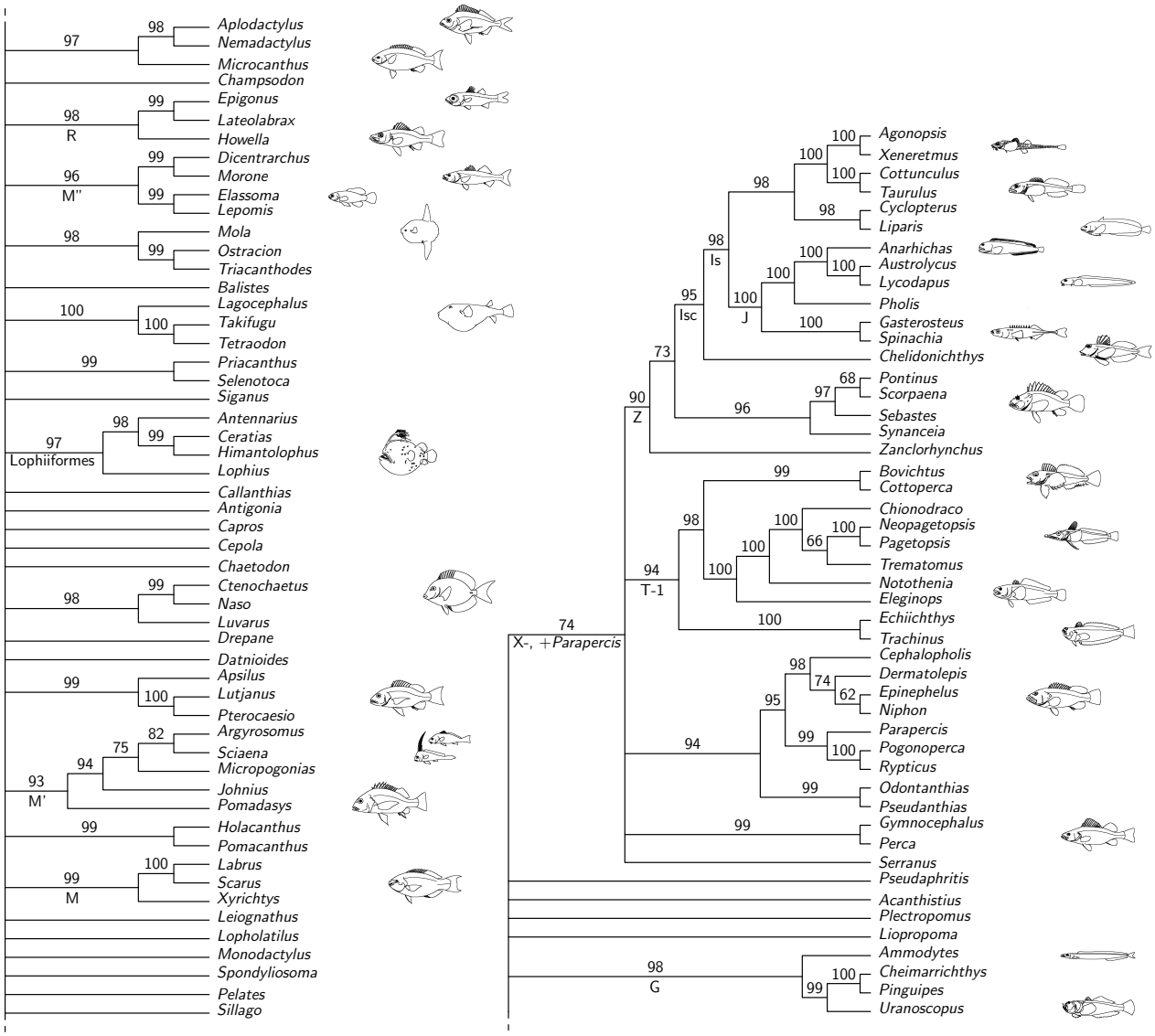


Figure 5 (continuing).

Table 3:

Last common ancestor to	Fig. 2	Fig. 3	Fig. 4	Fig. 5	Nomenclature (Chen *et al.* and this study)	Dettaï and Lecointre (2008)	Miya et al. (2003, 2005)	Mabuchi et al. (2007)	Kawahara et al. (2008)	Smith and Craig (2007)	New name
Zeioidei, Gadiformes	x	x	x	x	A	x	x	?	?	?	Zeioigadiformes
Zeioidei, Gadiformes, Polymixiiformes	?				O	x		?	?	?	
Lampridiformes, Percopsiformes	?	?	?	x		x		?	?	?	
Lampridiformes, Percopsiformes, Polymixiiformes	?	?	?	x				?	?	?	
Zeioidei, Gadiformes, Polymixiiformes, Lampridiformes, Percopsiformes	?	?	?	x	P	x		?	?	?	Paracanthomorpha
Trachichthyoidei, Berycoidei, Holocentroidei	x				B		x	?	?	?	Beryciformes (after chen et al., 2003)
P sister-group of the rest of acanthomorpha	?	x	x	x		x	?	?	?	?	
Ophidiiformes sister-group of non-P and non beryciform and non Stephanoberyciformes acanthomorphs		x	x	x	P'	?	x	x	x	?	
Mugiloidei, Atherinomorpha	?				C	?	?			?	
Blennioidei, Gobiesocoidei	?	x	x	x	D	x	x	x	x	?	Blenniiformes
Mugiloidei, Plesiopidae	?	x	x	x	Y	?	?	?	?	?	Stiassnyformes
Mugiloidei, Plesiopidae, Blenniiformes, Atherinomorpha, Cichlidae	x		x	x-1	Q	x	x?	x?	B?	x?	Parentiformes
Apogonidae, Gobioidei				x	W	?	?	?	?	?	
Syngnathoidei, Callionymoidei, Mullidae	x-1		x		E'	?	?	?	?	?	Longiformes
Centrolophidae, Bramidae, Nomeidae, Scombridae, Trichiuridae, Chiasmodontidae	x	x	x		U	?	?	?	?	?	
Stromateidae, Centrolophidae, Bramidae, Nomeidae, Scombridae, Trichiuriae, Chiasmodontidae	x	x	x	x	H	x	?	?	?	x?	Stromateoidei, new definition

Table 3: (continued)

Last common ancestor to	Fig. 2	Fig. 3	Fig. 4	Fig. 5	Nomenclature (Chen *et al.* and this study)	Dettaï and Lecointre (2008)	Miya et al. (2003, 2005)	Mabuchi et al. (2007)	Kawahara et al. (2008)	Smith and Craig (2007)	New name
Dactylopteridae, Aulostomidae, Macrorhamphosidae	?	x	x	x-1	E	?	?	?	x		Aulostomoidea
Stromateoidei + Aulostomoidea	?	x	x	x-1	S	?	?	?	?		Stromateiformes
Channidae, Anabantidae	x	x	x	x	f1	x	?	?	?	?	Labyrinthoidei
Symbranchidae, Mastacembelidae	?	x	x	x	f2	x	x	x	A?	?	Synbranchiformes
Channidae, Anabantidae, Mastacembelidae, Symbranchidae, Indostomidae	x	x	x	x	F	x	?	?	A?	?	Anabantiformes
Uranoscopidae, Ammodytidae, Cheimarrichthyidae, Pinguipedidae	x	x	x	x	G	x	?	?	?	x?	Paratrachinoidei
Sciaenidae, Haemulidae	x	x	?	x	M'	?	?	?	?		Sciaenoidei
Centrarchidae, Moronidae, Elassomatidae				x	M"	?	?	?	?		
Cottoidei, Zoarcoidei					I	x				x	
Zoarcoidei, Gasterosteidae		x	x	x	J	?	x	x	x		Zoarciformes
Cottoidei, Zoarcoidei, Gasterosteidae	x	x	x	x	Is	?	x	x	x	x	Cottimorpha
Cottoidei, Zoarcoidei, Gasterosteidae, Triglidae	x	x	x	x	Isc	?	x	?	?	x	Triglimorpha
Cottoidei, Zoarcoidei, Gasterosteidae, Triglidae, Scorpaenidae, Sebastidae, Synanceiidae, Congiopodidae	x		x	x	Z	?	?	?	?		
Notothenioidei, Percophidae	?	?	?				?	?	?	F	Notothenioidi (Smith and Craig, 2007)
Notothenioidi, Niphon, Acanthistius, Percidae							?	?	?	E	Percoidei
Notothenioidei, Trachinidae	x		x	x-1	T		?	?	?		
Notothenioidei, Percidae					K		?	?	?		

Table 3: (continued)

Last common ancestor to	Fig. 2	Fig. 3	Fig. 4	Fig. 5	Nomenclature (Chen *et al.* and this study)	Dettaï and Lecointre (2008)	Miya et al. (2003, 2005)	Mabuchi et al. (2007)	Kawahara et al. (2008)	Smith and Craig (2007)	New name
Notothenioidei, Percidae, Triglimorpha, Trachinidae, Scorpaenidae, Sebastidae, Synanceiidae, Serranidae, Congiopodidae	x	x	x	x-1	X	x	?	?	G?	x	Serraniformes
Carangoidei stem group of Echeneidae and Coryphaenidae	x	x	x	x	L'	x	?	?	?	?	
Sphyraenidae sister-group of L'		x	x	x	L"	?	?	?	?		
Pleuronectiformes, Centropomidae, Carangidae, Coryphaenidae, Menidae, Sphyraenidae, Polynemidae, Echeneidae, Toxotidae, Xiphiidae	x	x	x-1	x-1	L	x	x?	x?	E?	x?	Carangimorpha
Carangimorpha, Anabantiformes	x	x	x		V	?	?	?		?	
Tetraodontiformes, Lophiiformes, Caproidei, Elassomatidae, Acanthuridae, Siganidae, Pomacanthidae, Drepanidae, Chaetodontidae	?				N	x	x?	x?	H?	?	
Extended N	x+1	x-2	x-1			x?	x?	x?	H+G?		
Labridae, Scaridae	x	x	x	x	M	x	?	x	?	?	Labroidei *sensu stricto*
Percichthyidae, Epigonidae, Lateolabracidae	x		x	x	R	?	?	?	?	x	Epigonoidei

8. Résultats

8.1. Nouvelles données

L'obtention d'amplifiats séquençables a été relativement long car de nombreux extraits d'ADN « résistaient » à la PCR, soit qu'ils étaient de mauvaise qualité, soit que les amorces ne leur étaient pas bien adaptées. Un certain nombre de PCR n'ont finalement réussi qu'en utilisant un nombre de cycles inhabituellement long (jusqu'à 65 cycles).
371 séquences ont été obtenues, dont 349 entièrement nouvelles. Les numéros d'accès Genbank des séquences obtenues figurent dans le tableau 8.1, ainsi que les références des autres séquences utilisées.

Tableau 8.1. Références des séquences utilisées et/ou nouvellement obtenues. Les séquences en gras sont nouvelles, ou bien ont été mises à jour. Les séquences entièrement nouvelles sont celles dont le numéro d'accession commence par « EU6 ». Les séquences entre parenthèses n'ont pas été utilisées.

Espèce	Rhodopsine	MLL	IRBP	RNF213
Acanthistius brasilianus	–	**EU638024**	–	–
Aeoliscus strigatus	**EU637931**	–	**EU638100**	–
Agonopsis chiloensis	**EU637932**	**EU638025**	**EU638101**	**EU638167**
Alepocephalus antipodianus	**EU637933**	–	–	–
Ammodytes tobianus	AY141306	AY362234	**EU638102**	**EU638168**
Anableps anableps	**EU637935**	–	–	–
Anarhichas lupus	**EU637936**	**EU638026**	**EU638103**	**EU638169**
Antennarius striatus	AY368324	AY362215	DQ168037	–
Antigonia capros	**EU637937**	**EU638027**	**EU638104**	–
Aphanopus carbo	**EU637938**	**EU638028**	**EU638105**	**EU638170**
Aphredoderus sayanus	–	–	DQ168038	–
Apletodon dentatus	AY141274	AY362213	DQ168039	–
Aplodactylus punctatus	**EU637939**	–	–	–
Apogon quadrifasciatus	**EU637940**	–	–	**EU638171**
Apsilus fuscus	**EU637941**	**EU638029**	**EU638106**	–
Argyrosomus regius	**EU637942**	**EU638030**	**EU638107**	**EU638172**
Arnoglossus imperialis	AY141283	AY362228	–	–
Aspasma minima	**EU637943**	**EU638031**	**EU638108**	–
Assessor flavissimus	**EU637944**	**EU638032**	**EU638109**	**EU638173**
Aulostomus chinensis	AY141279	AY362226	DQ168040	–

Tableau 8.1. (Suite)

Espèce	Rhodopsine	MLL	IRBP	RNF213
Austrolycus depressiceps	AY141297	–	–	–
Balistes sp.	AF137212	–	–	–
Barbourisia rufa	AY368333	AY362264	DQ168041	–
Bathypterois dubius	**AY141257**	AY362219	DQ168042	–
Bedotia geayi	AY141267	AY362271	DQ168043	–
Belone belone	AY141268	AY362273	DQ168044	–
Beryx splendens	AY141265	AY362238	DQ168045	**EU638174**
Bothus podas	AY368313	**EU638033**	–	**EU638175**
Bovichtus variegatus	**AY141299**	AY362283	DQ168046	**EU638176**
Callanthias ruber	**EU637945**	**EU638034**	**EU638110**	–
Callionymus lyra	AY141270	AY362225	DQ168047	**EU638177**
Callionymus schaapii	**EU637946**	–	–	–
Capros aper	AY141262	AY362233	DQ168048	**EU638178**
Carapus boraborensis	–	–	–	**EU638179**
Cataetyx laticeps	**EU637947**	**EU638035**	–	–
Centropomus undecimalis	–	–	–	**EU638180**
Cephalopholis urodeta	–	**EU638036**	–	–
Cepola macrophthalma	**EU637948**	**EU638037**	**EU638111**	–
Ceratias holboelli	AY141263	AY362270	DQ168049	**EU638181**
Chaetodon semilarvatus	AY368312	AY362240	DQ168050	–
Champsodon snyderi	**EU637949**	**EU638038**	–	**EU638182**
Channa sp.	–	–	–	**EU638183**
Channa striata	AY141277	AY362241	DQ168051	–
Cheilopogon heterurus	**EU637950**	**EU638039**	**EU638113**	**EU638184**
Cheimarrichthys fosteri	**AY141307**	AY362229	DQ168052	**EU638185**
Chelidonichthys lucernus	AY141287	AY362284	DQ168053	**EU638186**
Chionodraco hamatus	–	AY362280	–	–
Chloroscombrus chrysurus	AY141313	AY362223	DQ168054	**EU638187**
Citharus linguatula	AY141323	AY362232	DQ168055	**EU638188**
Coryphaena equiselis	**EU637951**	**EU638040**	**EU638114**	**EU638189**
Coryphaena hippurus	–	–	DQ168056	–
Coryphaenoides rupestris	AY368319	**EU638041**	–	–
Cottoperca gobio	AY141300	–	–	–
Cottunculus thomsonii	AY368315	AY362260	–	–
Ctenochaetus sp.	–	–	–	**EU638190**
Ctenochaetus striatus	AY141320	AY362242	DQ168057	–
Ctenopoma sp.	AY141278	AY362210	DQ168058	**EU638191**
Cubiceps gracilis	**EU637952**	**EU638042**	**EU638115**	**EU638192**
Cyclopterus lumpus	AY368316	**AY362218**	**EU638116**	–
Dactylopterus volitans	AY141282	AY362243	DQ168059	–
Dascyllus trimaculatus	**EU637953**	**EU638043**	**EU638117**	**EU638193**
Datnioides polota	**EU637954**	**EU638044**	**EU638118**	**EU638194**
Dermatolepis dermatolepis	–	**EU638045**	–	–
Dicentrarchus labrax	–	–	**EU638119**	**EU638195**
Diretmoides sp.	–	AY362205	DQ168060	–
Drepane africana	AY141321	AY362244	DQ168061	**EU638196**

Tableau 8.1. (Suite)

Espèce	Rhodopsine	MLL	IRBP	RNF213
Echeneis naucrates	AY141315	AY362245	DQ168062	**EU638197**
Echiichthys vipera	**EU637955**	**EU638046**	**EU638120**	**EU638198**
Echiodon cryomargarites	**EU637956**	–	–	–
Elassoma zonatum	**EU637957**	–	DQ168063	–
Electrona antarctica	AY141258	AY36220	–	–
Eleginops maclovinus	AY141303	**EU638047**	**EU638121**	**EU638199**
Enchelyopus cimbrius	**EU637958**	–	–	–
Epigonus telescopus	**EU637959**	**EU638048**	**EU638122**	**EU638200**
Epinephelus aeneus	AY141291	**EU638049**	AY362227	**EU638201**
Favonigobius reichei	**EU637960**	–	–	–
Fistularia petimba	AY141324	–	–	**EU638202**
Forsterygion lapillum	AY141272	AY362276	DQ168065	**EU638203**
Gadus morhua	AF137211	**EU638050**	DQ168066	–
Gaidropsarus sp.	**EU637961**	–	–	–
Gaidropsarus novaezelandiae	–	**EU638051**	–	–
Gaidropsarus vulgaris	–	**(EU660039)**	DQ168067	–
Gasterosteus aculeatus	**EU637962**	**EU638052**	Ensembl	Ensembl
Gnathanodon speciosus	**EU637963**	**EU638053**	**EU638123**	**EU638204**
Gonostoma bathyphilum	AY141256	–	–	–
Grammicolepis brachiusculus	**EU637964**	**EU638054**	**EU638124**	–
Gymnocephalus cernuus	AY141296	AY362278	DQ168068	–
Halobatrachus didactylus	AY368323	AY362246	DQ168069	**EU638205**
Haplochromis sp.	AB084933	–	–	–
Haplochromis nubilus	–	–	DQ168070	–
Himantolophus groenlandicus	**EU637965**	**EU638055**	**EU638125**	–
Hippocampus guttulatus	AY368330	AY362216	**EU638126**	–
Holacanthus ciliaris	AY141322	AY362214	DQ168072	–
Holanthias chrysostictus	**AY141290**	**AY362209**	**DQ168073**	**EU638206**
Hoplostethus atlanticus	–	–	**EU638127**	**EU638207**
Hoplostethus mediterraneus	AY141264	AY362267	–	–
Howella brodiei	**EU637966**	**EU638056**	**EU638128**	**EU638208**
Indostomus paradoxus	**EU637967**	**EU638057**	–	**EU638209**
Johnius sp.	–	–	**EU638129**	–
Kali macrura	AY141308	AY362224	DQ168074	**EU638210**
Labrus bergylta	AY141318	AY362222	DQ168075	**EU638211**
Lagocephalus laevigatus	(AY141285)	AY362221	DQ168076	–
Lagocephalus lagocephalus	**EU637968**	–	–	**EU638212**
Lampris immaculatus	AY141259	–	DQ168077	–
Lamprogrammus shcherbachevi	**EU637969**	**EU638058**	**EU638130**	–
Lateolabrax japonicus	AY141293	AY362253	DQ168078	**EU638213**
Lates calcarifer	**EU637970**	**EU638059**	DQ168075	**EU638214**
Lates niloticus	**EU637971**	–	–	–
Leiognathus fasciatus	**EU637972**	**EU638060**	**EU638131**	–
Lepadogaster lepadogaster	AY141273	AY362247	DQ168080	**EU638215**
Lepomis gibbosus	AY742571	**EU638061**	**EU638132**	**EU638216**
Liopropoma fasciatum	–	**EU638062**	–	–

Tableau 8.1. (Suite)

Espèce	Rhodopsine	MLL	IRBP	RNF213
Liparis fabricii	AY368317	AY362235	DQ168081	–
Liza sp.	AY141266	AY362248	DQ168082	–
Lophius budegassa	–	–	–	**EU638217**
Lophius piscatorius	AY368325	AY362274	–	–
Lopholatilus chamaeleonticeps	**EU637973**	**EU638063**	**EU638133**	**EU638218**
Lutjanus sebae	**EU637974**	**EU638064**	**EU638134**	**EU638219**
Luvarus imperialis	**EU637975**	**EU638065**	**EU638135**	**EU638220**
Lycodapus antarcticus	**EU637976**	**EU638066**	**EU638136**	–
Macroramphosus scolopax	**AY141280**	AY362206	DQ168083	–
Mastacembelus erythrotaenia	**AY141275**	**AY362249**	**DQ168084**	–
Mene maculata	**AY141316**	**AY362250**	DQ168085	**EU638221**
Menidia menidia	**EU637977**	**EU638067**	**EU638137**	–
Merlangius merlangus	AY141260	–	–	–
Merluccius merluccius	–	**EU638068**	–	–
Microcanthus strigatus	**EU637978**	**EU638069**	**EU638138**	**EU638222**
Microchirus frechkopi	–	–	–	**EU638223**
Microchirus variegatus	AY141284	AY362275	DQ168086	–
Micropogonias sp.	–	–	–	**EU638224**
Micropogonias manni	**EU637979**	–	–	–
Mola mola	AF137215	AY362251	DQ168087	**EU638225**
Monodactylus sp.	**EU637980**	**EU638070**	**EU638139**	–
Monopterus albus	AY141276	AY362252	DQ168088	**EU638226**
Mora moro	AY368322	**EU638071**	DQ168089	**EU638227**
Morone saxatilis	**EU637981**	**EU638072**	**EU638140**	**EU638228**
Mullus surmuletus	**EU637982**	AY362231	DQ168090	**EU638229**
Muraenolepis marmoratus	–	**EU638073**	–	–
Myripristis sp.	**EU637983**	–	–	**EU638230**
Myripristis botche	–	AY362265	DQ168091	–
Naso lituratus	**EU637984**	**EU638074**	**EU638141**	–
Nemadactylus monodactylus	**EU637985**	**EU638075**	**EU638142**	**EU638231**
Neocyttus helgae	AY141261	**AY362288**	–	–
Neopagetopsis ionah	**EU637986**	AY362281	DQ16802	–
Nerophis lumbriciformis	**EU637987**	–	**EU638143**	**EU638232**
Nerophis ophidion	–	–	DQ168071	–
Niphon spinosus	**EU637934**	–	–	–
Notothenia coriiceps	AY141302	AY362282	DQ168093	–
Ophiocara porocephala	**EU637988**	–	–	–
Oryzias latipes	–	–	DQ168094	Ensembl
Ostracion sp.	AF137213	AY362207	DQ168095	–
Ostracion cubicus	–	–	–	**EU638233**
Pagetopsis macropterus	**EU637990**	**EU638076**	**EU638144**	**EU638235**
Pampus argenteus	AY141309	AY362220	DQ168096	**EU638236**
Parablennius gattorugine	AY141271	AY362255	DQ168097	**EU638237**
Parapercis clathrata	–	**EU638077**	–	**EU638238**
Pelates quadrilineatus	**EU637991**	–	–	–
Pentanemus quinquarius	AY141317	AY362272	DQ168098	**EU638239**

Tableau 8.1. (Suite)

Espèce	Rhodopsine	MLL	IRBP	RNF213
Perca fluviatilis	AY141295	AY362279	DQ168099	**EU638240**
Periophthalmus barbarus	**EU637992**	–	–	–
Pholis gunnellus	AY141298	AY362285	DQ168100	**EU638241**
Photoblepharon palpebratum	**EU637993**	AY362268	DQ168101	**EU638242**
Phycis phycis	**EU637994**	–	–	–
Pinguipes chilensis	**EU637989**	–	–	**EU638234**
Plectropomus leopardus	–	**EU638078**	–	–
Poecilia reticulata	Y11147	AY362203	DQ168102	**EU638243**
Pogonoperca punctata	AY141292	AY362256	DQ168103	**EU638244**
Polymixia nobilis	AY368320	AY362208	DQ168104	–
Pomacanthus maculosus	**EU637995**	**EU638079**	**EU638145**	**EU638245**
Pomadasys perotaei	–	AY362230	DQ168105	**EU638246**
Pomatoschistus sp.	–	**EU638080**	DQ168106	–
Pomatoschistus minutus	X62405	–	–	–
Pontinus longispinis	**EU637996**	**EU638081**	**EU638146**	**EU638247**
Priacanthus arenatus	**EU637997**	**EU638082**	**EU638147**	–
Psenopsis anomala	AY141310	AY362269	DQ168107	**EU638248**
Psettodes belcheri	EU637998	AY362259	DQ168108	**EU638249**
Pseudanthias squamipinnis	–	**EU638083**	–	–
Pseudaphritis urvillii	AY141301	–	–	–
Pterelectris zebra	**EU637999**	**EU638084**	–	–
Pterocaesio digramma	**EU638000**	**EU638085**	**EU638148**	**EU638250**
Pterycombus brama	**EU638001**	**EU638086**	**EU638149**	**EU638251**
Regalecus glesne	**AY368328**	AY362266	DQ168109	**EU638252**
Rondeletia sp.	AY368327	**EU638087**	DQ168110	–
Rypticus saponaceus	AY368329	**AY362257**	**DQ168111**	**EU638253**
Salaria pavo	Y18674	–	–	–
Scarus hoefleri	AY141319	AY362212	DQ168112	**EU638254**
Schedophilus medusophagus	**EU638003**	**EU660040**	**EU638151**	**EU638255**
Sciaena sp.	**EU638004**	–	–	–
Scomber japonicus	AY141311	AY362237	DQ168113	–
Scophthalmus rhombus	**EU638005**	–	**EU638152**	**EU638256**
Scorpaena onaria	AY141288	AY362236	DQ168114	**EU638257**
Sebastes sp.	–	–	–	**EU638258**
Selene dorsalis	**EU638006**	**EU638089**	**EU638153**	**EU638259**
Selenotoca multifasciata	**EU638002**	**EU638088**	**EU638150**	–
Serranus accraensis	AY141289	AY362202	DQ168115	**EU638260**
Siganus vulpinus	**EU638007**	**EU638090**	DQ168116	**EU638261**
Sillago sihama	**EU638008**	–	–	**EU638262**
Solea solea	**EU638009**	–	**DQ168117**	–
Sphaeramia nematoptera	**EU638010**	**EU638091**	**EU638154**	–
Sphyraena sphyraena	AY141312	AY362254	DQ168118	**EU638263**
Spinachia spinachia	AY141281	AY362261	–	**EU638264**
Spondyliosoma cantharus	–	**EU638092**	**EU638155**	**EU638265**
Syacium micrurum	AY368334	AY362262	DQ168119	**EU638266**
Synanceia verrucosa	**EU638011**	**EU638093**	**EU638156**	**EU638267**

Tableau 8.1. (Suite)

Espèce	Rhodopsine	MLL	IRBP	RNF213
Syngnathus typhle	AY368326	AY362211	DQ168120	–
Takifugu rubripes	–	Ensembl	Ensembl	Ensembl
Taurulus bubalis	U97275	AY362217	DQ168121	–
Tetraodon nigroviridis	Ensembl	Ensembl	Ensembl	Ensembl
Toxotes sp.	**EU638012**	**EU638094**	**EU638157**	–
Trachinotus ovatus	AY141314	AY362263	DQ168120	–
Trachinus draco	AY141304	AY362277	DQ168123	**EU638268**
Trachipterus arcticus	–	–	**EU638158**	–
Trachurus trachurus	**EU638013**	–	**EU638159**	**EU638269**
Trachyrincus murrayi	AY368318	**AY362289**	DQ168124	**EU638270**
Trematomus bernachii	**EU638014**	–	**EU638160**	**EU638271**
Triacanthodes sp.	AY368331	–	DQ168125	–
Triacanthodes anomalus	–	**EU638095**	–	**EU638272**
Trinectes maculatus	**EU638015**	**EU638096**	**EU638161**	**EU638273**
Tripterygion delaisi	**EU638016**	–	–	**EU638274**
Uranoscopus albesca	AY141305	AY362239	**DQ168126**	**EU638275**
Valenciennea strigata	**EU638017**	–	–	–
Xeneretmus latifrons	**EU638018**	**EU638097**	**EU638162**	–
Xiphias gladius	**EU638019**	**EU638098**	**EU638163**	**EU638276**
Xyrichtys novacula	**EU638020**	–	**EU638164**	**EU638277**
Zanclorhynchus spinifer	**EU638021**	–	**EU638165**	**EU638278**
Zenopsis conchifera	AY368314	AY362286	DQ168127	**EU638279**
Zeugopterus punctatus	**EU638022**	**EU638099**	**EU638166**	**EU638280**
Zeus faber	**EU638023**	**AY362287**	DQ168128	–

8.2. Résultat des analyses

8.2.1. Analyses primaires

Seuls les arbres obtenus pour les analyses des jeux de données élémentaires (Figures 8.1, 8.2, 8.3 et 8.4) et de la combinaison des quatre (8.5) sont présentés.

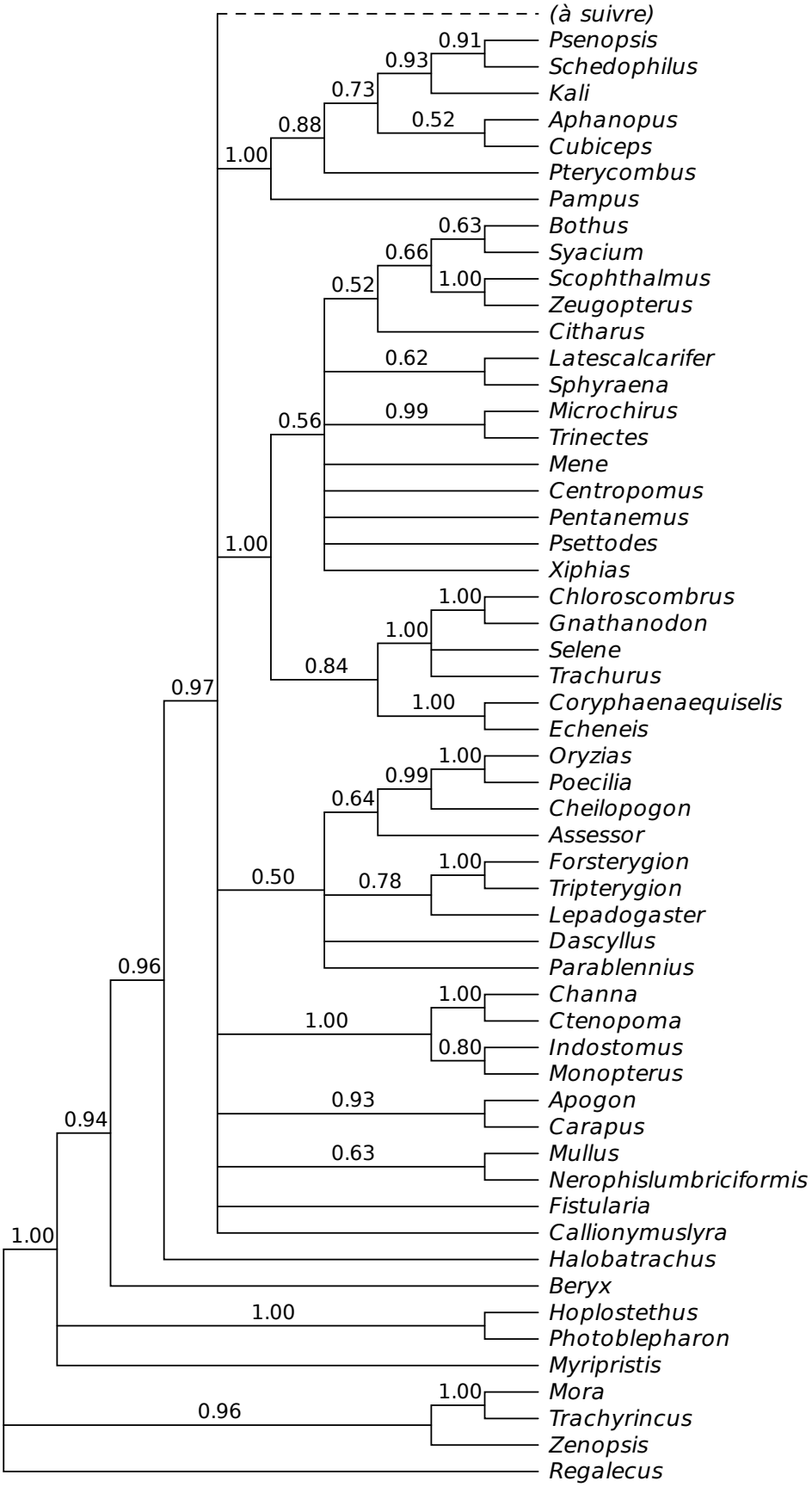
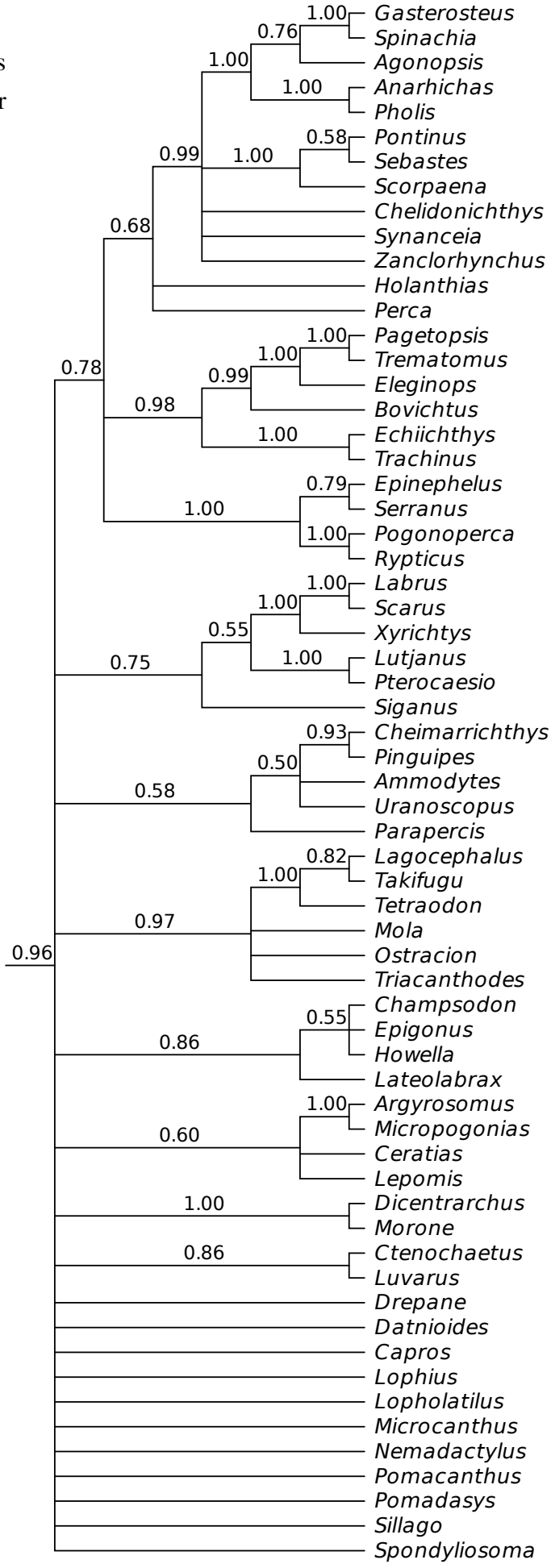


Figure 8.1. Consensus majoritaire des arbres échantillonnés par `MrBayes` pour RNF213. Les valeurs sur les branches sont les probabilités postérieures. Suite page suivante.

Figure 8.1. Consensus majoritaire des arbres échantillonnés par `MrBayes` pour RNF213 (suite).



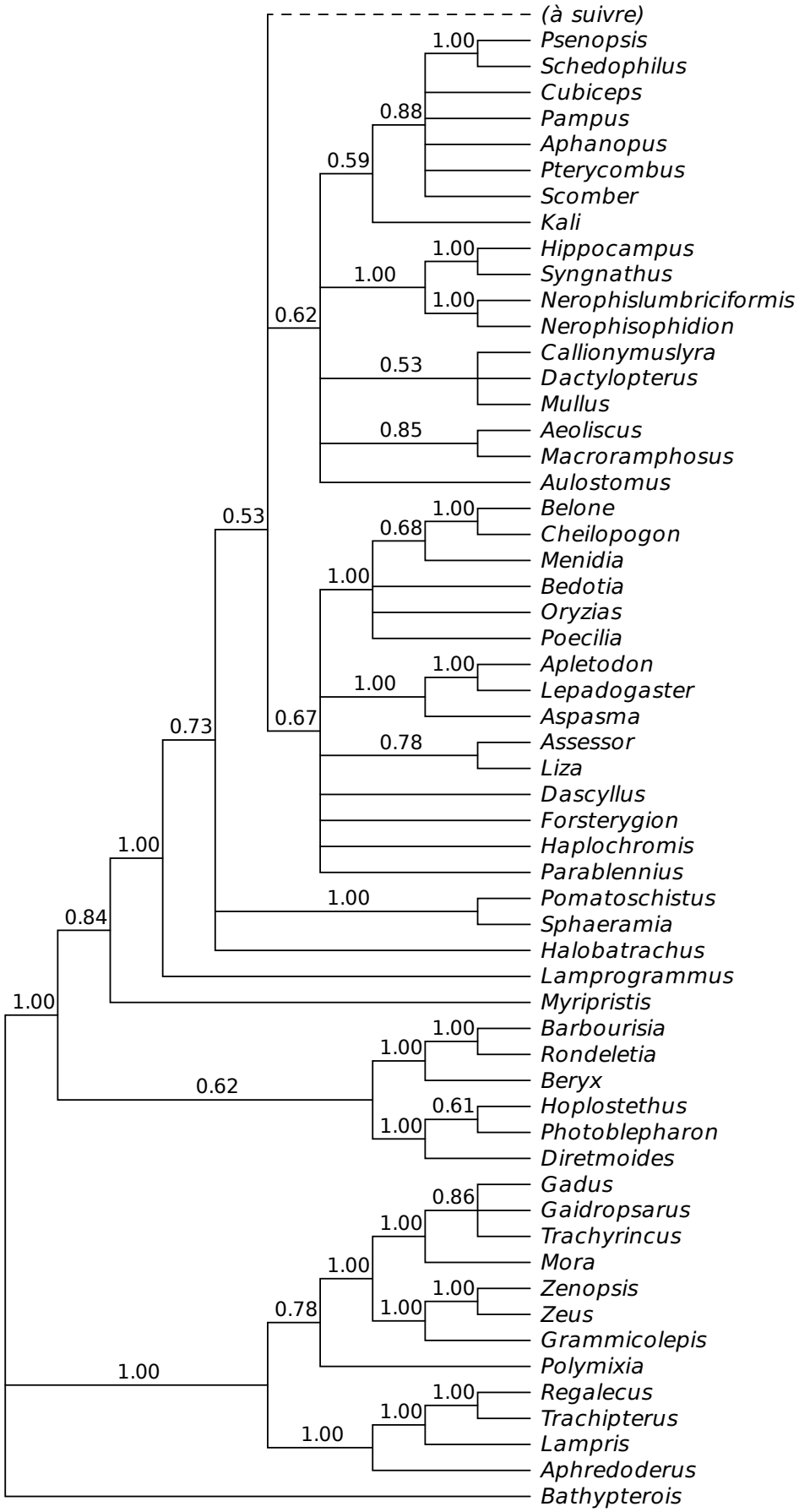


Figure 8.2. Consensus majoritaire des arbres échantillonnés par `MrBayes` pour IRBP. Les valeurs sur les branches sont les probabilités postérieures. Suite page suivante.

Figure 8.2. Consensus majoritaire des arbres échantillonnés par `MrBayes` pour IRBP (suite). Suite page suivante.

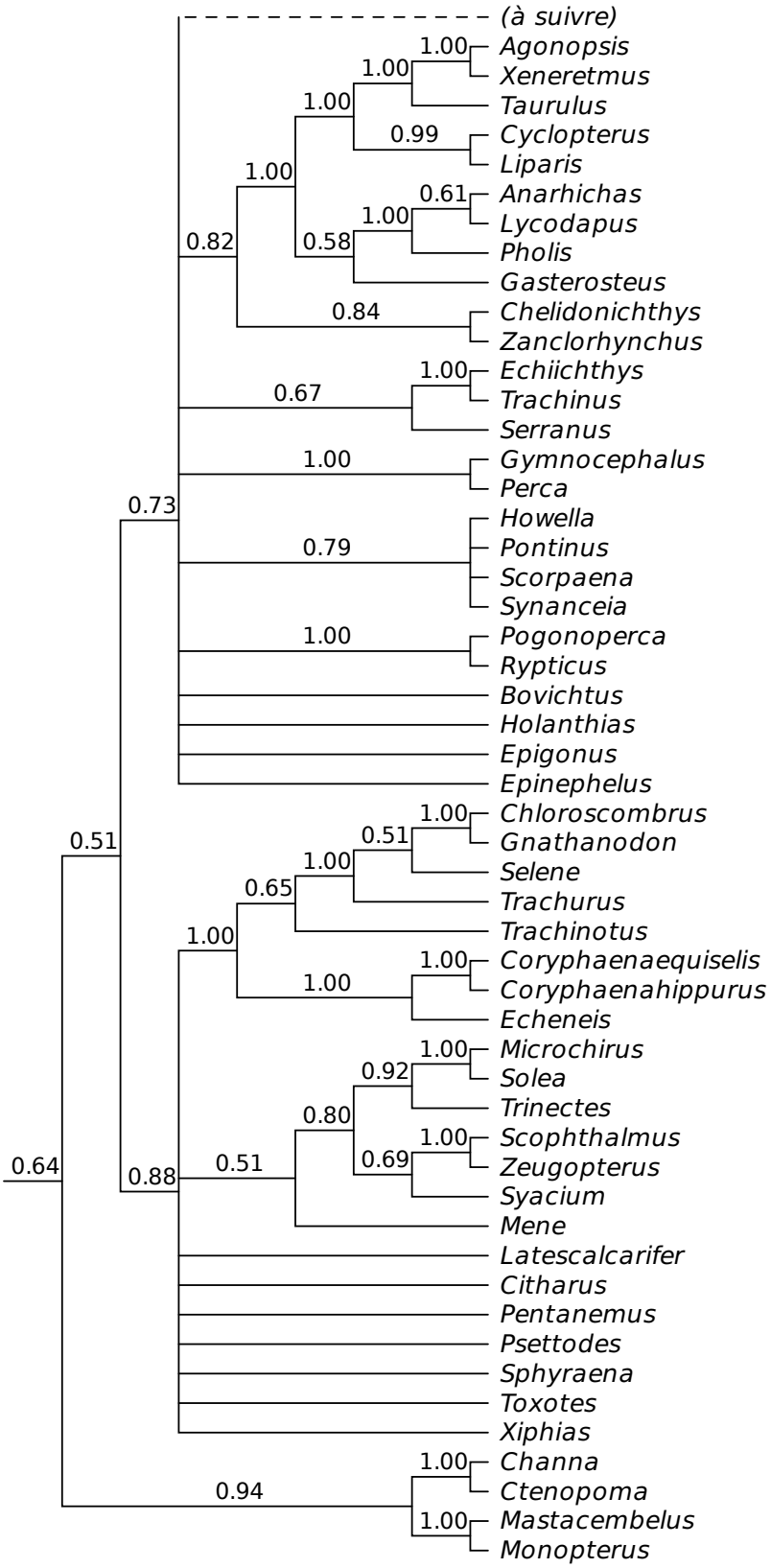


Figure 8.2. Consensus majoritaire des arbres échantillonnés par `MrBayes` pour IRBP (suite).

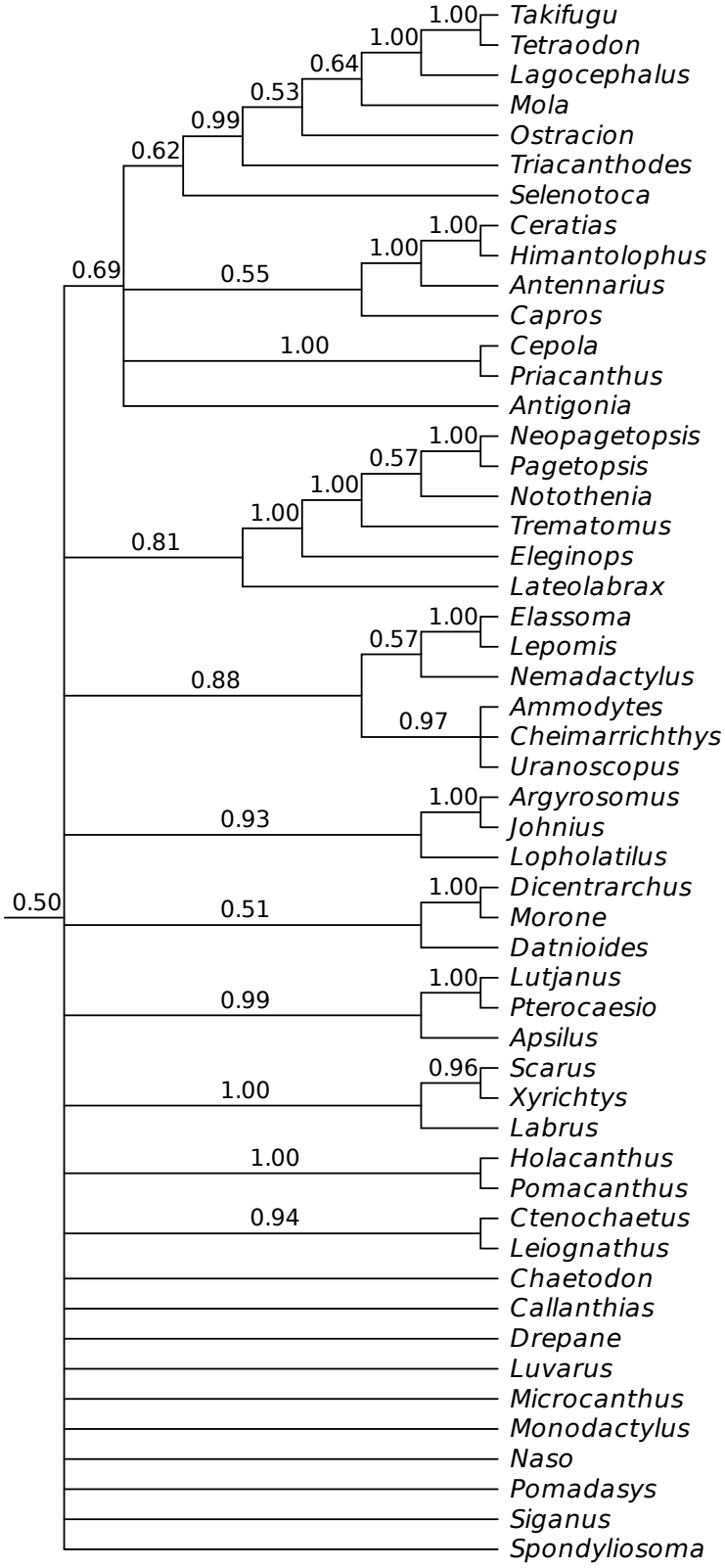
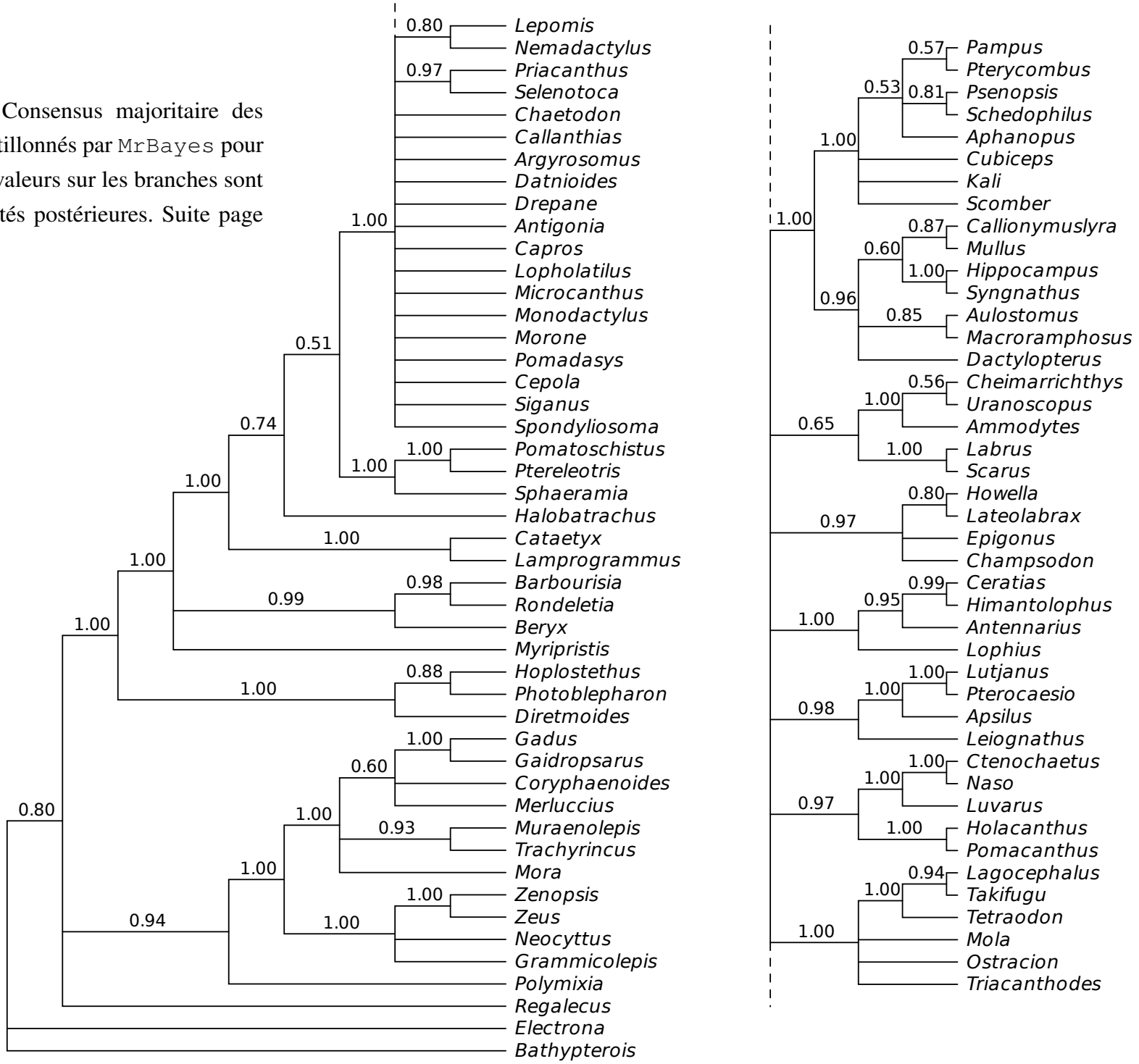


Figure 8.3. Consensus majoritaire des arbres échantillonnés par `MrBayes` pour MLL4. Les valeurs sur les branches sont les probabilités postérieures. Suite page suivante.



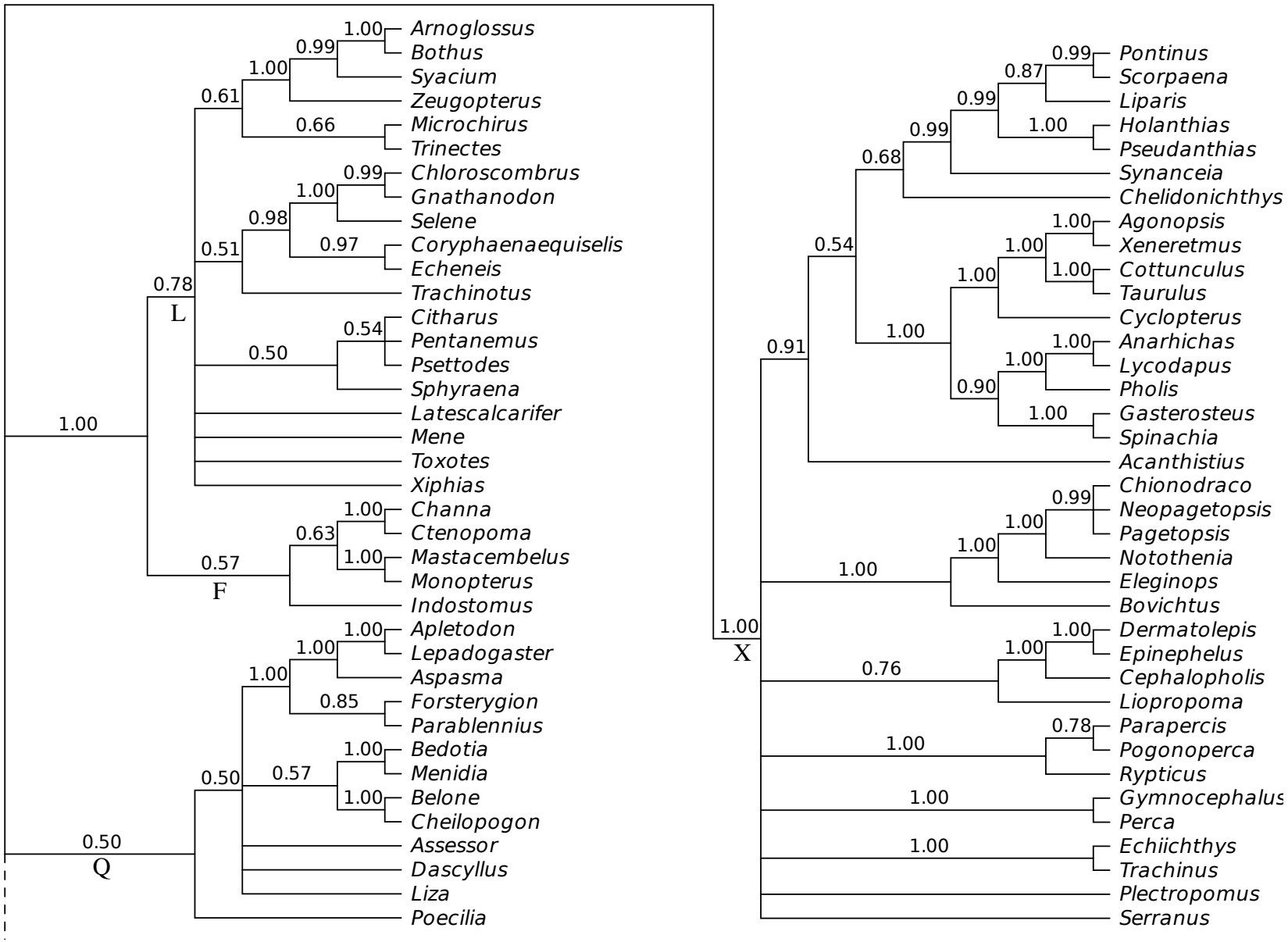


Figure 8.3. Consensus majoritaire des arbres échantillonnés par `MrBayes` pour MLL4 (suite). Clades Q (en bas à gauche), F (au milieu à gauche) et L (en haut à gauche) et X (à droite).

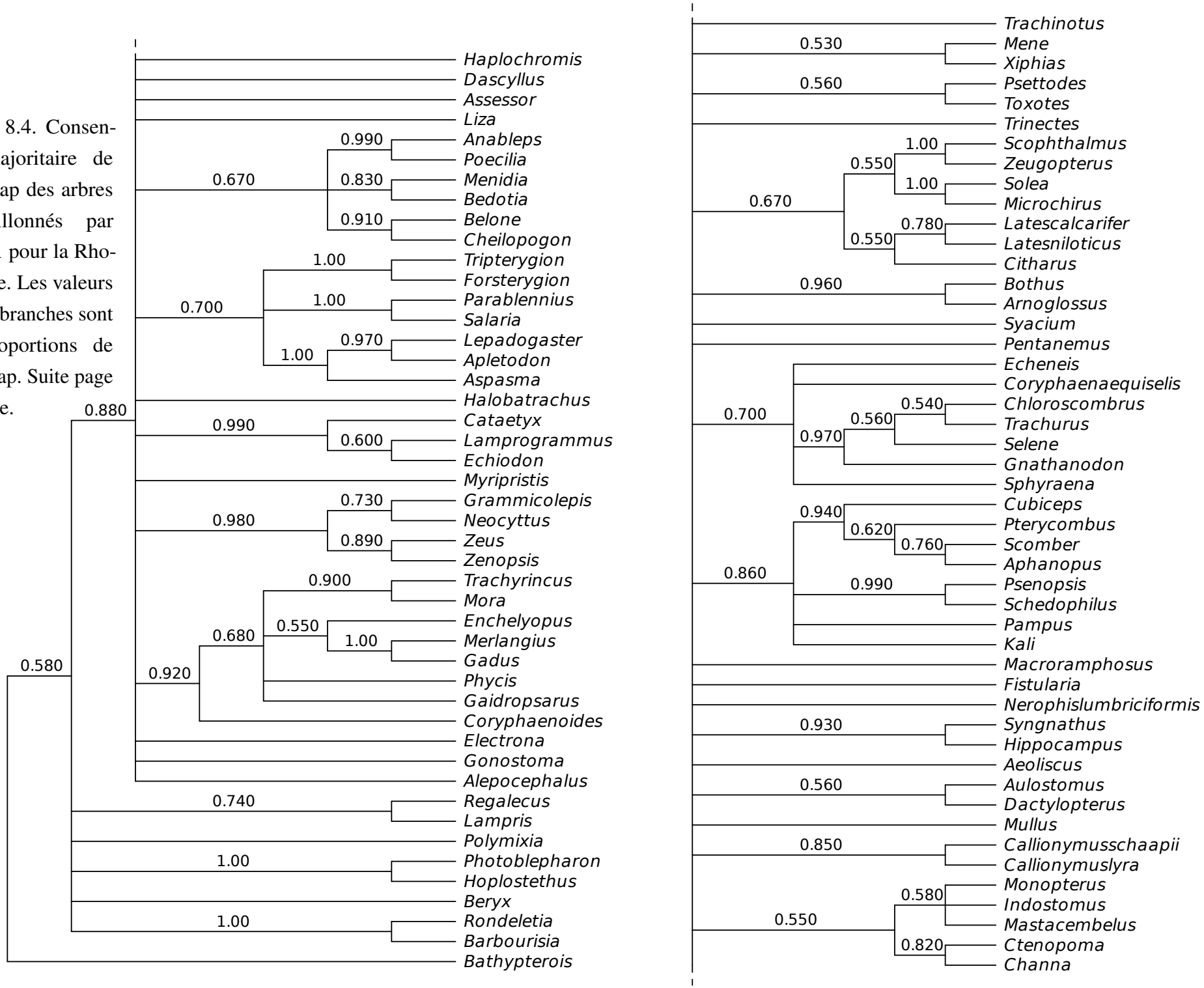
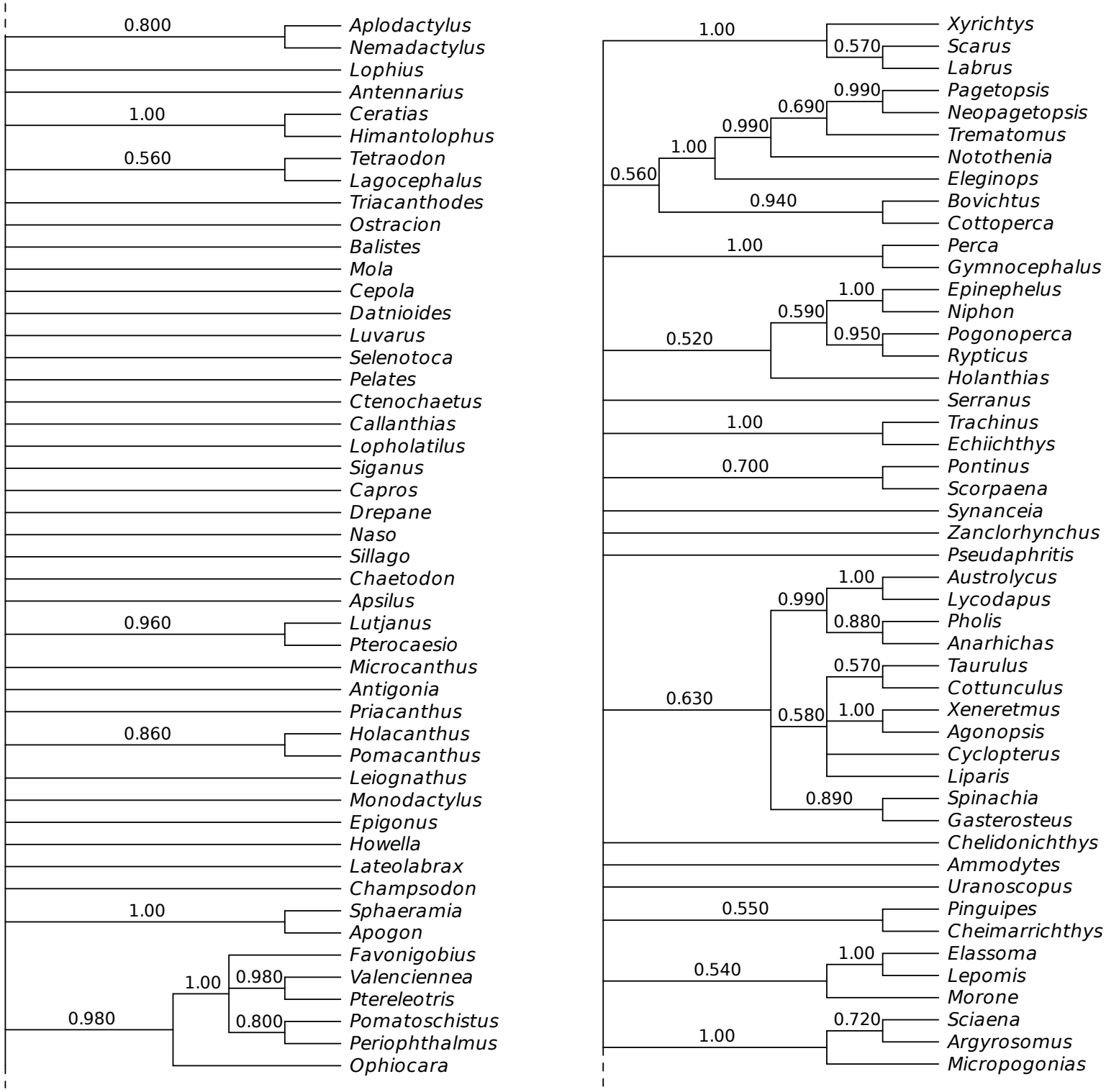


Figure 8.4. Consensus majoritaire de bootstrap des arbres échantillonnés par `Phyml` pour la Rhodopsine. Les valeurs sur les branches sont les proportions de bootstrap. Suite page suivante.

Figure 8.4. Consensus majoritaire de bootstrap des arbres échantillonnés par `Phyml` pour la Rhodopsine (suite).



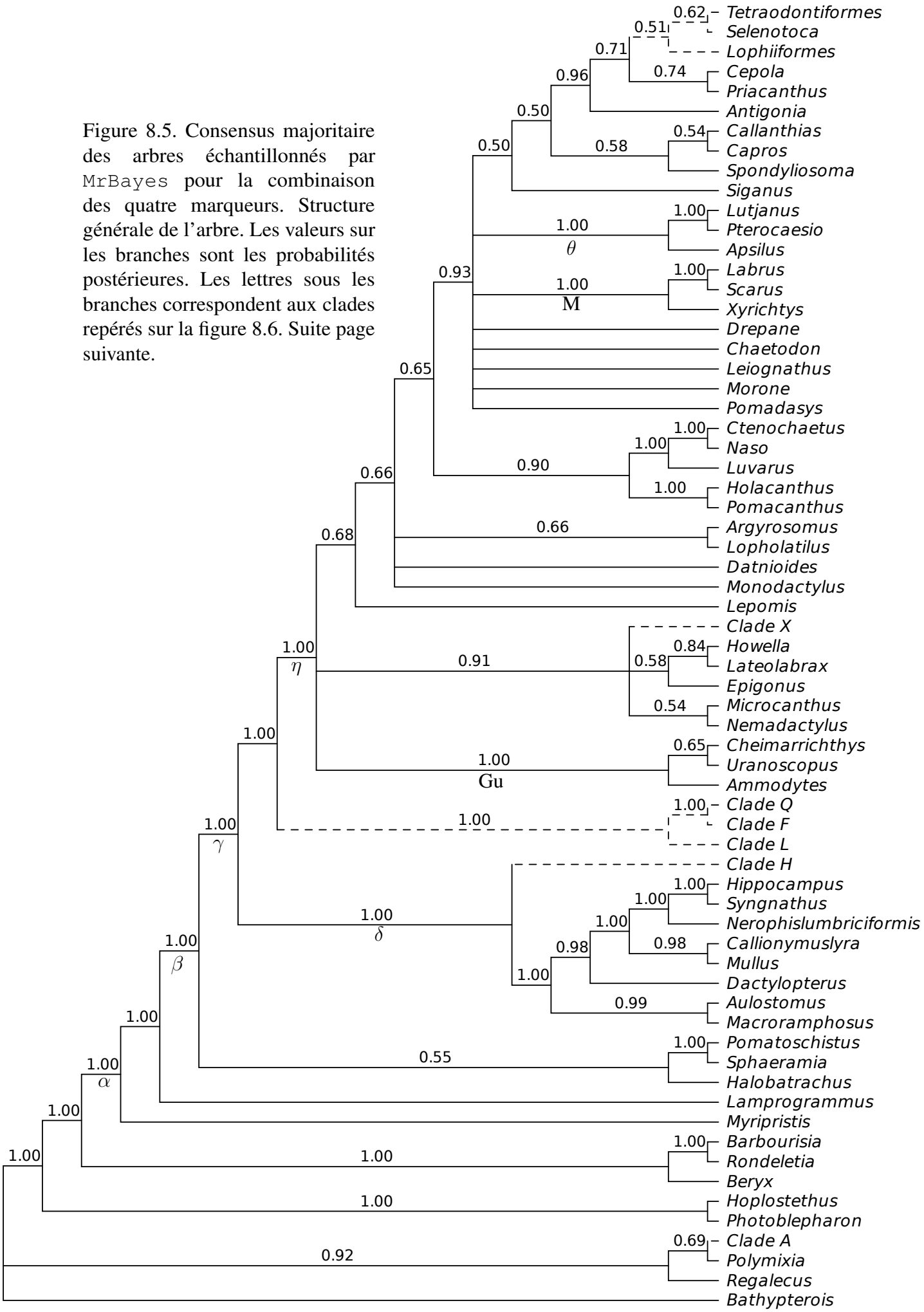


Figure 8.5. Consensus majoritaire des arbres échantillonnés par `MrBayes` pour la combinaison des quatre marqueurs. Structure générale de l'arbre. Les valeurs sur les branches sont les probabilités postérieures. Les lettres sous les branches correspondent aux clades repérés sur la figure 8.6. Suite page suivante.

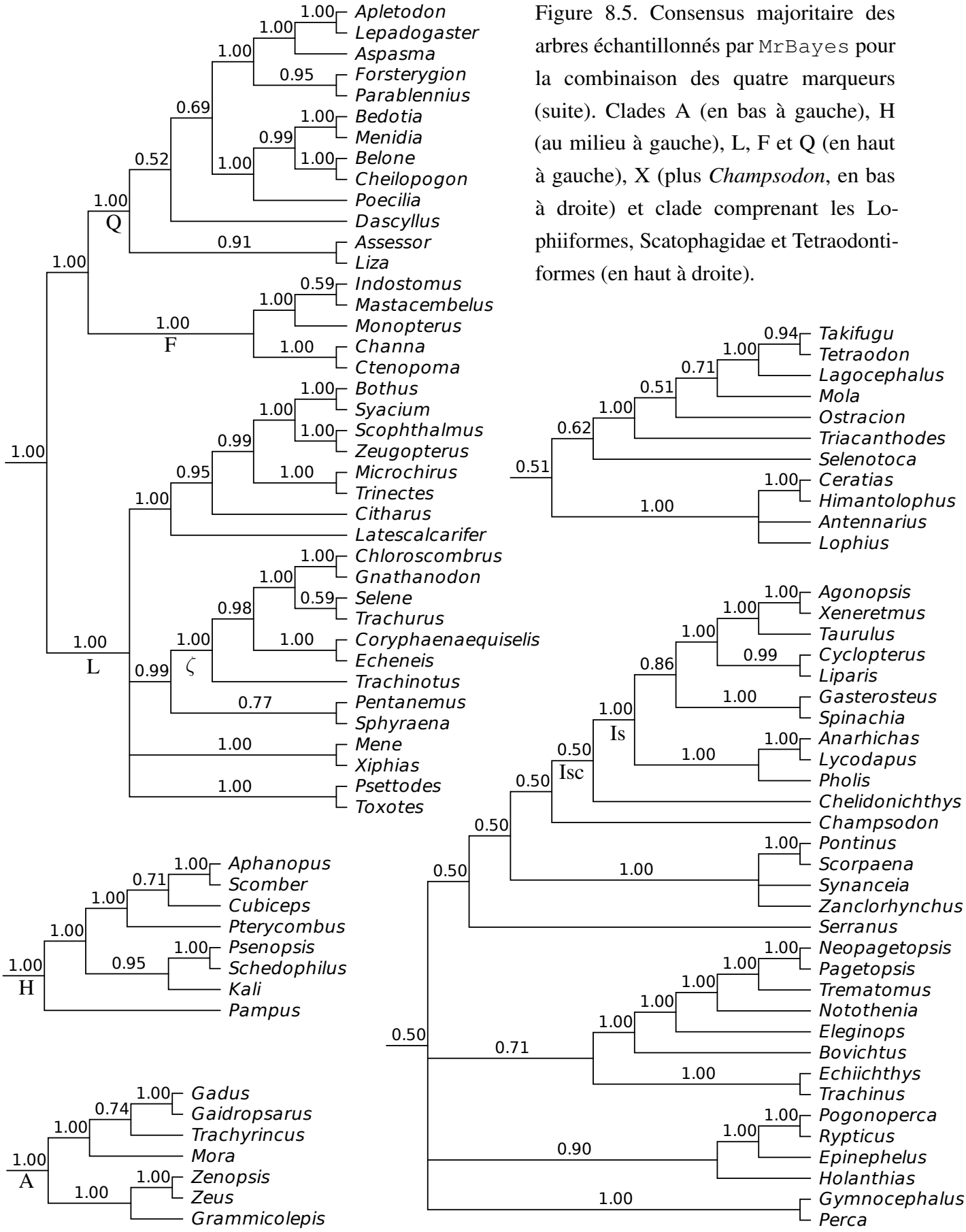


Figure 8.5. Consensus majoritaire des arbres échantillonnés par `MrBayes` pour la combinaison des quatre marqueurs (suite). Clades A (en bas à gauche), H (au milieu à gauche), L, F et Q (en haut à gauche), X (plus *Champsodon*, en bas à droite) et clade comprenant les Lophiiformes, Scatophagidae et Tetraodontiformes (en haut à droite).

8.2.2. Analyses secondaires

L'intersection de tous les domaines de validité permet le calcul d'un indice de fiabilité basé sur tous les jeux de données, mais pour un ensemble restreint de taxons (92). Un arbre de synthèse (Figure 8.6), portant les indices de répétition, a été construit dans cet ensemble selon la méthode exposée page 51, à partir des clades de fiabilité positive et compatibles entre eux enregistrés. Dans cet arbre, les clades X et N n'apparaissent pas.
Pour avoir un arbre de synthèse contenant tous les taxons, deux méthodes ont été utilisées. La méthode dérivée du MRP, exposée page 52 donne l'arbre de la figure 8.7. La méthode `CutMinKeepMax`, proposée page 55, donne l'arbre de la figure 8.8. Cette méthode échoue visiblement à placer les taxons présents dans un seul jeu de données.
Le tableau 8.2 met en évidence la correspondance entre les clades fiables de l'arbre de synthèse construit sur un échantillonnage restreint et ce qu'on observe dans les deux autres arbres de synthèse.

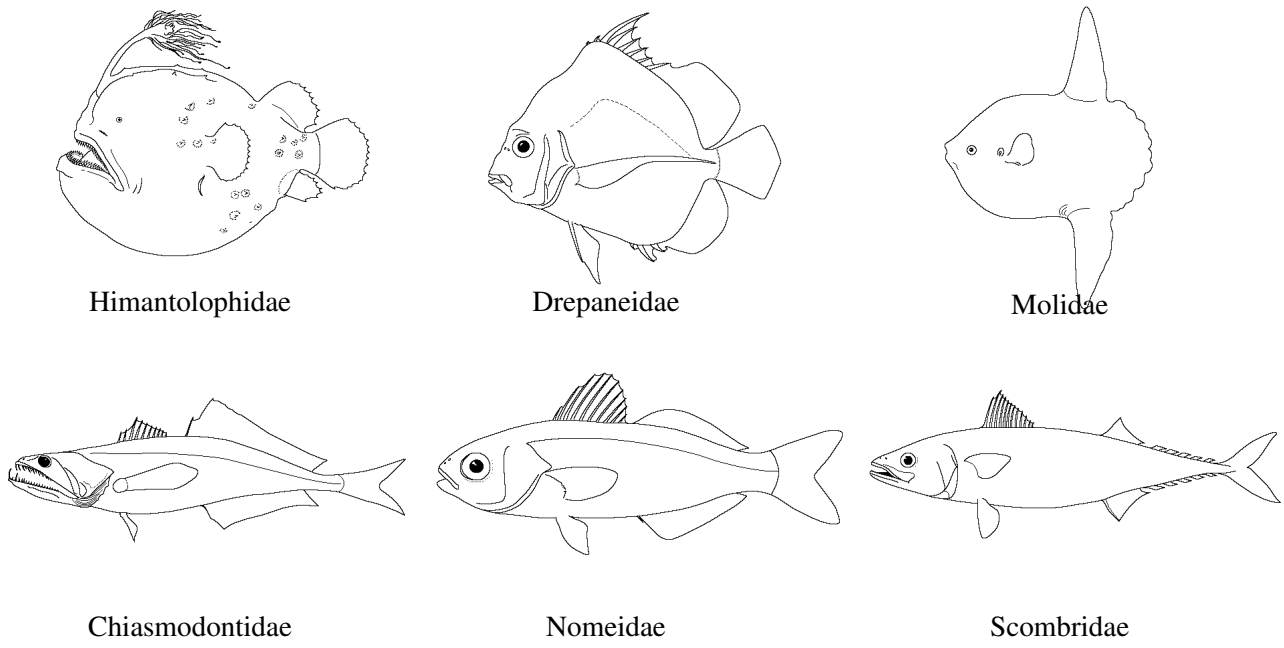


Figure 8.6. Arbre (page 170) de synthèse des clades d'indice de répétition positif dans l'intersection des domaines de validité de toutes les analyses primaires. Le premier nombre au dessus des branches est la somme maximale des valeurs de support obtenue sur des analyses de jeux de données indépendants, le deuxième nombre est l'indice de répétition calculé à partir de ces sommes de valeurs de support en prenant en compte les contradictions entre clades. L'arbre de synthèse a été construit à partir de ces indices selon la méthode exposée page 51. Les lettres grecques désignent de nouveaux clades fiables.

Figure 8.6.

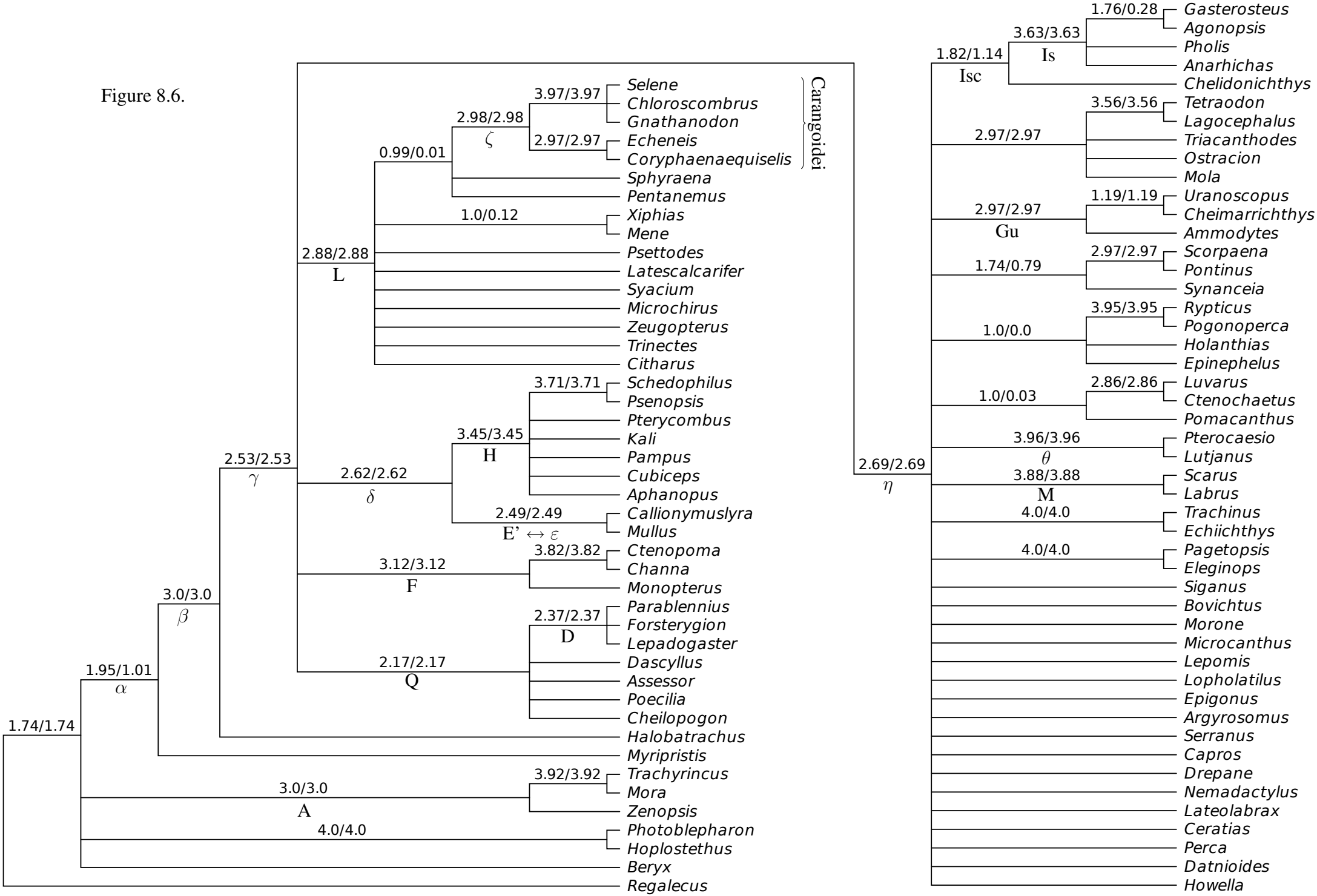

Carangoidei
Selene
Chloroscombrus
Gnathanodon
Echeneis
Coryphaenaequiselis
Sphyraena
Pentanemus
Xiphias
Mene
Psettodes
Latescalcarifer
Syacium
Microchirus
Zeugopterus
Trinectes
Citharus
Schedophilus
Psenopsis
Pterycombus
Kali
Pampus
Cubiceps
Aphanopus
Callionymuslyra
Mullus
Ctenopoma
Channa
Monopterus
Parablennius
Forsterygion
Lepadogaster
Dascyllus
Assessor
Poecilia
Cheilopogon
Halobatrachus
Myripristis
Trachyrincus
Mora
Zenopsis
Photoblepharon
Hoplostethus
Beryx
Regalecus
3.97/3.97
2.98/2.98
ζ
2.97/2.97
0.99/0.01
1.0/0.12
2.88/2.88
L
3.71/3.71
3.45/3.45
H
2.53/2.53
γ
2.62/2.62
δ
2.49/2.49
E' ↔ ε
3.82/3.82
3.12/3.12
F
3.0/3.0
β
2.37/2.37
D
2.17/2.17
Q
1.95/1.01
α
1.74/1.74
3.92/3.92
3.0/3.0
A
4.0/4.0
2.69/2.69
η
1.76/0.28
3.63/3.63
Is
1.82/1.14
Isc
Gasterosteus
Agonopsis
Pholis
Anarhichas
Chelidonichthys
3.56/3.56
Tetraodon
Lagocephalus
2.97/2.97
Triacanthodes
Ostracion
Mola
1.19/1.19
Uranoscopus
2.97/2.97
Cheimarrichthys
Gu
Ammodytes
2.97/2.97
Scorpaena
1.74/0.79
Pontinus
Synanceia
3.95/3.95
Rypticus
1.0/0.0
Pogonoperca
Holanthias
Epinephelus
2.86/2.86
Luvarus
1.0/0.03
Ctenochaetus
Pomacanthus
3.96/3.96
Pterocaesio
θ
Lutjanus
3.88/3.88
Scarus
M
Labrus
4.0/4.0
Trachinus
Echiichthys
4.0/4.0
Pagetopsis
Eleginops
Siganus
Bovichtus
Morone
Microcanthus
Lepomis
Lopholatilus
Epigonus
Argyrosomus
Serranus
Capros
Drepane
Nemadactylus
Lateolabrax
Ceratias
Perca
Datnioides
Howella

Figure 8.7. Arbre de synthèse construit par la méthode dérivée du MRP, en pondérant les clades par leur indice de répétion (voir page 52). Consensus majoritaire des arbres les plus parcimonieux obtenus par 10000 RAS+TBR et 10000 RAS+SPR. Suite page suivante.

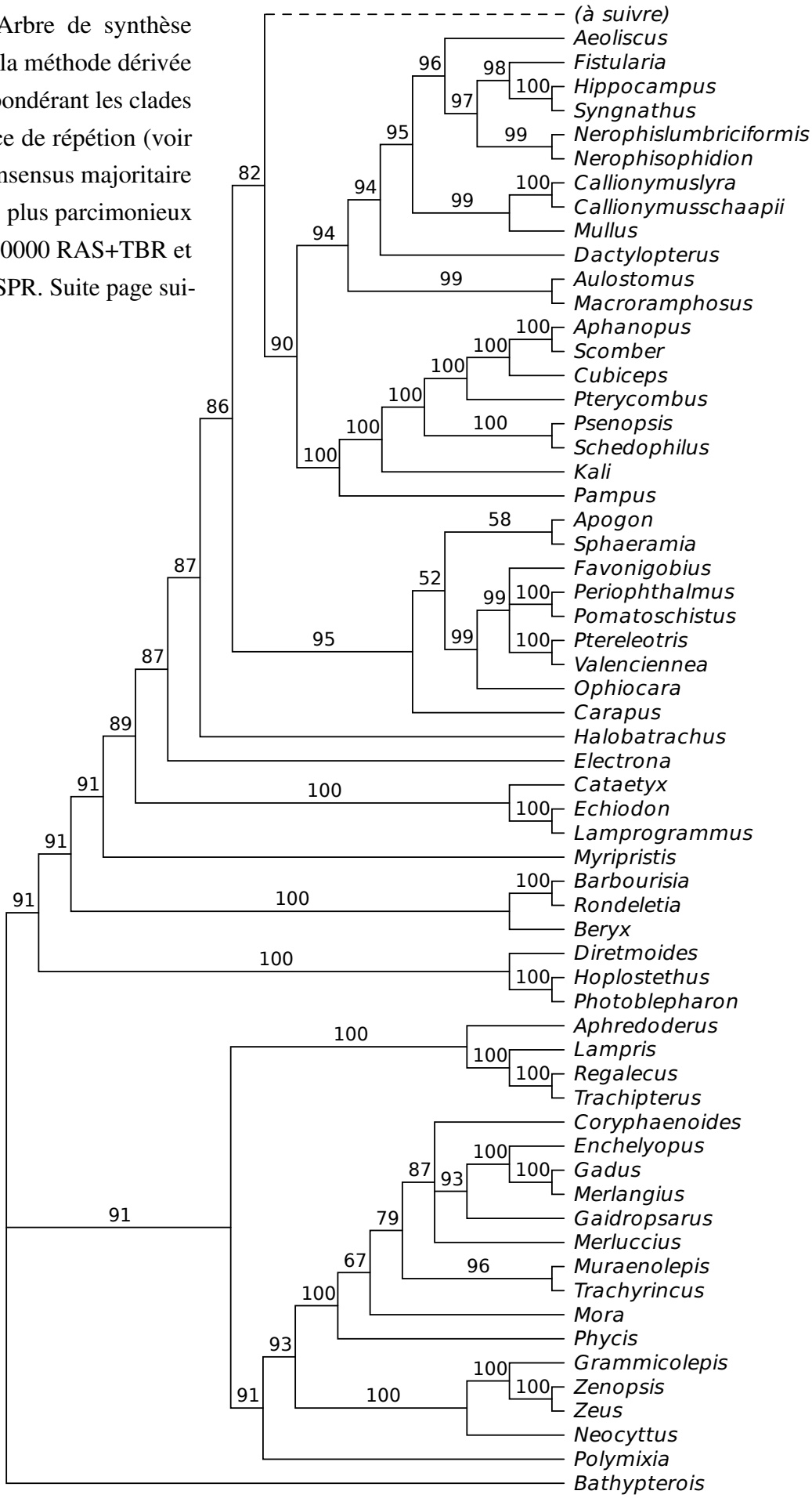
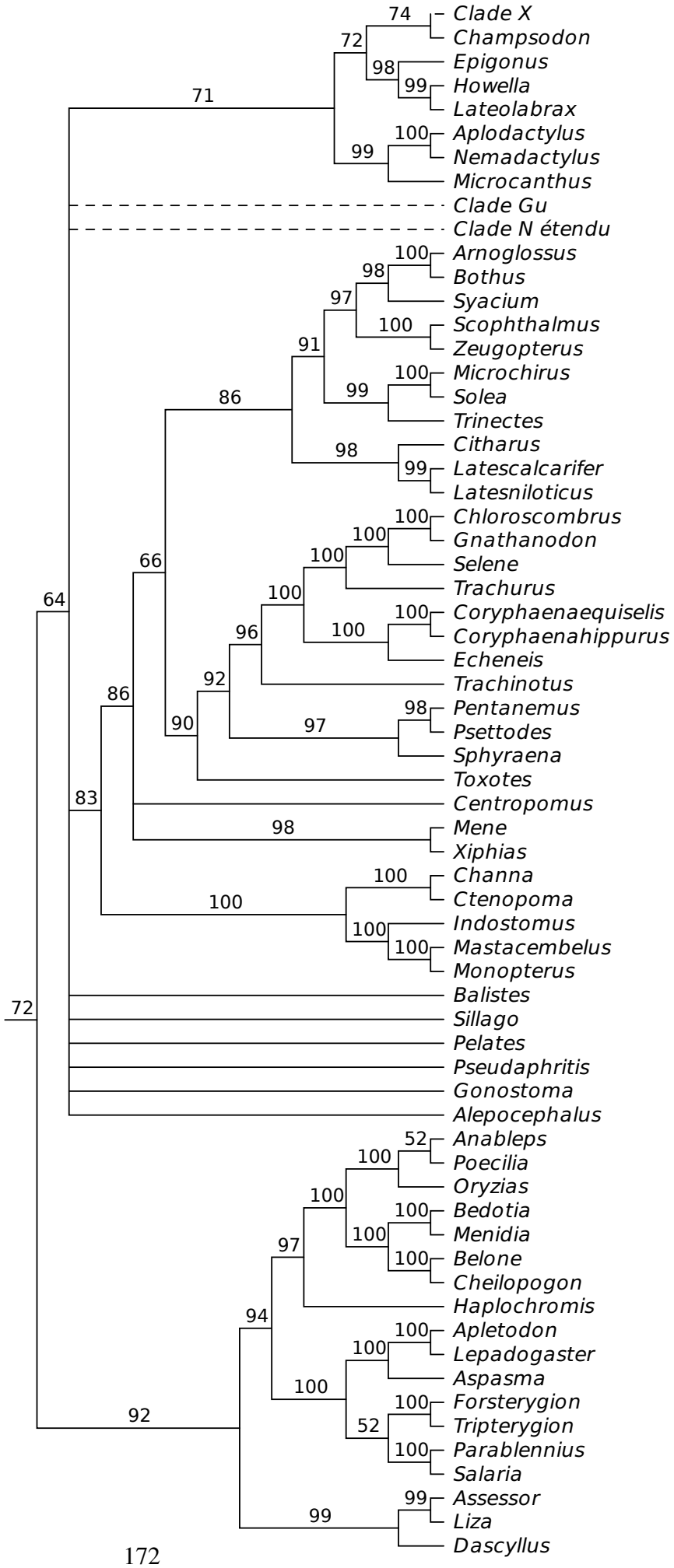


Figure 8.7. Arbre de synthèse construit par la méthode dérivée du MRP (suite). Suite page suivante (clades N étendu et Gu) et page 174 (clade X).



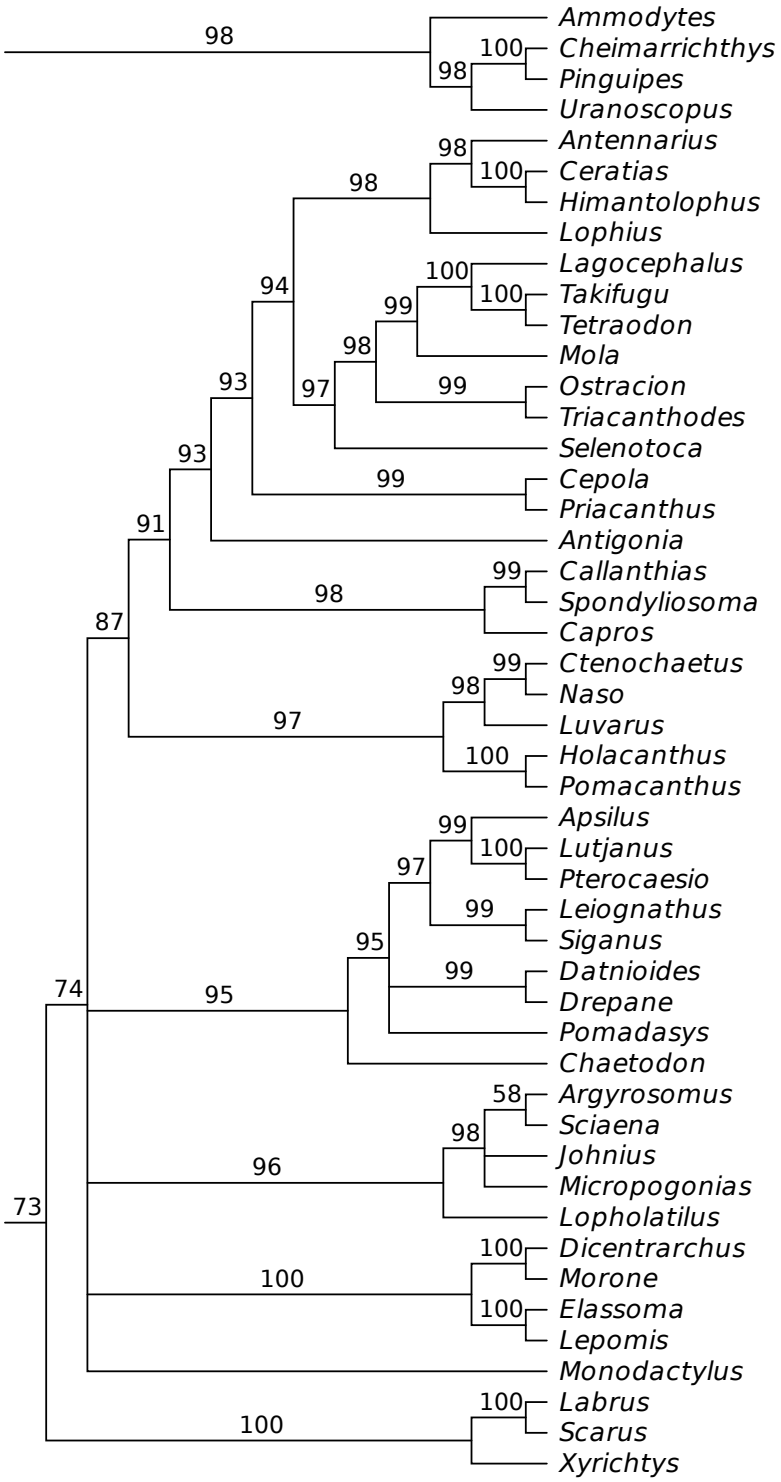


Figure 8.7. Arbre de synthèse construit par la méthode dérivée du MRP (suite). Clades N étendu (en bas) et Gu (en haut).

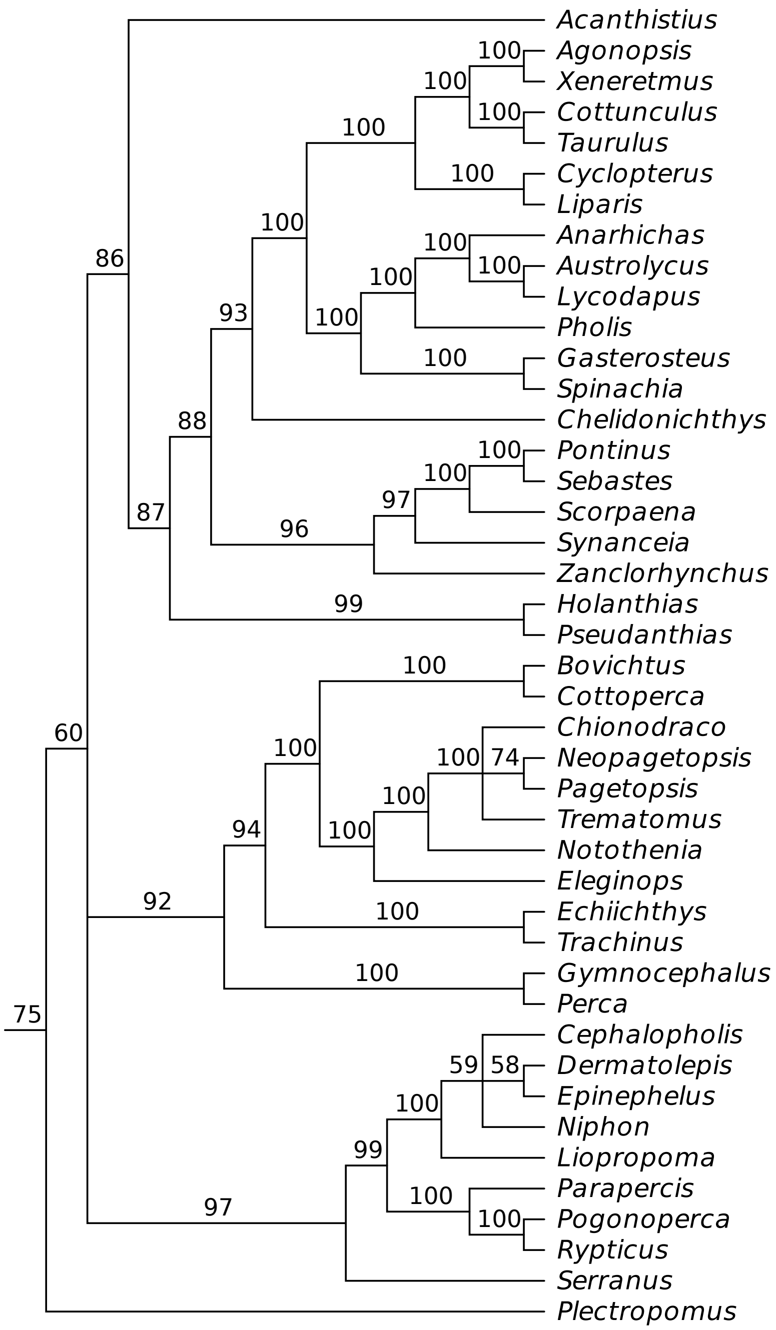
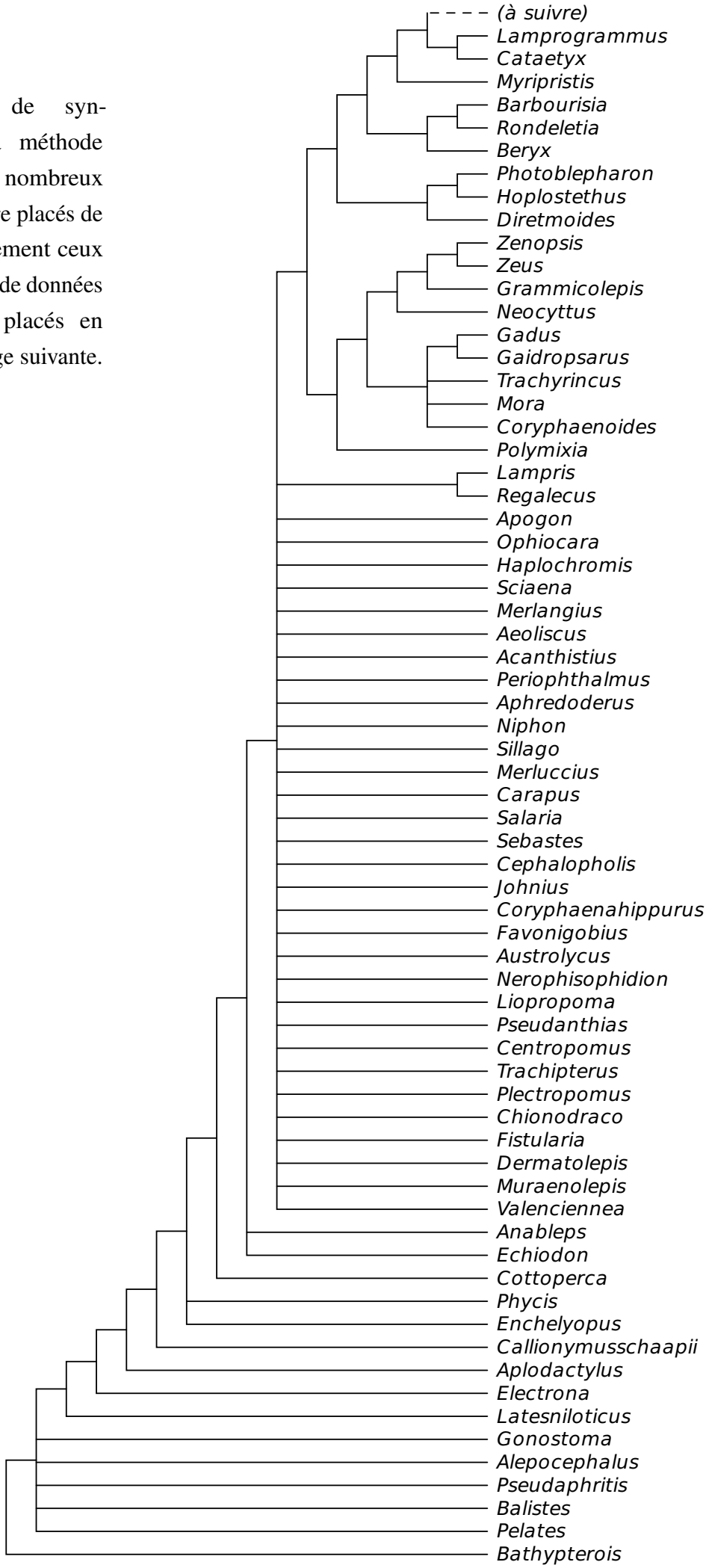


Figure 8.7. Arbre de synthèse construit par la méthode dérivée du MRP (suite). Clade X.

Figure 8.8. Arbre de synthèse construit par la méthode `CutMinKeepMax`. De nombreux taxons ne pouvant pas être placés de façon fiable, particulièrement ceux présents dans un seul jeu de données sont artefactuellement placés en position basale. Suite page suivante.



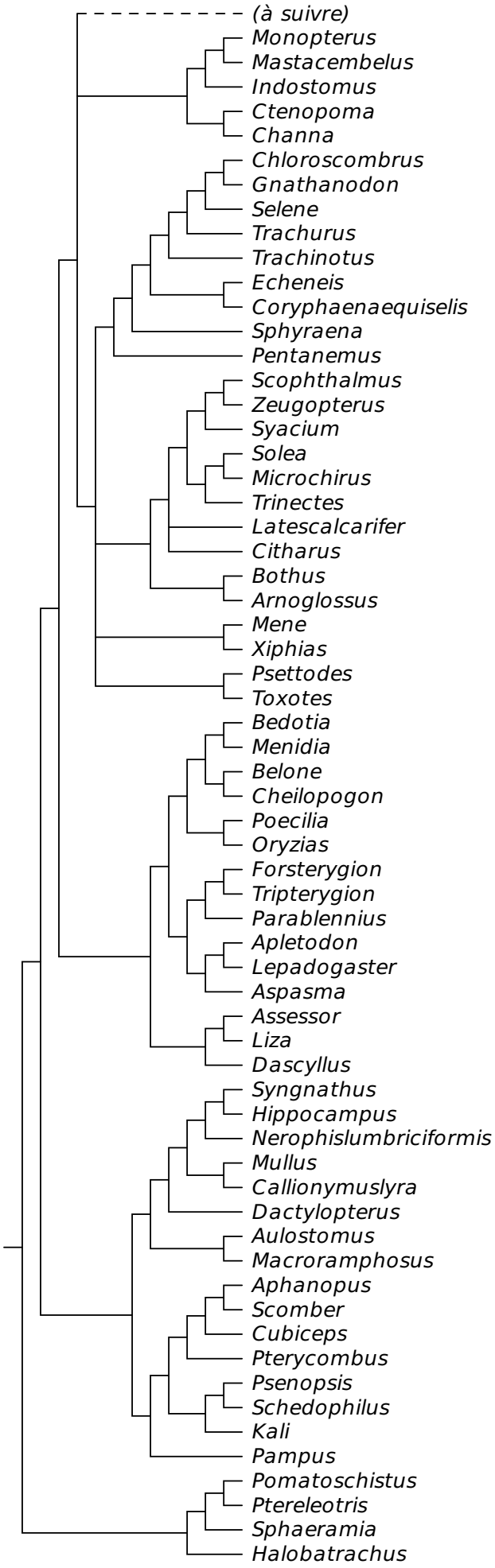


Figure 8.8. Arbre de synthèse construit par la méthode `CutMinKeepMax` (suite). Suite page suivante

Figure 8.8. Arbre de synthèse construit par la méthode `CutMinKeepMax` (suite).

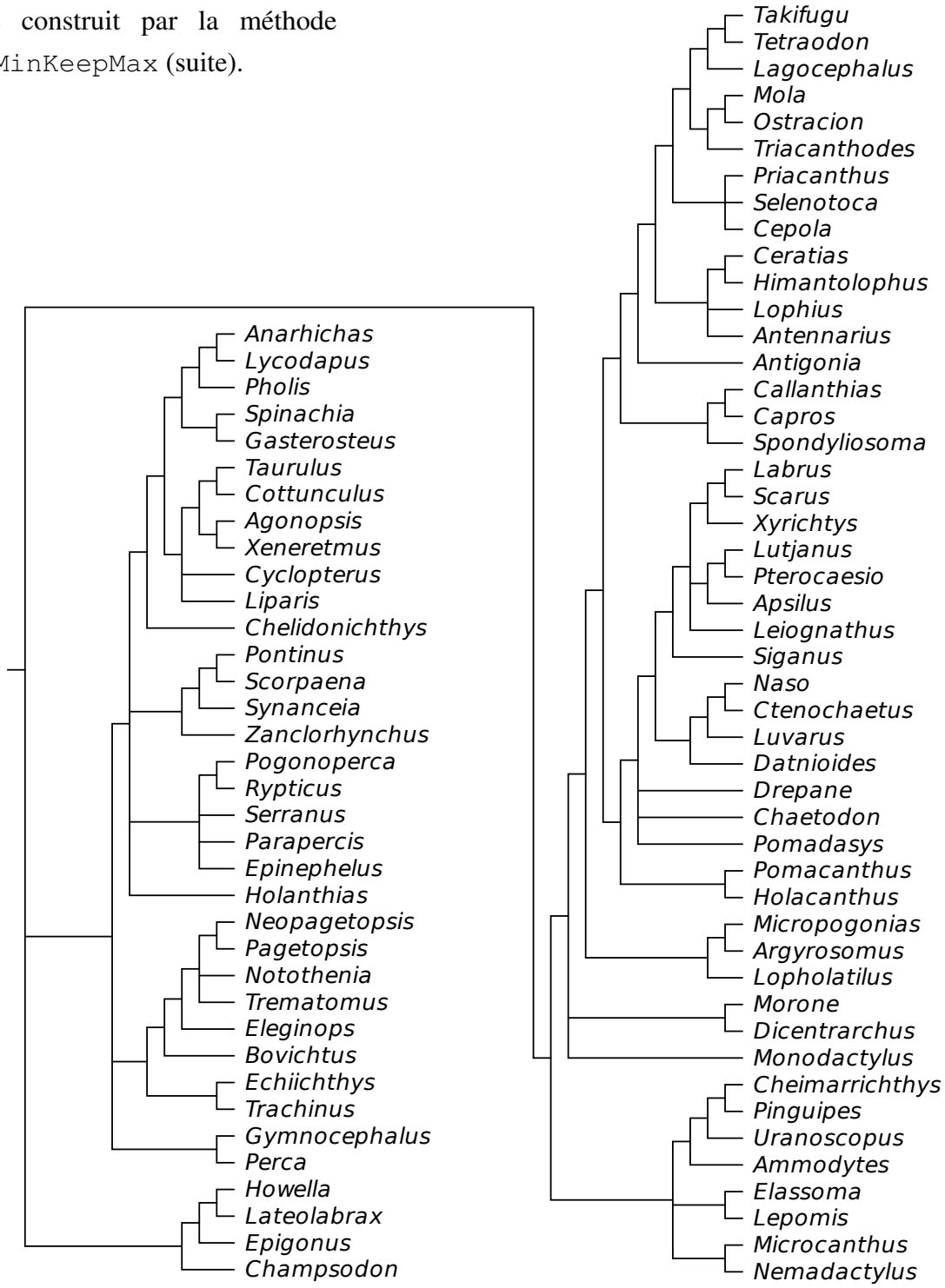


Tableau 8.2. Comparaison des topologies des arbres de synthèse. La comparaison est basée sur les clades fiables de la synthèse sur l'échantillonnage restreint (première colonne).

Figure 8.6 : Clades de fiabilité positive inter-compatibles.	Figure 8.7 : MRP.	Figure 8.8 : `CutMinKeepMax`.
Beryx, les Trachichthyoidei et le clade A sont en position basale par rapport à un clade contenant tout le reste (α).	Oui, mais *Lampris* est groupe-frère du clade A.	Oui (en l'absence des taxons rares rejetés en position basale ici).
Dans le reste, *Myripristis* est groupe-frère d'un clade dans lequel *Halobatrachus* est basal (β).	Oui (en l'absence d'Ophidiiformes et d'*Electrona* dans l'échantillonnage restreint).	Oui (en l'absence d'Ophidiiformes, de Gobioidei et d'Apogonidae dans l'échantillonnage restreint).
Le clade Q est reconnaissable, mais il manque un Cichlidae dans l'échantillonnage. À l'intérieur de ce clade le clade D (Gobioidei, Gobiesocoidei) apparaît.	Le clade Q est bien présent, le clade D également.	Le clade Q est bien présent, le clade D également.
Le clade F est présent, avec le clade f1 (*Channa*, *Ctenopoma*) en son sein.	Oui (avec *Indostomus* en plus, groupe-frère des Symbranchiformes).	Oui (avec *Indostomus* en plus, groupe-frère des Symbranchiformes).
Un clade (E',H) apparaît. DETTAÏ et LECOINTRE (2005) et DETTAÏ et LECOINTRE (2008) diffèrent dans leur définition de E'. C'est celle de 2008 qui est utilisée ici.	Oui (en l'absence de représentants du clade E dans l'échantillonnage restreint).	Oui (en l'absence de représentants du clade E dans l'échantillonnage restreint).

Tableau 8.2. (Suite)

Figure 8.6 : Clades de fiabilité positive intercompatibles.	Figure 8.7 : MRP.	Figure 8.8 : `CutMinKeepMax`.
Le clade L est présent, avec un sous-clade carangoïde (ζ, Carangoidei *sensu* JOHNSON 1993) ((*Coryphaena*, *Echeneis*), Carangidae), et sans que les Pleuronectiformes affirment leur monophylie.	On retrouve ce clade L, avec les mêmes relations internes, mais le Carangidae *Trachinotus*, non présent dans l'échantillonnage restreint, se place en position basale dans le sous-clade carangoïde au lieu de se placer au sein des Carangidae.	On retrouve ce clade L, avec les mêmes relations que dans l'échantillonnage restreint.
On trouve un clade regroupant les composantes des clades N et X et divers Perciformes (η). On pourrait appeller ce clade les Neoperciformes.	Ces Neoperciformes n'apparaissent pas ; ses principales composantes sont en polytomie avec un clade (F,L) et quelques taxons errant artefactuellement.	Ce clade est bien présent.
Le clade M (Labridae, Scaridae) est un constituant des Neoperciformes.	Oui.	Oui.
Les Caesionidae et les Lutjanidae forment un clade (θ).	Oui	Oui
Luvarus est groupe-frère des Acanthuridae.	Oui.	Oui.
Rypticus et *Pogonoperca* forment un clade.	Oui.	Oui.
Synanceia (Synanceiidae) est groupe-frère des Scorpaenidae.	On retrouve cette relation mais *Sebastes* (Sebastidae, absent de l'échantillonnage restreint) s'insère dans les Scorpaenidae et le tout a pour groupe frère *Zanclorhynchus* (Congiopodidae, également absent de l'échantillonnage restreint), formant ainsi un clade des Scorpaenoidei.	*Sebastes* est rejeté vers la base de l'arbre. Les relations des autres Scorpaenoidei sont identiques à celles qui figurent dans l'arbre de la figure 8.7.

Tableau 8.2. (Suite)

Figure 8.6 : Clades de fiabilité positive inter-compatibles.	Figure 8.7 : MRP.	Figure 8.8 : `CutMinKeepMax`.
Le clade Gu apparaît, dans lequel *Ammodytes* est groupe-frère de l'ensemble (*Cheimarrichthys*, *Uranoscopus*).	On retrouve le clade Gu (avec *Pinguipes* en plus, mais l'autre Pinguipedidae, *Parapercis* se place dans un clade de Serranidae).	On retrouve le clade Gu (avec *Pinguipes* en plus, mais l'autre Pinguipedidae, *Parapercis* se place dans un clade de Serranidae).
Le clade Is (Cottoidei, Zoarcoidei, Gasterosteidae est groupe-frère de *Chelidonichthys*, formant ainsi le clade Isc).	Oui.	Oui.

Troisième partie .

Discussion

9. Discussion des relations de parenté entre acanthomorphes

L'arbre de synthèse construit sur l'intersection de tous les domaines de validité constitue une référence topologique pour la discussion des arbres comprenant plus de taxons. En effet, contrairement aux autres, cet arbre de synthèse comporte des indices de répétition. La comparaison avec l'arbre muni d'indices permet de valider une partie des relations apparaissant dans les arbres de synthèse sans indices. Cette comparaison ne révèle pas de contradiction flagrante entre ces arbres (voir tableau 8.2). L'arbre construit selon la méthode MRP (figure 8.7) comprend une polytomie manifestement artefactuelle ; on y trouve notamment *Alepocephalus*, qui n'est pas un acanthomorphe, *Balistes* un Tetraodontiforme, et *Pseudaphritis*, un Bovichtidae. L'arbre obtenu par la méthode `CutMinKeepMax` (figure 8.8) comprend un grand nombre de taxons rejetés à la base, résultat d'un biais de la méthode discuté page 199. Pour discuter les relations de parenté entre acanthomorphes, il faut faire abstraction de ces taxons mal placés.

Avertissement à propos de « la base »

Il sera fréquemment fait usage dans la discussion d'expressions comme « en position basale » ou « à la base de ».

Ces expressions n'ont en toute rigueur de sens que relativement à un arbre donné, sur un ensemble de taxons bien défini. Pour prendre un exemple où ce problème est bien visible, on ne peut pas vraiment dire, dans l'absolu, que les Gadiformes sont en position basale dans le clade A (Gadiformes, Zeioidei), ni que ce sont les Zeioidei qui sont à la base de ce clade.

Par contre, si l'échantillonnage comprend deux Gadiformes et une dizaine de Zeioidei (en admettant que ces deux groupes sont bien groupes-frères dans l'arbre obtenu), dire que « les Gadiformes sont à la base du clade A » exprimera assez bien ce que l'on pourrait voir sur une représentation graphique de l'arbre.

Quand la discussion aborde les relations de parenté (et non plus un arbre précis) l'échantillonnage est beaucoup moins clairement défini, ce qui pose des problèmes de rigueur pour les expressions relatives à « la base ». Cependant le choix des noms de groupe, couplé à l'usage de ces expressions, permet une certaine souplesse descriptive. On pourra, au choix, dire que les Zeioidei sont en position basale dans le clade A (et il faudra alors imaginer un arbre contenant de nombreux Gadiformes formant un clade en position de groupe-frère d'un taxon « Zeioidei »), ou bien

que les Gadiformes sont à la base du clade A, auquel cas le lecteur pourra visualiser un taxon « Gadiformes » groupe-frère d'un clade constitué de nombreux Zeioidei.
Ces tournures seront donc utilisées pour leur commodité, et sous l'influence des arbres sur lesquels se base la discussion.

9.1. La base de l'arbre

Les relations à la base de l'arbre ne sont pas apparues assez répétées lors des études précédentes pour donner lieu à des clades désignés par des lettres. Dans l'échantillonnage restreint, trois clades à la base des acanthomorphes ont un indice de répétition supérieur à 1 (α, β et γ sur la figure 8.6). Les choses se compliquent (tout en restant compatibles avec ces trois clades) quand des Ophidiiformes, des Gobioidei, des Apogonidae et *Electrona* (qui n'est pas un acanthomorphe) sont présents :

– dans la figure 8.7, les Ophidiiformes s'insèrent entre *Myripristis* (Holocentroidei) et β, sauf *Carapus* qui est groupe-frère d'un clade (Apogonidae, Gobioidei), le tout formant un clade intercalé entre *Halobatrachus* et γ. *Electrona* s'insère entre les Ophidiiformes et *Halobatrachus* ;
– dans la figure 8.8, les deux Ophidiiformes qui ne sont pas rejetés à la base de l'arbre forment un clade qui s'intercale entre *Myripristis* et β. Le clade (Apogonidae, Gobioidei) est également présent, mais en groupe-frère de *Halobatrachus*. *Electrona* est en position plus basale.

Les deux méthodes de synthèse utilisant tous les taxons font donc apparaître un clade réunissant Gobioidei et Apogonidae. En outre, elle donnent toutes les deux *Polymixia* groupe-frère du clade A (clade O) et les Trachichthyoidei groupe-frère d'un clade ((*Beryx*, Stephanoberyciformes), α). Le clade O est donc confirmé. *Aphredoderus* (Percopsiformes) ne peut en revanche pas être placé pour le moment.

Le clade α est conforme au résultat obtenu sur des données morphologique par STIASSNY et MOORE (1992), à savoir la non monophylie des Beryciformes, les Holocentridae étant en position moins basale que les autres.

Comme ici, MIYA *et al.* (2005) obtiennent un clade (Ophidiiformes, β), mais avec *Carapus* faisant bien partie du clade des Ophidiiformes. Ils obtiennent en revanche un groupe monophylétique qu'ils nomment Berycomorpha contenant les Beryciformes et les Stephanoberyciformes. Leurs Batrachoidiformes sont groupe-frère du clade F, mais ils n'ont pas de Gobioidei ni d'Apogonidae dans leur échantillonnage. Ils placent les Percopsiformes en compagnie des Polymixiiformes, le tout étant groupe-frère du clade A.

SMITH et WHEELER (2006) ont un clade (Kurtoidei, (Apogonidae, Gobioidei)). De plus, SMITH et WHEELER (2004, p. 641) mentionnent une particularité de l'œuf qui se retrouve à la fois chez les Apogonidae, les Kurtoidei, les Gobioidei et de nombreux taxons du clade Q : les œufs

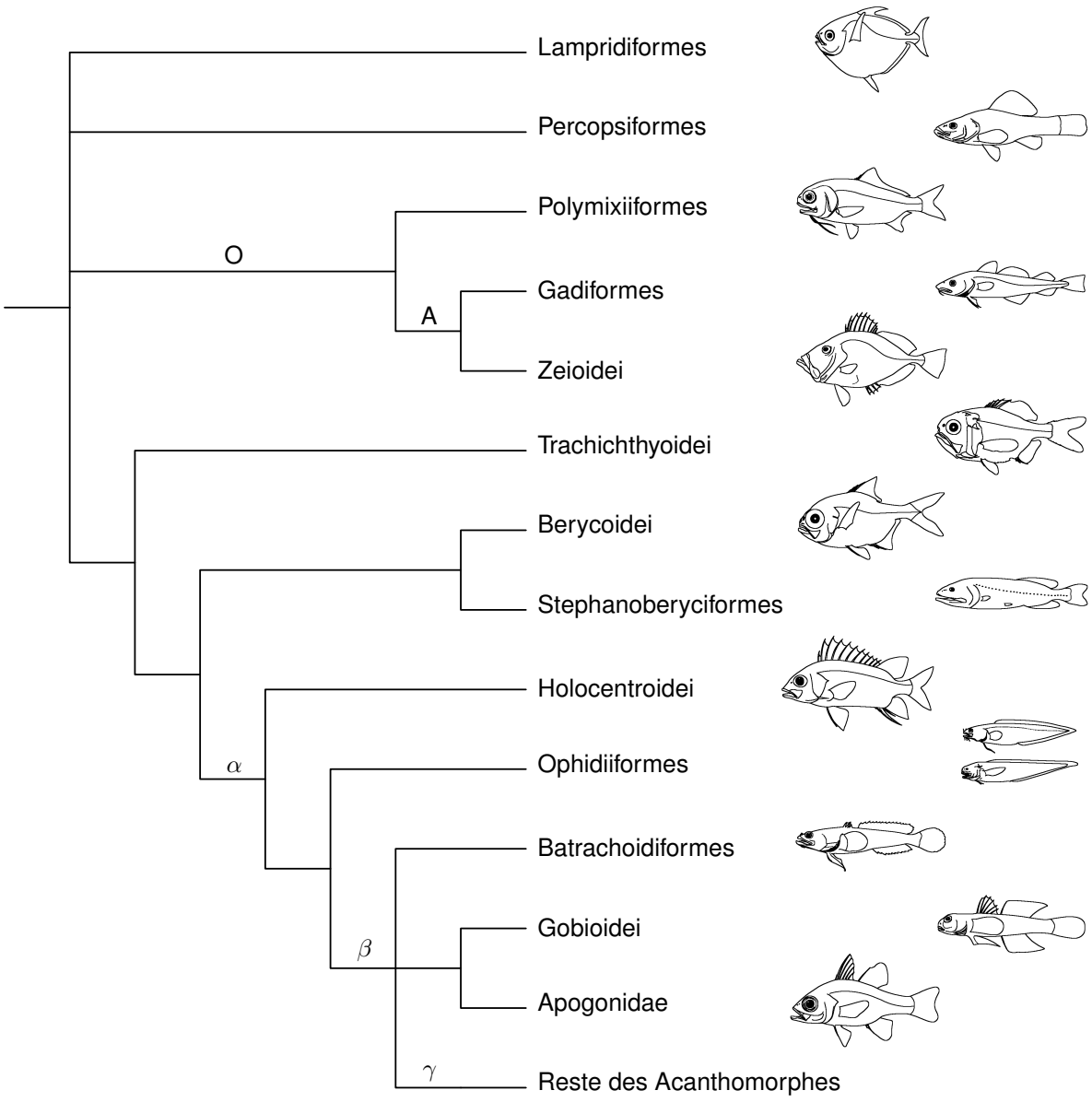


Figure 9.1. Résumé de la topologie de la base de l'arbre des acanthomorphes.

de ces taxons sont dotés de filaments adhésifs. L'hypothèse d'un clade comprenant Gobioidei, Kurtoidei et Apogonidae est probablement juste. En effet, JOHNSON (1993, p. 11) signale la présence de papilles sensorielles libres chez les Kurtoidei, les Gobioidei, les Apogonidae et les Champsodontidae ainsi que d'autres caractères suggérant fortement un rapprochement entre Kurtoidei et Apogonidae. PROKOFIEV (2006) rapproche également les Apogonidae des Kurtidae.

Si l'on met de côté les taxons rejetés à la base de l'arbre par la méthode `CutMinKeepMax` et que l'on néglige le comportement de *Carapus* et d'*Electrona* dans l'arbre obtenu par la méthode MRP, on peut résumer la base de l'arbre par la figure 9.1.

9.2. Les sous-clades de γ

Dans l'échantillonnage restreint, quelques grands clades déjà mis en évidence sont dotés d'un indice de répétition supérieur à 2 (le maximum possible est 4) : Q, D, F, E', H, L. À ces clades déjà répertoriés s'ajoutent trois nouveaux clades fiables (δ, ζ, η sur la figure 8.6) :

– δ regroupe les clades E' et H ;
– ζ est un sous-clade du clade L regroupant deux clades fiables, les Carangidae et (Echeneidae, Coryphaenidae). C'est ce que l'on peut appeler la composante carangoïde du clade L ;
– η regroupe de nombreux acanthomorphes, dont ceux qui composent les clades X, N et Gu. On pourrait nommer ce clade « Neoperciformes ».

Dans les arbres de synthèse comprenant tous les taxons, ces clades sont pour l'essentiel présents. Les relations entre ces clades peuvent par ailleurs être précisées.
Dans l'arbre de la figure 8.7, la topolgie du clade γ est ((H,(E+E')),(Q,((F,L),autres))), dans celui de la figure 8.8, ((H,(E+E')),(Q,(F,L,η))). La partie basale du clade γ est donc la même pour les deux méthodes (voir figure 9.2).

9.2.1. Le clade δ

Le clade le plus basal du clade γ est composé des taxons des clades H, E' et E. Dans ce clade, seul le clade H apparaît indiscutablement en tant que tel. Les taxons du clade E sont absents de l'échantillonnage restreint. Dans les deux autres arbres, les taxons des clade E et E' sont plus ou moins mélangés. Parmis les composantes du clade E', Callionymidae et Mullidae forment un clade fiable (ε) dans l'échantillonnage restreint[1], clade retrouvé dans tous les arbres. Dans les deux arbres de synthèse comprenant tous les taxons, ε est groupe-frère d'un ensemble comprenant les Syngnathidae (dont la monophylie n'est pas certaine ici), et peut-être *Fistularia* et *Aeoliscus*. Ce clade a pour groupe-frère *Dactylopterus*, le tout ayant pour groupe-frère un clade (*Aulostomus*, *Macroramphosus*).

[1] Dans l'échantillonnage restreint, *Callionymus* et *Mullus* sont les seuls représentants de E' ; il y a donc équivalence entre E' et ε dans la figure 8.6.

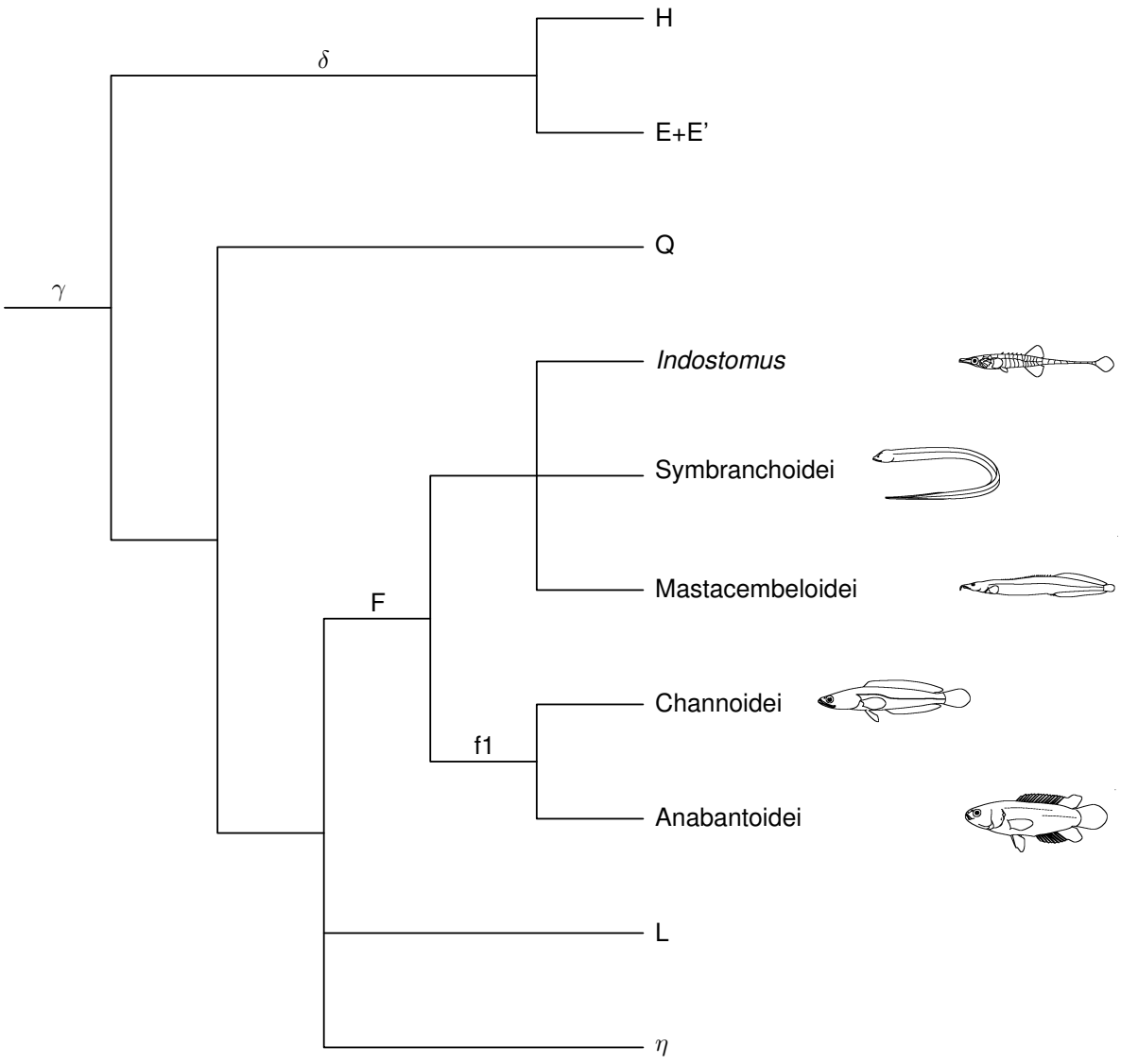


Figure 9.2. Résumé de la topologie du clade γ, avec le détail du clade F.

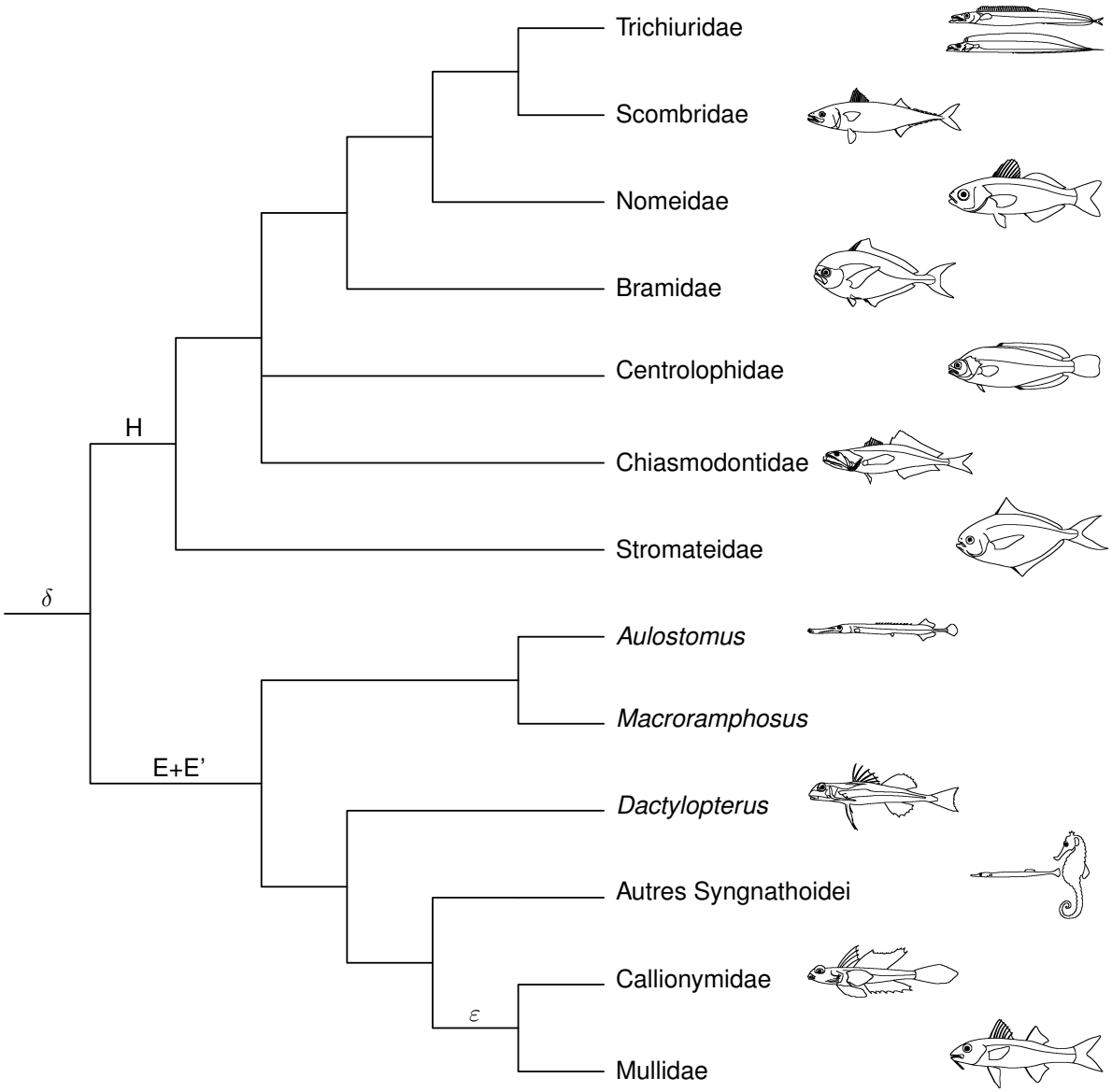


Figure 9.3. Résumé de la topologie du clade δ.

Le clade H comprend les Stromateoidei, dans lesquels s'insèrent un clade de deux Scombroidei (un Trichiuridae et un Scombridae), un Bramidae et le Chiasmodontidae *Kali*.
H est le groupe-frère de l'ensemble E + E', formant ainsi le clade fiable δ (figure 9.3).

CHEN *et al.* (2007) obtiennent un clade (H, Mullidae), équivalent (maigre, mais bien présent) du clade δ.
KAWAHARA *et al.* (2008), avec un échantillonnage conséquent parmis les Gasterosteiformes, obtiennent un clade constitué de l'ensemble des Syngnathoidei au sein duquel se logent les Dactylopteroidei. Leur échantillonnage ne comprend pas de Callionymoidei ni de Mullidae. Leurs représentants du clade H ne sont pas groupe-frère de cet ensemble, mais d'un clade constitué de leurs représentants du clade L.
SMITH et WHEELER (2006) obtiennent le clade H enrichi de quelques familles absentes de notre échantillonnage : Scombrolabracidae, Icosteidae, Pomatomidae, Arripidae, Gempylidae.

Il leur manque les Nomeidae, absents de leur échantillonnage. Un clade (Stromateidae, Scombrolabracidae) est groupe-frère du reste du clade H, ce qui concorde avec la position basale des Stromateidae obtenue avec nos données.
Les relations entre Stromateoidei obtenues ici contredisent celles obtenues sur une étude à petite échelle par DOIUCHI et NAKABO (2006) ; ces derniers placent les Centrolophidae en position plus basale que les Stromateidae.

Le groupe-frère de δ est composé de deux clades (figure 9.2) ; le clade Q et un clade comprenant les clades L, F et les composantes du clade η. Le clade η étant assez fiable dans l'arbre de synthèse construit sur l'échantillonnage restreint, on retiendra l'hypothèse d'un tel clade, composé d'au moins tous les taxons en faisant partie dans la figure 8.8 (certains des taxons rejetés à la base de cet arbre en font probablement également partie.).

9.2.2. Le clade Q

Dans l'échantillonnage restreint, seul le clade D (Blennioidei, Gobiesocoidei) est fiable dans le clade Q. Les deux synthèses comprenant tous les taxons présentent un clade Q nettement plus résolu (figure 9.4). Le clade D y a pour groupe-frère les Atherinomorpha (*sensu* ROSEN et PARENTI 1981), contrairement à l'hypothèse préférée par DETTAÏ (2004). Ceux-ci sont composés de deux clades. L'un regroupe les Cyprinodontiformes et les Adrianichthyidae, l'autre les Atheriniformes et un clade (Belonidae, Exocoetidae). Le clade Q comprend également les Plesiopidae, qui ont pour groupe-frère un clade (Pomacentridae, Mugilidae). La position des Cichlidae est incertaine ; ils sont rejetés à la base de l'arbre par la méthode `CutMinKeepMax`, mais *Haplochromis* fait bien partie du clade Q dans l'arbre de la figure 8.7.

Par rapport aux études précédentes, ce clade est enrichi des Plesiopidae et des Pomacentridae. SMITH et WHEELER (2006) obtiennent un clade contenant Atheriniformes, Cichlidae, Mugilidae, Cyprinodontiformes, Grammatidae, Pseudochromidae, Opisthognathidae, Blennioidei, mais les Gobiesocoidei, les Pomacentridae et les Beloniformes se groupent avec les Embiotocidae et *Indostomus* dans une autre partie de l'arbre. SMITH et WHEELER (2006) n'ont pas de Plesiopidae. SMITH et CRAIG (2007) trouvent un clade probablement équivalent au clade Q ; ils n'ont pas de Pomacentridae, de Mugiliformes ni de Gobiesocoidei, mais d'autres familles, absentes de notre échantillonnage, se placent en compagnie des Blennoidei, des Cichlidae, des Atheriniformes et des Plesiopidae ; les Pseudochromidae, les Opisthognathidae, les Grammatidae et les Pholidichthyidae. Ils mentionnent une caractéristique de l'œuf partagée par un certain nombre de ces taxons « pseudochromoïdes ». Les taxons du clade Q possèdent en effet pour la plupart des œufs adhésifs, caractère également présent chez les Apogonidae, les Kurtoidei et les Gobioidei (voir page 183 et également BREDER et ROSEN, 1966; MOOI, 1990).
Il est intéressant de noter que les Grammatidae comportaient autrefois des espèces qui ont ensuite rejoint les Plesiopidae ou les Cichlidae (JOHNSON 1993, p. 16).

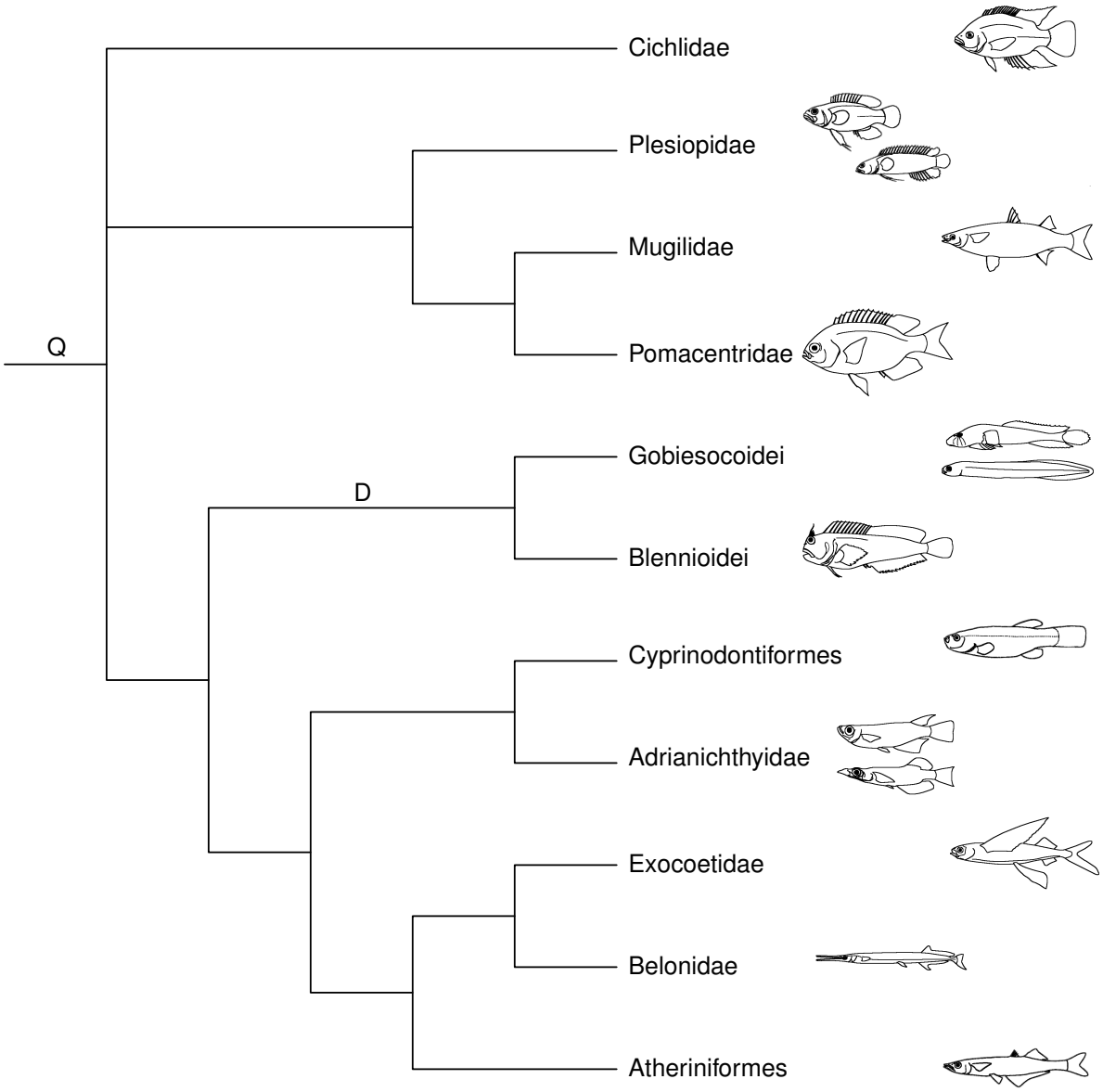


Figure 9.4. Résumé de la topologie du clade Q.

MABUCHI *et al.* (2007) confirment que les Embiotocidae et les Pseudochromidae font partie du clade Q (leur échantillonnage comprend la plupart des composantes de notre clade Q).
CHEN *et al.* (2007) obtiennent un clade Q dans lequel les Embiotocidae se placent en groupe-frère d'un clade (Pomacentridae, Mugilidae), ce qui, en l'absence d'Embiotocidae dans notre échantillonnage et de Plesiopidae dans le leur, est compatible. Dans l'autre composante de leur clade Q se trouvent les Cichlidae, groupe-frère d'un clade (Chandidae, Atherinomorpha). Les Chandidae sont donc probablement une composante de plus de ce grand clade Q.

9.2.3. Le clade F

Dans l'arbre de synthèse construit sur l'échantillonnage restreint, le clade F comprend les Symbranchiformes, groupe-frère du clade f1 (Channoidei, Anabantoidei). Dans les deux arbres

de synthèse construits sur l'ensemble des taxons disponibles, *Indostomus* s'insère dans ce clade, en groupe-frère des Symbranchiformes.

LAUDER et LIEM (1983, p. 178) font état d'un rapprochement entre Mastacembeloidei et Symbranchoidei ainsi que d'un rapprochement entre Channoidei et Anabantoidei. Ils rejettent ces deux hypothèses en concluant à une proximité entre Channoidei et Symbranchoidei. Apparemment, il y avait en fait du vrai dans les trois hypothèses.
L'appartenance d'*Indostomus* (Syngnathoidei ou Gasterosteoidei ?) a été très discutée. BRITZ et JOHNSON (2002), sur la base d'une étude morphologique du développement d'*Indostomus*, avaient conclu à un rapprochement avec les Gasterosteoidei. On peut toutefois remarquer qu'ils avaient pris la peine de signaler dans le résumé de leur article un caractère des rayons de la nageoire dorsale commun (entre autres taxons) à *Indostomus* et aux Mastacembelidae (les Symbranchoidei n'ont pas de rayons dans leur nageoire dorsale).
MIYA *et al.* (2003) obtenaient déjà un regroupement d'*Indostomus* avec les Symbranchiformes ; dans leur arbre (qui ne comprend ni Anabantoidei ni Channoidei), les Mastacembeloidei sont groupe-frère d'un clade (Symbranchoidei, *Indostomus*). En densifiant l'échantillonnage de MIYA *et al.* (2003, 2005) dans les Gasterosteiformes, KAWAHARA *et al.* (2008) confirment qu'*Indostomus* est bien éloigné des autres Gasterosteoidei.

La monophylie du clade f1 est fiable avec nos données. Par contre, le placement d'*Indostomus* est plus incertain ; aucun des jeux de données élémentaires ne donne la même relation que les arbres de synthèse. On retiendra un clade (*Indostomus*, Symbranchoidei, Mastacembeloidei) (figure 9.2).

9.2.4. Le clade L

Le clade L réunit les Pleuronectiformes, les Carangoidei *sensu* JOHNSON (1993), certains Scombroidei (Xiphiidae et Sphyraenidae), les Polynemidae, les Latidae, les Centropomidae, les Menidae et les Toxotidae. À l'intérieur de ce clade, les Carangoidei forment un clade fiable dans l'échantillonnage restreint, lui-même composé de deux clades fiables ; les Carangidae et les Echeneoidea *sensu* JOHNSON (1993) (Echeneidae, Coryphaenidae). Dans l'arbre de la figure 8.6, les Pleuronectiformes n'apparaissent pas monophylétiques, mais l'irrésolution observée reste compatible avec leur monophylie.
Dans les arbres de synthèse construits avec tous les taxons, *Psettodes* est éloigné des autres Pleuronectiformes, groupe-frère soit du Polynemidae *Pentanemus* (figure 8.7), soit de *Toxotes* (figure 8.8). *Lates* s'insère parmis les Pleuronectoidei. La monophylie des Carangidae est contredite dans l'arbre obtenu par la méthode MRP car *Trachinotus* se place à la base du clade ζ (Carangoidei). *Trachinotus* fait partie des taxons rejetés à la base de l'arbre par la méthode `CutMinKeepMax`.
Un clade (*Xiphias*, *Mene*) apparaît dans les trois arbres de synthèse, mais son indice de répéti-

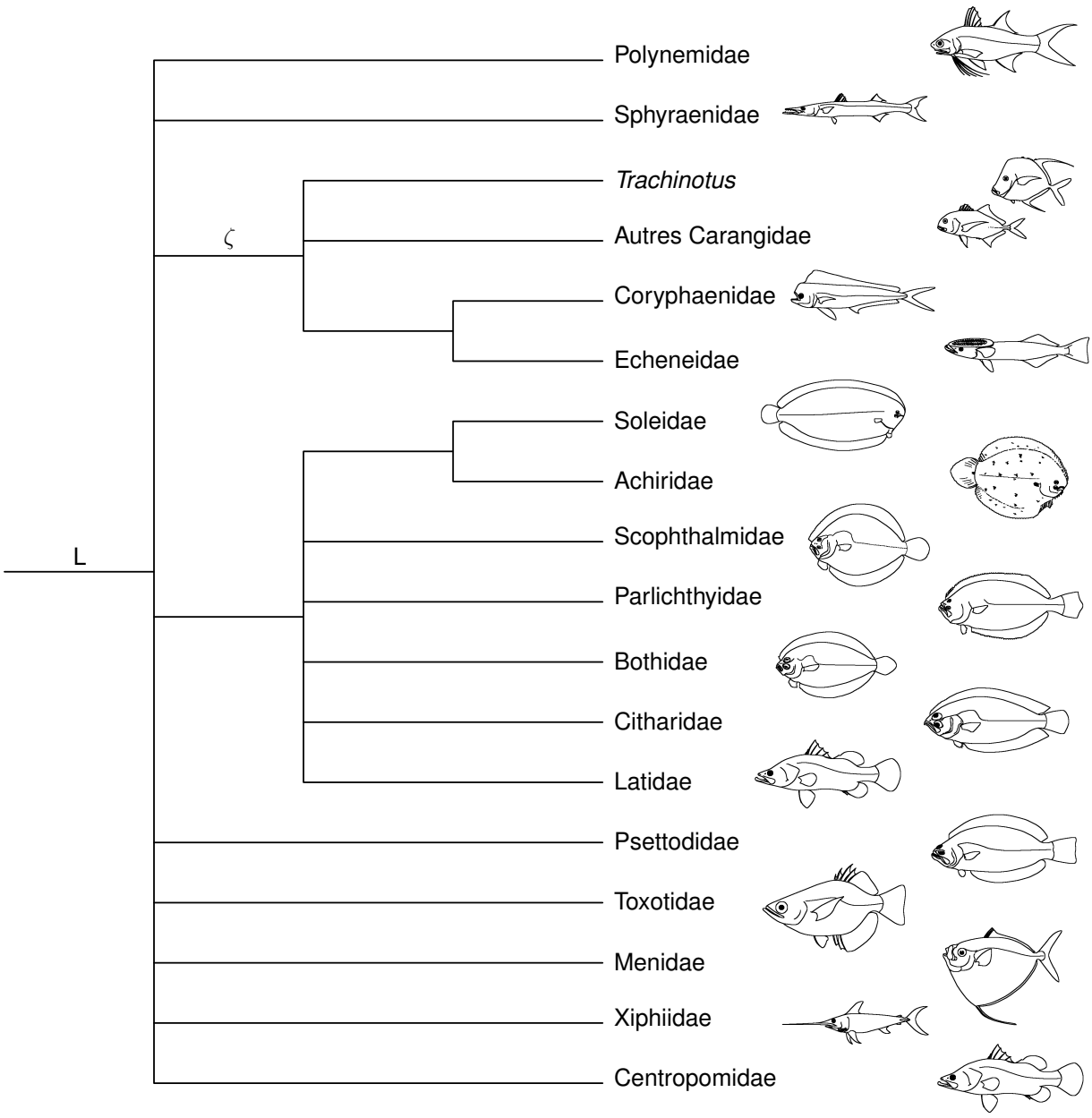


Figure 9.5. Résumé de la topologie du clade L.

tion est faible. *Sphyraena* et *Pentanemus* semblent se placer près des Carangoidei. Parmis les Pleuronectiformes, un clade (Soleidae, Achiridae) est obtenu par les deux méthodes de synthèse utilisant tous les taxons, de même qu'un clade comprenant les Pleuronectoidei et les Latidae.

Xiphias et *Sphyraena* sont séparés des autres Scombroidei (qui se trouvent dans le clade H). Cette spéaration a été également obtenue par ORRELL *et al.* (2006).
SMITH et WHEELER (2006) obtiennent un clade L assez complet ; ils n'ont pas de Toxotidae, mais ont un Leptobramidae et un Nematistiidae en plus par rapport à notre échantillonnage. Les Nematistiidae, membres des Carangoidei *sensu* JOHNSON (1993), ne se placent pas avec les Echeneoidea et les Carangidae (ces derniers étant paraphylétique en raison du placement des Echeneoidea en leur sein), mais avec les Leptobramidae et *Lates*. *Mene* et *Xiphias* ne sont pas regroupés. Comme dans les résultats de la présente thèse, les Centropomidae et les Latidae sont dans deux clades éloignés au sein du clade L. Cet éloignement entre *Lates* et

Centropomus est également en accord avec les résultats obtenus par OTERO (2004) sur des données morphologiques. Les Pleuronectoidei et les Psettodoidei ne sont pas groupe-frères.
CHEN *et al.* (2007) obtiennent un clade L dans lequel un clade (Latidae, Centropomidae) est groupe-frère des Pleuronectoidei, ces derniers n'étant en fait représentés que par un Soleidae et un Achiridae. L'autre composante de leur clade L comprend un clade (Menidae, Carangoidei) groupe-frère des Toxotidae. La composante scombroïde du clade L n'est pas représentée dans leur échantillonnage, ni les Polynemidae, ni les Psettodoidei.
MABUCHI *et al.* (2007) obtiennent un clade (Carangidae, Pleuronectoidei) ; ils n'ont pas d'autres représentants de notre clade L dans leur échantillonnage.

Les relations au sein du clade L restent encore assez peu détaillées, et la non-monophylie des Pleuronectiformes étonnante[2]. En attendant une étude plus intensive de ce clade, on peut résumer ce qui ressort de nos données par la figure 9.5.

9.2.5. Le clade η

Ce clade, n'apparaît pas dans la synthèse obtenue par la méthode MRP (figure 8.7). Cependant, sa fiabilité est bonne dans la synthèse construite sur l'échantillonnage restreint et il est obtenu lors de l'application de la méthode `CutMinKeepMax` (aux taxons rejetés à la base près), ainsi que dans les arbre des figures 8.1 (RNF213) et 8.2 (IRBP). Je suppose que s'il n'apparaît pas dans l'arbre de la figure 8.7, c'est en raison d'un placement artefactuel en polytomie de certaines de ses composantes. Le caractère artefactuel de cette polytomie est évident quand on constate qu'elle implique à la fois deux groupes extérieurs (*Gonostoma* et *Alepocephalus*), un Tetraodontiforme (appartenant normalement au clade N) et un Bovichtidae (Notothenioidei, donc appartenant au clade X). C'est la seule polytomie aussi manifestement artefactuelle reconnaissable dans cet arbre.

Le clade η contient les taxons placés précédemment dans le clade N (Tetraodontiformes, Lophiiformes, Acanthuroidei, Caproidei, Elassomatoidei, Chaetodontidae, Pomacanthidae et Drepaneidae), dans le clade M (Labridae et Scaridae ; c'est-à-dire la partie des Labroidei qui ne fait pas partie du clade Q), dans le clade Gu (Cheimarrichthyidae, Ammodytidae et Uranoscopidae), dans le clade X (Serranidae, Trachinidae, Percidae, Notothenioidei, Zoarcoidei, Gasterosteidae et tous les Scorpaeniformes sauf les Dactylopteroidei, qui se trouvent dans le clade δ). À ces taxons viennent s'ajouter les Pinguipedidae et les Champsodontidae ainsi que divers Percoidei (Moronidae, Centrarchidae, Lateolabracidae, Percichthyidae, Epigonidae, Callanthiidae, Priacanthidae, Malacanthidae, Leiognathidae, Lutjanidae, Caesionidae, Datnioididae, Haemulidae, Sparidae, Sciaenidae, Monodactylidae, Kyphosidae, Cheilodactylidae, Aplodactylidae et Cepolidae) dont une bonne partie gravitent autour du clade N[3].

[2]Elle est moins étonnante quand on sait qu'elle a été beaucoup mise en question par le passé (voir CHAPLEAU, 1993).

[3]Parmis les taxons présents dans l'échantillonnage de CHEN *et al.* (2007), les Haemulidae, les Sciaenidae, les

Sillago et *Pelates* sont absents de l'échantillonnage restreint, rejetés à la base de l'arbre par la méthode `CutMinKeepMax` et pris dans la polytomie artefactuelle de la figure 8.7. Leur appartenance est difficile à préciser. RNF213 place *Sillago* dans le clade η (voir page 159).

Si l'on néglige les taxons nouvellement échantillonnés dans cette thèse, le clade η apparaissait déjà dans l'analyse combinée de DETTAÏ et LECOINTRE (2005) (28S, 12S, 16S, Rhodopsine et MLL4) et dans DETTAÏ et LECOINTRE (2008) (IRBP seul). Il n'est donc pas surprenant de retrouver ce clade ici.
Ce clade a également un équivalent identifiable, avec un échantillonnage réduit à 19 taxons (dont un Emmelichthydae et un Hypoptychidae), chez MIYA *et al.* (2003, 2005), puis de manière un peu moins squelettique chez CHEN *et al.* (2007) (qui y inclut des Gerreidae), MABUCHI *et al.* (2007) (qui y incluent un Emmelichthydae, un Odacidae, un Hypoptychidae et un Trichodontidae) et KAWAHARA *et al.* (2008) (qui y incluent un Emmelichthydae, un un Hypoptychidae, un Aulorhynchidae et un Trichodontidae).

9.3. Les relations au sein du clade η

Dans l'échantillonnage restreint, l'intérieur du clade η est assez peu résolu. En particulier, les clade N et X, pourtant considérés comme fiables dans les études antérieures (DETTAÏ et LECOINTRE, 2004, 2005) n'apparaissent pas dans cet arbre.
Les principales composantes du clade η compatibles avec la synthèse sur l'échantillonnage restreint et apparaissant dans les deux arbres de synthèse construits sur tout l'échantillonnage sont les suivantes (figure 9.6) :

- un clade (Epigonidae, (Percichthyidae, Lateolabracidae)) ;
- un clade (Kyphosidae, Cheilodactylidae) auquel appartiennent peut-être aussi les Aplodactylidae, rejetés à la base par la méthode `CutMinKeepMax` ;
- le clade X ;
- le clade Gu, auquel *Pinguipes* a été ajouté, rassemblant un certain nombre de Trachinoidei ;
- un clade (Centrarchidae, Elassomatidae) ;
- le clade M (également présent sur la figure 8.6) ;
- un clade (Sciaenidae, Malacanthidae) ;
- un clade θ regroupant Lutjanidae et Caesionidae, également présent sur la figure 8.6 ;
- un clade (Acanthuridae, Luvaridae) ;
- un clade comprenant Lophiiformes, Tetraodontiformes, Caproidei, Scatophagidae, Cepolidae, Priacanthidae, Sparidae et Callanthiidae.

Lutjanidae et les Sparidae ont été ajoutés au clade N. Le statut des Moronidae est plus controversé.

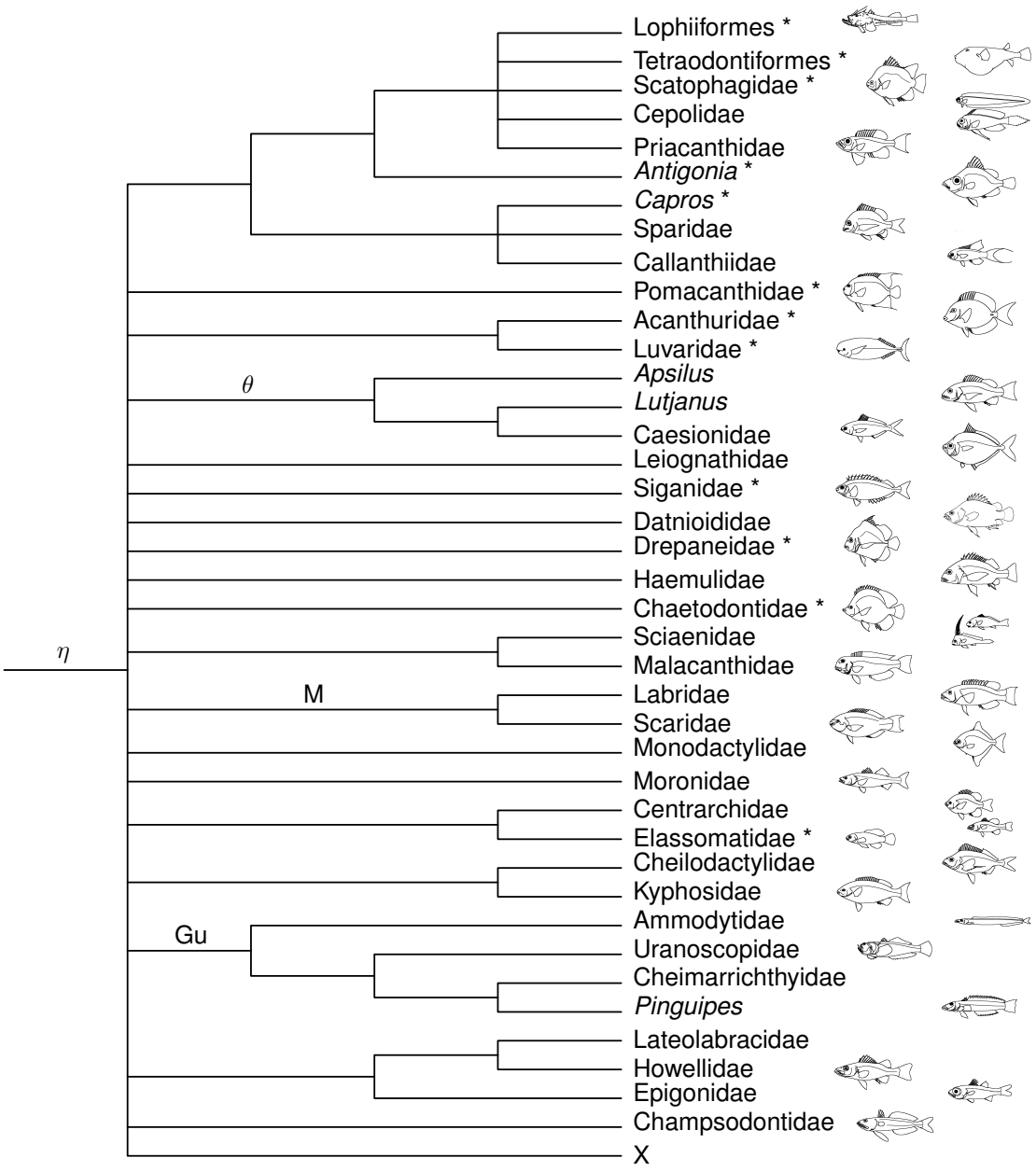


Figure 9.6. Résumé de la topologie du clade η. Les taxons constituant l'ancien clade N sont marqués d'un astérisque.

9.3.1. Les limites du clade N

Plusieurs composantes du clade η contiennent des taxons du clade N et des taxons nouveaux dans l'échantillonnage (Cepolidae, Priacanthidae, Sparidae, Callanthiidae, Centrarchidae). Si l'on considère que ces nouveaux taxons font également partie du clade N, la monophylie de ce clade n'est pas contredite.

Cependant, le manque de résolution de l'arbre de synthèse sur l'échantillonnage restreint et l'incompatibilité entre les résultats des deux arbres de synthèse sur l'échantillonnage étendu ne permettent pas de déterminer avec précision les limites de ce clade. En effet, si l'on se base sur la figure 8.7 (MRP), on peut considérer un clade N « étendu » contenant toutes les composantes du clade η sauf le clade Gu, le clade X, les Champsodontidae, le clade (Epigonidae, (Percichthyidae, Lateolabracidae)) et le clade (Kyphosidae, Cheilodactylidae). Si l'on se base sur l'arbre de la figure 8.8, il faut retirer de ce clade N le clade (Centrarchidae, Elassomatidae), ou bien, si l'on veut que le clade N contienne tous les taxons qu'il contenait jusqu'à présent, il faut y inclure également les clades Gu et (Kyphosidae, Cheilodactylidae).

Il semble donc bien y avoir au sein du clade η un clade contenant les taxons du clade N (sauf peut-être les Elassomatidae, ce qui est également suggéré par les résultats de CHEN *et al.*, 2007), du clade M et divers Percoidei dont aucun taxon du clade X, ni du clade (Epigonidae, (Percichthyidae, Lateolabracidae)), ni les Champsodontidae.

Au sein de ce clade « N étendu » aux limites floues, se distingue un clade « N restreint », centré autour des Lophiiformes, Tetraodontiformes, Caproidei et Scatophagidae. Ce clade « N restreint » est reconnaissable chez MABUCHI *et al.* (2007), qui y placent les Sparidae, les Caproidae, les Tetraodontiformes et les Lophiifiormes. Chez CHEN *et al.* (2007), les Sparidae se placent avec les Lutjanidae et les Haemulidae, mais on retrouve un clade contenant Scatophagidae, Tetraodontiformes, Lophiiformes et Caproidae. Chez HOLCROFT et WILEY (2008) apparaît également un clade contenant Lophiiformes, Tetraodontiformes, Scatophagidae et Caproidae, mais il contient aussi les Siganidae et, de même que chez CHEN *et al.* (2007) les Sparidae sont éloignés de ce clade, en compagnie des Lutjanidae[4]. Dans la présente thèse, les Lutjanidae sont rendus paraphylétiques par les Caesionidae (en accord avec JOHNSON, 1993, p. 20) dans les deux arbres de synthèse construits sur l'échantillonnage élargi. Le clade θ résultant de ce rapprochement se situe au sein du clade « N étendu », peut-être à proximité des Leiognathidae et des Siganidae. Une autre composante du clade « N étendu » est le clade contenant les Luvaridae et les Acanthuridae. Ce résultat, obtenu pour les deux méthodes de synthèse sur l'échantillonnage élargi, est concordant avec les résultats de HOLCROFT et WILEY (2008) et certaines hypothèses morphologiques (voir TYLER *et al.*, 1989). Le clade (Malacanthidae, Sciaenidae), incompatible avec les résultats de SMITH et CRAIG (2007), pourrait bien faire partie du clade « N étendu ».

[4]Il faut sans doute préciser ici que CHEN *et al.* (2007) et HOLCROFT et WILEY (2008) ont en commun le marqueur RAG1.

Le clade (Epigonidae, (Percichthyidae, Lateolabracidae)) apparaît dans les figures 8.7 et 8.8. Il est également obtenu par SMITH et CRAIG (2007). Ce clade, ainsi que les Champsodontidae (trouvés en compagnie des Epigonidae, des Percichthyidae et des Lateolabracidae par MLL4 et RNF213), semblent ne pas devoir être rapprochés du clade N. *Champsodon* est même inclus dans le clade X par l'analyse combinée totale (figure 8.5), et plus précisément en groupe-frère du clade Isc (Triglidae, (Zoarcoidei, (Cottoidei, Gasterosteidae))), le tout étant groupe-frère des Scorpaenoidei. Cette piste pour le placement des Champsodontidae est intéressante, bien que contredite au moins par MLL4 et RNF213. En effet, JOHNSON (1993, p. 14) signale qu'une particularité de l'insertion du ligament de Baudelot chez *Champsodon* est également observée chez certains taxons appartenant aux Scorpaenoidei, aux Cottoidei ou aux Zoarcoidei (voir aussi MOOI et JOHNSON, 1997).

Le clade Gu[5] regroupe une partie des Trachinoidei. Ce clade (*Ammodytes*, (*Uranoscopus*, *Cheimarrichthys*)), assez fiable dans l'échantillonnage restreint contient également *Pinguipes* dans les deux synthèses obtenues sur l'échantillonnage étendu. SMITH et CRAIG (2007) obtiennent un équivalent du clade Gu, contenant entre autres *Parapercis*, un Leptoscopidae (encore un membre des Trachinoidei), mais pas d'Uranoscopidae. Dans la synthèse obtenue par la méthode `CutMinKeepMax`, le clade Gu est placé à la base du clade N « étendu », en compagnie de certains taxons plus proches du clade X dans l'arbre obtenu par la méthode MRP (Kyphosidae et Cheilodactylidae) ou dans celui de CHEN *et al.* (2007) (*Elassoma*).

Le clade X est obtenu par les deux méthodes de synthèse, mais il contient le Pinguipedidae *Parapercis*. Il est probable que *Parapercis* se place en fait dans le clade Gu. C'est le cas pour l'analyse séparée de RNF213 (figure 8.1) et chez SMITH et CRAIG (2007). La séquence pour MLL4 de *Parapercis clathrata* est peut-être erronée. *Parapercis clathrata* a en effet posé des difficultés à l'amplification et très peu d'ADN était disponible. À l'exception de ce dernier cas, la composition du clade X est nettement mieux définie que celle du clade N. Le clade X est clairement distinct du clade N. Il a déjà été obtenu par plusieurs études indépendantes (DETTAÏ et LECOINTRE, 2004; SMITH et WHEELER, 2006; KAWAHARA *et al.*, 2008).

9.3.2. Les relations au sein du clade X

Le clade X n'apparaît pas dans la synthèse construite sur l'échantillonnage restreint, mais certaines de ses composantes se distinguent :

– un clade regroupant certains Serranidae (*Rypticus*, *Pogonoperca*, *Holanthias*, *Epinephelus*) ;
– les Scorpaenoidei ;
– les clades Is et Isc.

[5] Gu était initialement appelé G (CHEN *et al.*, 2003), puis, avec l'ajout d'*Uranoscopus*, a été renommé Gu (DETTAÏ et LECOINTRE, 2004). Cependant, il est à nouveau appelé G dans DETTAÏ et LECOINTRE (2005, 2008). J'ai préféré l'appellation la plus cohérente. Le clade G *sensu stricto* est rendu paraphylétique par la présence d'*Uranoscopus*. Par contre, le clade Gu, lui, est bien présent.

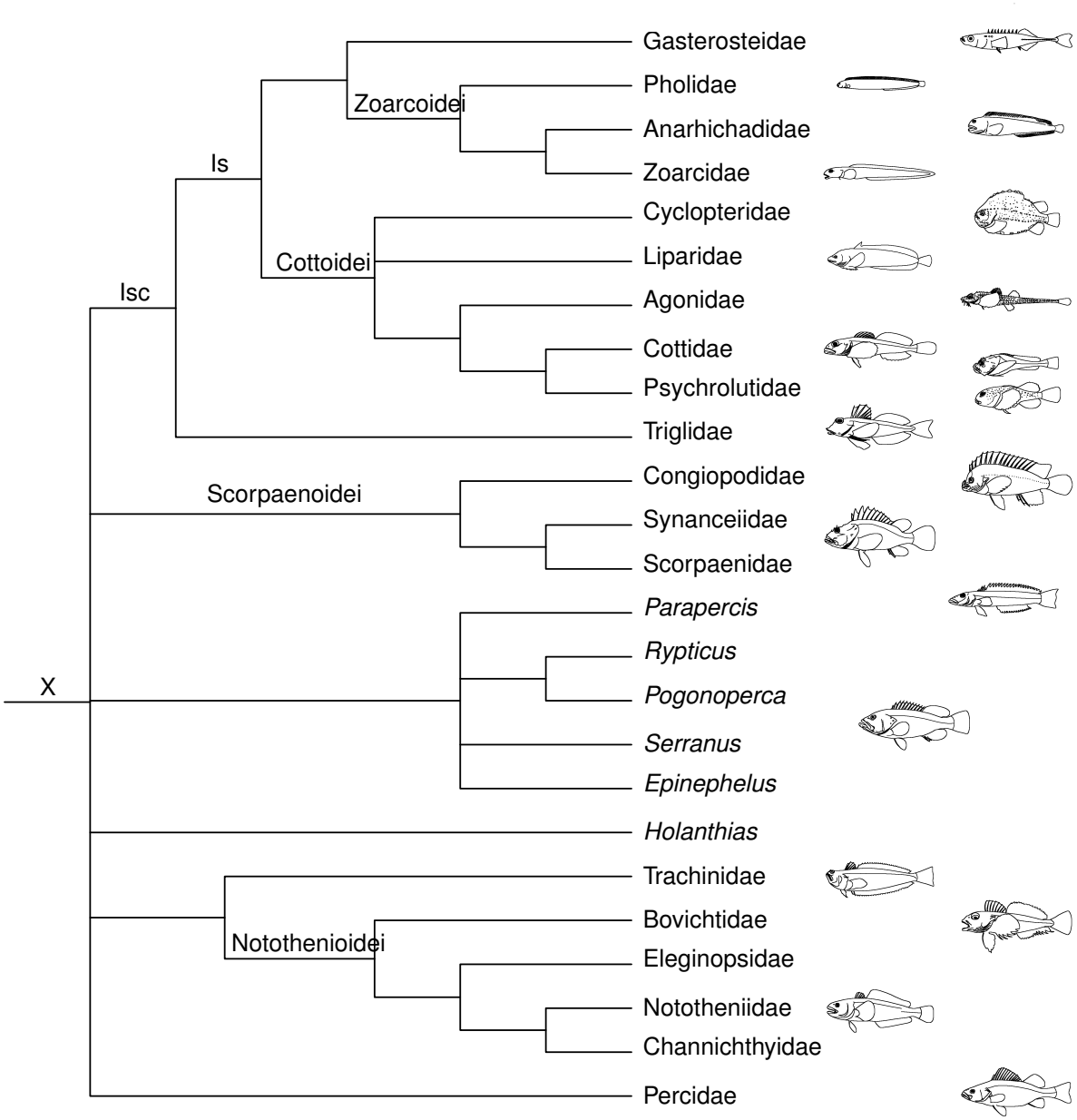


Figure 9.7. Résumé de la topologie du clade X.

Is et Isc ont une bonne fiabilité. À part le clade de Serranidae, on retrouve ces relations dans les synthèses sur l'échantillonnage étendu. Les deux arbres (figures 8.7 et 8.8) permettent d'apporter quelques précisions sur les relations entre taxons du clade X (figure 9.7).

Dans les deux arbres, les Gasterosteidae sont groupe-frère des Zoarcoidei (ce qui est cohérent avec les résultats de MIYA *et al.*, 2003, 2005), les Agonidae sont groupe-frère d'un clade (Psychrolutidae, Cottidae) et les Congiopodidae se placent à la base des Scorpaenoidei.
Un clade contenant Scorpaenoidei, Platycephaloidei, Cottoidei, Zoarcoidei et Gasterosteidae apparaît dans la synthèse par la méthode MRP. Ce clade est compatible avec les résultats de la méthode `CutMinKeepMax` ainsi qu'avec les résultats de KAWAHARA *et al.* (2008), qui ont également dans ce clade un Trichodontidae, un Hypoptychidae et un Aulorhynchidae. Ces deux dernier sont des Gasterosteoidei et forment un clade avec les Gasterosteidae, ce qui

complète le groupe-frère des Zoarcoidei. Les Trichodontidae sont des Trachinoidei que MOOI et JOHNSON (1997) proposent de reclasser dans les Scorpaenoidei. Dans leur phylogénie basée sur des caractères morphologiques, IMAMURA *et al.* (2005) placent les Trichodontidae dans les Cottoidei. Inclure un Trichodontidae dans notre échantillonnage permettrait de confirmer ou d'infirmer le placement par KAWAHARA *et al.* (2008) des Trichondontidae en groupe-frère des Cyclopteridae.
Un tel clade, auquel il faudrait peut-être ajouter les Champsodontidae (voir l'arbre de l'analyse combinée, page 168 et MOOI et JOHNSON, 1997) est une hypothèse intéressante à examiner plus en détail à l'avenir.

Au sein des Notothenioidei, pour les deux méthodes de synthèse sur l'échantillonnage étendu, les Bovichtidae sont groupe-frère du clade (Eleginopsidae, (Nototheniidae, Channichthyidae)) Ces relations sont conformes aux résultats plus détaillés obtenus par NEAR *et al.* (2004) et NEAR et CHENG (2008).
Dans les figures 8.7 et 8.8, le clade K (Trachinidae, Percidae) précédemment considéré comme fiable est contredit : Les Trachinidae sont groupe-frère des Notothenioidei. JOHNSON (1993, p. 13) signale qu'une des synapomorphies proposées pour les Trachinoidei est absente chez *Cheimarrichthys* et est en fait un caractère également présent chez les Bovichtidae et les Nototheniidae. SMITH et CRAIG (2007) trouvent au contraire que les Trachinidae sont plus proches de certains Serranidae (*Serranus*, *Pseudanthias*) alors que les Notothenioidei se trouvent dans le groupe-frère des Percidae en compagnie de *Niphon*, *Acanthistius* et *Bembrops*[6] (un Trachinoidei). Dans nos résultats, *Niphon* semble plutôt proche d'*Epinephelus*, *Pogonoperca* et *Rypticus* (figure 8.7). Cependant, la seule séquence utilisée pour *Niphon* est celle de la Rhodopsine, qui n'a été analysée qu'avec `Phyml` (figure 8.4).

Dans la figure 8.8, de nombreux Serranidae sont rejetés à la base de l'arbre (probablement parcequ'un seul marqueur a été séquencé pour ces Serranidae ; MLL4). Les relations entre Serranidae sont donc difficiles à établir. Il semble cependant qu'ils se répartissent en au moins deux groupes : sur la figure 8.7, *Holanthias* et *Pseudanthias* sont plus proches du clade Isc que d'*Epinephelus*, *Rypticus* et *Pogonoperca*. Les deux synthèses sur l'échantillonnage étendu regroupent ces trois taxons avec *Serranus* et *Parapercis* (mais voir page 196 à propos de ce dernier taxon).

[6]Un Percophidae (famille de *Bembrops*) a été collecté lors d'une mission pendant ma thèse, mais une erreur a eu lieu quelque part dans la chaîne qui va du terrain à la paillasse, et l'échantillon de tissus correspondant n'a pas pu être retrouvé. Cette anecdote illustre à quel point la possibilité d'une phylogénie moléculaire fiable dépend de la rigueur des procédures à de nombreuses étapes.

10. Nouvelles propositions méthodologiques

10.1. Mieux gérer les taxons rares

Certains taxons sont présents dans peu de jeux de données, que j'appellerais « taxons rares » ; les clades les contenant ne peuvent donc pas être répétés. Ceci semble causer des difficultés pour le placement de ces taxons dans l'arbre de synthèse. La méthode `CutMinKeepMax` y est très sensible (voir tous les taxons rejetés en position basale page 175) et celle dérivée du MRP l'est un peu également (voir sur la figure 8.7, page 172, le placement de *Balistes* ou de *Pseudaphritis* qui devraient faire partie des Tetraodontiformes et du clade X respectivement).
Lors de l'application de la méthode `CutMinKeepMax`, les sous-arbres sont progressivement séparés, avec la disparition des paires incluses impliquant des taxons appartenant à deux sous-parties distinctes de l'arbre. Les taxons rares, eux, sont impliqués dans des paires incluses en moyenne moins bien soutenues que celles n'impliquant que des taxons plus fréquents. Quand le graphe des taxons reste connexe malgré la suppression des arêtes les moins bien soutenues pour chaque taxon (*cut min*) et que l'heure est à la suppression de toutes les arêtes ayant globablement la moins bonne fiabilité, la dernière arête d'un taxon rare a plus de chance d'être éliminée qu'une autre puisqu'un tel taxon est impliqué dans peu de paires incluses, et en moyenne mal soutenues. La protection qui est accordée à ces arêtes par la règle « *keep max* »[1] n'est, en fin de compte, pas efficace.
Il faut peut-être se résoudre à ne pas placer certains taxons pour lesquels l'information phylogénétique disponible est insuffisante. À quelle étape les abandonner ? Lors des analyses primaires, ils apportent potentiellement une information utile et, si la proportion de données manquantes dans une combinaison n'est pas trop élévée, ils ne devraient pas avoir une influence trop néfaste sur le résultat. Lors du calcul de l'indice de répétition, les taxons rares ne sont déjà plus pris en compte que dans une faible proportion des domaines de validité ; ils n'ont une influence que sur les indices de répétition de bipartitions les impliquant. C'est donc postérieurement au calcul de l'indice qu'il faut se débarasser des taxons rares. L'exemple de l'arbre de synthèse dérivé du MRP montre qu'il est possible de placer apparemment raisonnablement un certain nombre de taxons rares. Il serait donc dommage d'abandonner tous les taxons présents dans un seul

[1] Voir la note page 57.

jeu de données. De plus des contradictions entre données peuvent également rendre incertain le placement de taxons plus fréquents. Ceci devrait se refléter dans les indices de fiabilité des clades contenant de tels taxons. Une procédure pour le choix des taxons à garder devrait donc se baser sur cette fiabilité. Par exemple on pourrait ne pas inclure les taxons n'étant présents dans aucun clade de fiabilité supérieure à un certain seuil. Ce seuil devrait être positif. En effet, un taxon rare sera probablement présent dans un seul jeu de données. Les clades le contenant ne seront donc pas contredits, ce qui devrait avoir pour conséquence un indice de répétition positif.

10.2. Utiliser la fiabilité pour contraindre des réanalyses

En supposant que l'indice de fiabilité permette de déterminer un certain nombre de clades fiables, on dispose alors de nouvelles connaissances qu'il serait légitime de vouloir prendre en compte. Si l'on a confiance dans la réalité de certains clades, que dire des résultats d'une analyse dans laquelle ces clades seraient contredits ? Ils devraient ne pas être considérés avec autant de confiance que d'autres résultats plus compatibles avec les connaissances nouvelles. Une manière de prendre en compte ce fait serait de réanalyser les données en contraignant les topologies possibles de sorte qu'elles soient toujours compatibles avec les clades précédemment déterminés comme fiables. On « éviterait » en quelque sorte que les données soient interprétées inutilement dans un sens incompatible avec la « réalité ». Une telle réanalyse des données pourrait permettre la mise en évidence de nouveaux clades fiables qui étaient auparavant « cachés » du fait de l'existence d'une « erreur » dans l'interprétation des données. On obtiendrait potentiellement une amélioration des connaissances grâce à une mise en cohérence de l'interprétation des données. L'efficacité de cette manière de procéder est cependant à nuancer. Premièrement, si les clades initialement considérés comme fiables ne l'étaient pas, contraindre les réanalyses par ces clades risquerait au contraire d'induire un renforcement dans le sens d'une mauvaise interprétation des données. Le choix des clades fiables à utiliser comme contraintes doit donc être plutôt conservatif. Deuxièmement, contraindre des données trompeuses à dire la vérité produit-il l'effet attendu ? Le signal véritablement phylogénétique est-il vraiment renforcé ? On est ici amené à faire le même pari que celui qui motive l'analyse combinée des données.

10.3. Utiliser la fiabilité comme critère d'optimalité

On dispose d'un indice de fiabilité pour les clades. Ne pourrait-on pas en dériver un indice de fiabilité pour un arbre ? Un tel indice de fiabilité serait un critère de sélection idéal lors d'une recherche d'arbres et pourrait être utilisé en conjonction avec l'arsenal d'heuristiques d'exploration de l'espace des arbres déjà utilisé avec d'autres critères d'optimalité, comme la

vraisemblance ou la parcimonie. Pour qu'une telle méthode soit applicable, il faut trouver une façon d'attribuer une fiabilité à un arbre qui soit calculable efficacement.
Écartons tout d'abord la transposition directe de l'indice de répétition exposé précédemment : avoir un nombre suffisant de jeux de données pour pouvoir espérer observer la répétition à l'identique d'un arbre un nombre significatif de fois est illusoire dès que le nombre de taxons en jeu est un tant soit peu intéressant biologiquement. L'expérience montre qu'un arbre riche en taxons est rarement obtenu plusieurs fois à l'identique lors d'analyses de données indépendantes. Il faut donc se baser sur la fiabilité d'éléments d'information moins complexes, comme les clades ou les triplets, calculée selon la méthode développée dans la présente thèse, puis utiliser les fiabilités ainsi obtenues pour calculer la fiabilité de chaque arbre rencontré lors de l'exploration de l'espace des arbres.

10.3.1. Quel type d'éléments utiliser ?

Lors d'une exploration de l'espace des arbres, les arbres rencontrés n'ont pas forcément que des clades dont on a déjà calculé la fiabilité. Calculer la fiabilité de chaque nouveau clade rencontré (en considérant que son nombre d'occurrences est nul et en cherchant ses contradicteurs parmis les clades déjà évalués) prendrait sans doute trop de temps, et garder en mémoire la fiabilité de tous ces clades prendrait beaucoup de place ; si n est le nombre de taxons, le nombre de clades possibles est de l'ordre de 2^n. De plus, on l'a vu précédemment, l'utilisation de clades pose problème quand les jeux de données n'ont pas exactement les mêmes taxons[2]. Il faut donc utiliser des éléments d'information plus élémentaires.

Chaque clade observé lors de l'analyse de fiabilité induit un certain nombre de triplets : $(abcd(efg))$ induit par exemple $(a|ef)$, $(a|eg)$, $(a|fg)$, $(b|ef)$, *etc.* On pourrait définir la fiabilité d'un triplet comme étant égale à la fiabilité du plus fiable clade l'induisant. Si aucun clade n'induit ce triplet, c'est qu'il est induit par un clade ayant 0 occurrences : sa fiabilité est donc égale à l'opposé du nombre d'occurrences du plus fiable clade induisant un des deux autres triplets comprenant les mêmes taxons[3].
Combien y a-t-il de triplets possibles ? Pour chaque combinaison de 3 taxons, il y a 3 triplets différents. Il y a donc $3 \times C_3^n = \frac{n!}{2(n-3)!}$ triplets possibles c'est-à-dire de l'ordre de n^3 ce qui est déjà plus raisonnable que le nombre de clades possibles, et qui permet de s'affranchir des différences de domaines de validité. Si l'on voulait stocker en mémoire tous les triplets possibles avec leurs indices de fiabilité, il faudrait donc au moins de l'ordre de n^4 bits[4], ce qui pour 100 taxons représente de l'ordre de quelques centaines de Mo, mais déjà une dizaine de Go pour 200

[2] L'arbre à qui attribuer un indice porte sur l'union des domaines de validité des clades dont on a calculé la fiabilité.
[3] Ceci n'est valable bien sûr que dans le cas où l'indice de fiabilité utilisé est l'indice de répétition proposé dans LI et LECOINTRE (in press). Si l'indice est basé sur des probabilités postérieures, il suffit de remplacer les occurrences par des occurrences partielles, c'est-à-dire les probabilités postérieures.
[4] Un triplet prend de l'ordre de n bits en mémoire (voir en annexe page 216), et un indice de fiabilité prend quelques bits (typiquement 32, mais on pourrait se contenter de moins de précision).

taxons. On se trouve donc confronté aux limites des moyens informatiques disponibles dans un laboratoire de biologie normal en 2008. Oublions pour le moment ces contraintes techniques et supposons qu'on veuille définir la fiabilité d'un arbre en se basant sur celle de ses triplets.

10.3.2. Comment passer de la fiabilité des triplets à celle de l'arbre ?

La fiabilité d'un arbre pourrait se calculer en faisant la moyenne des fiabilités des triplets le composant. Plus un arbre est complexe, plus il contient de triplets, plus il est difficile d'obtenir une moyenne élevée. Au contraire, un arbre peu résolu mais ne comportant que des clades très fiables sera lui-même très fiable. Le problème qui se pose alors est de trouver un compromis entre fiabilité et contenu en information. Il faudrait probablement n'examiner que des arbres entièrement résolus, retenir celui qui a la meilleure fiabilité, puis en retirer les clades n'induisant que des triplets peu fiables.

10.4. Placer des indices sur l'arbre de synthèse

Les méthodes décrites en 4.2 se contentent de donner une topologie pour un arbre de synthèse. Pour mieux rendre compte de l'analyse de fiabilité, il serait bon de pouvoir distinguer les clades d'un tel arbre selon leur fiabilité.
Si l'arbre de synthèse ne contient que des clades dont on a déterminé la fiabilité (cas d'un recouvrement parfait des échantillonnages taxinomiques, section 4.1), la solution est triviale ; il suffit de reporter les indices sur l'arbre.
Dans le cas plus général, quand l'arbre synthétisant les clades fiables est constitué de taxons qui ne sont pas dans les domaines de validité de tous les indices de fiabilité, ces indices ne peuvent pas être attribués à des clades de l'arbre de synthèse.

On a vu en 10.3.1 qu'on pouvait facilement attribuer une fiabilité à un triplet. On a ensuite vu en 10.3.2 qu'on pouvait utiliser la fiabilité des triplets pour définir la fiabilité d'un arbre. En s'inspirant de cette méthode, on pourrait attribuer une fiabilité aux clades de l'arbre de synthèse qui serait, pour un clade donné, la moyenne des fiabilités des triplets induits par ce clade. Les propriétés de cette moyenne restent à étudier.

Une autre approche, apparentée à celle de BREMER (1988), serait de construire des arbres de synthèse selon une méthode donnée, en n'incluant à chaque fois que des clades dont la fiabilité serait au dessus d'un certain seuil. Si les arbres obtenus pour les différents seuils sont compatibles et de plus en plus résolus au fur et à mesure que le seuil de fiabilité diminue, on peut estimer que la fiabilité d'un clade de l'arbre est supérieure au seuil le plus haut avec lequel ce clade apparaît dans l'arbre et inférieure au seuil le plus bas avec lequel ce clade n'apparaît pas.

11. Discussion générale sur la fiabilité

L'idée de mettre une valeur numérique à une fiabilité est présente chez CARNAP (1995, p. 22) ; il parle de « degré de confirmation ». Malheureusement, son concept est un peu trop abstrait pour être facilement appliqué. La fiabilité est donc le plus souvent une notion subjective ; un individu a l'intuition qu'une chose est vraie, intuition d'autant plus forte que des faits observés et cohérents avec la chose en question sont nombreux. Si l'on veut une fiabilité plus objective, qui puisse être partagée par des individus, leur servir d'argument dans une discussion scientifique, il faut nécessairement passer du stade psychologique à un stade formalisé. Formaliser, c'est — au moins implicitement — passer par une phase d'énumération et de sélection de certaines propriétés dont on veut rendre compte. Plus on veut obtenir quelque chose d'applicable à une situation concrète, plus il faut un cahier des charges précis, et plus on est obligé de simplifier la réalité. Une fois un cahier des charges défini, on peut être confronté à des obstacles lors de l'implémentation, qui obligent à faire de nouvelles simplifications. Il faut donc s'attendre à ce qu'un indice de fiabilité chiffré soit biaisé d'une façon ou d'une autre. De plus, l'acceptation des résultats par une communauté scientifique ne se fera probablement jamais sur la seule base d'un indice de fiabilité tel que proposé dans cette thèse. La connaissance admise vient de la reproduction de résultats par des équipes indépendantes ; c'est un aspect sociologique qui est difficile à prendre en compte dans un indice de fiabilité.

Conclusion

Conclusions méthodologiques

Après avoir baigné quatre ans dans la congruence taxinomique, il est surprenant de constater que SMITH et CRAIG (2007) tirent des conclusions taxinomiques sur la base d'une analyse combinée, à grande échelle, en parcimonie : il est probable que leurs Moronoidei doivent être remaniés par la suite, ce qu'ils admettent (p. 51).
À quoi sert donc d'avoir un bel arbre bien résolu, si la fiabilité des relations présentées est inconnue ? Les conclusions taxinomiques ne devraient pas êtres tirées sans avoir une idée de cette fiabilité. Ceci passe par une comparaison des résultats de différentes analyses : une méta-analyse. On obtient plus d'informations en comparant les arbres qu'en se contentant de chercher le « meilleur » arbre.
L'indice de répétition proposé dans cette thèse est un moyen de mener une méta-analyse de manière interne à une équipe de recherche, élaborant un ensemble cohérent de jeux de données. Il formalise la fiabilité d'une manière raisonnablement utilisable dans un certain nombre de situations. Dans le cas où les échantillonnages taxinomiques des différents jeux de données indépendants sont les mêmes, l'indice et la méthode de synthèse proposée page 51 sont assez satisfaisants. Dans le cas — le plus courant — où les échantillonnages taxinomiques des jeux de données ne se recouvrent que partiellement, il manque une méthode de synthèse opérationnelle donnant une fiabilité aux clades de l'arbre de synthèse. Il a donc fallu jongler entre un arbre comportant des indices de répétition mais portant sur un ensemble restreint de taxons (figure 8.6) et des arbres de synthèse prenant en compte la fiabilité des clades et portant sur tout l'échantillonnage, mais ne permettant pas de distinguer parmis les clades présents lesquels sont les plus fiables (figures 8.7 et 8.8). La proposition inspirée de l'indice de BREMER (1988), page 202, est la plus facilement implémentable à l'heure actuelle.

Disposer d'une indication de la fiabilité (sous forme d'indice ou simplement par la présence du clade dans un arbre de synthèse construit pour inclure en priorité des clades fiables) permet de sélectionner une partie des résultats sur lesquels porter son attention, et ce, de manière un peu plus confortable que *via* la construction à la main d'un tableau des clades répétés. Ainsi, des nombreuses hypothèses morphologiques et moléculaires proposées dans la littérature, seul un petit nombre a été relevé dans la discussion (chapitre 9).

Conclusions phylogénétiques

De l'intérêt d'élargir l'échantillonnage taxinomique

L'extension de l'échantillonnage taxinomique, même faite un peu au hasard et au gré des opportunités, permet de mettre de l'ordre dans la phylogénie. Je dirais que cette non-stratégie d'échantillonnage est utile à l'échelle de la phylogénie des acanthomorphes, parceque le champ à couvrir est encore bien clairsemé. Ainsi, en intégrant des Apogonidae dans l'échantillonnage, à l'occasion d'un échange avec Leo `Smith` et du passage au laboratoire d'une chercheuse de Singapour, on a pu mettre en évidence un clade répété (Gobioidei, Apogonidae) qui s'est trouvé avoir déjà été obtenu par SMITH et CRAIG (2007), sans attirer particulièrement leur attention, et pour lequel une synapomorphie potentielle a été trouvée dans la discussion de la position phylogénétique des Kurtoidei par JOHNSON (1993, p. 11).
On peut constater que, comme ci-dessus, une partie des résultats obtenus est une confirmation indépendante de résultats déjà obtenus ou suggérés par d'autres équipes :
– clade N « restreint » (MABUCHI *et al.*, 2007) ;
– clade (Epigonidae, (Percichthyidae, Lateolabracidae)) (SMITH et CRAIG, 2007) ;
– clade η (CHEN *et al.*, 2007; MABUCHI *et al.*, 2007) ;
– clade ζ (Carangoidei) (JOHNSON, 1993) ;
– présence d'*Indostomus* dans le clade F (MIYA *et al.*, 2003, 2005; KAWAHARA *et al.*, 2008) ;
– clade (Pomacentridae, Mugilidae) (CHEN *et al.*, 2007) ;
– clade (Ophidiiformes, β) (MIYA *et al.*, 2005).
Pour la plupart, les nouveaux clades ci-dessus sont également dûs à l'ajout de nouveaux taxons (nombreux nouveaux taxons dans le clade η, *Epigonus*, *Howella*, *Gnathanodon*, *Selene*, *Indostomus*, *Dascyllus*, *Lamprogrammus*, *Cataetyx*).
L'ajout de nouveaux taxons a également tout simplement permis de placer ces taxons. Ainsi, les Plesiopidae forment probablement un clade avec les Pomacentridae et les Mugilidae, les Bramidae appartiennent au clade H, les Cepolidae sont probablement proches des Tetraodontiformes et des Lophiiformes.
Un problème lié à l'augmentation de l'échantillonnage taxinomique s'est tout de même fait sentir : la difficulté qu'il y a à placer dans l'arbre de synthèse un taxon dont une seule séquence est disponible (*Pelates*, par exemple). On peut aussi voir ce problème comme un problème lié à la méthode de synthèse.

De l'intérêt de savoir quels clades sont fiables

Certaines des publications mentionnées ci-dessus ne sont pas consacrées spécifiquement à la phylogénie des acanthomorphes en général, mais sont centrées autour d'un thème plus ciblé annoncé dans le titre de la publication : les Batrachoidiformes pour MIYA *et al.* (2005),

les Serranidae pour SMITH et CRAIG (2007), les Gerreidae pour CHEN *et al.* (2007), les Labroidei pour MABUCHI *et al.* (2007), les Gasterosteiformes pour KAWAHARA *et al.* (2008). Les échantillonnages taxinomiques de ces publications se prêtent pourtant relativement bien à une étude générale de la phylogénie des acanthomorphes. Une des explications possible à ce phénomène pourrait être le manque d'outils permettant de distinguer les clades fiables des autres. Ainsi, les auteurs seraient contraints de laisser un peu de côté de nombreux clades obtenus, ne sachant pas trop qu'en penser. Centrer une publication autour d'un thème précis évite de devoir attacher trop d'importance à des clades hors-thème. Cette explication n'est sans doute pas la principale : disposer d'un thème précis permet souvent d'attirer l'attention sur une question biologique qui augmente l'intérêt de la publication. Se contenter du *pattern* sans évoquer de question biologique risque de rendre la publication austère pour de nombreux lecteurs. C'est d'ailleurs peut-être le cas pour cette thèse. Il serait intéressant de revenir sur des aspects biologiques comme l'acquisition de la morphologie des poissons plats au sein du clade L, ou comme l'étude de l'évolution des modes de respiration au sein du clade F (ce clade contient un certain nombre de poissons prélevant de l'air en surface).

De l'intérêt du nouveau marqueur

L'apport du nouveau marqueur RNF213 n'est pas très évident ; il aurait fallu refaire l'analyse de fiabilité sans prendre en compte RNF213 pour le tester dans les règles de l'art. Cependant, si le jeu de données se comporte convenablement à l'échelle étudiée, son effet sur l'indice de répétition est utile. La figure 8.1 montre que le nouveau marqueur RNF213 permet l'obtention d'un nombre de clades fiables loin d'être ridicule ; on reconnaît par exemple le clade X, le clade M, le clade Gu, le clade η, le clade L, le clade Q, le clade H et le clade A.
Introduire de nouveaux marqueurs est donc potentiellement une bonne chose, mais il n'est pas sûr que n'importe quel marqueur ait un effet significatif et positif sur les résultats. En outre, plus le nombre de marqueurs augmente, moins il est facile de boucher les trous dans l'échantillonnage. Un autre problème qui se pose quand le nombre de gènes est élevé est la capacité limitée du matériel informatique. Quand le séquençage est difficile et que les ordinateurs sont lents, il est sage de bien choisir les maqueurs utilisés.

Conclusions techniques

Une grande part du travail du phylogénéticien consiste en des tâches répétitives. Et encore, il ne cherche plus les arbres à la main ! L'outil informatique joue donc potentiellement un rôle essentiel dans ce travail. Rechercher les données déjà disponibles, garder une trace des amplifications à faire, des séquences à obtenir, nettoyer des séquences et les aligner, regrouper des matrices et les analyser, comparer les résultats, écrire un article. Toutes ces étapes peuvent être au moins

en partie assistées par ordinateur, et assez souvent pourraient l'être plus qu'elles ne le sont. La raison de cet usage encore limité de l'informatique tient au fait que l'utilisation d'un ordinateur est en fait plus difficile qu'on voudrait le croire. Mettre réellement l'ordinateur à son service nécessite un investissement en temps qui n'est pas fait assez souvent. Une fois cet investissement fait, on réalise plus concrètement les possibilités qu'offre l'informatique[1].

L'aspect informatique le plus marquant utilisé lors de cette thèse est la programmation. Apprendre le langage `Python` lors d'un module d'école doctorale m'a permis d'être rapidement en mesure d'automatiser le calcul des indices de fiabilité[2], ce qui est un point crucial sans lequel mon travail méthodologique aurait eu peu d'utilité pratique. Il faut cependant préciser que le programme `rely.py` n'est pas très ergonomique et probablement pas programmé dans le meilleur des styles. Il doit pouvoir être utilisé par un phylogénéticien motivé et intéressé par l'informatique, mais du travail reste à faire pour le rendre facilement utilisable par la plupart des personnes potentiellement intéressées par la méthode.

[1] On peut par exemple imaginer une suite de programmes qui assistent le chercheur dans son travail de la sélection des données à utiliser et dans la soumission des séquences jusqu'à la préparation de la publication en passant par l'analyse des données. Notamment, préparer le tableau des données utilisées *devrait* être automatisé et couplé à la soumission des matrices de données dans `TreeBase`, si l'éditeur exige une telle soumission.

[2] Savoir un peu programmer m'a également permis d'automatiser une méthode de *supertree* dont la publication est soumise (ROPIQUET *et al.*, soumis). Programmer est en outre une activité ludique et intellectuellement stimulante, du moins pour le peu que j'ai expérimenté.

Bibliographie

ADAMS E. (1986). N-trees as nestings : Complexity, similarity, and consensus. *Journal of Classification*, **3** : 299–317.

AHO A., SAGIV Y., SZYMANSKI T. et ULLMAN J. (1981). Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. *SIAM Journal on Computing*, **10**(3) : 405–421. doi :10.1137/0210030. URL `http://dx.doi.org/10.1137/0210030`.

BAUM B. et RAGAN M. (2004). The MRP method. Dans *Phylogenetic supertrees : combining information to reveal the tree of life* (O. BININDA-EDMONDS, réd.), p. 17–34. Kluver Academic, Dordrecht, the Netherlands.

BERRY V. et GASCUEL O. (2000). Inferring evolutionary trees with strong combinatorial evidence. *Theoretical Computer Science*, **240** : 271–298.

BREDER C.M. et ROSEN D.E. (1966). *Modes of reproduction in fishes*. The Natural History Press, New York.

BREMER K. (1988). The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. *Evolution*, **42**(4) : 795–803. doi :10.2307/2408870. URL `http://dx.doi.org/10.2307/2408870`.

BRINKMANN H., VAN DER GIEZEN M., ZHOU Y., PONCELIN DE RAUCOURT G. et PHILIPPE H. (2005). An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. *Systematic Biology*, **54**(5) : 743–757.

BRITZ R. et JOHNSON D. (2002). "paradox lost" : skeletal ontogeny of *Indostomus paradoxus*, and its significance for the phylogenetic relationships of Indostomidae (Teleostei, Gasterosteiformes). *American Museum Novitates*, (3383) : 1–43. URL `http://hdl.handle.net/2246/2872`.

BULL J., HUELSENBECK J., CUNNINGHAM C., SWOFFORD D. et WADDELL P. (1993). Partitioning and combining data in phylogenetic analysis. *Systematic Biology*, **42**(3) : 384–397.

CARNAP R. (1950). *Logical Foundations of Probability*. University of Chicago Press, Chicago.

CARNAP R. (1995). *An introduction to the philosophy of science*. Dover.

CHAPLEAU F. (1993). Pleuronectiform relationships : a cladistic reassessment. *Bulletin of Marine Science*, **52**(1) : 516–540.

CHEN W.J. (2001). *La répétitivité des clades comme critère de fiabilité : application à la phylogénie des Acanthomorpha (Teleostei) et des Notothenioidei (acanthomorphes antarctiques)*. Thèse de doctorat, Université Paris VI Pierre et Marie Curie.

CHEN W.J., BONILLO C. et LECOINTRE G. (2003). Repeatability of clades as a criterion of reliability : a case study for molecular phylogeny of Acanthomorpha (Teleostei) with larger number of taxa. *Molecular Phylogenetics and Evolution*, **26** : 262–288.

CHEN W.J., RUIZ-CARUS R. et ORTÍ G. (2007). Relationships among four genera of mojarras (Teleostei : Perciformes : Gerreidae) from the western Atlantic and their tentative placement among percomorph fishes. *Journal of Fish Biology*, **70** : 202–218. doi :10.1111/j.1095-8649.2007.01395.x. URL http://dx.doi.org/10.1111/j.1095-8649.2007.01395.x.

COOPER A. et PENNY D. (1997). Mass survival of birds across the cretaceous-tertiary boundary : molecular evidence. *Science*, **275**(5303) : 1109–1113. doi :10.1126/science.275.5303.1109. URL http://dx.doi.org/10.1126/science.275.5303.1109.

DEBRY R.W. et SAGEL R.M. (2001). Phylogeny of Rodentia (Mammalia) inferred from the nuclear gene IRBP. *Molecular Phylogenetics and Evolution*, **19**(2) : 290–301. doi : 10.1006/mpev.2001.0945. URL http://dx.doi.org/10.1006/mpev.2001.0945.

DETTAÏ A. (2004). *La phylogénie des Acanthomorpha (Teleostei) inférée par l'étude de la congruence taxinomique*. Thèse de doctorat, Université Paris VI Pierre et Marie Curie.

DETTAÏ A. et LECOINTRE G. (2004). In search of nothothenioid (Teleostei) relatives. *Antarctic Science*, **16**(1) : 71–85. doi :10.1017/S0954102004. URL http://dx.doi.org/10.1017/S0954102004.

DETTAÏ A. et LECOINTRE G. (2005). Further support for the clades obtained by multiple molecular phylogenies in the acanthomorph bush. *Comptes Rendus Biologies*, **328** : 674–689.

DETTAÏ A. et LECOINTRE G. (2008). New insights into the organization and evolution of vertebrate IRBP genes and utility of IRBP gene sequences for the phylogenetic study of the Acanthomorpha (Actinopterygii : Teleostei). *Molecular Phylogenetics and Evolution*, **48**(1) : 258–269. doi :10.1016/j.ympev.2008.04.003. URL http://dx.doi.org/10.1016/j.ympev.2008.04.003.

DOIUCHI R. et NAKABO T. (2006). Molecular phylogeny of the stromateoid fishes (Teleostei : Perciformes) inferred from mitochondrial DNA sequences and compared with morphology-based hypotheses. *Molecular Phylogenetics and Evolution*, **39** : 111–123. doi :10.1016/j.ympev.2005.10.007. URL http://dx.doi.org/10.1016/j.ympev.2005.10.007.

DOYLE J. (1992). Gene trees and species trees : molecular systematics as one-character taxonomy. *Systematic Botany*, **17**(1) : 144–163.

FELSENSTEIN J. (1985). Confidence limits on phylogenies : an approach using the bootstrap. *Evolution*, **39**(4) : 783–791. doi :10.2307/2408678. URL http://dx.doi.org/10.2307/2408678.

FROESE R. et PAULY D. (2006). Fishbase. World Wide Web electronic publication. www.fishbase.org, version (06/2006). URL www.fishbase.org.

FRY B., VIDAL N., NORMAN J., VONK F., SCHREIB H., RAMJAN R., KURUPPU S., FUNG K., HEDGES B., RICHARDSON M., HODGSON W., IGNJATOVIC V., SUMMERHAYES R. et KOCHVA E. (2005). Early evolution of the venom system in lizards and snakes. *Nature*, **439** : 584–588.

GRAYBEAL A. (1994). Evaluating phylogenetic utility of genes : a search for genes informative about deep divergences among vertebrates. *Systematic Biology*, **43**(2) : 174–193. doi : 10.2307/2413460. URL http://dx.doi.org/10.2307/2413460.

GUINDON S. et GASCUEL O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, **52**(5) : 696–704. doi :10.1080/10635150390235520. URL http://dx.doi.org/10.1080/10635150390235520.

HARRISON G., MCLENACHAN P., PHILLIPS M., SLACK K., COOPER A. et PENNY D. (2004). Four new avian mitochondrial genomes help get to basic evolutionary questions in the late Cretaceous. *Molecular Biology and Evolution*, **21**(6) : 974–983. doi :10.1093/molbev/msh065. URL http://dx.doi.org/10.1093/molbev/msh065.

HILLIS D. et BULL J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology*, **42**(2) : 182–192.

HOLCROFT Nancy I. (2004). A molecular test of alternative hypotheses of tetraodontiform (Acanthomorpha : Tetraodontiformes) sister group relationships using data from the RAG1 gene. *Molecular Phylogenetics and Evolution*, **32** : 749–760. doi :10.1016/j.ympev.2004.04.002. URL http://dx.doi.org/10.1016/j.ympev.2004.04.002.

HOLCROFT Nancy I. (2005). A molecular analysis of the interrelationships of tetraodontiform fishes (Acanthomorpha : Tetraodontiformes). *Molecular Phylogenetics and Evolution*, **34** :

525–544. doi :10.1016/j.ympev.2004.11.003. URL http://dx.doi.org/10.1016/j.ympev.2004.11.003.

HOLCROFT N.I. et WILEY E.O. (2008). Acanthuroid relationships revisited : a new nuclear gene-based analysis that incorporates tetraodontiform representatives. *Ichthyological Research*, **Online** : 1–10. doi :10.1007/s10228-007-0026-x. URL http://dx.doi.org/10.1007/s10228-007-0026-x.

IMAMURA H., SHIRAI S. et YABE M. (2005). Phylogenetic position of the family Trichodontidae (Teleostei : Perciformes), with a revised classification of the perciform suborder Cottoidei. *Ichthyological Research*, **52** : 264–274. doi :10.1007/s10228-005-0282-6. URL http://dx.doi.org/10.1007/s10228-005-0282-6.

IMAMURA H. et YABE M. (2002). Demise of the Scorpaeniformes (Actinopterygii : Percomorpha) : an alternative phylogenetic hypothesis. *Bulletin of Fisheries Sciences, Hokkaido University*, **53**(3) : 107–128.

INOUE J., MIYA M., TSUKAMOTO K. et NISHIDA M. (2003). Basal actinopterygian relationships : a mitogenomic perspective on the phylogeny of the "ancient fish". *Molecular Phylogenetics and Evolution*, **26**(1) : 110–120. URL http://dx.doi.org/10.1016/S1055-7903(02)00331-7.

JOHNSON D. (1993). Percomorph phylogeny : progress and problems. *Bulletin of Marine Science*, **52**(1) : 3–28.

JOHNSON D. et PATTERSON C. (1993). Percomorph phylogeny : a survey of acanthomorphs and a new proposal. *Bulletin of Marine Science*, **52**(1) : 554–626.

KAWAHARA R., MIYA M., MABUCHI K., LAVOUÉ S., INOUE J., SATOH T., KAWAGUCHI A. et NISHIDA M. (2008). Interrelationships of the 11 gasterosteiform families (sticklebacks, pipefishes, and their relatives) : A new perspective based on whole mitogenomes sequences from 75 higher teleosts. *Molecular Phylogenetics and Evolution*, **46** : 224–236. doi :10.1016/j.ympev.2007.07.009. URL http://dx.doi.org/10.1016/j.ympev.2007.07.009.

KLUGE A. (1989). A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae, Serpentes). *Systematic Zoology*, **38** : 7–25.

LAUDER G. et LIEM K. (1983). The evolution and interrelationships of the actinopterygian fishes. *Bulletin of the Museum of Comparative Zoology*, **150**(3) : 95–197.

LECOINTRE G. et DELEPORTE P. (2005). Total evidence requires exclusion of phylogenetically misleading data. *Zoologica Scripta*, **34**(1) : 101–117.

LI B. et LECOINTRE G. (2005). Towards a reliability index for clades : An application on acanthomorph teleosts. Dans *85th Annual Meeting ASIH, Tampa, Florida, July 6-11, 2005*. American Society of Ichthyologists and Herpetologists.

LI B. et LECOINTRE G. (in press). Formalizing reliability in the taxonomic congruence approach. Article accepté par *Zoologica Scripta*. doi :10.1111/j.1463-6409.2008.00361.x. URL `http://dx.doi.org/10.1111/j.1463-6409.2008.00361.x`.

MABUCHI K., MIYA M., AZUMA Y. et NISHIDA M. (2007). Independent evolution of the specialized pharyngeal jaw apparatus in cichlid and labrid fishes. *BMC Evolutionary Biology*, **7**(10) : 1–12. doi :10.1186/1471-2148-7-10. URL `http://dx.doi.org/10.1186/1471-2148-7-10`.

MADDISON W. (1997). Gene trees in species trees. *Systematic Biology*, **46**(3) : 523–536.

MAYR G. (2007). Avian higher-level phylogeny : well-supported clades and what we can learn from a phylogenetic analysis of 2954 morphological characters. *Journal of Zoological Systematics and Evolutionary Research*, **46**(1) : 63–72. URL `http://dx.doi.org/10.1111/j.1439-0469.2007.00433.x`.

MICKEVICH M. (1978). Taxonomic congruence. *Systematic Zoology*, **27** : 143–158.

MIYA M., KAWAGUCHI A. et NISHIDA M. (2001). Mitogenomic exploration of higher teleostean phylogenies : A case study for moderate-scale evolutionary genomics with 38 newly determined complete mitochondrial DNA sequences. *Molecular Biology and Evolution*, **18**(11) : 1993–2009.

MIYA M., SATOH T. et NISHIDA M. (2005). The phylogenetic position of toadfishes (order Batrachoidiformes) in the higher ray-finned fish as inferred from partitioned Bayesian analysis of 102 whole mitochondrial genome sequences. *Biological Journal of the Linnean Society*, **85** : 289–306.

MIYA M., TAKESHIMA H., ENDO H., ISHIGURO N., INOUE J., MUKAI T., SATOH T., YAMAGUCHI M., KAWAGUCHI A., MABUCHI K., SHIRAI S. et NISHIDA M. (2003). Major patterns of higher teleostean phylogenies : a new perspective based on 100 complete mitochondrial DNA sequences. *Molecular Phylogenetics and Evolution*, **26** : 121–138.

MIYAMOTO M. et FITCH W. (1995). Testing species phylogenies and phylogenetic methods with congruence. *Systematic Biology*, **44**(1) : 64–75.

MOOI R. et JOHNSON D. (1997). Dismantling the Trachinoidei : evidence of a scorpaenoid relationship for the Champsodontidae. *Ichthyological Research*, **44**(2) : 143–176.

MOOI R.D. (1990). Egg surface morphology of pseudochromoids (Perciformes : Percoidei), with comments on its phylogenetic implications. *Copeia*, (2) : 455–475.

MULLIS K.B. et FALOONA F.A. (1987). Specific synthesis of DNA in vitro via polymerase-catalyzed chain reaction. *Methods in Enzymology*, **155** : 335–350.

NEAR T.J. et CHENG C.H.C. (2008). Phylogenetics of notothenioid fishes (Teleostei : Acanthomorpha) : Inferences from mitochondrial and nuclear gene sequences. *Molecular Phylogenetics and Evolution*, **47**(2) : 832–840. doi :10.1016/j.ympev.2007.11.027. URL http://dx.doi.org/10.1016/j.ympev.2007.11.027.

NEAR T.J., PESAVENTO J.J. et CHENG C.H.C. (2004). Phylogenetic investigations of Antarctic notothenioid fishes (Perciformes : Notothenioidei) using complete gene sequences of the mitochondrial encoded 16S rRNA. *Molecular Phylogenetics and Evolution*, **32**(3) : 881–891. doi :10.1016/j.ympev.2004.01.002. URL http://dx.doi.org/10.1016/j.ympev.2004.01.002.

NELSON J. (2006). *Fishes of the World*. John Wiley and Sons, Inc., Hoboken, New Jersey, 4e éd.

NG M.P. et WORMALD N. (1996). Reconstruction of rooted trees from subtrees. *Discrete Applied Mathematics*, **69** : 19–31.

ORRELL T., COLLETTE B. et JOHNSON D. (2006). Molecular data support separate scombroid and xiphioid clades. *Bulletin of Marine Science*, **79**(3) : 505–519.

OTERO O. (2004). Anatomy, systematics and phylogeny of both recent and fossil latid fishes (Teleostei, Perciformes, Latidae). *Zoological Journal of the Linnean Society*, **141** : 81–133.

PAGE R. (2002). Modified mincut supertrees. Dans *Algorithms in Bioinformatics : Second International Workshop, Wabi 2002, Rome, Italy, September 17-21, 2002* (R. GUIGÓ et D. GUSFIELD, réds.), p. 537–551. Springer-Verlag Telos.

PATTERSON C. (1993). An overview of the early fossil record of acanthomorphs. *Bulletin of Marine Science*, **52**(1) : 29–59.

PHILIPPE H. et DOUZERY E. (1994). The pitfalls of molecular phylogeny based on four species, as illustrated by the Cetacea/Artiodactyla relationships. *Journal of Mammalian Evolution*, **2**(2) : 133–152.

PROKOFIEV A.M. (2006). A new genus of cardinalfishes (Perciformes : Apogonidae) from the south China sea, with a discussion of the relationships between the families Apogonidae and Kurtidae. *Journal of Ichthyology*, **46**(4) : 279–291. doi :10.1134/S0032945206040011. URL http://dx.doi.org/10.1134/S0032945206040011.

RAMBAUT A. (2002). *Se-Al, Sequence Alignment Editor, version 2.0a11*. Department of Zoology, University of Oxford, South Parks Road, Oxford, OX1 3PS, UK.

ROPIQUET A., LI B. et HASSANIN A. (soumis). Supertri : a new approach based on branch support analyses of multiple independent molecular data sets for identifying robust phylogenetic hypotheses. Application to the family Bovidae. Article soumis à *Molecular Pylogenetics and Evolution*.

ROSEN D.E. (1973). Interrelationships of higher euteleostean fishes. Dans *Interrelationships of fishes* (P. GREENWOOD, R. MILES et C. PATTERSON, réds.), supplement number 1 to the *Zoological Journal of the Linnean Society* **53**, p. 397–513. Academic Press.

ROSEN D.E. et PARENTI L.R. (1981). Relationships of *Oryzias*, and the groups of atherinomorph fishes. *American Museum Novitates*, (2719) : 1–25. URL http://hdl.handle.net/2246/5335.

SAINT K.M., AUSTIN C.C., DONNELLAN S.C. et HUTCHINSON M.N. (1998). C-mos, a nuclear marker useful for squamate phylogenetic analysis. *Molecular Phylogenetics and Evolution*, **10**(2) : 259–263. doi :10.1006/mpev.1998.0515. URL http://dx.doi.org/10.1006/mpev.1998.0515.

SANGER F., NICKLEN S. et COULSON A.R. (1977). DNA sequencing with chain terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, **74**(12) : 5463–5467.

SEMPLE C. et STEEL M. (2000). A supertree method for rooted trees. *Discrete Applied Mathematics*, **105** : 147–158. URL http://dx.doi.org/10.1016/S0166-218X(00)00202-X.

SMITH L. et CRAIG M. (2007). Casting the percomorph net widely : the importance of broad taxonomic sampling in the search for the placement of serranid and percid fishes. *Copeia*, (1) : 35–55.

SMITH L. et WHEELER W. (2004). Polyphyly of the mail-cheeked fishes (Teleostei : Scorpaeniformes) : evidence from mitochondrial and nuclear sequence data. *Molecular Phylogenetics and Evolution*, **32** : 627–646.

SMITH L. et WHEELER W. (2006). Venom evolution widespread in fishes : a phylogenetic road map for the bioprospecting of piscine venoms. *Journal of Heredity*, **97**(3) : 206–217. URL http://dx.doi.org/10.1093/jhered/esj034.

SPRINGER M., STANHOPE M., MADSEN O. et DE JONG W. (2004). Molecules consolidate the placental tree. *Trends in Ecology and Evolution*, **19**(8) : 430–438. URL http://dx.doi.org/10.1016/j.tree.2004.05.006.

STIASSNY M. (1986). The limits and relationships of the acanthomorph teleosts. *Journal of Zoology. Series B*, **1**(2) : 411–460.

STIASSNY M. et MOORE J. (1992). A review of the pelvic girdle of acanthomorph fishes, with comments on hypotheses of acanthomorph intrarelationships. *Zoological Journal of the Linnean Society*, **104** : 209–242.

STREELMAN J.T. et KARL S.A. (1997). Reconstructing labroid evolution with single-copy nuclear DNA. *Proceedings of the Royal Society of London Series B-Biological Sciences*, **264** : 1011–1020.

TYLER J.C., JOHNSON D.G., NAKAMURA I. et COLLETTE B.B. (1989). Morphology of *Luvarus imperialis* (Luvaridae), with a phylogenetic analysis of the Acanthuroidei (Pisces). *Smithsonian Contributions to Zoology*, **485** : 1–78. URL `http://www.sil.si.edu/smithsoniancontributions/Zoology/`.

WIENS J. (1998). The accuracy of methods for coding and sampling higher-level taxa for phylogenetic analysis : a simulation study. *Systematic Biology*, **47**(3) : 397–413.

WIENS J. et REEDER T. (1995). Combining data sets with different numbers of taxa for phylogenetic analysis. *Systematic Biology*, **44**(4) : 548–558. URL `http://dx.doi.org/10.2307/2413660`.

WILEY E.O., JOHNSON G.D. et DIMMICK W.W. (2000). The interrelationships of acanthomorph fishes : a total evidence approach using molecular and morphological data. *Biochemical Systematics and Ecology*, **28** : 319–350.

WILKINSON M. (1994). Common cladistic information and its consensus representation : reduced Adams and reduced cladistic consensus trees and profiles. *Systematic Biology*, **43**(3) : 343–368.

WILKINSON M., PISANI D., COTTON J. et CORFE I. (2005). Measuring support and finding unsupported relationships in supertrees. *Systematic Biology*, **54**(5) : 823–831.

YAMANOUE Y., MIYA M., MATSUURA K., YAGISHITA N., MABUCHI K., SAKAI H., KATOH M. et NISHIDA M. (2007). Phylogenetic position of tetraodontiform fishes within the higher teleosts : Bayesian inferences based on 44 whole mitochondrial genome sequences. *Molecular Phylogenetics and Evolution*, **45** : 89–101. doi :10.1016/j.ympev.2007.03.008. URL `http://dx.doi.org/10.1016/j.ympev.2007.03.008`.

Annexe

Représentation informatique des relations phylogénétiques

Soit n le nombre de taxons. Un clade peut être représenté comme un vecteur de n bits ; 1 si le taxon est dans le clade, 0 sinon. Une représentation assez commode à manipuler est de coder les taxons comme des puissances de 2 : Le premier taxon vaut 1, le deuxième vaut 2, le troisième vaut 4, ... le i-ième vaut 2^{i-1}. Les clades sont codés comme étant la somme des valeurs des taxons constituant leur partie interne. Exemple : Si l'on considère les taxons a à e, dans l'ordre alphabétique, $(abd(ce))$ vaut $(0, 0, 1, 0, 1)$ dans la première représentation et 20 dans la seconde ($2^2 + 2^4$).

Ces représentations supposent implicitement un domaine de validité bien défini et immuable. Pour pouvoir varier le domaine de validité, il faut dans le premier type de représentation un vecteur d'éléments à 3 états : dans le clade, hors du clade ou hors du domaine de validité, ce qui n'est pas très adapté au monde binaire de l'informatique habituelle. Avec le deuxième type de représentation, il faut associer à la valeur de la partie interne du clade une autre valeur ; soit celle du domaine de validité, soit celle de la partie externe. Les relations partielles comme les triplets s'accordent bien avec la représentation de type « (partie interne, partie externe) ». Avec ce type de représentation, la place prise en mémoire par une relation phylogénétique est donc au minimum de $2n$ bits.

Résumé

Si le but de la reconstruction phylogénétique est d'avoir une idée des relations de parenté réelles entre les êtres vivants, il est bon de ne pas se contenter d'un simple arbre obtenu par l'analyse combinée d'un ensemble de données. En effet, même des clades robustes apparaissant dans un tel arbre peuvent ne pas être fiables. La confiance dans une affirmation phylogénétique ne peut émerger qu'après une comparaison de résultats obtenus par des données indépendantes.
Dans un premier temps, la présente thèse propose de mesurer la fiabilité d'un clade à partir d'un indice de répétition prenant en compte le nombre d'occurrences obtenues pour ce clade sur un ensemble d'analyses de données indépendantes, c'est-à-dire peu susceptibles de donner lieu aux mêmes biais de reconstruction. Plus un clade est obtenu un nombre élevé de fois de cette façon, plus il peut être considéré comme fiable. Il est également tenu compte de la présence ou non de clades eux-mêmes répétés et incompatibles avec le clade d'intérêt. Plus un clade est contredit par un clade possédant un grand nombre d'occurrences, moins il doit être considéré comme fiable.
Dans une deuxième partie, l'indice de répétition est calculé à partir d'une série d'analyses mettant en jeu environ 200 taxons et basées sur quatre marqueurs nucléaires : Rhodopsine, MLL4, IRBP et RNF213 (ce dernier étant utilisé ici pour la première fois). Ces marqueurs sont analysés suivant des méthodes probabilistes, séparément et en combinaisons de 2, 3 ou 4, ce qui permet de bénéficier des avantages de l'analyse combinée tout en ayant des séries de résultats indépendants à comparer.
Les résultats de l'analyse de fiabilité sont ensuite synthétisés sous forme d'arbres incluant en priorité les clades les plus fiables, suivant des méthodes gérant de plusieurs façons les différences d'échantillonnages taxinomiques entre les jeux de données.
Les arbres de synthèse obtenus permettent de préciser la structure de la phylogénie des téléostéens acanthomorphes (Actinopterygii : Teleostei). De nouveaux clades fiables sont identifiés à plusieurs niveaux de résolution, et de nouveaux taxons sont placés dans la phylogénie des téléostéens acanthomorphes.

Abstract

The goal of a phylogeny is usually to get an idea of the real relationships between living organisms. It is thus not a good idea to be satisfied with the tree obtained by a simple combined analysis of the data. Indeed, even robust clades in such a tree may not be reliable. Confidence about a phylogenetic statement can only stem from a comparison between results obtained from independent data.
This thesis begins by proposing a repetition index to measure reliability. This index takes into account the number of occurrences for a clade over a set of trees obtained from independent data, that is, data unlikely to be subject to the same reconstruction biaises. The more a clade occurs this way, the more it can be considered reliable. The index also takes into account clades that are not compatible with the clade under focus and that are also repeated. The highest the number of occurrences of a clade's contradictor, the less reliable it should be considered.
In the second part of the thesis, the repetition index is calculated from a series of analyses involving about two hundred taxa. The analyses are primarily based on four nuclear markers: Rhodopsin, MLL4, IRBP and RNF213 (the latter being used for the first time). These markers are analysed under probabilistic models, separately and in combinations of 2, 3 or 4. This allows to take advantage of combined analysis while still being able to compare sets of independent results.
The results are then summarized into trees including mostly reliable clades. Several ways of dealing with the differences between the taxonomic samplings of the datasets are used.
The synthesis trees allow to refine the structure of the acantomorph (Actinopterygii: Teleostei) phylogeny. New reliable clades are identified, at several resolution levels, and new taxa are placed into the phylogeny of acanthomorph teleosts.