Codage et traitement phylogénétique des caractères structuraux de génomes entiers

Cyril Gallut

UNIVERSITE PIERRE & MARIE CURIE
LA SCIENCE A PARIS

École doctorale Diversité du vivant **UFR des Sciences de la Vie**

Codage et traitement phylogénétique de caractères structuraux de génomes entiers

THÈSE

présentée et soutenue publiquement le 18 décembre 2001

pour l'obtention du

Doctorat de l'Université Pierre et Marie Curie – Paris 6

(spécialité Sciences de la Vie)

par

Cyril GALLUT

Composition du jury

Président :	Dominique HIGUET
Rapporteurs :	Douglas EERNISSE Manolo GOUY
Examinateurs :	Véronique BARRIEL Pascal TASSY Régine VIGNES-LEBBE
Directeur de thèse :	Gabriel GACHELIN

LABORATOIRE INFORMATIQUE ET SYSTÉMATIQUE

À Jacques,

Remerciements

Je remercie très chaleureusement Gabriel GACHELIN, Véronique BARRIEL et Régine VIGNES-LEBBE pour m'avoir soutenu et guidé dans ce travail.

Je remercie sincèrement Manolo GOUY et Douglas EERNISSE d'avoir accepté d'être mes rapporteurs et Dominique HIGUET ainsi que Pascal TASSY pour avoir accepté de faire partie de mon jury.

Je suis très reconnaissant à Simon TILLIER de m'avoir accueilli au Service de Systématique Moléculaire et à Régine VIGNES-LEBBE de m'avoir fait une place au Laboratoire Informatique et Systématique.

Un grand merci à Gauthier DOBIGNY pour ses conseils et ses photos ainsi qu'à Vitaly VOLOBOUEV et Vladimir ANISKIN pour m'avoir donné accès aux données sur les chromosomes du genre *Mastomys*. Merci aussi à Émile LECOMPTE qui m'a informé de ses résultats de phylogénie moléculaire.

Pour les nombreuses discussions passionnantes que nous avons eu ensembles, je remercie Alexandre HASSANIN, Cyrille D'HAESE et Monique MASSELOT.

Merci à Christiane DELARBRE et Annie TILLIER pour les trésors de patience dont elles ont su faire preuve alors que je m'exerçais à la paillasse.

Je remercie vivement tous ceux qui m'ont aidé quotidiennement par leur amitié et leurs conseils, dans le désordre : Jean BROUTIN, Bernard LEPEN, Catherine REEB, Maihdi EL FASSI FHIRI, Mélanie PACES-FESSY, Maité LUSQUINHOS, Robert BOSSY, Guillaume ROUSSE, Jérome BARBEAU, Guillaume SAUVENAY et Laurent GUEGUEN.

Je tiens enfin à exprimer toute ma gratitude à Jacques LEBBE, Régine VIGNES-LEBBE, Véronique BARRIEL et Pascal TASSY qui, par leur enthousiasme débordant et leur passion pour la systématique m'ont transmis le virus de la recherche.

Table des matières

Table des figures

Liste des tableaux

Introduction

Les grandes avancées techniques qu'a connues la biologie moléculaire ces dernières années ont fourni à la communauté scientifique et aux biologistes en particulier une énorme quantité de données. La bioinformatique est née de la nécessité de stocker efficacement, de gérer, de traiter et de mettre à disposition ces données et les informations qu'il est possible d'en tirer. Elle intègre des compétences provenant de différentes disciplines, qui vont de la biologie aux mathématiques en passant par les statistiques et l'informatique. La plupart des travaux de bioinformatique portent sur l'analyse de séquences pour la recherche de gènes, de motifs, l'analyse de familles multigéniques ou la prédiction de structure de protéines. Cependant, ses champs d'investigation ne se limitent pas à ce domaine. La bioinformatique s'intéresse également à l'analyse d'images pour l'imagerie médicale ou à la systématique (voir Lebbe (1996) pour l'intérêt de l'informatique en systématique) ou bien encore à la génomique au sens large.

Pour étudier le génome, la génomique embrasse différents domaines qui vont de l'analyse de séquences à l'étude des processus d'expression des gènes en passant par la cartographie comparée de génomes. Cette dernière trouve ses racines dans des disciplines, qui vont de la cytogénétique aux mathématiques combinatoires. Elle cherche à répondre à plusieurs questions d'intérêt biologique majeur :

- quelle est l'origine des réarrangements chromosomiques (origine biologique et évolutive) ?
- quelles informations peut-on tirer de la connaissance d'un génome donné pour l'étude de génomes apparentés ?

– comment utiliser les réarrangements observés entre génomes à des fins phylogénétiques ?

Les premières cartographies comparées remontent aux cartes génétiques obtenues à partir d'études de recombinaison de marqueurs chez les drosophiles. Les premières cartes physiques ayant nécessité une technologie plus lourde, il fallut attendre la découverte de l'hybridation cellulaire dans les années cinquante pour que ce domaine se développe vraiment. Ces études répondent généralement à la nécessité de localiser les facteurs génétiques de maladies héréditaires et de comprendre leur transmission. De façon plus générale, elles ont permis d'envisager les mécanismes responsables des réarrangements observés entre génomes. Ces réarrangements sont à l'origine des travaux empiriques et formels qui ont pour but l'étude de l'évolution du génome. Ces travaux connaissent une expansion très forte avec l'avènement du séquençage systématique de génomes complets, que ce soit au niveau des génomes de virus et d'organites (les plus nombreux de par leur petite taille) ou les génomes procaryotes et nucléaires chez les eucaryotes. Les travaux visant à utiliser la cartographie comparée dans un but phylogénétique sont plus rares. Les premières tentatives d'utilisation de l'ordre des gènes remontent au début des années 90 (Sankoff *et al.*, 1992) et la plupart se fondent sur la détection d'événements évolutifs pour reconstruire l'arbre évolutif.

Le but de notre travail est d'explorer les possibilités de reconstruction phylogénétique à partir de la comparaison de génomes dans le cadre de la méthode cladistique. L'aspect le plus important est de fonder la reconstruction sur l'organisation des génomes sans envisager les événements de remaniements *a priori*.

Nous verrons dans un premier temps quelles sont les techniques employées en cartographie génomique et le type de cartes qui en résultent. Cela peut être la cytogénétique avec l'étude de caryotypes ou la cartographie génétique au sens large ou bien encore le séquençage de génomes entiers. Puis, nous verrons les méthodes développées pour l'étude comparée de cartes génomiques, comme par exemple la recherche de la série minimale d'inversions entre deux chromosomes.

Dans un deuxième chapitre, nous préciserons le cadre de notre travail, tant sur le plan phylogénétique que sur celui de l'organisation du génome et de la manière d'en tirer des caractères.

Nous présenterons, dans le troisième chapitre, différents codages originaux de l'organisation du génome permettant d'inférer une phylogénie par la méthode de parcimonie. La mise en œuvre et les implications de ces codages seront étudiées en détail grâce à des exemples théoriques. Nous envisagerons ensuite l'évolution des réarrangements chromosomiques à partir des résultats obtenus.

Enfin, dans le quatrième chapitre nous appliquerons les codages proposés à deux jeux de données différents : le génome mitochondrial des métazoaires et les chromosomes du genre *Mastomys* (petits rongeurs africains).

1

Contexte actuel de la génomique comparative

1.1 Cartographie du génome

De nombreux types de données sont à notre disposition pour aborder la comparaison de génomes dans leur intégralité. Les données les plus traditionnelles sont les caryotypes et les cartes génétiques et physiques. L'accomplissement d'un travail énorme de séquençage a fourni, ces dernières années, l'accès à la séquence complète de nombreux génomes. Ces différents types de données présentent des niveaux de résolution variables. Les caryotypes ne fournissent pas d'information sur la localisation des gènes sur les chromosomes mais sur l'organisation physique de l'ADN. Les cartes génétiques et physiques permettent d'avoir une information plus précise notamment sur la localisation de gènes ou de marqueurs sur les chromosomes. Néanmoins, l'orientation des gènes localisés n'est pas toujours connue. Enfin grâce aux séquences complètes nous avons accès à une information plus détaillée sur la position des gènes, leur orientation etc.

1.1.1 Carte génétique

Une carte génétique représente la succession de locus le long d'un chromosome donné. Les cartes génétiques sont construites à partir des données obtenues par l'étude des crossing-over. Lors de la méiose, les allèles associés à des locus liés physiquement sur le chromosome ne peuvent pas être répartis indépendamment entre les gamètes. Le crossing-over peut briser ce lien, les allèles sont alors répartis différemment de l'association parentale. Le taux de recombinaison est de 50% pour les locus non liés. Le taux de recombinaison entre deux locus liés est proportionnel à la distance génétique entre ces deux locus. En réalité cette distance, exprimée en centiMorgan (1 cM = 1% de recombinaison), n'est pas complètement proportionnelle à la distance physique entre les deux locus car il existe des point chauds où les crossing-over sont plus fréquents, ce qui augmente artificiellement la distance génétique entre les locus. Cette distance est aussi appelée *distance de liaison*. Le terme de *carte de liaison* est d'ailleurs mieux approprié que celui de *carte génétique*, qui prend à l'heure actuelle une acception plus large. De nombreux types de marqueurs sont utilisés pour construire une carte de liaison, des gènes, des sites de polymorphisme de longueur de fragment de restriction (RFLP) ou des microsatellites. Sur ce type de cartes, l'orientation des gènes est bien entendue inconnue.

1.1.2 Carte physique

La cartographie physique est un complément important de la cartographie de liaison. La réalisation d'une carte physique nécessite la création d'une banque génomique de fragments chromosomiques recouvrant. Les fragments sont obtenus par la digestion de l'ADN d'un organisme par des enzymes de restriction. Les fragments sont ensuite intégrés dans un chromosome de levure ou de bactérie (YAC pour *Yeast Artificial Chromosome* ou BAC pour *Bacterial Artificial Chromosome*). Les fragments sont amplifiés par multiplication des levures ou des bactéries hôtes et la succession des fragments est déterminée par recherche de leurs recouvrements grâce au séquençage de leurs extrémités. C'est un travail long et difficile. Il est possible de caracté-

riser les fragments en utilisant divers marqueurs qui servent de point d'ancrage pour les cartes. L'orientation de ces marqueurs n'est généralement pas connue.

1.1.3 Caryotype

Un caryotype est une présentation ordonnée des chromosomes où ils sont disposés par paires et triés par ordre décroissant de taille, sauf pour les chromosomes sexuels (figure 1.1). Pour obtenir les chromosomes, on cultive des cellules en synchronisant leur cycle puis on les fixe, généralement au moment de la métaphase, là où les chromosomes, constitués de deux chromatides sœurs, sont les plus condensés et enfin, ils sont colorés. Il existe différentes techniques de coloration qui permettent de révéler des zones particulières du chromosome en fonction de leurs caractéristiques. Ces zones sont mises en évidences et forment des motifs de bandes le long du chromosome. Ainsi, chaque type de bande correspond à des zones aux caractéristiques particulières, les bandes G correspondent à des zones pauvres en gènes et riches en GC. Les bandes R sont spécifiques de zones riches en gènes et en AT. Bandes G et bandes R, mettent en évidence des zones aux caractéristiques opposées ; ainsi un caryotype en bande G donne un profil de motifs de bandes inversé par rapport à celui d'un caryotype en bande R. De plus, les bandes R donnent accès à des informations sur la réplication du chromosome. Les bandes C permettent de distinguer des zones d'ADN hautement répété, tel que le centromère par exemple ou des blocs de transposons.

À partir de caryotypes obtenus avec différentes colorations, nous avons accès à des informations sur la composition (en bases, en gènes etc.), sur la structure physique ainsi que sur le fonctionnement des chromosomes. On distingue différents types de chromosome selon la position de la région spécialisée qui relie les deux chromatides sœurs entre elles, le *centromère*. Les chromosomes qui ont un centromère en position médiane ou intermédiaire sont appelés chromosomes *métacentriques* ou *submétacentriques*. Les chromosomes dont le centromère se situe à l'extrémité du chromosome sont appelés *acrocentriques*.

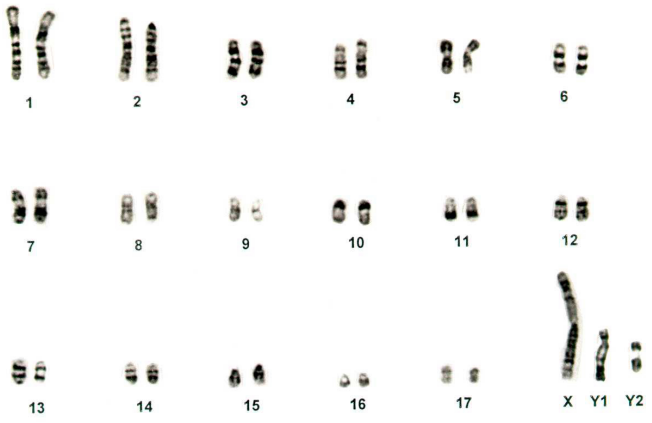


FIGURE 1.1 – Caryotype en bandes G de *Taterillus gracilis*, mâle (2n=37, 17 paires d'autosomes et 3 chromosomes sexuels). Photo Gauthier DOBIGNY.

Les techniques récentes de cytogénétique, telle que le *tri par flux* par exemple, permettent une étude plus précise de la morphologie du chromosome et se sont avérées utiles pour permettre une description standardisée des chromosomes d'espèces au « caryotype compliqué », comme celui du chien par exemple, qui a 78 paires de chromosomes dont beaucoup sont petits et difficilement distinguables. Le tri par flux permet de séparer les chromosomes et de les étudier individuellement ou de les utiliser comme sondes pour la cartographie comparée.

Il est possible d'établir des homologies entre caryotypes d'espèces différentes à partir des profils chromosomiques de motifs de bandes. Les régions présentant les mêmes motifs étant supposés homologues.

1.1.4 Cartographie comparée

Si la comparaison de caryotypes permet d'établir directement des homologies, la réalisation de cartes comparées entre différentes espèces nécessite l'utilisation de techniques permettant la localisation de marqueurs chez une espèce cible, marqueurs dont la localisation est connue chez l'espèce de référence. L'espèce de référence est le plus souvent l'homme mais cela peut être l'une des espèces modèles dont le séquençage complet est en cours ou achevé (*Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster* ou *Saccharomyces cerevisiæ* par

exemple). La cartographie très précise d'espèces de références, entre autre grâce au séquençage complet, permet d'envisager la réalisation de cartes beaucoup plus précises pour les autres espèces. Citons par exemple les techniques appelées *Zoo-FISH* et *FISH* (Hybridation Fluorescente In Situ), parfois appelées *peinture sur chromosome*. Avec la technique de Zoo-FISH on utilise chaque chromosome d'une espèce de référence comme sonde que l'on hybride avec les chromosomes de l'espèce cible. En repérant le(s) chromosome(s) avec le(s)quel s'est hybridé chaque chromosome de l'espèce de référence, il est possible de déterminer les fragments de chromosomes homologues entre les deux espèces. Avec la Zoo-FISH on ne connaît pas l'ordre des gènes sur les différents fragments, ce que permet en revanche la technique FISH. Dans cette dernière on utilise comme sonde des marqueurs localisés sur les chromosomes de l'espèce de référence que l'on hybride avec les chromosomes de l'espèce cible. Elle permet donc de délimiter des fragments homologues entre les deux espèces et de connaître l'ordre des marqueurs utilisés sur ces fragments et ainsi de les orienter. Le nombre de fragments obtenu n'est pas forcément le même si on inverse l'espèce de référence et l'espèce cible. L'hybridation de marqueurs humains chez le chien a fourni 70 segments homologues alors que l'hybridation de marqueurs du chien chez l'homme a fourni 90 segments homologues. Ce résultat contradictoire est dû à la complexité du caryotype du chien qui complique les analyses. De plus, tous les marqueurs d'une espèce ne s'hybrident pas forcément chez l'espèce cible.

La plupart des cartes comparées concernent des espèces d'intérêt médical ou économique ou encore des animaux de compagnie. Par exemple, un effort particulier est réalisé pour la cartographie des génomes de céréales (voir la page [1] du *French Cereal Mapping Network*). Néanmoins, quelques travaux s'intéressent à un éventail phylogénétique plus large, comme les mammifères dans leur ensemble. Nadeau et Sankoff (1997) ont montré la nécessité d'avoir un grand nombre de gènes homologues localisés chez de nombreuses espèces afin d'établir des cartes comparées suffisamment précises pour pouvoir retracer l'évolution des mammifères. Ces gènes servent alors de points d'ancrage au sein des différents génomes afin de pouvoir délimiter

[1] http://grain.jouy.inra.fr/FCMN/fcmn.html

les segments homologues. Lyons *et al.* (1997) ont développé un ensemble de plus de trois cents marqueurs ou *CATS* (Comparative Anchor Tagged Sequences) communs à plusieurs ordres de mammifères. Pour l'instant, les informations comparables sur une large échelle restent limitées mais tout porte à croire que les prochaines années seront riches en données nouvelles. Cela rend d'autant plus important et nécessaire le développement de méthodes appropriées à ce nouveau type de données.

1.1.5 Synténie

Lorsque la localisation des marqueurs au sein des chromosomes demeure colinéaire à l'intérieur de segments homologues, on parle de *synténie* et de *microsynténie* à l'échelle du centi-Morgan. La synténie n'implique pas forcément que l'ordre soit strictement conservé, d'autant plus que l'ordre n'est pas toujours connu. Mesurer le taux de conservation de la synténie au sein de différentes lignées est un problème ardu. En effet, les techniques de comparaison ne sont pas indépendantes et de nombreux marqueurs ne sont pas encore localisés. Ehrlich *et al.* (1997) ont proposé une méthode pour évaluer la conservation de la synténie (connue ou inconnue) et ont montré que 65% de la synténie entre l'homme et la souris est déjà identifiée. L'évaluation précise de la synténie est d'une importance cruciale pour la recherche de gènes candidats, en médecine ou en agronomie par exemple. Bien évidemment, elle présente aussi un intérêt majeur pour l'étude des réarrangements chromosomiques et de l'évolution du génome.

Il existe peu d'exemples de synténie sur une large échelle évolutive et les plus connus sont ceux des gènes *hox* et des gènes de la famille des histones. Dans ces deux cas, les gènes ont une origine commune mais d'autres exemples impliquent des gènes qui n'ont pas de similarité, ni du point de vue de leur séquence, ni du point de vue de leur fonction. Des études récentes (Dandekar *et al.*, 1998 ; Overbeek *et al.*, 1999 ; Huynen *et al.*, 2000 ; Wolf *et al.*, 2001) ont montré que ces gènes synténiques codent pour des protéines qui interagissent physiquement au cours de leur fonctionnement. Cette interaction conduirait à une coévolution des gènes qui maintiendrait la synténie, même entre des espèces très éloignées. Ce résultat permet d'envisager

la prédiction de la fonction de protéines inconnues grâce à leur environnement génomique, mais bien sûr, seulement lorsque leurs gènes sont situés dans un bloc synténique. Chez les bactéries, la synténie à grande échelle phylogénétique concerne environ 10% des gènes (Wolf *et al.*, 2001) et ce sont très souvent des gènes impliqués dans des opérons. Il faut préciser que le nombre de gènes qui ont des homologues identifiés chez d'autres espèces est faible (moins de 50%) et qu'une part importante des gènes n'a pas encore de fonction connue.

1.1.6 Génomes complets

La séquence nucléotidique complète de près d'un millier [2] de génomes est aujourd'hui identifiée. Ces séquences appartiennent majoritairement à des génomes de virus mais il y a aussi de nombreuses séquences de bactéries et archaebactéries ainsi que d'eucaryotes (séquences de génomes nucléaires et d'organites).

Le génome des virus peut être constitué d'une ou plusieurs molécules d'ADN ou d'ARN, simple ou double brin, circulaire ou linéaire. On dénombre 638 génomes complets de virus, la taille de ces génomes variant de 300 pb à 300 kb. Le génome des bactéries et des archaebactéries est constitué d'un chromosome circulaire, et éventuellement de petits chromosomes, qui ne sont pas toujours présents : les plasmides. On dispose de 41 génomes bactériens complets, d'une taille allant de 400 kb à 7 Mb et il y a 10 génomes complets d'archaebactérie dont la taille varie de 1,5 Mb à 3 Mb. Chez les eucaryotes, le génome est constitué de plusieurs chromosomes linéaires, associés en paires de chromosomes « homologues » dans les cellules diploïdes et situés dans le noyau. Lorsque le sexe a un déterminisme chromosomique, le couple de chromosomes sexuels est constitué d'un chromosome X et d'un chromosome Y, ces deux chromosomes sont partiellement homologues. Les sexes sont déterminés soit par la présence du couple XX, soit par la présence XY. Certains organites, comme les plastes et les mitochondries, possèdent un génome indépendant de celui du noyau. Chez les eucaryotes, 21 génomes chloroplastiques com-

[2] Données provenant de GENBANK (http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html), octobre 2001.

plets ont été déterminés (de 35 kb à 200 kb) ainsi que 224 génomes mitochondriaux (de 6 kb à 370 kb). Le génome nucléaire complet est connu pour quatre espèces :

- *Arabidopsis thaliana* 125 Mb, 2n = 10,
- *Caenorhabditis elegans* 97 Mb, 2n = 12,
- *Drosophila melanogaster* 132,7, 2n = 8,
- *Saccharomyces cerevisiæ* 12 Mb, 2n = 32.

Enfin, le séquençage complet du génome nucléaire de plusieurs espèces est en cours : *Danio rerio*, *Homo sapiens*, *Leishmania major*, *Mus musculus*, *Oryza sativa*, *Plasmodium falciparum*, *Zea mays*.

1.2 Comparaison de génomes

L'étude des réarrangements chromosomiques a été initiée par Dobzhansky et Sturtevant (1938) à la fin des années 1930, qui ont identifié 17 inversions au sein des génomes de *Drosophila*. Une longue tradition d'étude des remaniements chromosomiques s'est développée en génétique depuis cette époque. L'évolution chromosomique chez les mammifères a été très étudiée notamment, grâce à la comparaison de caryotypes. Ces études ont montré que le nombre et le type de remaniements chromosomiques varient beaucoup en fonction des taxons étudiés. Chez les mammifères il est fréquent que l'évolution morphologique soit beaucoup moins rapide que l'évolution des réarrangements chromosomiques observés sur les caryotypes, ce qui conduit à des espèces indifférenciées sur le plan morphologique. Dans ce cas, la comparaison de caryotypes s'est révélée très efficace pour distinguer ces espèces entre elles. De la caractérisation d'espèces à l'étude de leurs relations de parentés, il n'y a qu'un pas. La phylogénie chromosomique fonde les groupements d'espèces sur le partage de remaniements chromosomiques particuliers. La plupart du temps, la recherche de ces remaniements se base sur l'expertise du chercheur. Le premier article de bioinformatique sur l'étude des remaniements chromosomiques a été écrit par Watterson *et al.* au début des années 1980 (1982). Par ailleurs, en partant

du très petit nombre de gènes (83) dont la localisation chromosomique était connue, à la fois chez la souris et chez l'homme, Nadeau et Taylor (1984) ont proposé une estimation du nombre de segments chromosomiques conservés depuis la séparation entre les lignées de la souris et de l'homme. Leur estimation à 173 $\pm$39 segments, s'est avérée être remarquablement précise (Sankoff *et al.*, 2000b). L'article de Nadeau et Taylor est une charnière dans l'étude des réarrangements chromosomiques. Depuis, de nombreuses publications présentent des algorithmes très complexes permettant de passer d'un génome à un autre avec une série minimale d'événements de réarrangement.

Avec l'accroissement très important des données de cartographie comparative, de cartes génétiques, physiques et même de séquences complètes, la recherche en génomique comparative a connu, ces dix dernières années, un essor considérable. Le récent ouvrage de Sankoff et Nadeau (2000) permet de se faire une idée très précise de l'état de la recherche dans le domaine.

1.2.1 Opérateurs de réarrangement

Les réarrangements, qui peuvent affecter les chromosomes au cours de l'évolution, se classent en différentes catégories. Chaque catégorie de réarrangement est associée à un opérateur, un opérateur représentant une opération qui modifie tout ou partie du chromosome. Les principaux opérateurs sont :

- l'inversion,
- la transposition,
- la transinversion,
- le gain–perte,
- la translocation,
- la fusion,
- la fission.

Seuls les trois derniers opérateurs représentent des opérations qui s'appliquent à deux chromosomes. Les autres représentent des opérations qui concernent une portion de chromosome.

Nous n'aborderons pas ici les mécanismes biologiques qui sont à l'origine des réarrangements mais simplement le principe de fonctionnement de chaque opérateur. L'opérateur d'inversion inverse l'ordre et l'orientation des éléments d'un fragment de chromosome. La transposition déplace un fragment d'un point A vers un point B. La transinversion est une transposition doublée d'une inversion, c'est-à-dire que le fragment déplacé se réinsère dans le sens opposé à celui qu'il avait au départ. L'opérateur gain–perte implique l'insertion ou la délétion d'un gène ou d'un fragment de chromosome. Nous considérerons le gain et la perte comme un seul et même opérateur bien que ces deux types d'événements impliquent des mécanismes biologiques différents. La translocation transfert un fragment d'un chromosome à un autre et la translocation réciproque échange deux fragments entre deux chromosomes. La fusion réunit deux chromosomes en un seul tandis que la fission scinde un chromosome en deux.

1.2.2 Tri par inversion ou par transposition

En simplifiant, les réarrangements de génomes peuvent être modélisés par un problème de combinatoire qui revient à trouver la plus petite série d'inversions nécessaire pour transformer un génome en un autre (Pevzner, 2000, chap. 10). L'ordre des gènes sur un chromosome est représenté par une permutation : $\pi = \pi_1\pi_2\ldots\pi_n$. Une inversion $\rho(i,j)$ a pour effet d'inverser l'ordre des gènes $\pi_i\ldots\pi_j$ et transforme $\pi = \pi_1\ldots\pi_{i-1}\pi_i\ldots\pi_j\pi_{j+1}\ldots\pi_n$ en $\pi\cdot\rho(i,j) = \pi_1\ldots\pi_{i-1}\pi_j\ldots\pi_i\pi_{j+1}\ldots\pi_n$. La figure 1.2 présente un exemple de transformation d'un chromosome A en un chromosome B par inversions successives (ici cinq événements).

Étant donné deux permutations π et σ (deux chromosomes), le problème est de trouver la série d'inversions $\rho_1\rho_2\ldots\rho_t$ telle que $\pi\cdot\rho_1\cdot\rho_2\ldots\cdot\rho_t = \sigma$ et que t soit minimale, t étant la *distance d'inversion* entre π et σ. Cela revient à *trier* la permutation π pour obtenir la σ. Par exemple, prenons tous les trèfles dans un jeu de cartes et mélangeons-les, les cartes se retrouvent dans un ordre quelconque, ce qui constitue une permutation. Maintenant, inversons successivement l'ordre de série de cartes jusqu'à ce que l'ensemble soit dans l'ordre croissant (as, deux, trois, quatre etc.), c'est-à-dire la *permutation identité*. Cela constitue ce que l'on ap-

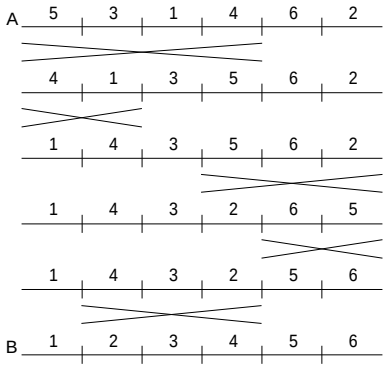


FIGURE 1.2 – Transformation par inversions successives du chromosome A pour obtenir le chromosome B. Les croix indiquent les inversions.

pelle un *tri par inversions*. Pour les chromosomes, cela équivaut à trier les gènes d'un des deux chromosomes pour qu'ils se retrouvent dans l'ordre de l'autre chromosome. Caprara (1997) a montré que le tri d'une permutation par inversion est un problème NP-difficile, ce qui signifie que le temps de calcul augmente extrêmement vite à mesure que le nombre d'éléments de la permutation augmente. Dans le cas, biologiquement plus réaliste, où l'orientation de chaque gène du chromosome est représentée par un signe + ou -, ce que l'on appellera une *permutation signée*, le tri par inversion a une complexité polynomiale (Hannenhalli et Pevzner, 1999). Dans ce cas, le calcul de la distance d'inversion entre deux chromosomes dont les gènes sont orientés (signés) prend un temps polynomial avec la taille des données. Le logiciel **signed_dist** [3] permet de calculer la *distance d'inversion* entre deux chromosomes signés.

Pour passer d'un chromosome à un autre, il est possible de rechercher le nombre minimal d'événements impliquant la transposition plutôt que l'inversion. Le principe est le même : il s'agit de *trier par transposition* l'ordre d'une permutation dans l'ordre de la permutation identité. Le calcul de la distance de transposition entre deux permutations a été envisagé par Bafna et Pevzner (1995) et si la complexité du problème n'a pas encore été démontrée, il est probablement NP-complet (Sankoff et Blanchette, 1999). Dans certaines lignées évolutives, les gènes sont tous transcrits à partir du même brin et dans ce cas, le tri par transposition est beaucoup

[3] http://www-hto.usc.edu/plain/people/Hannenhalli.html

plus adapté, le phénomène d'inversion étant probablement très rare et même inexistant chez ces organismes.

Le séquençage complet d'un génome permet de connaître l'orientation des gènes et donc de construire une représentation du chromosome par une permutation signée. C'est le cas pour les génomes de petite taille, les virus, les chloroplastes, les mitochondries et même éventuellement les génomes bactériens. Cependant, le séquençage de génomes complets reste pour l'instant un processus long et très coûteux. Les données les plus fréquemment disponibles pour les grands génomes sont les cartes physiques et/ou génétiques mais surtout les caryotypes. Dans le cas des cartes physiques, la localisation des gènes sur les chromosomes est connue ; en revanche, leur orientation ne l'est généralement pas, ce qui conduit à une représentation du chromosome par une permutation non signée. L'étude des remaniements est alors très difficile. Il est possible de contourner la difficulté en construisant une représentation à un niveau d'organisation plus élevé : on peut utiliser la position relative des gènes au sein d'un groupe de liaison pour orienter le segment chromosomique délimité par le groupe de liaison. Le chromosome est alors représenté par une permutation signée dont les éléments de base ne sont plus les gènes mais des segments orientés de chromosome. Le principe est le même avec les caryotypes, il est possible d'identifier des fragments homologues de chromosomes entre génomes par leur succession de bandes. L'orientation de ces fragments se fait de façon relative par comparaison avec les autres génomes. Le chromosome est ici représenté par une permutation signée dont les éléments constitutifs sont des fragments homologues de chromosomes caractérisés par leurs motifs de bandes.

1.2.3 Analyses de génomes multi-chromosomiques

La transformation d'un génome en un autre, lorsqu'il s'agit de génomes multi-chromosomes comme c'est le cas pour ceux que nous venons d'évoquer, est extrêmement difficile. Au problème de tri par inversions, s'ajoute le problème de réarrangements entre chromosomes. Dans le cas de remaniements multi-chromosomiques, les opérateurs principaux sont : l'inversion, la translocation, la fusion et la fission. L'un des problèmes qui rend ce type d'analyse plus difficile

est la résolution des cartes disponibles. Si les cartes comparées de l'homme et de la souris sont très bien documentées (puisque l'on dispose de plusieurs centaines de gènes localisés dans les deux génomes), la résolution des cartes disponibles dans les autres groupes est nettement moins bonne. Le nombre de gènes localisés à la fois dans plusieurs génomes est faible, ce qui rend l'étude des remaniements entre ces génomes d'autant plus compliquée. Il est peu probable que l'effort fourni pour obtenir les cartes comparées de l'homme et de la souris soit répété pour de nombreuses autres espèces, en dehors des espèces d'intérêt agronomique et des animaux de compagnie. Dans ces conditions, il est difficile d'envisager des études phylogénétiques, vu le nombre d'espèces pour lesquelles les cartes sont suffisamment documentées. La technique récente de *peinture sur chromosome* permet de localiser facilement la position un gène d'une espèce dans le génome d'une autre, permettant ainsi de construire des ordres de gènes comparés entre espèces sans avoir recours à la réalisation de cartes génétiques et physiques. Chez les mammifères, la peinture sur chromosome commence à fournir un jeu de données suffisamment important pour pouvoir envisager des études phylogénétiques à partir des réarrangements chromosomiques au sein de ce groupe. Kececioglu et Ravi (1995) sont les premiers à s'être intéressés au calcul d'une distance entre génomes multi-chromosomiques. Ils ont développé une distance basée sur la translocation. La complexité du calcul de cette distance demeure inconnue pour l'instant.

Les objectifs principaux de la comparaison de génomes sont la localisation de gènes d'intérêt médical ou agronomique, la reconstruction de génomes ancestraux ou l'étude structurale fonctionnelle et évolutive des génomes. Mais il est aussi passionnant d'imaginer retracer l'histoire évolutive des espèces à partir de la comparaison de leurs génomes. La première phylogénie basée sur l'ordre des gènes a été publiée par Sankoff *et al.* (1992). Les auteurs ont comparé l'ordre des gènes du génome mitochondrial d'espèces de champignons et d'animaux. Ils ont utilisé une distance d'édition basée sur l'ensemble minimal d'événements de remaniements chromosomiques d'inversions, de transpositions et de gain–perte nécessaire pour passer de l'ordre des gènes d'une espèce à l'ordre des gènes d'une autre. Ils ont calculé cette distance

pour chaque couple d'espèces de leur échantillon et obtenu une matrice de distances deux à deux. Ils ont utilisé le critère des *moindres carrés pondérés* pour choisir l'arbre qui correspondait le mieux à la matrice de distances. Chacun des opérateurs utilisés (inversion, transposition et gain–perte) a reçu le même poids dans le calcul de la distance d'édition. Ce calcul a été réalisé au moyen d'une heuristique que les auteurs ont développé et implémenté dans un logiciel appelé **derange**. Blanchette *et al.* (1996) ont montré que la pondération la plus adaptée à ce type de données revenait à donner un poids de 2 aux transpositions et transinversions par rapport aux inversions. Pour cela, ils ont étudié la stabilité des résultats obtenus en faisant varier les poids donnés aux différents opérateurs. Ils ont considéré que la transposition est équivalente à la transinversion car à partir du moment où un fragment d'ADN s'est détaché, il peut se réinsérer dans un sens ou dans l'autre avec la même probabilité. Ils ont par ailleurs exploré l'influence de la taille des fragments impliqués dans les événements de réarrangements en donnant aux opérateurs un poids proportionnel à la taille (en paires de bases) des fragments impliqués. Ils ont pu montrer que dans le cas du génome mitochondrial de deux champignons, la distance obtenue est peu sensible aux poids élevés, c'est-à-dire que les réarrangements qui ont eu lieu entre ces deux espèces ont probablement impliqué de petits fragments. Le logiciel **derange2** [4] permet de calculer des distances d'édition entre deux ordres de gènes, enrichies de pondérations entre les opérateurs d'inversion, de transposition et de transinversion.

1.2.4 Points de cassure

Watterson *et al.* (1982) ont introduit la notion de point de cassure ou *breakpoint*, pour la comparaison de deux ordres de gènes. Un point de cassure est un couple de gènes qui sont adjacents dans un ordre et qui ne sont pas adjacents dans l'autre ordre. Intuitivement nous pouvons remarquer qu'une inversion d'un bloc de gènes peut au maximum éliminer deux points de cassure. La distance d'inversion entre deux ordres a et b est donc supérieure ou égale à deux fois le nombre de points de cassure entre ces deux ordres. La *distance de points de cassure* qui

[4] http://www.cd.washington.edu/homes/blanchette/blanchem/software.html

est beaucoup plus simple à calculer que les distances d'édition, a donc été proposée pour la phylogénie à partir de génomes multiples. Blanchette *et al.* (1999) ont montré qu'il y a, dans certains cas, une relation presque linéaire entre la distance de points de cassure et la distance d'édition associant inversion, transposition et transinversion. Dans ce cas l'arbre minimisant le nombre de points de cassure doit être proche de celui minimisant le nombre de réarrangements. La recherche du minimum de points de cassure, qui est NP-difficile pour trois taxons, revient à reconstituer un génome intermédiaire qui minimise le nombre de points de cassure entre les trois taxons auxquels il est relié. Sankoff et Blanchette ont montré que ce problème peut se réduire au problème du voyageur de commerce et proposé une heuristique qu'ils ont appliquée au génome mitochondrial des métazoaires. Ils calculent un génome « intermédiaire » pour chaque groupe de trois taxons sur l'arbre puis par itérations successives ils améliorent les génomes intermédiaires pour diminuer le nombre de points de cassure global sur l'arbre. Ils évaluent ainsi chaque topologie. Par conséquent, cette heuristique a une complexité exponentielle avec le nombre de gènes et le nombre de génomes, le calcul devient donc très rapidement impossible. Partant de cette constatation et du fait qu'il existe de nombreuses heuristiques performantes en parcimonie, Cosner *et al.* (2000b) ont proposé d'utiliser un codage en présence/absence des points de cassure signés qu'ils ont appelé « Maximum Parsimony on Binary Encodings of genomes » (MBPE). Ils traitent la matrice obtenue avec les algorithmes classiques de parcimonie et utilisent les arbres obtenus comme point de départ pour la recherche de l'arbre minimisant le nombre de points de cassure. Le codage MBPE leur sert d'« heuristique » pour l'analyse des points de cassure. Les auteurs ont comparé leur stratégie d'analyse avec les distances d'édition et l'analyse de points de cassure avec des données expérimentales et des données simulées (Cosner *et al.*, 2000a). Ils ont montré que leur approche permet d'obtenir des résultats satisfaisants et ce dans un temps raisonnable (l'analyse des points de cassure s'étant révélée impossible avec leurs données).

Par ailleurs, Sankoff *et al.* (2000a) ont proposé une distance de points de cassure normalisée pour tenir compte des contenus divergents en gènes ainsi que du problème des gènes dupliqués.

Le calcul des distances d'édition (distances d'inversion, de transposition etc.) ou de points de cassure a pour prérequis que les génomes aient le même contenu en gènes. Pour les gènes dupliqués, les auteurs ont retenu deux stratégies : s'il est possible de distinguer les deux copies, par leurs positions ou leurs séquences par exemple, seule la copie qui induit le moins de changement de points cassure est conservée ; s'il n'est pas possible de distinguer les deux copies, elles sont éliminées des génomes. Pour calculer la distance de points de cassure entre des génomes dont le contenu en gènes diffère, Sankoff *et al.* ont choisi de « normer » les génomes, c'est-à-dire d'éliminer les gènes présents dans un génome et absents dans l'autre ; ils calculent ainsi des *points de cassure induits* entre génomes réduits. La mesure des points de cassure induits est très peu sensible aux données manquantes. Sankoff et Blanchette (1999) ont introduit un codage similaire basé sur la présence/absence de paires de gènes non signées ou *Adjacencies Parsimony*.

1.2.5 Maximum de vraisemblance et contenu génique

En phylogénie moléculaire classique, la méthode du *maximum de vraisemblance* est couramment utilisée pour estimer la qualité d'arbres évolutifs sur la base de modèles de substitutions élaborés. Devant le succès de cette méthode, Dicks (2000) a commencé à explorer ses possibilités pour la phylogénie chromosomique. Le principe est d'évaluer la vraisemblance de différentes topologies à partir de modèles d'évolution des remaniements chromosomiques.

Les études les plus récentes de phylogénie génomique fondent leurs analyses sur le contenu en gènes du génome des organismes étudiés (Snel *et al.*, 1999 ; Tekaia *et al.*, 1999 ; Fitz-Gibbon et House, 1999 ; Gu, 2000). La phylogénie du vivant, malgré les premiers résultats obtenus avec l'ARNr, demeure mal résolue (Tekaia *et al.*, 1999). La non congruence des résultats obtenus avec des gènes différents a même poussé certains auteurs à émettre l'hypothèse qu'un arbre unique n'est pas forcément la bonne représentation de la phylogénie du vivant (Doolittle, 2000). Le taux de transfert de gènes entre micro-organismes étant élevé, des gènes ayant été transférés sont susceptibles de raconter une histoire évolutive différente de celle des autres

gènes. Les auteurs de ces études récentes ont construit un ensemble de familles multigéniques à partir de plusieurs micro-organismes et ont reconstruit les relations de parentés de ces espèces sur la base de la présence/absence de ces familles. Utiliser la présence/absence de familles multigéniques plutôt que la présence/absence de simples gènes permet de se libérer du problème d'identification des orthologues et des différences importantes en contenu de gènes, la présence d'une famille de gènes gommant l'absence de tel ou tel gène. En fait, cela permet de se placer à un niveau d'organisation supérieur à celui du gène, ce qui rend les analyses plus robustes. Les résultats montrent une très bonne congruence avec les résultats obtenus avec l'ARNr et sont fortement robustes. En effet, la topologie obtenue est peu sensible aux variations des paramètres de construction des familles de gènes. Ces études montrent l'intérêt des analyses regroupant des données de génomes complets pour la phylogénie « profonde ». Albà *et al.* (2001) ont réalisé le même type d'analyse sur 19 génomes complets d'herpesvirus.

2

Cadre de l'étude

2.1 Méthodes de reconstruction phylogénétiques

Le terme de phylogénie a été introduit par Ernst HÆCKEL en 1866 pour définir le concept de « formation des espèces animales et végétales au cours du temps ». Dans son acception moderne, le terme de phylogénie recouvre le concept de « relations de parentés entre taxons ». La phylogénie, domaine de la biologie évolutive, s'intéresse donc à la recherche des relations de parenté entre taxons, de rang spécifique ou supra spécifique. Les relations entre taxons sont classiquement représentées par un arbre, figure en deux dimensions constituée de nœuds, de branches et d'une racine. Il existe deux types de nœuds : les nœuds internes et les nœuds terminaux. Les nœuds internes représentent les regroupements de taxons et les nœuds terminaux (ou feuilles) représentent les taxons. Les branches relient les nœuds entre eux, et tous les nœuds de l'arbre sont connectés. Sur un arbre, il n'existe qu'un seul chemin pour aller d'un nœud à autre nœud, à la différence d'un réseau. L'arbre phylogénétique comporte une dimension temporelle qui permet d'évaluer le temps de divergence entre les taxons, le point de départ étant représenté par la racine.

Le principe de la reconstruction phylogénétique est de choisir la meilleure hypothèse de relations de parentés, parmi un ensemble d'hypothèses alternatives et selon un critère prédéfini.

Les hypothèses alternatives sont représentées par des arbres de topologies différentes. Chaque taxon est décrit par une combinaison d'états de caractères qui permet de le caractériser. Un caractère en phylogénie représente un attribut, généralement morphologique ou moléculaire, observable sur tout ou partie des taxons étudiés et dont on peut faire l'hypothèse qu'il est le résultat d'une ascendance commune. Le caractère identifié est considéré comme homologue, l'homologie étant un concept fondamental en biologie comparative. Un caractère peut prendre plusieurs formes, chacune de ces formes étant ce que l'on appelle un état de caractère. Par exemple, le caractère « couleur des yeux » peut prendre les formes « vert », « gris » ou « orange » qui sont trois états différents de ce caractère. Le critère de sélection de la meilleure hypothèse repose sur des mesures portant sur les caractères, qui permettent de quantifier la parenté entre taxons. Le critère de sélection varie selon la méthode envisagée, méthodes regroupées en trois catégories :

- méthodes de distances,
- méthode de maximum de vraisemblance,
- méthode de parcimonie.

Les *méthodes de distances* impliquent le calcul d'une distance entre chaque paire de taxons présents dans l'analyse. L'ensemble de ces distances est regroupé dans une matrice de distances deux à deux. Le choix de l'arbre qui représente le mieux cette matrice peut se faire de deux manières, soit en cherchant parmi l'ensemble des arbres, celui qui minimise la déformation de la matrice, par la méthode des moindres carrés par exemple, soit en utilisant un algorithme d'agrégation des taxons pour construire directement un arbre à partir de la matrice. Remarquons que dans ce dernier cas, les hypothèses alternatives ne sont pas envisagées. Cette méthode n'est plus utilisée à l'heure actuelle qu'en phylogénie moléculaire où le choix d'une distance donnée peut être justifié par un modèle évolutif applicable aux données étudiées.

La *méthode de maximum de vraisemblance*, utilisée en phylogénie moléculaire, établit au préalable un modèle probabiliste d'évolution des caractères. Ce modèle peut être plus ou moins complexe et ainsi prendre en compte un nombre plus ou moins important de facteurs différents.

Il établit un ensemble de probabilités qui correspondent aux différentes classes de substitutions envisagées. Ainsi, pour chaque classe de transformation d'un état de caractères en un autre, le modèle définit une probabilité. Pour un caractère donné, une transformation correspond au passage d'un état x vers un état y, passage auquel une probabilité est assignée. Sur chaque arbre possible il convient ensuite de déterminer pour chaque caractère la séquence de transformations d'états de caractère qui maximise la somme des probabilités. La vraisemblance d'un arbre est donc la somme des probabilités calculées pour chacun des caractères. L'arbre qui sera retenu est celui dont la vraisemblance est maximale.

La *méthode de parcimonie* considère chaque caractère individuellement. Sur un arbre donné, les transformations d'états de caractères sont placées sur les branches de façon à ce que le nombre total de transformations soit minimal. La somme des transformations de chacun des caractères représente la longueur totale de l'arbre. C'est cette longueur qui sert de critère de choix entre différents arbres possibles. La méthode de parcimonie retient donc le ou les arbres qui présentent la longueur minimale. La parcimonie cherche à minimiser le nombre de transformations sur un arbre et entre les différents arbres possibles. Lorsque la transformation implique le passage d'un état *plésiomorphe* (ancestral) vers un état *apomorphe* (dérivé) on parle de *cladistique*. La cladistique fonde les regroupements de taxons sur le partage d'un état apomorphe ou *synapomorphie*. Maximiser le nombre de synapomorphies sur un arbre revient à minimiser le nombre de transformations. Les regroupements de taxons fondés sur des synapomorphies sont considérés comme des groupes *monophylétiques* ou *clades*. Les regroupements basés sur le partage d'états plésiomorphes ou *synplésiomorphie* sont des groupes *paraphylétiques*.

Quelle que soit la méthode de reconstruction, le nombre d'arbres possibles augmente de façon exponentielle avec le nombre de taxons. Dès lors que le nombre de taxons dépasse dix il n'est plus possible d'évaluer tous les arbres. Dans ces conditions, il devient nécessaire d'utiliser un algorithme heuristique, c'est-à-dire un algorithme qui n'envisage pas tous les cas possibles mais qui permet d'obtenir un résultat proche de l'optimal dans un temps raisonnable.

La majorité des études visant à reconstruire une phylogénie à partir de la comparaison de génomes se situe dans le cadre de la méthode de distances (voir chap. 1). Ces études sont basées sur le calcul de distances impliquant différents types de réarrangements (inversion, transposition etc.). Une seule étude se place dans le cadre du maximum de vraisemblance (Dicks, 2000). Peu d'études se placent dans le cadre de la parcimonie et ce n'est pas leur objectif principal (Sankoff et Blanchette, 1999 ; Cosner *et al.*, 2000b). Notons qu'au début de ce travail aucune étude, dans le cadre de la parcimonie, ne présentait une approche formalisée. C'est ce qui nous a amené à nous engager dans cette voie. Nous proposons une approche formalisée pour reconstruire les relations de parenté entre taxons à partir de l'organisation de leur génome dans un contexte cladistique.

2.2 Recherche d'homologies dans l'organisation du génome

Nous considérerons ici le génome selon le point de vue de son organisation structurale. Ainsi nous ne tiendrons pas compte de son organisation fonctionnelle telle que l'état de condensation de la chromatine au cours des différentes phases du cycle cellulaire. En effet ces différentes phases et l'état du génome qui leur est associé, dépendent de la physiologie de la cellule. Les traits d'organisation du génome liés au fonctionnement de la cellule ne seront pas considérés comme caractères phylogénétiques. Bien qu'il soit possible d'envisager des caractères physiologiques dans un cadre phylogénétique, ce n'est pas le propos de notre travail. Nous nous focaliserons uniquement sur l'organisation du génome et sur les caractères pouvant être utilisés dans ce contexte.

Notre premier objectif est de parvenir à une description indépendante de chaque taxon étudié grâce à un ensemble de caractères établis par la comparaison conjointe de tous les taxons. Cette démarche s'articule en deux phases, une première phase d'exploration de l'organisation du génome des taxons étudiés afin d'établir une série d'hypothèses d'homologie permettant la définition d'un ensemble de caractères et une deuxième phase de description de chacun des

taxons par le truchement des caractères établis au préalable. Cette deuxième phase revient à déterminer pour chaque taxon la forme (ou état de caractère) que revêt chaque caractère.

En cladistique la définition d'un caractère repose sur une hypothèse d'homologie ou *homologie primaire* (de Pinna, 1991). La définition d'un caractère est intrinsèquement liée à l'identification d'attributs homologues, observés chez les organismes étudiés. De la délimitation précise des attributs en caractères et états de caractères dépend le résultat de l'analyse. Une attention précise doit donc être portée à l'identification de ces derniers.

L'homologie primaire, souvent identifiée à l'homologie de OWEN, dérivée du *principe des connexions* de Geoffroy ST HILAIRE, se fonde (en morphologie comme en moléculaire) sur la position constante d'une structure variable au sein d'un plan d'organisation partagé par plusieurs organismes. Dans le cas de la génomique comparative ce principe est utilisé efficacement de façon inférentielle pour la localisation de gènes homologues chez des espèces proches grâce à la synténie. Il est néanmoins très important de faire ici une distinction majeure avec le principe des connexions dans la pratique de la génomique comparative : en effet les homologies primaires ne sont pas établies sur la base des connexions entre gènes mais sur la base de leur homologie de structure interne *i.e.* sur leur « homologie » de séquence (bien souvent d'ailleurs uniquement sur la base d'un pourcentage d'identité de séquence). Il faut reconnaître que d'un point de vue opérationnel, c'est très efficace et de plus il est difficile de tester – vérifier – l'homologie primaire de ces séquences (phylogénie conjointe de gènes et de taxons). Contrairement au *principe des connexions*, les connexions entre structures ne sont plus utilisées ici pour établir l'homologie primaire mais comme éléments variables servant à définir les états de caractères. Le principe est en quelque sorte « inversé » : les gènes, éléments fixes partagés par les taxons, servent à définir les homologies primaires et ce sont les plans d'organisation variés qui environnent ces gènes qui constituent les états de caractères.

2.2.1 Morphologie du génome

L'organisation du génome dépend bien sûr de l'espèce considérée mais aussi de la technique utilisée pour l'étudier, comme nous l'avons vu au chapitre précédent. La représentation que l'on peut se faire d'un génome n'est pas la même si l'on dispose de cartes génétiques ou physiques, de caryotypes ou de séquences nucléotidiques complètes. Néanmoins, il est possible de mettre en évidence des attributs communs permettant de caractériser l'organisation du génome ou sa « morphologie ». Cela nécessite de gommer les aspects spécifiques à chaque représentation de manière à élaborer une représentation générique qui servira de canevas pour la recherche d'homologies. Une fois établies les hypothèses d'homologies grâce à ce canevas générique, la mise en œuvre d'applications particulières nécessite des adaptations mais l'approche demeure identique.

Le génome est structuré en trois niveaux d'organisation :

– chromosome(s),

– unité fonctionnelle,

– séquence.

Le niveau le plus élevé est caractérisé par le nombre de molécules d'ADN constituant le génome (un ou plusieurs chromosomes), la forme de ces molécules, leur taille et même le cas échéant leur localisation cellulaire. Le deuxième niveau d'organisation représente l'ensemble des unités fonctionnelles que l'on peut rencontrer dans un génome. Le gène est le meilleur exemple d'unité fonctionnelle mais tout élément constitutif du génome peut prétendre à ce titre. Citons par exemple, un centromère, une origine de réplication, un opéron ou bien encore un segment homologue identifié par la technique de peinture sur chromosome. Le niveau de l'unité fonctionnelle se distingue par le nombre, la forme, la taille, la fonction, la position et l'orientation des unités rencontrées. Enfin, le dernier niveau est celui de la séquence nucléotidique, caractérisée par sa composition en bases et l'agencement de ces dernières. Parmi ces éléments et leurs caractéristiques, certains peuvent faire l'objet d'une hypothèse d'homologie entre les taxons étudiés et représente des caractères potentiels.

2.2.2 Niveau chromosome

Les chromosomes peuvent revêtir des formes différentes : circulaire ou linéaire, métacentrique ou acrocentrique. La forme du chromosome peut être utilisée comme caractère, les différentes formes représentant les différents états de ce caractère. Le nombre de chromosomes peut aussi être considéré comme un caractère si on peut faire l'hypothèse que si deux génomes ont le même nombre de chromosomes cela découle d'une ascendance commune. Par exemple, deux taxons dont les génomes contiennent une même série de segments homologues non agrégés de la même façon, peuvent avoir un nombre différent de chromosomes. Dans ce cas le nombre de chromosome peut être considéré en tant que caractère. Dans le cas de la taille du chromosome, il nous semble que l'hypothèse d'ascendance commune est difficile à justifier, en effet de nombreux événements différents sont susceptibles de conduire au même résultat brouillant le signal phylogénétique. Dans le cadre d'une analyse regroupant des génomes de localisations cellulaires différentes pour les mêmes taxons, génome nucléaire et mitochondrial par exemple, la localisation représente un caractère à prendre en compte.

2.2.3 Niveau unité fonctionnelle

Sous le terme d'« unité fonctionnelle », nous regroupons les éléments, au sens large, qui constituent le génome. Seules les unités fonctionnelles dont on peut faire l'hypothèse qu'elles sont issues d'une ascendance commune constituent des caractères potentiels. Voici quelques exemples d'unités fonctionnelles qui ne sont pas de bons candidats pour être utilisées comme caractères :

- les régions d'hétérochromatine dont l'évolution est très rapide,
- l'origine de réplication du génome mitochondrial dont seule la fonction est conservée,
- les séquences non codantes,
- les gènes paralogues.

Les gènes paralogues sont issus d'un événement de duplication antérieur à l'événement de spéciation qui a conduit aux espèces qui les portent. Une phylogénie basée sur des gènes paralogues ne retrace pas l'histoire des espèces mais celle des gènes. Par conséquent les gènes paralogues ne peuvent pas être utilisés pour reconstruire une phylogénie de taxons, contrairement aux gènes orthologues, qui sont issus d'un événement de spéciation.

Les unités fonctionnelles telles que les gènes orthologues ou les segments homologues de chromosomes, pour lesquelles on peut faire l'hypothèse qu'elles sont issues d'une ascendance commune, sont des caractères potentiels dont les caractéristiques représentent les états possibles. Les unités fonctionnelles sont caractérisées par leur nombre, leur taille, leur fonction, leur position et leur orientation.

Le nombre d'unités fonctionnelles, contrairement au nombre de chromosomes, est plus difficilement utilisable comme caractère. Comme pour la taille, un même nombre d'unités peut être le résultat d'une succession d'événements différents et donc ne pas être le reflet de l'ascendance commune de deux taxons. La « Présence Absence » de chaque unité, considérée individuellement, nous semble mieux refléter une possible origine commune et ce faisant, nous semble être un meilleur caractère.

La fonction d'une unité fonctionnelle, si elle peut être interprétée en termes d'homologie, peut être utilisée comme caractère. Certains gènes homologues peuvent ne pas exercer la même fonction chez deux espèces distinctes. Un même gène peut avoir été recruté pour des fonctions différentes au cours de l'évolution. Une fonction en tant que telle, peut être considérée comme un caractère par certains auteurs. Ainsi, Albà *et al.* (2001) ont utilisé la présence/absence de différentes fonctions (réplication de l'ADN, transport de protéines etc.) comme caractères pour la phylogénie d'herpesvirus. Il nous semble que si une fonction ne peut pas être rattachée à une unité fonctionnelle homologue chez les différents taxons étudiés, elle peut difficilement être utilisée comme caractère.

La position d'une unité fonctionnelle est une caractéristique dont on peut raisonnablement faire l'hypothèse qu'elle provient d'une origine commune. La localisation des unités fonction-

nelles le long du chromosome constitue un ordre, or à part pour de très petits génomes, le nombre d'ordres possibles est immense. Ce faisant, il est très peu probable que deux taxons aient obtenu par hasard le même ordre d'unités fonctionnelles sur leur chromosome. L'orientation des unités fonctionnelles peut aussi être utilisée comme caractère. Pour l'établir l'orientation d'un gène il faut déterminer à partir de quel brin il est transcrit. Dans le cas d'un chromosome circulaire, cela nécessite l'identification des deux brins et l'établissement de l'homologie entre les brins des différentes espèces. Par exemple, les deux brins du génome mitochondrial des vertébrés possèdent des poids moléculaires différents qui les classent en brin lourd et brin léger. Cette caractéristique permet d'identifier l'homologie des brins de la molécule chez les vertébrés et ainsi, il est possible de connaître de façon absolue l'orientation d'un même gène chez différentes espèces. S'il n'existe pas de critère absolu pour identifier les brins de plusieurs espèces, il devient alors nécessaire de traiter l'orientation des gènes de façon relative *i.e.* par rapport à celles des gènes adjacents (même orientation ou orientation opposée). De la même manière l'orientation de segments homologues de chromosomes établis à partir de caryotypes ne peut être traitée que de façon relative. La technique de peinture sur chromosome permet de contourner cette difficulté : en localisant des marqueurs sur ces segments, il est possible de les orienter.

2.2.4 Niveau séquence

Le dernier niveau d'organisation du génome est celui de la séquence nucléotidique. Cette dernière se caractérise par la succession des nucléotides le long de la molécule. L'utilisation des données de séquences d'ADN ou d'acides aminés est très répandue en phylogénie moléculaire et fait l'objet d'une littérature extrêmement fournie. Par conséquent nous ne développerons pas plus cette partie.

2.2.5 Conclusion

Des caractères établis à partir de niveaux d'organisation différents peuvent éventuellement être redondants, c'est-à-dire qu'ils peuvent recouvrir la même information, pour tout ou partie. Il est important de ne pas utiliser des caractères redondants qui biaiseraient le résultat. Par exemple, si on considère séparement le nombre de chromosomes et l'ordre sur ces chromosomes, on a dans une certaine mesure, la même information est présente deux fois car avec le codage de l'ordre on peut déduire le nombre de chromosomes.

Comme nous l'avons vu au chapitre précédent (page 14), la position et l'orientation des unités fonctionnelles sur le chromosome peuvent être représentées par une permutation signée. Pour ne pas introduire d'hypothèses *a priori* sur les événements évolutifs, nous considérerons la position et l'orientation des unités fonctionnelles sur le chromosome comme un ordre. Chaque ordre sera traité comme un élément de description du génome (voir chapitre suivant).

Grâce au canevas que nous venons de présenter, nous pouvons établir un ensemble de caractères à partir desquels il est possible de décrire indépendamment l'organisation du génome de chaque taxon étudié, conformément à notre objectif. Cette description se fait sans avoir recours à une comparaison deux à deux des taxons et sans faire d'hypothèse sur les événements évolutifs. Comme l'ont décrit Sankoff et Nadeau (2000, page 87), notre approche consiste à convertir l'ordre des gènes sous une forme utilisable par la méthode de parcimonie.

3

De l’organisation du génome à la phylogénie

Au chapitre précédent nous avons proposé un canevas de description de l’organisation du génome. À partir de ce canevas nous établissons un ensemble de caractères permettant la description des taxons étudiés. La deuxième phase de notre approche consiste à coder l’état de caractère porté par chaque taxon pour chacun des caractères établis lors de la première phase. Chaque taxon est ainsi représenté par une combinaison d’états de caractères qui lui est propre. L’étape de codage fournit une matrice taxons/caractères qui sera traitée par la méthode de parcimonie lors de l’étape d’analyse.

Pour des observations identiques, il existe différents codages possibles qui vont du codage entièrement « binaire » au codage totalement en états multiples, avec tous les intermédiaires possibles. Un codage entièrement binaire divise les observations en une série de caractères comportant deux états différents, par exemple :

- caractère 1 :
  - présence d’une fosse,
  - absence d’une fosse.
- caractère 2 :

  - grande fosse,
  - petite fosse.

Pleijel (1995) propose un codage binaire particulier où chaque état est considéré indépendamment par le truchement de sa présence et donc de son absence, c'est un codage entièrement basé la présence/absence. Le codage totalement en états multiples réunit dans une même série de transformations tous les états d'un caractère :

- caractère 1 :
  - absence d'une fosse.
  - grande fosse,
  - petite fosse.

Dans ce cas, l'absence du caractère est elle-même considérée comme un état de caractère à part entière et par conséquent peut être codée comme un état supplémentaire au même titre que les autres états. Les codages intermédiaires allient codages binaires et en états multiples dans la même matrice. Ces différents codages présentent chacun des avantages et des inconvénients notamment pour le codage de données inapplicables (Strong et Lipscomb, 1999). Lorsqu'un caractère est absent chez un taxon, le codage des états de ce caractère pose problème. Le codage entièrement en présence/absence de Pleijel contourne le problème puisque c'est la présence/absence de chaque état qui est codée, mais ce codage introduit une redondance. En effet, la disparition d'un état est corrélée à l'apparition d'un autre état. Le codage en états multiples complet code les données manquantes comme absentes, ce qui peut conduire à des regroupements de taxons basés sur le partage d'une absence qui est en réalité une donnée non applicable (Strong et Lipscomb, 1999). Nous avons donc choisi de coder les données inapplicables par un point d'interrogation (codage réducteur), qui est la solution qui présente le moins d'inconvénients (Barriel et Tassy, 1993 ; Strong et Lipscomb, 1999).

Notre schéma de description de l'organisation du génome permet d'établir des caractères à partir de différents niveaux d'organisation du génome. Nous nous intéresserons ici plus parti-

culièrement au codage de l'ordre des unités fonctionnelles sur les chromosomes (voir page 29), ce qui est assimilable au codage de l'ordre des gènes.

3.1 Codage de l'ordre des gènes ou données comparables

Nous proposons deux codages originaux, le codage « Position relative » et le codage « Jonctions » et nous présenterons aussi le codage de Cosner *et al.* (2000b) dans un but de comparaison. Ces codages peuvent s'appliquer à un ou plusieurs chromosomes, linéaires ou circulaires avec le cas échéant des adaptations qui seront précisées. Nous présenterons, pour chacun des codages « Position relative » et codage « Jonctions », deux options alternatives.

3.1.1 Codage « Position relative »

Le composant du génome pris comme élément central de ce codage est l'unité fonctionnelle, généralement le gène. Les descripteurs de l'unité fonctionnelle sont sa position et son orientation sur le chromosome. La position de chaque unité fonctionnelle est codée par un caractère à états multiples où chaque position de l'unité fonctionnelle observée au sein des génomes étudiés représentant un état du caractère. La polarité de chaque unité fonctionnelle, lorsque cela a une signification biologique, est codée par un caractère binaire. L'orientation sur le chromosome peut être « directe » ou « opposée ». Nous avons choisi de retenir un codage à états multiples intermédiaire et non complets (cf. discussion plus haut page 33) ; c'est pour cette raison que nous avons introduit un caractère supplémentaire de « Présence Absence » de l'unité fonctionnelle pour les cas où cette dernière n'est pas présente dans tous les génomes pris en compte. La description de ce codage a fait l'objet d'une publication (Gallut et Barriel, 2001) et il est comparé avec le codage en « jonction » (§ 3.1.2) dans Gallut *et al.* (2000). Les paragraphes qui suivent décrivent chaque caractère utilisé et les implications de ce codage.

3.1.1.1 Caractères

a Présence Absence du gène

Lorsqu'une unité fonctionnelle est absente chez certains taxons nous utilisons un caractère binaire « Présence Absence du gène ». Ce caractère prend la valeur 1 lorsque l'unité est présente et la valeur 0 pour les génomes où elle est absente [5].

Par exemple dans la série de génomes de la figure 3.3 (page 66) le gène « D » est absent chez les taxons 7 et 8 et présent chez les autres. Pour décrire cela, nous introduisons un caractère « Présence Absence du gène D », codé en binaire. Ce caractère prend la valeur 0 pour les taxons 7 et 8, puisqu'il est absent chez ces taxons et prend la valeur 1 pour les autres taxons. Pour les génomes où cette unité est absente, la position et la polarité de l'unité sont codées par un point d'interrogation « ? », afin de ne pas arbitrairement ajouter des pas.

b Polarité du gène

Le brin d'ADN sur lequel est portée l'unité fonctionnelle est codé par un caractère binaire « Polarité ». Les unités portées par le brin 5'-3' sont codées « plus » alors que les unités portées par le brin opposé sont codées « moins ». On a donc un caractère de « Polarité » pour chaque unité présente au moins une fois au sein des génomes étudiés.

Par exemple, toujours pour les génomes de la figure 3.3 (page 66), la polarité des gènes du taxon 7 est décrite par autant de caractères de polarité, dont le codage est représenté dans le tableau 3.1. Comme le gène « D » est absent chez le taxon 7, sa polarité est codée par un point d'interrogation.

Gène	A	B	C	D	E	F	G	H	I	J	K	L
Taxon 7	+	+	-	Abs	-	+	-	+	-	+	+	+
Codage	1	1	0	?	0	1	0	1	0	1	1	1

TABLEAU 3.1 – Codage de la polarité des gènes du taxon 7 de la figure 3.3.

[5] Les valeurs 1 et 0 sont choisies par conventions et n'ont que la signification qu'on leur donne ; en particulier cela ne présume pas de l'orientation des transformations.

c Position relative du gène

Pour chacune des unités fonctionnelles identifiées au sein des génomes étudiés et considérées comme homologues pour cet ensemble taxonomique, nous utilisons un caractère de « Position relative ». Cette catégorie de caractère est codée en caractères à états multiples. L'ensemble des états du caractère « Position relative » d'une unité fonctionnelle est constitué de chacune des positions où se trouve cette unité dans les différents génomes. Un caractère de « Position relative » peut donc avoir, au maximum, un nombre d'états égal aux nombres de génomes étudiés et au minimum une seule position, c'est-à-dire de 1 à n, n étant le nombre de génomes. Afin de faciliter l'exposé, nous utiliserons le terme de *gène* mais que nous considérerons ici comme synonyme d'*unité fonctionnelle*.

La position d'une unité fonctionnelle au sein d'un génome est représentée par le couple d'unités qui encadrent celle-ci. La position d'un gène g_i dans un génome G est représentée par le couple de gènes $\{g_{i-1}, g_{i+1}\}$ qui l'encadre au sein de G. Les différentes positions de g_i au sein des génomes constituent les états du caractère « Position relative du gène g_i ».

Prenons par exemple le gène « C » dans la série de génomes de la figure 3.3 (page 66), sa position dans le génome du *Taxon 1* est le couple $\{A,\ B\}$. L'ensemble $\mathcal{R}$ ci-dessous contient tous les couples de gènes représentant les différentes positions de « C » :

$$\mathcal{R} = \{\{A,\ B\},\ \{B,\ J\},\ \{E,\ F\},\ \{B,\ D\},\ \{B,\ F\},\ \{B,\ E\},\ \{B,\ K\}\}$$

$\mathcal{R}$ est le référentiel du caractère « Position relative du gène C », c'est-à-dire que $\mathcal{R}$ est l'ensemble des états de caractères pour ce caractère. Le tableau 3.2 présente le codage du gène « C ».

	Taxon 1	Taxon 2	Taxon 3	Taxon 4	Taxon 5	Taxon 6	Taxon 7	Taxon 8
État de caractère	{A, B}	{B, J}	{E, F}	{B, D}	{B, F}	{B, D}	{B, E}	{B, K}
Codage	0	1	2	3	4	3	5	6

TABLEAU 3.2 – Codage de la position relative du gène « C » de la figure 3.3.

Le gène « C » pourrait avoir de 1 à 8 états différents étant donné qu'il y a huit génomes ; ici le gène « C » a sept états puisqu'il se trouve à la même position pour les taxons 4 et 6.

Il peut arriver qu'un gène « X » soit entouré au sein d'un génome, par le couple de gènes (A,B) et au sein d'un autre génome par le couple (B,A). Nous avons la possibilité de ne pas faire la distinction entre ces deux états et de les considérer comme un seul et même état. C'est ce que nous appellerons l'option AB=BA. Notons que dans ce cas, les états de caractères sont des **paires** de gènes où l'ordre n'a pas de signification. Ainsi, que le gène « A » se trouve avant ou après le gène « X » et que le gène « B » se trouve à l'inverse après ou avant le gène « X » ne change rien, cela sera codé comme un seul et même état. Mais nous avons aussi la possibilité de considérer que ce sont deux états distincts et de les coder différemment. C'est ce que nous appellerons l'option AB≠BA. Dans ce cas, les états de caractères sont considérés comme des **couples** et l'ordre des gènes dans ces couples est important puisque le couple (A,B) est différent du couple (B,A).

Par exemple le gène « I » est encadré par le couple (H,G) pour le taxon 1 et par le couple (G,H) pour le taxon 8. Avec l'option AB=BA, le caractère « position relative du gène I » a le même état de caractère pour les taxons 1 et 8, c'est-à-dire la paire (G,H). Avec l'option AB≠BA, le caractère « position relative du gène I » a deux états différents pour ces deux taxons, *i.e.* les couples (H,G) et (G,H). Ces deux options seront envisagées successivement.

3.1.1.2 Implications du codage « Position relative »

a Modèle impliqué

Bien que nous ayons défini le codage « Position relative » de façon à ce qu'il soit le plus indépendant possible d'hypothèses évolutives *a priori*, il implique malgré tout un modèle évolutif. Comme chaque gène ou « unité homologue » est considéré indépendamment, ce modèle implique que chaque gène peut se déplacer, se retourner, s'insérer ou se perdre indépendamment des autres. Ce modèle, où chaque gène peut évoluer indépendamment, se conforme très bien au modèle « Duplication / Pertes aléatoires » proposé par Macey *et al.* (1997, 1998). Le modèle

« Duplication / Pertes aléatoires » suppose une première étape de duplication d'un fragment d'ordre de gènes puis une deuxième étape de perte aléatoire de l'une des deux copies de chacune des paires de gènes dupliqués, voir la figure 3.1. L'une des deux copies de chaque gène peut être perdue aléatoirement, conduisant éventuellement à un nouvel ordre de gènes par rapport à l'ordre avant la duplication. Une duplication anti-sens entraîne l'inversion des gènes dupliqués. L'ordre final de gènes peut donc non seulement être modifié mais intégrer en plus des gènes inversés. Il est possible d'envisager toute sorte de « Duplication / Pertes aléatoires » permettant d'expliquer le passage d'un ordre à un autre. Le modèle « Duplication / Pertes aléatoires » implique donc que chaque gène puisse se déplacer dans l'ordre et s'inverser indépendamment des autres gènes. Ajoutons à cela la possibilité d'insérer ou de perdre des gènes et nous retrouvons le modèle impliqué par le codage « Position relative ».

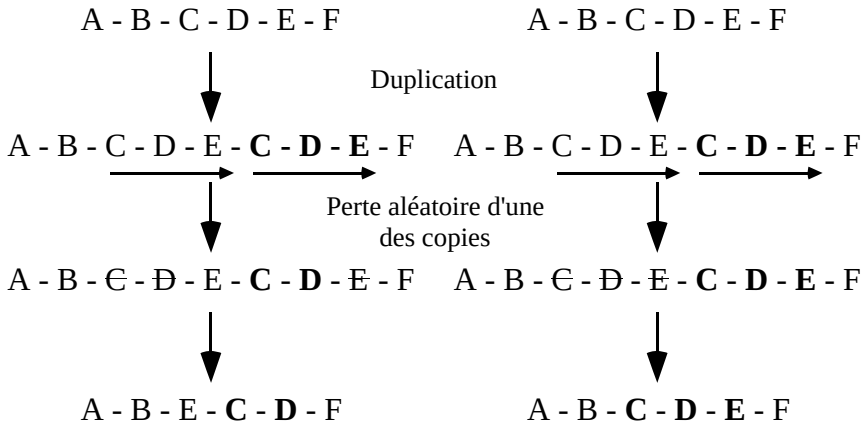


FIGURE 3.1 – Illustration du modèle « Duplication / Pertes aléatoires » (Macey *et al.*, 1997, 1998). Duplication d'un fragment d'ordre de gènes suivie de la perte aléatoire de l'une des deux copies de chacune des paires de gènes dupliqués. À gauche série de pertes conduisant à un nouvel ordre de gènes, à droite : série de pertes conduisant à l'ordre de gènes initial.

b Coût des opérateurs

Afin de mieux cerner les implications évolutives du codage « Positions relatives », nous avons calculé le coût (en nombre de pas) de différents opérateurs. Ces opérateurs (voir § 1.2.1, page 13) sont l'inversion, la transposition, la transinversion et le gain–perte. Nous avons retenu les opérateurs utilisés dans le cadre de l'exemple théorique de génomes circulaires que nous

avons réalisé pour tester les différents codages (voir § 3.4 et la figure 3.3 qui illustre ces génomes théoriques). Nous avons pris en compte pour les calculs, le nombre de gènes sur lesquels s'appliquent ces opérateurs (par exemple l'inversion d'un, deux ou n gènes à la fois). Nous avons également pris en compte, pour la transposition et la transinversion, la distance entre le point de départ et le point d'arrivée, mesurée en nombre de gènes, par exemple une transposition de n gènes sur une distance de 4, 5 ou p gènes. Le coût en fonction de n et de p, des différents opérateurs, impliqués par le codage « Position relatives » est présenté dans le tableau 3.3 pour l'option AB=BA et dans le tableau 3.4 pour l'option AB≠BA.

b.1 Inversion

Prenons tout d'abord comme exemple l'inversion d'un seul gène. Les taxons 4 et 6 (voir figure 3.3) diffèrent par l'inversion du gène « I ». Cette inversion ne change, ni la position relative des gènes adjacents (H et J), ni celle du gène « I », seule sa polarité est changée. C'est-à-dire que seul le caractère « Polarité du gène I » change d'état. L'inversion d'un gène ne coûte *a priori* qu'un pas.

Considérons maintenant l'inversion de plusieurs gènes. Les taxons 6 et 7 diffèrent par l'inversion du bloc de trois gènes G, H et I. Cette inversion implique le changement de position relative des gènes se trouvant aux limites de l'inversion (F G I et J) et le changement de polarité des trois gènes inversés. Avec l'option AB=BA, le gène « H » qui se trouve au centre du bloc inversé ne change pas de position relative, il est entouré par (G,I) ou par (I,G) ce qui revient au même. Par contre, avec l'option AB≠BA le gène « H » change de position relative. L'inversion de trois gènes coûte donc 7 pas avec l'option AB=BA (4 de position relative et 3 de polarité) et 8 avec l'option AB≠BA (5 de position relative et 3 de polarité).

b.2 Transposition

Dans le cas d'un événement de transposition la polarité des gènes ne change jamais, seule change la position relative de certains gènes. Cependant, notons que la transposition n'a pas

le même coût si elle a lieu sur une distance d'un gène ou sur une distance supérieure à un gène. Comme aucun gène n'est inversé par la transposition, il ne peut pas y avoir de couple de gènes dont l'ordre soit opposé. Cela explique également que les coûts de la transposition soient identiques avec les deux options.

La transposition du gène « C » entre les taxons 1 et 4 a lieu sur une distance d'un gène. La position relative du gène transposé et celles des gènes qui l'encadrent aux points de départ et d'arrivée (gènes A, B et D) sont changées. Notons que le gène « B », par-dessus lequel « C » est transposé, fait partie des gènes qui encadrent « C » aux points de départ et d'arrivée. Cela explique pourquoi il n'y a que trois gènes impliqués pour les points d'arrivée et de départ. Le coût d'une transposition d'un gène sur une distance d'un gène est donc de 4 pas (1 pas pour chacun des caractères « Position relative » des gènes C, A, B et D).

La transposition des gènes H et I sur une distance d'un gène (gène « G ») entre les taxons 1 et 4 coûte 5 pas. La position des gènes qui encadrent les points de départ et d'arrivée (F, G et J), et celle des gènes qui se trouvent aux extrémités du bloc de gènes transposés (H et I) ne sont pas les mêmes chez les deux taxons. Ainsi, la transposition de plusieurs gènes sur une distance d'un gène coûte toujours 5 pas puisqu'elle implique le changement de position relative des quatre gènes qui encadrent les points de départ et d'arrivée (dont un est commun aux deux) et le changement de position relative des deux gènes qui se trouvent aux extrémités du bloc transposé. Les gènes qui se trouvent à l'intérieur du bloc transposé ne changent pas de position relative.

Une transposition sur une distance supérieure à un gène coûte un pas de plus puisqu'il n'y a pas de gène commun aux points de départ et d'arrivée. Par exemple, la transposition du gène « E » sur une distance de deux gènes, entre les taxons 3 et 5 coûte 5 pas *a priori*, car la position relative des gènes E, C, B, A et D est différente.

Enfin, la transposition du bloc de gènes ABC sur une distance de deux gènes (D et E), entre les taxons 5 et 6, coûte 6 pas. Notons simplement que la position relative du gène « B » – qui se trouve à l'intérieur du bloc ABC – ne change pas.

Position relatives AB=BA						
	Inversion	Transposition sur		Transinversion sur		Gain–Perte
		$p = 1$ gène	$p > 1$ gènes	$p = 1$ gène	$p > 1$ gènes	
1 gène	1	4	5	5	6	3
2 gènes	6	5	6	**6**	8	4
3 gènes	**7**	5	6	**7**	**9**	5
4 gènes	**8**	5	6	**8**	**10**	6
5 gènes	**9**	5	6	**9**	**11**	7

TABLEAU 3.3 – Coûts (en nombre pas) de différents opérateurs avec le codage « Positions relatives » et l'option « AB=BA ».

Position relatives AB≠BA						
	Inversion	Transposition sur		Transinversion sur		Gain–Perte
		$p = 1$ gène	$p > 1$ gènes	$p = 1$ gène	$p > 1$ gènes	
1 gène	1	4	5	5	6	3
2 gènes	6	5	6	**7**	8	4
3 gènes	**8**	5	6	**9**	**10**	5
4 gènes	**10**	5	6	**11**	**12**	6
5 gènes	**12**	5	6	**13**	**14**	7

TABLEAU 3.4 – Coûts (en nombre pas) de différents opérateurs avec le codage « Positions relatives » et l'option AB≠BA. En gras : coûts différents de ceux avec l'option AB=BA.

b.3 Transinversion

Étudions les coûts de la transinversion. La transinversion du gène « J » entre les taxons 1 et 2 sur une distance d'un gène implique le changement de polarité de « J » et le changement de position relative des gènes J, G, A et C. Cette transinversion coûte donc 5 pas quelque soit l'option.

La transinversion du bloc de gènes IH sur une distance d'un gène (G) entre les taxons 3 et 5 coûte 6 pas avec l'option AB=BA : 2 pas pour le changement de polarité des gènes transinversés I et H, 1 pas pour le changement de position relative de I et enfin 3 pas pour le changement de position relative des gènes qui encadrent le bloc HI aux points de départ et d'arrivée (F, G et J) sachant que « G » se trouve à la fois aux points de départ et d'arrivée. Le gène « H » est entouré par le couple (I,G) dans le taxon 3 et par le couple (G,I) dans le taxon 5. Avec l'option AB=BA, ces deux couples sont considérés comme identiques ce qui fait H ne change pas de position. Par

contre, H change de position avec l'option AB≠BA. Par conséquent, la transinversion de HI sur une distance d'un gène coûte 7 pas avec l'option AB≠BA.

La transinversion sur une distance supérieure à un gène implique le changement de polarité des gènes transinversés, le changement de position relative des deux gènes aux extrémités du bloc transinversé (1 s'il n'y a qu'un gène), des deux gènes au point de départ et des deux gènes au point d'arrivée. Les gènes qui se trouvent à l'intérieur du bloc changent aussi de position relative avec l'option AB≠BA, puisqu'après transinversion ils se retrouvent entourés par le couple de gènes opposés. Par exemple, la transinversion du gène « K » sur une distance de deux gènes, entre les taxons 7 et 8, coûte 6 pas avec les deux options, 1 pour le changement de polarité de « K », 1 pour son changement de position relative et 4 pour le changement de position relative des gènes L, B, C et E. La transinversion sur une distance supérieure à un gène, a un coût plus élevé avec l'option AB≠BA, à partir du moment où un gène peut se retrouver entouré par un couple de gènes et son opposé après inversion. Ce n'est vrai qu'à partir du moment où le bloc transinversé comporte au moins trois gènes.

b.4 Gain–perte

Le gain–perte d'un bloc d'un ou plusieurs gènes implique le changement du caractère « Présence Absence du gène » du ou des gènes du bloc et le changement de position relative des deux gènes qui entourent le bloc. Le changement de position du ou des gènes du bloc n'augmente pas le coût car la position des gènes absents est codée par « ? ». Le passage du « ? » à un autre état ne coûte rien. Le gain ou la perte de « D » entre les taxons 6 et 7 coûte 3 pas, 1 pour la « Présence Absence du gène D » et 2 pour le changement de position relative des gènes C et E. De la même manière, le gain ou perte des gènes LK entre les taxons 6 et 7 coûte 4 pas, 2 pour la « Présence Absence » de L et K et 2 pour la position relative de A et B.

b.5 Formalisation

Les coûts des différents opérateurs en fonction du nombre de gènes impliqués n et de la distance p (en nombre de gènes) avec le codage « Position relative » et l'option AB=BA, peuvent être calculés avec les équations suivantes :

- Inversion : coût = $[\ n\ (si\ n = 1)]\ ou\ [\ n + 4\ (si\ n > 1)]$
- Transposition : coût = $2 + \min\{2, n\} + \min\{2, p\}$
- Transinversion : coût = $[\ n + 4\ (si\ p = 1)]$ ou $[\ n + 4 + \min\{2, n\}\ (si\ p > 1)]$
- Gain–perte : coût = $n + 2$

Les coûts des différents opérateurs en fonction du nombre de gènes impliqués n et de la distance p (en nombre de gènes) avec le codage « Position relative » et l'option AB$\neq$BA, peuvent être calculés avec les équations suivantes :

- Inversion : coût = $[\ n\ (si\ n = 1)]\ ou\ [\ 2n + 2\ (si\ n > 1)]$
- Transposition : coût = $2 + \min\{2, n\} + \min\{2, p\}$
- Transinversion : coût = $[\ 2n + 3\ (si\ p = 1)]\ ou\ 2n + 4\ (si\ p > 1)]$
- Gain–perte : coût = $n + 2$

Quelle que soit l'option, le coût de la transposition et celui du gain–perte sont les mêmes. La transposition coûte 4, 5 ou 6 pas alors que le coût du gain–perte commence à 3 pas et augmente de 1 pas par gène supplémentaire. Pour les deux options, il y a une différence de 5 pas entre le coût de l'inversion d'un seul gène et celle de deux gènes. Cette différence importante s'explique par le fait que l'inversion de plus d'un gène implique non seulement des changements d'orientations mais aussi des changements de positions relatives. Cela montre également qu'il peut y avoir un lien entre caractères de position et de polarité.

Les coûts de l'inversion et de la transinversion sont différents entre les deux options à partir du moment où l'événement concerne suffisamment de gènes pour que cela implique des couples opposés de type (A,B) et (B,A). Le coût n'augmente que de 1 pas avec l'option AB=BA et de

2 pas avec l'option AB≠BA. Cette différence est due à la prise en compte des couples opposés par l'option AB≠BA, qui donne un coût supplémentaire à l'inversion, coût supplémentaire qui est répercuté sur le coût de la transinversion.

3.1.2 Codage « Jonctions »

Avec le codage « Jonctions », l'élément central est à la fois l'unité fonctionnelle mais aussi la distribution des unités fonctionnelles le long du chromosome. Les descripteurs de cette distribution sont les jonctions entre gènes adjacents. Chaque jonction est considérée comme un état codé en « Présence Absence ». Ici, l'ordre des gènes est codé grâce aux jonctions et non grâce à la position de chaque gène pris individuellement. Par contre, la polarité et la « Présence Absence » des gènes sont traitées individuellement. Ces trois types de descripteurs sont codés par des caractères binaires.

3.1.2.1 Caractères

a Présence Absence du gène

Pour chacun des gènes qui ne sont pas présents dans tous les génomes, nous introduisons un caractère de « Présence Absence du gène ». Ce caractère prend la valeur « 1 » pour les taxons où le gène est présent et la valeur « 0 » pour les taxons où le gène est absent. Par exemple, dans la série de génomes de la figure 3.3 (page 66) le gène « K » est présent chez les taxons 7 et 8, mais il est absent chez tous les autres. Nous introduisons pour ce gène un caractère binaire « Présence Absence du gène K ». Ce caractère prend la valeur « 0 » pour les taxons 1 à 6 et « 1 » pour les taxons 7 et 8.

b Polarité du gène

L'orientation de chacun des gènes est décrite par un caractère « Polarité du gène » codé en binaire. Les gènes transcrits dont l'orientation est directe sont codés « 1 » alors que les gènes orientés dans le sens opposé sont codés « 0 ». Ainsi, la matrice comprend autant de caractères

de polarité qu'il y a de gènes différents dans l'échantillon. Ces caractères peuvent avoir un ou deux états différents. Le gène « A » des génomes de la figure 3.3 est toujours orienté dans le sens positif, nous utilisons donc un caractère « polarité du gène A » qui, évidement, n'a qu'un seul état de caractère, c'est-à-dire « 1 ». C'est un caractère constant.

Lorsqu'un gène est absent d'un génome, en plus du caractère « Présence Absence du gène » pour ce gène, sa polarité est codée par un point d'interrogation. Ainsi, la polarité du gène « K » est codée par un point d'interrogation pour les taxons 1 à 6.

c Présence Absence d'une jonction

Une jonction est définie par le contact – ou connexion – entre deux gènes contigus. Par exemple, on peut tirer de la série de gènes X-A-Y les deux jonctions suivantes : « X / A » et « A / Y ». Pour un échantillon donné, la liste complète et non redondante des jonctions est établie à partir de l'ensemble des génomes de cet échantillon. Ensuite, pour chacune des jonctions de cette liste nous établissons un caractère binaire « Présence Absence d'une jonction ». La distribution des gènes d'un génome est ainsi représentée par une série binaire de « 1 » pour la présence d'une jonction et de « 0 » pour l'absence. Dans le cas d'un chromosome circulaire contenant n gènes la série contient n « 1 » alors que la série d'un chromosome linéaire contient $n-1$ « 1 ». Dans le cas où tous les génomes seraient identiques, il y aurait n jonctions (génomes circulaires) ou $n-1$ jonctions (génomes linéaires). À l'opposé, si tous les génomes étaient différents il y aurait le nombre de jonctions par génomes multiplié par le nombre de génomes. Entre ces deux extrêmes, il est nécessaire de comparer tous les génomes pour connaître le nombre de jonctions. Évidemment il y a autant de caractères de « Présence Absence d'une jonction » qu'il y a de jonctions au sein de l'ensemble des génomes étudiés.

Lorsqu'un gène est absent d'un génome, les jonctions qui contiennent ce gène sont codées par un point d'interrogation. Par exemple, les jonctions qui contiennent le gène « D » seront codées « ? » pour les taxons 7 et 8. Dans la plupart des cas, cela ne change rien étant donné qu'une jonction qui contient un gène absent d'un génome ne peut pas être présente dans ce génome.

Cependant, lors du gain ou de la perte d'un ou de plusieurs gènes, cela a une conséquence sur le coût de l'événement (voir § b, page 48).

Le tableau 3.5 présente le codage « jonction » du taxon 7. Pour chaque jonction présente chez le taxon 7, le caractère de « Présence Absence d'une jonction » prend la valeur 1 pour les jonctions présentes chez le taxon 7 et 0 pour les jonctions absentes.

Jonctions	A/B	A/C	A/D	A/E	A/G	A/H	A/J	A/L	B/C	B/D	B/E
Taxon 7							oui	oui	oui		
Codage	0	0	?	0	0	0	1	1	1	?	0
...	B/K	B/L	C/D	C/E	C/F	C/J	C/K	D/E	D/J	E/F	E/K
...	oui			oui						oui	
...	1	0	?	1	0	0	0	?	?	1	0
...	F/G	F/H	F/I	F/J	G/H	G/I	G/J	H/I	I/J	K/L	
...			oui		oui		oui	oui		oui	
...	0	0	1	0	1	0	1	1	0	1	

TABLEAU 3.5 – Codage « Jonctions » du taxon 7 de la figure 3.3. Codé avec l'option AB=BA.

La liste complète des jonctions peut comporter des jonctions de la même paire de gènes mais dont l'ordre est opposé, telles que les jonctions « A / B » et « B / A ». Nous pouvons considérer ces deux types de jonctions comme un seul et même état et dans ce cas c'est ce que nous appellerons l'option « AB=BA ». Mais nous pouvons aussi les considérer comme deux états différents, ce qui correspond à l'option « AB≠BA ». Avec l'option AB=BA la liste complète des jonctions ne comprend qu'une seule des deux jonctions pour chacune des paires de jonctions opposées. Par exemple les génomes de la figure 3.3 contiennent entre autres les jonctions « A / J » et « J / A », « B / C » et « C / B ». Avec l'option AB=BA la liste complète des jonctions, établie à partir de ces génomes, ne comprend que deux des quatre jonctions, soit par exemple « A / J » et « B / C ». La jonction « A / J » représente à la fois la jonction « A / J » mais aussi la jonction « J / A » et de la même manière la jonction « B / C » représente à la fois les deux jonctions « B / C » et « C / B ». Avec l'option AB≠BA, la liste complète contient toutes les jonctions quelles qu'elles soient. Ainsi les quatre jonctions « A / J », « J / A », « B / C » et « C / B » sont retenues pour le codage.

3.1.2.2 Implications du codage « Jonctions »

a Modèle impliqué

Le codage « Jonctions » cherche, autant que possible, à s'émanciper de toute hypothèse évolutive *a priori*. Cependant, le codage associé à la méthode cladistique implique un modèle évolutif. Le principe de la parcimonie est de minimiser le nombre d'événements sur l'arbre. Ici, les événements que l'on cherche à minimiser sont de trois types :

- la rupture et la création de jonctions,
- le changement de polarité des gènes,
- le gain ou la perte de gènes.

Les événements dont nous cherchons à minimiser le nombre sur l'arbre ne sont en aucun cas ceux impliquant les opérateurs d'inversion de transposition et de transinversion. Néanmoins, pour évaluer l'impact du modèle impliqué par le codage « Jonctions » sur ces opérateurs, nous avons calculé leurs coûts.

b Coût des opérateurs

Nous avons calculé le coût (en nombre de pas) de différents opérateurs pour mieux cerner les implications évolutives du codage « Jonctions » et le comparer avec celui des autres codages. Ces opérateurs (voir § 1.2.1 (page 13) et la figure 3.3) sont l'inversion, la transposition, la transinversion et le gain ou la perte. Les calculs tiennent compte du nombre de gènes n impliqués dans l'événement, et de la distance p (en nombre de gènes) sur laquelle ont lieu la transposition et la transinversion. Ces coûts sont présentés dans les tableaux 3.6 pour l'option AB=BA et 3.7 pour l'option AB$\neq$BA.

b.1 Inversion

L'inversion d'un gène, comme celle du gène « I » entre les taxons 4 et 6 (voir figure 3.3), coûte 1 pas. Il n'y a ni création ni rupture de jonction suite à cette inversion mais le caractère « Polarité du gène I » change d'état. L'inversion de plusieurs gènes implique un coût plus im-

portant. Par exemple, l'inversion du bloc GHI entre les taxons 6 et 7 provoque le changement de polarité des gènes G, H et I, la rupture des jonctions « F / G » et « I / J » ainsi que la création des jonctions « F / I » et « G / J ». Les caractères « Présence Absence d'une jonction » des jonctions « F / G », « I / J », « F / I » et « G / J » changent d'état, ce qui implique au total un coût de sept pas pour l'inversion de trois gènes avec l'option AB=BA. Avec l'option opposée, les jonctions internes au bloc GHI sont elles aussi changées. La jonction « G / H » devient « H / G » et la jonction « H / I » devient « I / H » ce qui ajoute quatre pas supplémentaires au coût de l'événement. L'inversion de trois gènes avec le codage « Jonctions » et l'option AB≠BA coûte par conséquent 11 pas.

b.2 Transposition

La transposition implique la rupture de deux jonctions au point de départ et d'une jonction au point d'arrivée ainsi que la création d'une jonction au point de départ et de deux autres au point d'arrivée. Par exemple, la transposition du bloc de gènes ABC entre les taxons 5 et 6 sur une distance de deux gènes (D et E) implique la rupture des jonctions « E / A » et « C / F » et la création de la jonction « E / F » d'une part et la rupture de la jonction « J / D » et la création des jonctions « J / A » et « C / D » d'autre part. Seul le coût de la transposition d'un gène, sur une distance d'un gène, avec l'option AB=BA déroge à cette règle. Par exemple, la transposition du gène « C » entre les taxons 1 et 4 ne coûte que 4 pas, en effet parmi les six jonctions touchées, les deux jonctions « C / B » et « B / C » sont considérées comme identiques. Évidemment la rupture de l'une et la création de l'autre n'entraînent pas de coût supplémentaire. La transposition a un coût de 6 pas *a priori* avec l'option AB≠BA.

b.3 Transinversion

La transinversion du gène « J » entre les taxons 1 et 2, sur une distance d'un gène, implique le changement de polarité de « J » et la rupture des jonctions « G / J » « J / A » et « A / C » et la création des jonctions « G / A », « A / J » et « J / C ». Avec l'option AB≠BA cela implique

Jonctions AB=BA						
	Inversion	Transposition sur		Transinversion sur		Gain–Perte
		$p = 1$ gène	$p > 1$ gènes	$p = 1$ gène	$p > 1$ gènes	
1 gène	1	**4**	6	**5**	7	2
2 gènes	**6**	6	6	**6**	**8**	3
3 gènes	**7**	6	6	**7**	**9**	4
4 gènes	**8**	6	6	**8**	**10**	5
5 gènes	**9**	6	6	**9**	**11**	6

TABLEAU 3.6 – Coûts de différents opérateurs avec le codage « Jonctions » et l'option AB=BA.

Jonctions AB≠BA						
	Inversion	Transposition sur		Transinversion sur		Gain–Perte
		$p = 1$ gène	$p > 1$ gènes	$p = 1$ gène	$p > 1$ gènes	
1 gène	1	**6**	6	**7**	7	2
2 gènes	**8**	6	6	**10**	**10**	3
3 gènes	**11**	6	6	**13**	**13**	4
4 gènes	**14**	6	6	**16**	**16**	5
5 gènes	**17**	6	6	**19**	**19**	6

TABLEAU 3.7 – Coûts de différents opérateurs avec le codage « Jonctions » et l'option AB≠BA.

un coût de 7 pas réduit à 5 pas avec l'option AB=BA, étant donné que les jonctions « J / A » et « A / J » sont considérées comme identiques. Après transinversion, le gène se trouvant à la fin du bloc transinversé, se retrouve au début du bloc. Cela entraîne, dans le cas d'une transinversion d'une distance d'un gène, que ce gène soit toujours en contact avec le gène « sauté » par la transinversion, ce qui donne avec l'exemple précédent les jonctions « J / A » et « A / J ». Avec l'option AB=BA cela implique un coût inférieur de 2 pas à celui du coût d'une transinversion sur une distance supérieure à un gène. Par exemple, la transinversion du gène « K », sur une distance de deux gènes, entre les taxons 7 et 8 coûte 7 pas.

Avec l'option AB≠BA, les coûts de la transinversion sont les mêmes quelque soit la distance ; dans l'exemple précédent les jonctions « J / A » et « A / J » sont distinctes. Par rapport aux coûts avec l'option AB=BA, il faut ajouter 2 pas pour chacune des jonctions internes au bloc transinversé, car elles produisent des jonctions opposées après événement. La transinversion du bloc GJ entre les taxons 7 et 8 coûte 2 pas de plus qu'avec l'option AB=BA à cause de la jonction « G / J » qui devient « J / G ».

b.4 Gain–perte

Le gain ou la perte d'un bloc d'un ou plusieurs gènes implique le changement du caractère « Présence Absence du gène » du ou des gènes du bloc et le changement d'état du caractère « Présence Absence d'une jonction » de la jonction des gènes qui bordent le bloc. Il y a création de cette jonction en cas de perte et rupture de cette même jonction en cas de gain. Le gain ou la perte des gènes LK entre les taxons 6 et 7 coûte 3 pas, 2 pas pour la « Présence Absence » de L et K et 1 pas pour la jonction « A / B ». Les autres jonctions impliquées dans l'événement (« A / L », « L / K » et « K / B ») sont codées « ? » pour le taxon 6, or le « ? » n'est pas considéré comme un état supplémentaire, ce qui n'augmente pas le coût du gain–perte des gènes LK.

b.5 Formalisation

Les coûts des différents opérateurs en fonction du nombre de gènes impliqués n et de la distance p (en nombre de gènes) avec le codage « Jonctions » et l'option AB=BA, peuvent être calculés avec les équations suivantes :

- Inversion : coût = $[\ n\ (si\ n = 1)]\ ou\ [\ n + 4\ (si\ n > 1)]$
- Transposition : coût = $[\ 4\ (si\ n\ et\ p = 1)]\ ou\ [\ 6\ (si\ n\ et/ou\ p > 1)]$
- Transinversion : coût = $[\ n + 4\ (si\ p = 1)]\ ou\ [\ n + 6\ (si\ p > 1)]$
- Gain–perte : coût = $n + 1$

Les coûts des différents opérateurs en fonction du nombre de gènes impliqués n et de la distance p (en nombre de gènes) avec le codage « Jonctions » et l'option AB≠BA, peuvent être calculés avec les équations suivantes :

- Inversion : coût = $[\ n\ (si\ n = 1)]\ ou\ [\ n + 4 + 2(n - 1)\ (si\ n > 1)]$
- Transposition : coût = 6
- Transinversion : coût = $n + 2(n - 1) + 6$
- Gain–perte : coût = $n + 1$

Que ce soit avec l'option AB=BA ou l'option AB≠BA, le gain–perte a le même coût, ce dernier augmente de 1 pas par gène supplémentaire. De la même manière, la transposition coûte 6 pas avec les deux options, sauf le coût de la transposition d'un gène sur une distance d'un gène avec l'option AB=BA qui coûte 4 pas.

Avec les deux options il y a une différence de 5 pas entre le coût de l'inversion d'un seul gène et deux gènes. L'inversion de deux gènes ou plus implique le changement de la création et la rupture de certaines jonctions, c'est ce qui explique cette différence importante. Cela montre également qu'il peut y avoir un lien entre caractères de « Présence Absence d'une jonction » et caractères « Polarité du gène ».

L'inversion et la transposition n'ont pas le même coût avec les deux options. Avec l'option AB=BA, le coût de ces deux opérateurs augmente de 1 pas par gène supplémentaire. Par contre, il augmente de 3 pas avec l'option AB≠BA. Avec cette dernière, la prise en compte des options opposées augmente beaucoup le coût de l'inversion, et ce surcoût est reporté sur le coût de la transposition. Cela donne, avec l'option AB≠BA, à l'inversion et la transposition un coût nettement plus élevé que celui des autres opérateurs.

3.1.3 Codage « Jonctions signées »

Ce codage correspond au codage « Maximum Parsimony on Binary Encodings of genomes » de Cosner *et al.* (2000b). Nous l'appellerons codage « Jonctions signées » car il est basé sur un principe proche de celui du codage « Jonctions ». Avec ce codage, seul l'ordre des gènes est pris en compte par le biais des jonctions signées et il n'y a pas d'autre descripteur utilisé. Nous faisons ici une description du codage « Jonctions signées » selon la même approche que celle utilisée pour les deux codages que nous proposons, afin d'en étudier les implications et de le comparer aux codages « Position relative » et « Jonctions ».

3.1.3.1 Caractères

a Présence Absence d'une jonction signée

Dans ce codage le seul descripteur de l'ordre des gènes est la jonction. Le signe des gènes est intégré aux jonctions afin de prendre en compte l'orientation. Nous utilisons donc des jonctions signées. Une jonction signée représente une paire de gènes $\{+g_i, +g_j\}$ consécutifs dans le génome (les indices i et j sont arbitraires). La paire $\{+g_i, +g_j\}$ est équivalente à la paire $\{-g_j, -g_i\}$. Chaque jonction signée est codée par un caractère binaire « Présence Absence d'une jonction signée » qui prend la valeur « 0 » si elle est absente du génome et la valeur « 1 » si elle est présente.

Ce codage ne comprend pas d'autre type de caractère, en effet la « Présence Absence » des jonctions signées suffit à décrire complètement un ordre de gènes. À titre d'exemple voir le tableau 3.8 qui reprend le codage du taxon 7 mais avec les jonctions signées.

Jonction signée	+A/+B	+A/-C	-A/+D	-A/+E	-A/-G	-A/-H	+A/-J	-A/-J
Taxon 7								oui
Codage	0	0	0	0	0	0	0	1
…	+A/+L	+B/-C	-B/+C	+B/-D	+B/-E	-B/-K	-B/-L	-C/-D
…	oui	oui				oui		
…	1	1	0	0	0	1	0	0
…	+C/+E	-C/-E	-C/+F	+C/+J	-C/-K	-D/-E	+D/-J	-E/+F
…		oui						oui
…	0	1	0	0	0	0	0	1
…	+E/+K	+F/+G	+F/-H	+F/-I	+F/-J	+G/-H	-G/-H	+G/-I
…				oui		oui		
…	0	0	0	1	0	1	0	0
…	-G/+I	+G/+J	-G/+J	-H/+I	-H/-I	+I/+J	-I/+J	-K/-L
…			oui	oui				oui
…	0	0	1	1	0	0	0	1

TABLEAU 3.8 – Codage « Jonctions signées » du taxon 7 de la figure 3.3.

3.1.3.2 Implications du codage « Jonctions signées »

a Modèle impliqué

Des trois codages, le codage « Jonctions signées » est celui dont les implications évolutives sont les plus simples. Le codage « Jonctions signées » n'implique qu'un seul type d'événement : la rupture et la création de jonctions signées. Chaque jonction signée peut être cassée ou créée indépendamment des autres, même si la rupture d'une jonction signée entraîne forcément la création d'une autre jonction signée, mais de n'importe quelle autre.

b Coût des opérateurs

Nous avons calculé le coût de l'inversion, la transposition, la transinversion et le gain ou la perte d'un ou plusieurs gènes à la fois, avec le codage « Jonctions signées ». Ces opérateurs sont décrits dans le paragraphe 1.2.1, et illustrés sur la figure 3.3. Les calculs tiennent compte du nombre de gènes n impliqués dans l'événement, et de la distance p (en nombre de gènes) sur laquelle ont lieu la transposition et la transinversion. Ces coûts sont présentés dans le tableau 3.9.

b.1 Inversion

Le coût de l'inversion d'un ou plusieurs gènes avec le codage « Jonctions signées » est constant. Par exemple, l'inversion du gène « I » entre les taxons 4 et 6 (voir figure 3.3) coûte 4 pas, 2 pour le changement d'état des caractères « Présence Absence d'une jonction signée » des jonctions « -H / -I » et « -I / +J » et 2 pour le changement d'état des caractères « Présence Absence d'une jonction signée » des jonctions « -H / +I » et « +I / +J ». De la même manière, l'inversion du bloc GHI entre les taxons 6 et 7 implique le changement d'état des caractères « Présence Absence d'une jonction signée » des jonctions « +F / +G », « +I / +J », « +F / -I » et « -G / +J ». Les jonctions opposées telles que « -H / +I » et « -I / +H » sont considérées comme identiques par le codage « Jonctions signées ». Dans ces conditions le changement des jonctions internes du bloc inversé n'est pas pris en compte, et donc quel que soit le nombre de gènes inversés le coût de l'inversion est de 4 pas.

b.2 Transposition

La transposition implique la rupture de deux jonctions signées au point de départ et d'une jonction signée au point d'arrivée ainsi que la création d'une jonction signée au point de départ et de deux autres au point d'arrivée et ce quel que soit le nombre de gènes impliqués et la distance sur laquelle a lieu la transposition. Par exemple, la transposition du bloc de gènes ABC entre les taxons 5 et 6 sur une distance de deux gènes (D et E) implique d'une part la rupture des jonctions « -E / +A » et « -C / +F » et d'autre part la création de la jonction « -E / +F » et la rupture de la jonction « +J / -D » et la création des jonctions « +J / +A » et « -C / -D ». De plus, comme les jonctions internes du bloc transposé ne changent pas, la transposition coûte toujours 6 pas avec le codage « Jonctions signées ».

Jonctions Signées						
	Inversion	Transposition sur		Transinversion sur		Gain–Perte
		$p = 1$ gène	$p > 1$ gènes	$p = 1$ gène	$p > 1$ gènes	
1 gène	4	6	6	6	6	3
2 gènes	4	6	6	6	6	4
3 gènes	4	6	6	6	6	5
4 gènes	4	6	6	6	6	6
5 gènes	4	6	6	6	6	7

TABLEAU 3.9 – Coûts de différents opérateurs avec le codage « Jonctions signées ».

b.3 Transinversion

Avec le codage « Jonctions signées » la transinversion, à l'instar de la transposition, coûte toujours 6 pas. La transinversion est une transposition doublée d'une inversion ; comme l'inversion – avec le codage « Jonctions signées » – n'implique que des changements au niveau des extrémités du bloc inversé, la transinversion se comporte alors comme une transposition. Par exemple, la transinversion des gènes GJ entre les taxons 7 et 8 entraîne le changement d'état des caractères de « Présence Absence d'une jonction signée » des jonctions « +H / -G », « +J / +A », « +F / -I », « +F / -J », « +G / -I » et « +H / +A » soit un coût total de 6 pas *a priori*.

b.4 Gain–perte

Le gain ou la perte d'un bloc d'un ou plusieurs gènes implique le changement d'état des caractères de « Présence Absence d'une jonction signée » des jonctions aux extrémités et à l'intérieur du bloc perdu. Le gain ou la perte des gènes LK entre les taxons 6 et 7 coûte 4 pas, 1 pas pour chacune des jonctions signées suivantes : « +A / +B », « +A / +L », « +L / +K » et « +K / +B ». Avec le codage « Jonctions signées », le gain–perte est le seul opérateur dont le coût n'est pas constant.

b.5 Formalisation

Les coûts des différents opérateurs en fonction du nombre de gènes impliqués n et de la distance p (en nombre de gènes) avec le codage « Jonctions signées » peuvent se résumer avec les équations suivantes :

- Inversion : coût = 4
- Transposition : coût = 6
- Transinversion : coût = 6
- Gain–perte : coût = $n + 2$

Ce codage entraîne très peu de différence de coût entre les différents opérateurs, ce qui n'était pas le cas pour les deux autres codages.

3.2 Analyse

L'étape de codage des génomes fournit une matrice « taxon / caractères » semblable à celles utilisées généralement en parcimonie. Cette matrice peut ensuite être traitée par les logiciels de parcimonie standards. Nous allons néanmoins voir quelques points spécifiques à notre approche.

Avec le codage « Position relative » le nombre d'états par caractère « Position d'un gène » peut, dans certains cas, être élevé, c'est-à-dire supérieur à vingt. Le choix du logiciel de recons-

truction devient alors limité puisque seul PAUP* (Swofford, 1998) est capable de traiter des matrices avec un nombre d'états élevés (>10), il autorise jusqu'à trente-deux états. Au-delà de cette valeur il n'y a, jusqu'à présent, aucun recours possible.

Le codage en états multiples, retenu pour coder la position des gènes dans le codage « Position relative », peut sérieusement augmenter les temps de calcul par rapport au codage complètement binaire, et ce dès lors que la taille du jeu de données est importante. Pour le même ensemble de génomes, le temps de calcul peut varier de quelques heures avec le codage « Jonctions » à plusieurs jours avec le codage « Position relative ».

Quelque soit le codage, « Position relative », « Jonctions » ou « Jonctions signées », nous utilisons une pondération nulle, c'est-à-dire que chaque caractère a un poids équivalent dans l'analyse. Il serait difficile, à notre avis, de justifier l'utilisation d'un autre schéma de pondération. En effet, donner des poids différents à certains caractères par rapport à d'autres, se justifie par un modèle évolutif sous jacent. Pour les séquences nucléiques par exemple, les schémas de pondération utilisés classiquement sont basés sur la position des sites dans la structure secondaire de la molécule (pondérations entre caractères) ou bien sur un modèle de substitution donnant plus de poids aux transversions par rapport aux transitions (pondération des transformations). Dans le cas de réarrangement d'ordre de gènes, les modèles envisageables donnent des poids aux différents types de réarrangements les uns par rapport aux autres. Pour pouvoir donner des poids aux événements il faut d'abord les identifier. L'identification des événements peut se faire *a priori*, mais cela nécessiterait ensuite de coder ces événements, or ce n'est pas l'objectif de notre approche, bien au contraire puisque nous cherchons justement à ne pas identifier les événements *a priori*. L'identification des événements pourrait se faire au cours l'analyse, mais cela n'est pas possible dans le cadre de la méthode de parcimonie. Il est possible par contre, de donner des poids différents à chaque catégorie de caractères, ce qui permettrait de jouer sur les implications *a priori* du codage, c'est-à-dire sur les coûts *a priori* des différents opérateurs. Néanmoins, chaque caractère est indépendant durant l'analyse, ce qui implique que les clades obtenus ne seront pas forcément soutenus par des événements du type attendu. Il est également

possible de donner des poids différents à chaque type de transformation. Par exemple, un poids de 1 pour le passage d'une position relative A vers une position relative B et un poids de 10 pour la transformation inverse. Ce type de pondération nous semble encore plus difficile à justifier. Par conséquent, nous utilisons une pondération équivalente pour tous les caractères et toutes les transformations.

3.3 Retour aux caractères

L'un des avantages de la méthode cladistique est de permettre le « retour aux caractères », c'est-à-dire l'étude des transformations d'états de caractères sur le cladogramme obtenu à partir des données. Cela permet de déterminer quelles sont les transformations ou synapomorphies qui soutiennent les clades présents sur l'arbre. On peut distinguer deux types différents de transformation, les transformations non ambiguës d'une part et les transformations ambiguës d'autre part. Les transformations ambiguës correspondent à des événements qui ne peuvent pas être attribués à une branche précise de l'arbre mais à un ensemble de branches, contrairement aux transformations non ambiguës qui, elles, sont attribuées sans ambiguïté à une branche précise. Le placement des transformations ambiguës sur certaines branches de l'arbre correspond à ce que l'on appelle l'*optimisation*. L'optimisation des transformations se fait *a posteriori* et ne change en aucun cas la longueur de l'arbre. Il est classique de distinguer deux optimisations (Farris, 1972 ; Swofford et Maddison, 1987). L'optimisation dite ACCTRAN pour « accelerated transformations » place les transformations ambiguës le plus tôt possible le long des branches internes, ce qui a pour conséquence de favoriser globalement les réversions. L'optimisation dite DELTRAN pour « delayed transformations » place les transformations ambiguës le plus tard possible le long des branches internes, ce qui a pour conséquence de favoriser globalement les convergences. Notons que la notion de transformation ambiguë ou non ambiguë est indépendante de la notion d'homoplasie et de synapomorphie « unique ». C'est-à-dire qu'une transformation peut être non ambiguë sur une branche mais ne pas être unique sur l'arbre et à l'opposé

une transformation peut être ambiguë mais unique *i.e.* n'être présente une seule fois sur l'arbre quelque soit la branche où elle sera placée sur l'arbre par l'optimisation. Le choix de l'optimisation a donc une conséquence importante sur l'interprétation de l'évolution des caractères sur l'arbre.

3.3.1 Reconstitution des génomes ancestraux

Pour étudier l'évolution des caractères sur l'arbre obtenu à partir de l'analyse de l'organisation du génome, nous avons cherché à reconstituer des génomes complets aux nœuds internes. Il est communément admis que, sur un cladogramme, les ancêtres hypothétiques communs n'ont pas besoin d'être « viables ». En effet les ancêtres hypothétiques communs représentent une combinaison d'états de caractères et en aucun cas un organisme complet. Ceci étant dit, les taxons terminaux représentent ici des génomes dont l'organisation est (espérons-nous !) complètement décrite. Il n'est donc pas aberrant de prétendre retrouver des ancêtres hypothétiques communs « viables », c'est-à-dire des génomes complets !

Pour reconstituer les génomes au niveau des nœuds internes sur un arbre, il est possible d'utiliser des méthodes qui partent des génomes en position terminale et reconstituent les génomes internes de proche en proche (Blanchette *et al.* (1999) ;Cosner *et al.* (2000a, § 2.3)). Mais ce qui nous intéresse c'est de profiter du retour aux caractères pour en tirer l'information nécessaire à la reconstitution de génomes ancestraux.

De façon pratique, les logiciels de reconstruction phylogénétique fournissent la liste des transformations sur l'arbre en fonction d'une optimisation donnée (ACCTRAN, DELTRAN etc.) mais ils ne fournissent pas la liste de toutes les combinaisons de transformations sur l'arbre. Il est important de préciser que le fait de fixer une transformation ambiguë sur une branche donnée va restreindre les possibilités de fixation des autres transformations impliquant les états du même caractère. Bien évidemment, chaque caractère est optimisé indépendamment des autres et l'optimisation des transformations d'un caractère n'a strictement aucune influence sur l'optimisation des transformations des autres caractères. Toutes les possibilités d'optimisation d'un

caractère peuvent être combinées avec toutes celles des autres caractères conduisant à un très grand nombre de combinaisons différentes. Il n'est donc pas surprenant que les logiciels ne fournissent pas cette information qui serait probablement longue à calculer et dont le volume serait très important. Par contre, ils fournissent une information qui découle de la précédente (bien que ce ne soit pas un corollaire), c'est la liste des états possibles de chaque caractère à chacun des nœuds internes. Au niveau des nœuds internes, chaque caractère peut avoir un ou plusieurs états possibles. Choisir une optimisation donnée revient à assigner un unique état (parmi ceux possibles) à chaque caractère et ce pour chacun des nœuds. Une transformation implique une relation entre les états d'un même caractère de part et d'autre d'une branche, or en assignant un état à chaque caractère au niveau des nœuds nous ne tenons pas compte de cette relation impliquée par la transformation. Cela a pour conséquence, dans certains cas, d'induire un nombre de pas total sur l'arbre supérieur à la longueur de celui-ci. Les combinaisons d'états de caractères assignées aux différents nœuds de l'arbre doivent donc être compatibles avec la longueur de l'arbre.

Que ce soit avec le codage « Position relative », le codage « Jonctions » ou « Jonctions signées », la position des gènes est codée par des états qui impliquent d'autres gènes, position de « A » entre « F » et « G » ou jonctions « F / A » et « A / G », par exemple. Par conséquent la position d'un gène, au niveau d'un nœud interne, implique la position des gènes qui lui sont adjacents. Nous pouvons donc envisager de reconstituer de proche en proche un génome complet en choisissant une série de positions – ou états de caractères – compatibles les unes avec les autres. Il suffit ensuite de recommencer l'opération pour chaque nœud interne. Nous aurons fixé, à la fin, un état de caractère pour chaque caractère à chaque nœud interne. Il reste ensuite à vérifier que cette combinaison de série d'états de caractères est compatible avec la longueur de l'arbre. Chaque combinaison compatible avec la longueur de l'arbre correspond à une optimisation particulière.

Prenons tout d'abord le cas des codages « Jonctions » et « Jonctions signées ». Pour pouvoir reconstituer un génome il faut qu'il y ait autant de jonctions « présentes » qu'il y a de gènes

« présents » au niveau du nœud. Malheureusement il y a toujours moins de jonctions présentes que de gènes, il s'avère donc impossible de reconstituer des génomes complets sur les arbres obtenus avec les codages « Jonctions » et « Jonctions signées ».

Avec le codage « Position relative », chaque gène présent à un nœud a forcément au moins une position possible, puisque celle-ci est codée par un caractère à états multiples. La reconstitution d'un génome complet consiste donc à choisir une position pour chaque gène présent qui soit compatible avec celle de ses voisins. La figure 3.2 présente un exemple de reconstitution de génome à un nœud interne. Chaque gène a une ou plusieurs positions possibles. Un génome complètement reconstitué correspond à une série de positions, compatibles les unes avec les autres, chaque gène se voyant ainsi attribuer une position unique. Il peut arriver qu'il y ait plusieurs séries de positions compatibles avec un génome complet mais il peut aussi arriver qu'il n'y en ait aucune. Ce qui signifie que certaines positions sont incompatibles entre elles. Une fois que l'ordre des gènes a été reconstitué il faut assigner à chaque gène une polarité, le gène pouvant avoir le sens positif ou négatif ou bien les deux ; dans ce cas, nous reconstituons deux génomes qui diffèrent seulement par la polarité de ce gène.

A	F	O	P	T	W	P	A	W	O	F	T
PW	WP	WF	AF	FP	FO	AF	**PW**	FO	**WF**	WP	**FP**
PO	AT		TA		AO	**TA**	PO	**AO**		AT	
	OT		PO		AT	PO		AT		**OT**	
			WO			WO					

FIGURE 3.2 – Exemple de reconstitution d'un génome ancestral avec le codage « Position relative ». La ligne du haut contient les gènes, les lignes inférieures contiennent les positions possibles des gènes. La partie de droite montre une série de positions compatibles les unes avec les autres et compatibles avec un génome complet, l'ordre des gènes a été changé de façon à correspondre à l'ordre impliqué par les différentes positions choisies.

c Algorithme

Voici l'algorithme de reconstitution des génomes circulaires ancestraux aux nœuds internes pour les arbres obtenus avec le codage « Position relative » et l'option AB≠BA. La reconstitution est faite indépendamment pour chaque nœud et pour un nœud donné nous partons de

la liste des états possibles pour chacun des caractères. Cela correspond à explorer l'arbre des combinaisons de positions de gènes. Nous utilisons une exploration en profondeur avec retour en arrière lorsqu'il n'y a plus de possibilité.

Nous disposons à chaque nœud de la liste des états possibles de tous les caractères. C'est-à-dire les états possibles des caractères « Position relative du gène », « Orientation du gène » et « Présence Absence du gène » ce qui nous permet de connaître pour chaque nœud :

- l'ensemble des gènes présents : $\mathcal{P}$,
- l'ensemble des positions possibles pour chaque gène g présent :
  $\mathcal{B}_g = \{(p_1, s_1), \dots (p_i, s_i), \dots (p_n, s_n)\}$,
- le ou les sens possibles de chaque gène présent.

La position d'un gène g est représentée par un couple (p, s), p étant le gène précédent de g et s étant le suivant.

Nous notons :

- $\mathcal{T}$ l'ensemble des génomes reconstruits à un nœud.
- $\mathcal{G}$ le génome en cours de construction.
  $\mathcal{G}$ est une liste de gènes, **dernier**($\mathcal{G}$) renvoie le dernier gène de cette liste, réciproquement **premier**($\mathcal{G}$) renvoie le premier et **long**($\mathcal{G}$) donne le nombre de gènes de $\mathcal{G}$.
  $\mathcal{G} \cdot g_i$ ajoute le gène g_i à la fin de la liste de gènes $\mathcal{G}$.
- $\mathcal{L}_g \subset \mathcal{B}_g$ l'ensemble des positions acceptables pour le gène g à un nœud tel que
  $\forall (p_i, s_i) \in \mathcal{L}_g,\ p_i \in \mathcal{P}$ et $s_i \in \mathcal{P}$

Nous définissons :

– Une fonction d'initialisation qui sélectionne un gène présent g et renvoie l'ensemble $\mathcal{L}_g$ des positions acceptables pour ce gène.

Initialisation $\longrightarrow \mathcal{L}_g = \{(p_1, s_1), \dots (p_i, s_i), \dots (p_n, s_n)\}$ tel que :

$g \in \mathcal{P}$ et $\forall g_i \in \mathcal{P}, |\mathcal{L}_g| \leq |\mathcal{L}_{g_i}|$

– Une fonction de choix des gènes possibles à ajouter, lors d'une étape donnée, à un génome en cours de reconstitution.

Choix_contraint($\mathcal{G}$) $\longrightarrow \mathcal{M} \subset \mathcal{P}$ tel que :

si **dernier**($\mathcal{G}$) = **premier**($\mathcal{G}$) alors $\mathcal{M} = \emptyset$

sinon :

$k \leftarrow$ **avant_dernier**($\mathcal{G}$)

$d \leftarrow$ **dernier**($\mathcal{G}$)

$\mathcal{M} = \{s_1, \ldots, s_i, \ldots, s_n\}$ tel que $\forall s_i, (k, s_i) \in \mathcal{L}_d$

Algorithme :

$\mathcal{L} \longleftarrow$ **Initialisation**

$\mathcal{T} = \emptyset$

$\mathcal{G} = [\,]$

Pour tout $(p_i, s_i) \in \mathcal{L}$:

reconstitution($[p_i, g, s_i]$) Avec la fonction récursive de reconstitution suivante :

reconstitution($\mathcal{G}$)

$\mathcal{M} \longleftarrow$ **Choix_contraint**($\mathcal{G}$)

pour tout $g_i \in \mathcal{M}$:

si $g_i =$ **premier**($\mathcal{G}$)

si **long**($\mathcal{G}$) = $|\mathcal{P}|$

alors $\mathcal{T} = \mathcal{T} \cup \mathcal{G}$

sinon rien

sinon

si $g_i \in \mathcal{G}$ alors rien

sinon **reconstitution**($\mathcal{G} \cdot g_i$)

Le signe de chaque gène du ou des génomes reconstitués est attribué en fonction de son orientation au nœud, si les deux orientations sont possibles, deux génomes, qui ne diffèrent que

par le signe de ce gène, sont reconstitués. Si plusieurs gènes ont les deux orientations possibles cela fait autant de combinaisons.

Avec l'option AB=BA, il n'y a pas d'ordre dans les positions, c'est-à-dire que pour un gène g la position (p_i, s_i) implique que p_i et s_i soient interchangeables. Ainsi les deux gènes de la position doivent être testés pour savoir s'ils sont compatibles avec le gène précédent. Pour le reste l'algorithme est le même.

Cet algorithme de reconstitution des génomes ancestraux ne fonctionne que pour des génomes constitués d'un seul chromosome circulaire, il ne fonctionne pas pour les génomes composés de chromosomes multiples et ou linéaires. Il a été implémenté dans le programme **Recons** [6] qui permet de reconstituer les génomes ancestraux avec le codage « Position relative » et les deux options AB=BA et AB$\neq$BA. La reconstitution pour un arbre donné se fait en deux étapes : le calcul des états possibles aux nœuds de cet arbre par le logiciel PAUP*, le résultat est sauvegardé dans un fichier texte qui est ensuite traité par **Recons**. Le facteur limitant de cette reconstitution n'est pas la reconstitution elle-même mais le calcul des états possibles aux nœuds et les temps d'écriture et de lecture du fichier contenant les états possibles. Pour un arbre d'une cinquantaine de taxons l'opération dure entre une et deux minutes.

3.3.2 Reconstruction des événements

Une fois que nous avons reconstitué un génome à chacun des nœuds internes d'un arbre, il devient possible de réinterpréter les événements évolutifs sur cet arbre en comparant chaque paire de génomes entourant une branche. Cette réinterprétation peut se faire à travers le filtre d'un modèle évolutif donné, par exemple l'ensemble des inversions et transpositions nécessaires pour passer d'un génome à un autre. Il peut y avoir plusieurs réinterprétations possibles en fonction du modèle choisi, chacune impliquant un ensemble d'événements sur l'arbre. Cela permet d'évaluer l'évolution des génomes dans le cadre impliqué par le codage. Voir l'exemple de la figure 3.6, page 73.

[6] http://lis.snv.jussieu.fr/~gallut/These/recons.html

3.4 Exemples théoriques

Nous avons réalisé un ensemble de génomes circulaires artificiels pour tester et illustrer l'utilisation des trois codages. Ils sont constitués de dix ou onze gènes arrangés dans un ordre différent à chaque fois. L'ordre des gènes sur ces génomes a été établi de façon à ce qu'ils diffèrent par une série d'événements mettant en œuvre les opérateurs d'inversion, de transposition, de transinversion et de gain–perte. Ils n'ont pas été réalisés pour correspondre à un arbre évolutif mais pour comparer les coûts des opérateurs impliqués *a priori* et *a posteriori*. Ces génomes et les événements qui les lient sont représentés sur la figure 3.3.

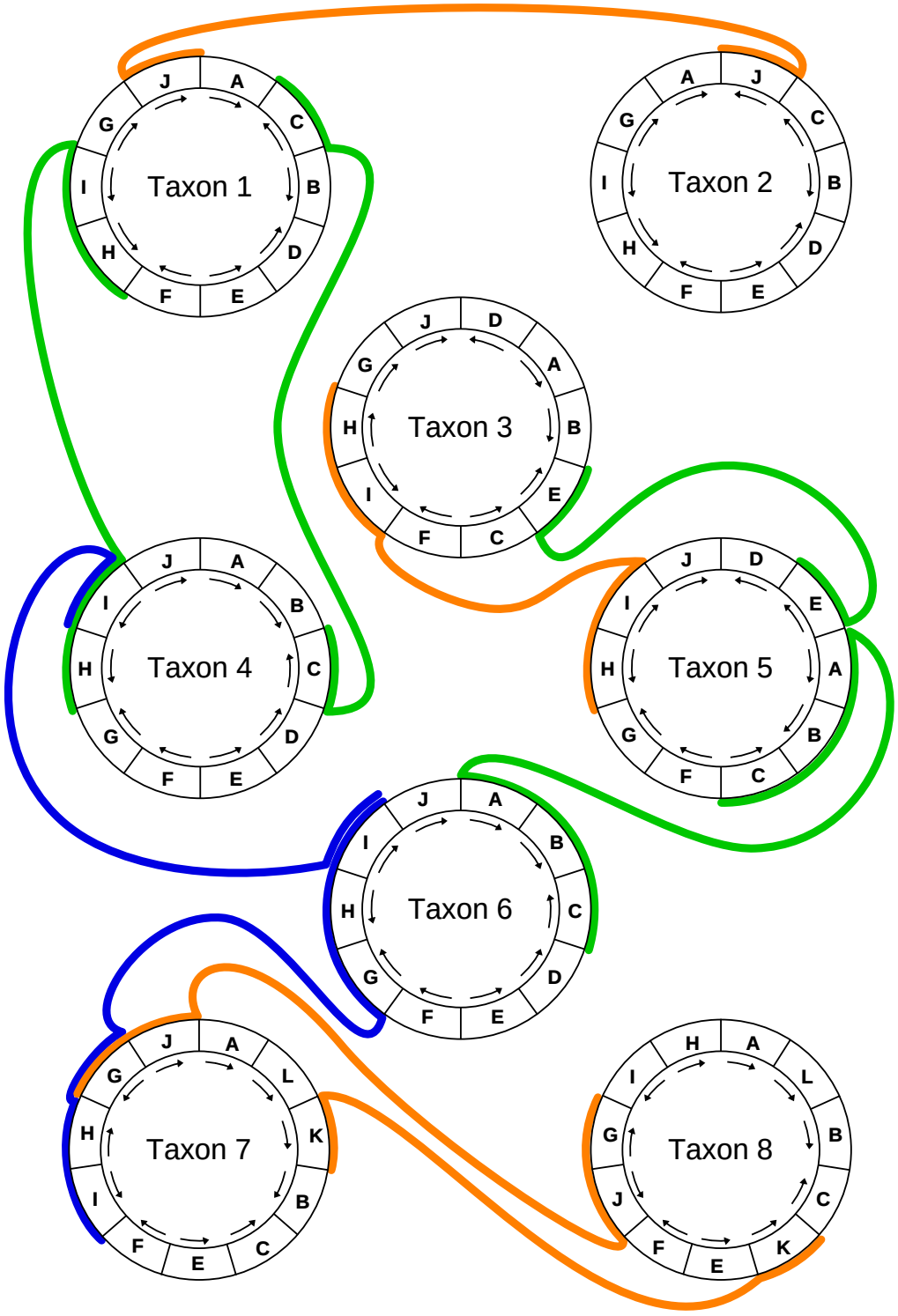


FIGURE 3.3 – Génomes circulaires théoriques, avec différents opérateurs. Les gènes sont notés de A à L, l'orientation de chaque gène est représentée par les flèches intérieures. En bleu : inversion ; en vert : transposition ; en orange : transinversion. Gain–perte du gène D et du bloc de gènes LK entre les taxons 6 et 7.

3.4.1 « Position relative »

Nous avons codé les huit génomes théoriques de la figure 3.3 avec le codage « Position relative » pour les deux options AB=BA et AB≠BA.

3.4.1.1 Option AB=BA

Avec l'option AB=BA, la matrice (voir en annexe, page 171) est constituée de 8 taxons et 27 caractères, soit 12 caractères de polarité, 3 caractères « Présence Absence du gène » pour les gènes D, K et L qui sont parfois absents et 12 caractères de position pour chacun des gènes de A à L. Parmi ces derniers, 7 caractères sont informatifs et aucun n'est constant (ce qui signifie qu'aucun gène n'a la même position pour tous les taxons).

L'analyse de cette matrice (recherche exhaustive) fournit un seul arbre de 53 pas avec un indice de cohérence de 0,9434 et un indice de rétention de 0,8235 (figure 3.4). La topologie de cet arbre ne sera pas discutée étant donné que les génomes n'ont pas été construits sur la base de relations phylogénétiques mais pour permettre d'évaluer les implications du codage.

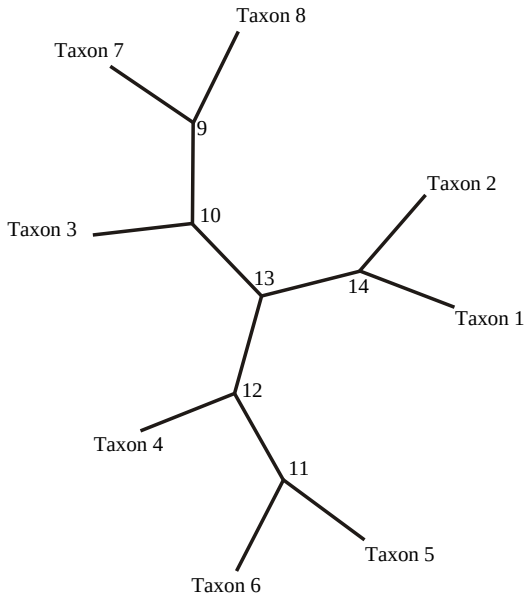


FIGURE 3.4 – Arbre le plus parcimonieux obtenu à partir de la matrice théorique avec le codage « Position relative » et l'option AB=BA. Longueur = 53 pas, IC = 0,9434 IR = 0,8235. Les nœuds internes sont numérotés à partir de 9 pour ne pas être confondus avec les numéros de taxons.

Nous avons reconstitué à l'aide du programme **Recons** les génomes ancestraux sur cet arbre. Il a été possible de reconstituer quatre génomes différents au nœud 9, deux génomes au nœud 10, un seul génome aux nœuds 11 et 12, deux au nœud 13 et un seul au nœud 14. Les génomes reconstitués à chacun des nœuds sont présentés dans le tableau 3.10. Parmi ces génomes reconstitués, certains sont identiques à l'un des taxons, voire similaires *i.e.* qu'ils ont le même ordre de gènes mais tous les gènes n'ont pas forcément la même polarité. D'autres génomes, par contre, présentent un ordre de gènes nouveau. Dans ce cas, il s'agit bien évidemment d'une mosaïque de fragments d'ordre de gènes provenant des taxons terminaux. Trois génomes reconstitués sont identiques à celui d'un taxon terminal : le génome reconstitué au nœud 11 est identique au taxon 6, le génome reconstitué au nœud 12 est identique au taxon 4, le génome reconstitué au nœud 14 est identique au taxon 1. Les deux premiers génomes du nœud 9 ont le même ordre de gènes que le taxon 7 mais la polarité du gène « G » est différente pour les deux génomes ainsi que celle du gène « K » pour le deuxième génome. Les autres génomes reconstitués ont un nouvel ordre de gènes, par exemple, le troisième génome du nœud 9 est une mosaïque de l'ordre des gènes des taxons 7 et 8.

Nœud 9	: +A	+L	**+K**	**+B**	**-C**	-E	+F	-I	+H	+G	+J	≈Taxon 7
	: +A	+L	**-K**	**+B**	**-C**	-E	+F	-I	+H	+G	+J	≈Taxon 7
	: +A	+L	**+B**	**-C**	**+K**	-E	+F	-I	+H	+G	+J	Original
	: +A	+L	**+B**	**-C**	**-K**	-E	+F	-I	+H	+G	+J	Original
Nœud 10	: +A	**-C**	**+B**	-D	-E	+F	-I	+H	+G	+J		Original
	: +A	**+B**	**-C**	-D	-E	+F	-I	+H	+G	+J		Original
Nœud 11	: +A	+B	-C	-D	-E	+F	+G	-H	+I	+J		=Taxon 6
Nœud 12	: +A	+B	-C	-D	-E	+F	+G	-H	-I	+J		=Taxon 4
Nœud 13	: +A	**-C**	**+B**	-D	-E	+F	-I	-H	+G	+J		Original
	: +A	**+B**	**-C**	-D	-E	+F	-I	-H	+G	+J		Original
Nœud 14	: +A	-C	+B	-D	-E	+F	-H	-I	+G	+J		=Taxon 1

TABLEAU 3.10 – Ancêtres communs hypothétiques reconstitués aux nœuds internes de l'arbre obtenu avec le codage « Position relative » et l'option AB=BA (voir la figure 3.4). Chaque gène est représenté par une lettre et sa polarité est représentée par le signe + ou -. La dernière colonne indique si le génome reconstitué est identique à celui d'un taxon terminal ou si il est nouveau. ≈ signifie que l'ordre des gènes est le même mais que la polarité d'un ou de plusieurs gènes est différente.

	Nœud 9	**Nœud 10**	**Nœud 11**	**Nœud 12**	**Nœud 13**	**Nœud 14**	**Coût**
1	Génome 1	Génome 1	Génome 1	Génome 1	Génome 1	Génome 1	53 pas
2	Génome 1	Génome 1	Génome 1	Génome 1	Génome 2	Génome 1	57 pas
3	Génome 1	Génome 2	Génome 1	Génome 1	Génome 1	Génome 1	57 pas
4	Génome 1	Génome 2	Génome 1	Génome 1	Génome 2	Génome 1	53 pas
5	Génome 2	Génome 1	Génome 1	Génome 1	Génome 1	Génome 1	53 pas
6	Génome 2	Génome 1	Génome 1	Génome 1	Génome 2	Génome 1	57 pas
7	Génome 2	Génome 2	Génome 1	Génome 1	Génome 1	Génome 1	57 pas
8	Génome 2	Génome 2	Génome 1	Génome 1	Génome 2	Génome 1	53 pas
9	Génome 3	Génome 1	Génome 1	Génome 1	Génome 1	Génome 1	53 pas
10	Génome 3	Génome 1	Génome 1	Génome 1	Génome 2	Génome 1	57 pas
11	Génome 3	Génome 2	Génome 1	Génome 1	Génome 1	Génome 1	57 pas
12	Génome 3	Génome 2	Génome 1	Génome 1	Génome 2	Génome 1	53 pas
13	Génome 4	Génome 1	Génome 1	Génome 1	Génome 1	Génome 1	53 pas
14	Génome 4	Génome 1	Génome 1	Génome 1	Génome 2	Génome 1	57 pas
15	Génome 4	Génome 2	Génome 1	Génome 1	Génome 1	Génome 1	57 pas
16	Génome 4	Génome 2	Génome 1	Génome 1	Génome 2	Génome 1	53 pas

TABLEAU 3.11 – Combinaisons des génomes reconstitués sur l'arbre obtenu avec le codage « Position relative » et l'option AB=BA, figure 3.4. Les génomes sont présentés dans le tableau 3.10.

Avec quatre génomes reconstitués au nœud 9, deux au nœud 10 et deux au nœud 13, il y a seize combinaisons possibles. Par exemple la première combinaison revient à choisir le premier AHC du nœud 9 associé au premier du nœud 10 et au premier du nœud 13 ainsi que l'AHC de tous les autres nœuds. La deuxième combinaison correspond au deuxième AHC du nœud 9 et au premier de chacun des autres nœuds et ainsi de suite pour les autres combinaisons. Ces seize combinaisons correspondent à seize scénarios évolutifs différents (de longueur 53 et 57 pas) et sont présentées dans le tableau 3.11.

La figure 3.5 présente la première des seize combinaisons – elle regroupe le premier génome reconstitué de chaque nœud interne ainsi que les génomes de départ – sur l'arbre le plus court.

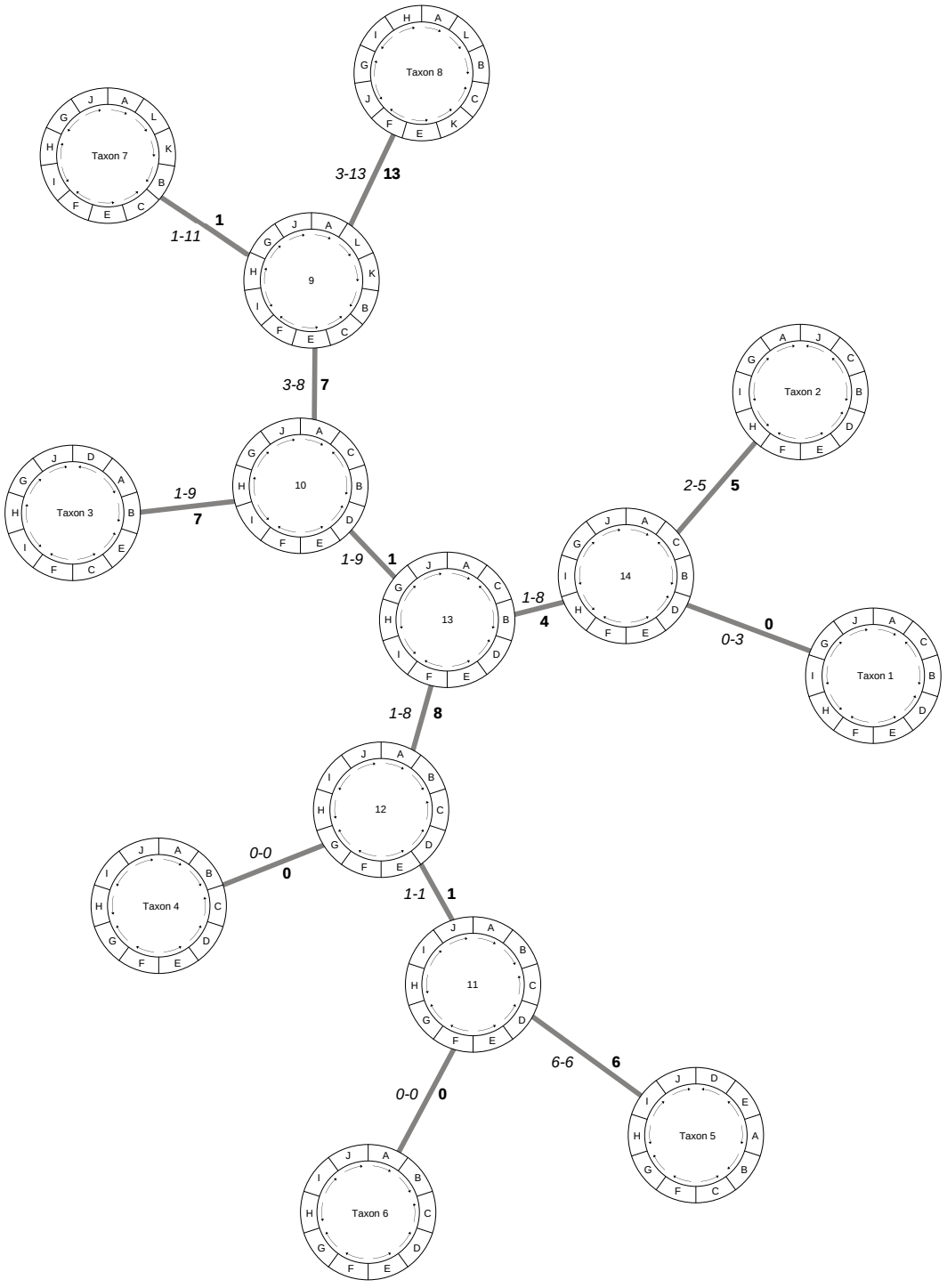


FIGURE 3.5 – Combinaison d'ancêtres hypothétiques communs pour l'arbre de la figure 3.4. Cette combinaison coûte 53 pas, ce qui correspond à longueur de l'arbre. En italique : nombre minimal et maximal de pas ; en gras : nombre de pas impliqués par la combinaison choisie.

Choisir sur l'arbre une combinaison d'AHC revient à fixer les transformations de caractères ambiguës sur certaines branches c'est-à-dire forcer une optimisation particulière (voir § 3.3.1, page 59). La figure 3.5 montre pour chaque branche le nombre minimal et maximal de transformations. Quelle que soit l'optimisation choisie, le nombre de transformations ne pourra se situer en dehors de ces valeurs. Ainsi pour la branche qui relie le nœud 12 au nœud 13 le nombre de transformations sera forcément compris entre 1 et 8. Dans le cas de la première combinaison, les génomes reconstitués aux nœuds 12 et 13 impliquent huit transformations de caractères. Ainsi une combinaison d'AHC correspond à un nombre de transformations global sur l'arbre. Pour que la combinaison soit « valide », ce nombre total de transformations ne doit pas dépasser la longueur de l'arbre. Ici le nombre de pas impliqué est de 53 ce qui est égal à la longueur. Parmi les seize combinaisons, seules huit sont compatibles avec la longueur de l'arbre. Les huit autres coûtent 57 pas. Comme elles nécessitent un nombre de transformations supérieur au nombre de pas de l'arbre, elles ne peuvent pas être retenues.

Nous pouvons remarquer que sur les branches qui relient le taxon 1 au nœud 14 d'une part, le taxon 4 au nœud 12 d'autre part et enfin le taxon 6 au nœud 11, le nombre de pas impliqués par la combinaison choisie est de 0. Cela signifie qu'il n'y a aucune transformation le long de ces trois branches. C'est-à-dire que les génomes reconstitués aux nœuds 11,12 et 14 sont respectivement identiques aux taxons 1, 4 et 6. En effet s'il n'y a pas de changement d'état de caractère le long d'une branche, les génomes sont forcément identiques.

Sur un arbre avec une combinaison d'ancêtres hypothétiques communs compatible avec la longueur de l'arbre, il est possible de reconstruire une série d'événements évolutifs, chacun correspondant à une ou plusieurs transformations. La figure 3.6 montre certains réarrangements reconstruits à partir de la combinaison de la figure 3.5. Certains des réarrangements supposés *a priori* (voir la figure 3.3) sont retrouvés sur l'arbre alors que d'autres sont scindés en plusieurs événements. Par exemple, la transposition des gènes ABC entre le taxon 5 et le taxon 6 se retrouve sur l'arbre en seul événement placé entre le taxon 5 et l'ancêtre hypothétique commun

aux taxons 5 et 6 (égal au taxon 6). En revanche, la transinversion des gènes HI entre les taxons 3 et 5 est divisée en quatre événements indépendants :

- inversion du gène « I » entre les AHC (ancêtre hypothétique commun) 11 et 12,
- transposition du gène « H » entre les AHC 12 et 13,
- transposition du gène « I » entre les AHC 12 et 13,
- inversion du gène « H » entre les AHC 13 10.

Ces réarrangements ne sont pas retrouvés parce que les taxons entre lesquelles ils ont lieu ne sont pas étroitement apparentés sur l'arbre, et le chemin qui relie ces taxons sur l'arbre comprend des transformations d'états de caractères dispersées sur différentes branches, impliquant plusieurs événements séparés. C'est le cas de la transinversion des gènes HI entre les taxons 3 et 5 qui est divisée en quatre événements distincts et indépendants. D'autre part, du fait du modèle impliqué par le codage (voir § a, page 38) chaque gène peut changer de position et d'orientation indépendamment des autres gènes. Cela entraîne que, pour les réarrangements impliquant plusieurs gènes, ces derniers ne subiront pas forcément de transformations sur les mêmes branches, avec pour conséquence que les réarrangements supposés *a priori* se retrouvent divisés en différents événements. Ces différents événements correspondent à une ou plusieurs des transformations d'états de caractères des gènes impliqués par les réarrangements supposés *a priori*. Dans le tableau 3.12 nous pouvons remarquer que parmi les réarrangements retrouvés en un seul événement, seule la transposition des gènes ABC implique plusieurs gènes, et qu'elle implique deux taxons qui se trouvent en position de groupes frères sur l'arbre. Tous les autres réarrangements impliquant plusieurs gènes sont retrouvés scindés en plusieurs événements (confère tableau 3.13). Inversement : le gain–perte du gène « D » entre les taxons 6 et 7 est le seul réarrangement impliquant un seul gène qui ne soit pas retrouvé en un seul événement. Ceci est dû aux faits que : 1^{o}) les taxons 6 et 7 ne sont pas proches sur l'arbre, 2^{o}) les gènes qui entourent « D » dans le taxon 6 sont impliqués dans d'autres événements. Par exemple le gène « C » est transposé deux fois – sur le chemin entre les taxons 6 et 7 – une fois entre les AHC 12 et 13 et une fois entre les AHC 10 et 9.

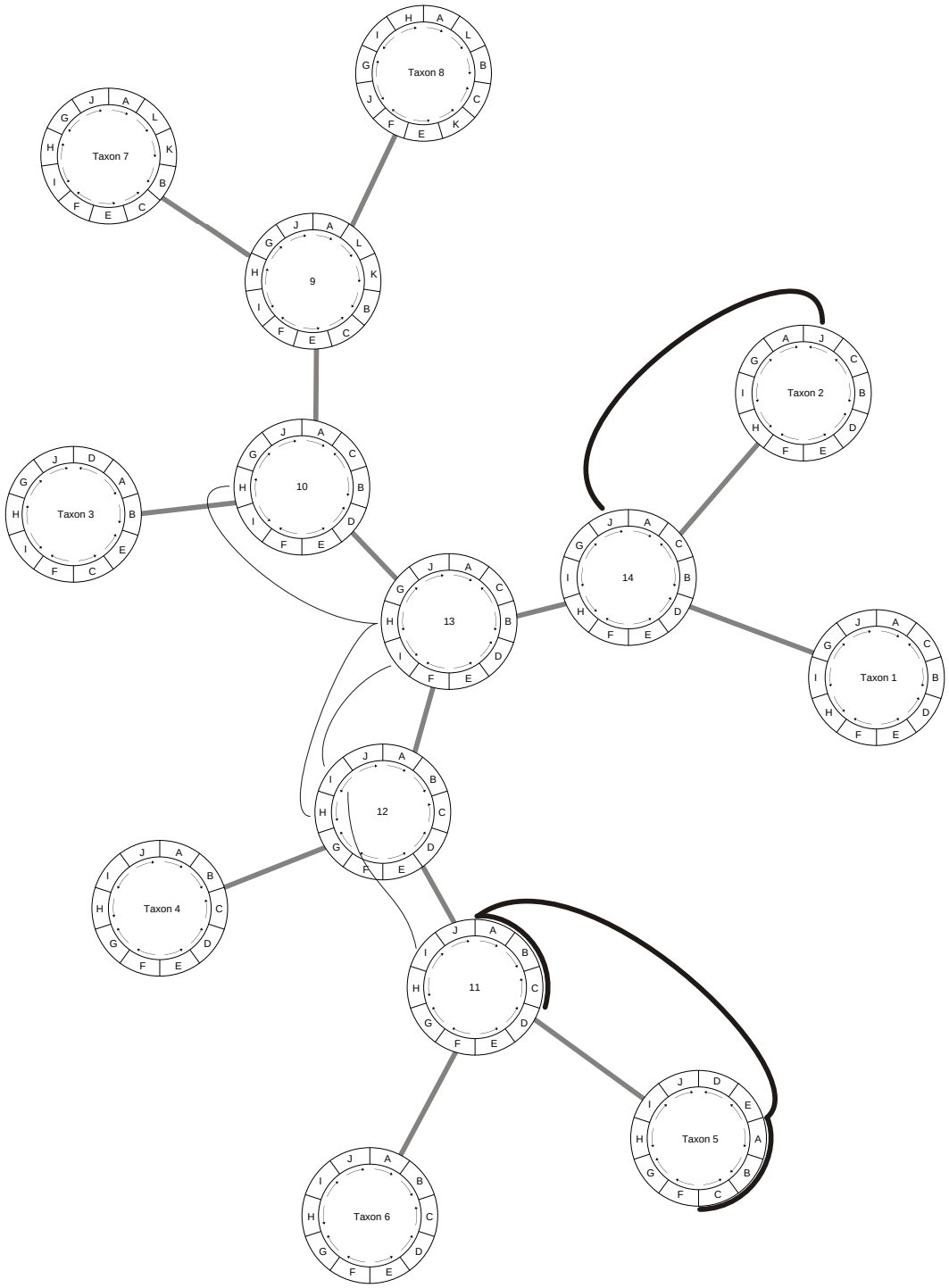


FIGURE 3.6 – Reconstruction d'événements évolutifs sur l'arbre de la figure 3.4 avec la première combinaison d'ancêtres hypothétiques communs. En gras : événements envisagés *a priori* (voir figure 3.3) retrouvés en un seul événement. Filet : événement non envisagé *a priori*.

Le codage entraîne, pour chaque type de réarrangements, un coût *a priori* en nombre de pas (voir le tableau 3.3). Les réarrangements qui se retrouvent en un seul événement sur l'arbre coûtent forcément le même nombre de pas *a priori* qu'*a posteriori*, par exemple la transposition des gènes ABC entre les taxons 5 et 6. Par contre les réarrangements qui sont scindés en plusieurs événements sur l'arbre peuvent *a posteriori* coûter le même nombre de pas ou bien coûter un nombre de pas supérieur. Ainsi la transinversion de GJ entre les taxons 7 et 8 coûte le même nombre de pas (6 pas) alors que la transposition des gènes HI entre les taxons 1 et 4 ne coûte que 5 pas *a priori* et 8 pas sur l'arbre (voir tableaux 3.12 et 3.13). Dans ce dernier exemple, pour passer du taxon 4 au taxon 1 sur l'arbre, les gènes HI sont transposés indépendamment entre les AHC 12 et 13 et se retrouvent alors dans un ordre différent. Puis le gène « H » est de nouveau transposé entre les AHC 13 et 14 pour se retrouver avant le gène « I ». Cela implique donc trois événements différents et un nombre de pas supérieur à celui supposé *a priori*. La somme des coûts des événements *a posteriori* est de 79 pas ce qui est bien supérieur

Réarrangement	**Coût *a priori***	**Coût sur l'arbre**
Transinversion du gène J taxon 1 $\leftrightarrow$ taxon 2 (sur $p = 1$)	5 pas	5 pas
Transposition du gène C taxon 1 $\leftrightarrow$ taxon 4 (sur $p = 1$)	4 pas	4 pas
Inversion du gène I taxon 4 $\leftrightarrow$ taxon 6	1 pas	1 pas
Transposition des gènes ABC taxon 5 $\leftrightarrow$ taxon 6 (sur $p = 2$)	6 pas	6 pas (cf. figure 3.6)
Transinversion du gène K taxon 7 $\leftrightarrow$ taxon 8 (sur $p = 2$)	6 pas	6 pas

TABLEAU 3.12 – Réarrangements supposés *a priori* (voir la figure 3.3 page 66) et retrouvés en un seul événement *a posteriori* sur la première combinaison de génomes ancestraux (voir figures 3.5 et 3.6).

à la longueur de l'arbre (53 pas). Cela s'explique par le fait que de nombreuses transformations d'états de caractères sont partagées par plusieurs de ces événements. Par exemple, la transposition du gène « I » entre les AHC 12 et 13 est impliquée dans la transposition des gènes HI entre les taxons 1 et 4 et la transinversion des gènes HI entre les taxons 3 et 5 comme nous l'avons

vu précédemment. Même si ces événements sont invoqués plusieurs fois pour expliquer *a posteriori* les réarrangements sur l'arbre leur coût n'est comptabilisé qu'une fois pour calculer la longueur de l'arbre.

Réarrangement	**Coût *a priori***	**Coût sur l'arbre**	**Nombre d'événements**
Transposition des gènes IH taxon 1 $\leftrightarrow$ taxon 4 (sur $p = 1$)	**5 pas**	**8 pas**	3
Transinversion des gènes IH taxon 3 $\leftrightarrow$ taxon 5 (sur $p = 1$)	6 pas	6 pas	4 (cf. texte et figure 3.6)
Transposition du gène E taxon 3 $\leftrightarrow$ taxon 5 (sur $p = 2$)	**5 pas**	**17 pas**	4
Inversion des gènes IHG taxon 6 $\leftrightarrow$ taxon 7	7 pas	7 pas	5
Transinversion des gènes GJ taxon 7 $\leftrightarrow$ taxon 8 (sur $p = 2$)	8 pas	8 pas	3
Gain–Perte du gène D taxon 6 $\leftrightarrow$ taxon 7	**3 pas**	**5 pas**	2 (cf. texte)
Gain–Perte des gènes KL taxon 6 $\leftrightarrow$ taxon 7	**4 pas**	**6 pas**	2

TABLEAU 3.13 – Réarrangements supposés *a priori* (voir la figure 3.3 page 66) et retrouvés scindés en plusieurs événements *a posteriori* sur la première combinaison de génomes ancestraux (voir figures 3.5 et 3.6).

3.4.1.2 Option AB≠BA

Nous avons codé les huit génomes théoriques de la figure 3.3 avec le codage « Position relative » et l'option AB≠BA. Dans ce cas, la matrice (voir en annexe, page 172) est constituée de 3 taxons et 27 caractères, soit 12 caractères de polarité, 3 caractères de « Présence Absence du gène » pour les gènes D K et L qui sont parfois absents et 12 caractères « Position relative du gène » pour chacun des gènes de A à L. Parmi ces derniers, sept caractères sont informatifs et aucun n'est constant. Le nombre de caractères est le même qu'avec l'option AB=BA mais ici lorsqu'un gène est encadré par un couple (A,B) et par le couple opposé (B,A) cela correspond à deux états différents. La matrice avec l'option AB≠BA contient par conséquent des états de caractère supplémentaires. Pour le reste, elle est en tout point identique à la matrice avec l'option

AB=BA. L'analyse de cette matrice (recherche exhaustive) a fourni trois arbres équiparcimonieux de 54 pas (soit 1 pas de plus qu'avec l'option AB=BA) avec un indice de cohérence de 0,9815 et un indice de rétention de 0,9375.

Ces trois arbres diffèrent seulement par la position du taxon 2 et le consensus strict est présenté dans la figure 3.7. Nous ne discuterons pas de la topologie de cet arbre étant donné que les huit génomes théoriques n'ont pas été réalisés avec une histoire phylogénétique sous jacente. Notons cependant que l'un des trois arbres est identique à l'arbre obtenu avec l'option AB=BA et que le consensus ne diffère de ce dernier que par l'absence du nœud 14 (voir figure 3.4).

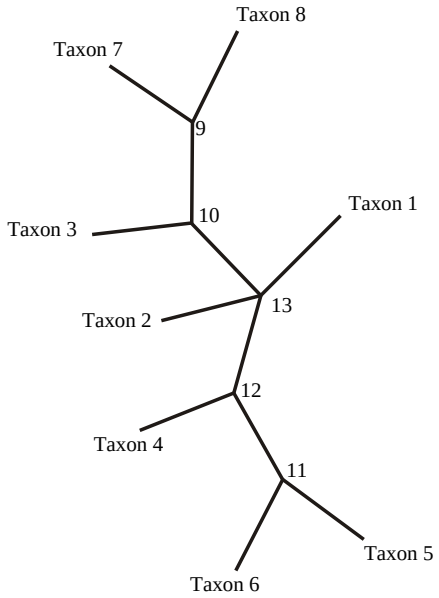


FIGURE 3.7 – Arbre de consensus strict des trois arbres équiparcimonieux obtenus à partir de la matrice théorique avec le codage « Position relative » et option AB≠BA. Longueur de l'arbre le plus court égale à 54 pas, IC = 0,9815 IR = 0,9375.

Nous avons reconstitué à l'aide du programme **Recons** les génomes ancestraux sur les trois arbres équiparcimonieux de même que sur le consensus. Sur ce dernier, il a été possible de reconstituer 4 génomes différents au nœud 9 et 1 génome à chacun des autres nœuds, ces génomes reconstitués sont présentés dans le tableau 3.14. Trois de ces derniers sont identiques à celui d'un taxon terminal, le génome reconstitué au nœud 11 est identique au taxon 6, le génome reconstitué au nœud 12 est identique au taxon 4, le génome reconstitué au nœud 13 est identique au taxon 1. Le génome reconstitué au nœud 10 présente un ordre de gènes original, identique

au premier génome reconstitué au nœud 10 de l'arbre obtenu avec l'option AB=BA (voir tableau 3.10). Les quatre génomes reconstitués au nœud 9 sont les mêmes que ceux reconstitués au nœud 9 de l'arbre obtenu avec l'option AB=BA.

Nœud 9	: +A	+L	**+K**	**+B**	**-C**	-E	+F	-I	+H	+G	+J	≈Taxon 7
	: +A	+L	**-K**	**+B**	**-C**	-E	+F	-I	+H	+G	+J	≈Taxon 7
	: +A	+L	**+B**	**-C**	**+K**	-E	+F	-I	+H	+G	+J	Original
	: +A	+L	**+B**	**-C**	**-K**	-E	+F	-I	+H	+G	+J	Original
Nœud 10	: +A	**-C**	**+B**	-D	-E	+F	-I	+H	+G	+J		Original
Nœud 11	: +A	+B	-C	-D	-E	+F	+G	-H	+I	+J		=Taxon 6
Nœud 12	: +A	+B	-C	-D	-E	+F	+G	-H	-I	+J		=Taxon 4
Nœud 13	: +A	-C	+B	-D	-E	+F	-H	-I	+G	+J		=Taxon 1

TABLEAU 3.14 – Ancêtres communs hypothétiques reconstitués aux nœuds internes de l'arbre consensus obtenu avec le codage « Position relative » avec l'option AB≠BA (voir la figure 3.7). Chaque gène est représenté par une lettre et sa polarité est représentée par le signe + ou -. La dernière colonne indique si le génome reconstitué est identique à celui d'un taxon terminal ou si il est nouveau. ≈ signifie que l'ordre des gènes est le même mais que la polarité d'un ou de plusieurs gènes est différente.

Sur l'arbre correspondant à l'arbre obtenu avec l'option AB=BA, les génomes reconstitués sont identiques à ceux reconstitués sur l'arbre obtenu avec l'option AB=BA mais deux génomes supplémentaires ont été reconstitués au nœud 13. L'un est identique au taxon 1 et l'autre présente un ordre de gènes complètement nouveau. Cela montre qu'il est possible de reconstituer plus de génomes avec l'option AB≠BA qu'avec l'option AB=BA. Cela montre également qu'il est possible d'avoir le même génome reconstitué à deux nœuds internes différents. En effet, sur cet arbre les nœuds 13 et 14 ont un génome reconstitué identique au taxon 1. Cela implique qu'il n'y a aucune transformation sur les branches reliant le taxon 1 au nœud 14 et le nœud 14 au nœud 13. Le nombre de génomes reconstitués sur cet arbre implique 32 combinaisons différentes de génomes aux nœuds internes, deux fois plus que sur l'arbre obtenu avec l'option AB=BA puisqu'il y a deux génomes supplémentaires reconstitués au nœud 13. Parmi ces 32 combinaisons, 16 impliquent une longueur de 54 pas, compatible avec la longueur de l'arbre, alors que les 16 autres impliquent une longueur de 58 pas. Seules 4 des 16 combinaisons compatibles avec la longueur de l'arbre présentent le « taxon 1 » reconstitué aux nœuds internes 13

et 14. Cela signifie que non seulement il est possible d'avoir le même génome reconstitué à deux nœuds internes différents mais en plus cela n'est pas incompatible avec la longueur de l'arbre.

3.4.2 « Jonctions »

Les huit génomes de la figure 3.3 sont ici codés et analysés avec le codage « Jonctions », soit avec l'option AB=BA soit avec l'option AB≠BA.

Comme nous l'avons vu au § 3.3.1 sur la reconstitution de génomes aux nœuds internes, il n'est pas possible de reconstituer des génomes à partir de la seule information disponible avec le codage « Jonctions », que ce soit avec l'option AB=BA ou avec l'option AB≠BA.

3.4.2.1 Option AB=BA

La matrice avec l'option AB=BA (voir en annexe, page 172) comprend 47 caractères binaires, dont 12 caractères de polarité, 3 caractères de « Présence Absence du gène » et 32 caractères de « Présence Absence d'une jonction ». Parmi les 12 caractères de polarité, 3 sont informatifs et 7 sont constants (ceux des gènes A, B, C, D, E, F et L). Les 3 caractères « Présence Absence du gène » sont informatifs. Parmi les 32 caractères « Présence Absence d'une jonction », 17 sont informatifs et 2 caractères sont constants (la jonction « I / H » ou « H / I » est commune à tous les taxons et la jonction « A / L » est présente chez tous les taxons qui ont le gène « L » *i.e.* les taxons 7 et 8, les autres taxons ont un point d'interrogation pour cette jonction). L'analyse exhaustive de cette matrice a produit deux arbres équiparcimonieux d'une longueur de 50 pas avec un IC de 0,7600 et IR de 0,6129. La topologie du consensus strict de ces arbres est présentée figure 3.8.

3.4.2.2 Option AB≠BA

Nous avons codé les huit génomes théoriques de la figure 3.3 avec le codage « Jonctions » et l'option AB≠BA, la matrice (voir en annexe, page 173) est composée de 54 caractères binaires,

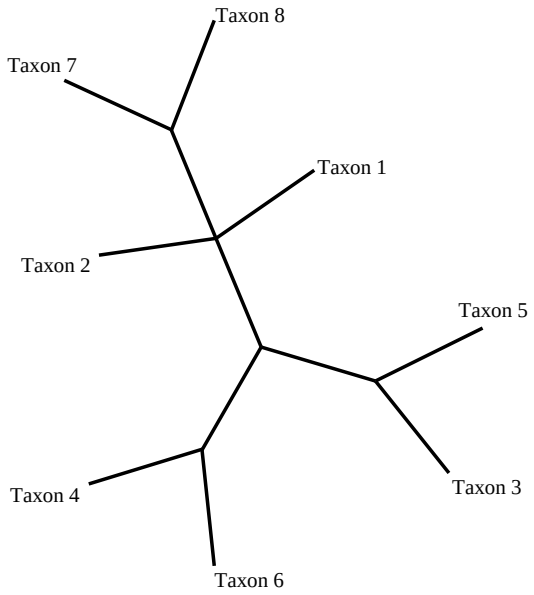


FIGURE 3.8 – Arbre de consensus strict des deux arbres équiparcimonieux obtenus à partir de la matrice théorique codée avec le codage « Jonctions » et l'option AB=BA. Longueur de l'arbre le plus court égale à 50 pas, IC = 0,7600 IR = 0,6129.

dont 12 caractères de polarité, 3 caractères « Présence Absence du gène » et 39 caractères de « Présence Absence d'une jonction ». Soit, 7 caractères « Présence Absence d'une jonction » supplémentaires par rapport au codage « Jonctions » avec l'option AB=BA. Cela signifie que les jonctions opposées qui étaient considérées comme une seule jonction avec l'option AB=BA, sont ici considérées indépendamment. Les jonctions « I / H » et « H / I » sont représentées ici par 2 caractères « Présence Absence d'une jonction ». Parmi les 12 caractères « Polarité de gène », 3 sont informatifs et 7 sont constants (ceux des gènes A, B, C, D, E, F et L). Les trois caractères « Présence Absence du gène » sont informatifs. Parmi les 39 caractères « Présence Absence d'une jonction », 19 caractères sont informatifs et 1 caractère est constant (celui la jonction « A / L »). L'analyse exhaustive de cette matrice a fourni deux arbres d'une longueur de 59 pas avec un IC et un IR de respectivement 0,7797 et 0,6486. La topologie du consensus strict de ces arbres est présentée figure 3.9.

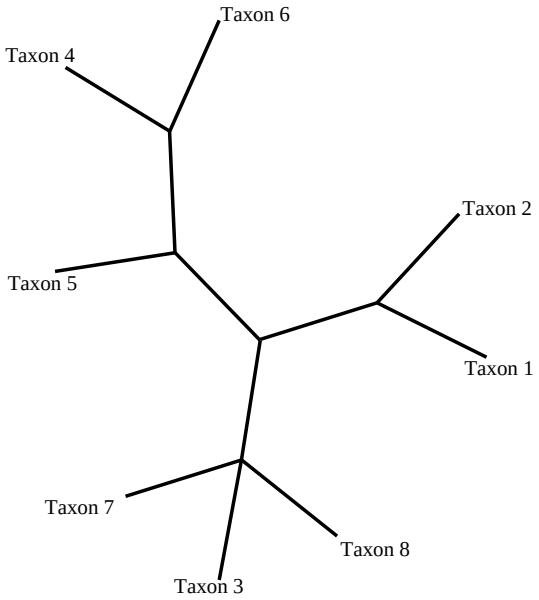


FIGURE 3.9 – Arbre de consensus strict des deux arbres équiparcimonieux obtenus à partir de la matrice théorique codée avec le codage « Jonctions » et l'option AB≠BA. Longueur de l'arbre le plus court égale à 59 pas, IC = 0,7797 IR = 0,6486.

3.4.3 « Jonctions signées »

Les génomes de la figure 3.3 sont maintenant codés avec le codage « Jonctions signées », la matrice (voir en annexe, page 174) comprend 8 taxons et 40 caractères binaires de « Présence Absence d'une jonction signée », dont 21 sont informatifs et aucun n'est constant. L'analyse exhaustive de cette matrice a produit un arbre d'une longueur de 52 pas avec un IC de 0,7692 et IR de 0,6250. La topologie de cet arbre est présentée figure 3.10.

Comme avec le codage « Jonctions », le « Jonctions signées » ne fournit pas assez d'informations au niveau des nœuds internes pour pouvoir reconstituer des génomes ancestraux. Il n'a donc pas été possible de reconstituer les génomes des ancêtres hypothétiques communs sur cet arbre.

3.4.4 Comparaisons des résultats

L'analyse d'un même jeu de données théoriques avec les différents codages est intéressante pour pouvoir comparer leur comportement. Le tableau synoptique 3.15 présente le nombre

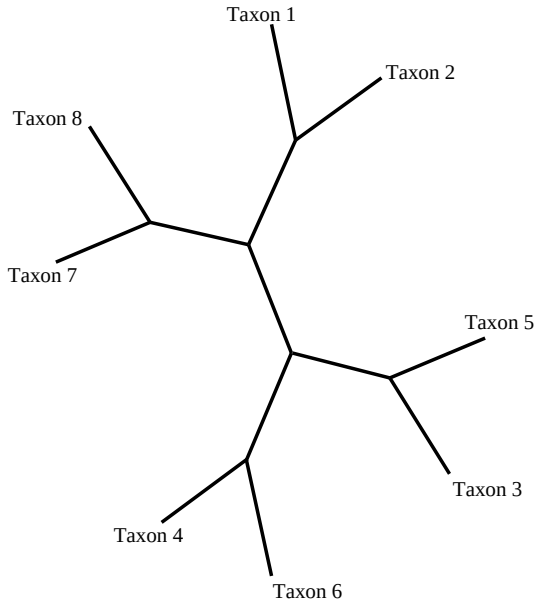


FIGURE 3.10 – Arbre le plus court obtenu à partir de la matrice théorique codées avec le codage « Jonctions signées ». Longueur de l'arbre : 52 pas, IC = 0,7692 IR = 0,6250.

d'arbres obtenus avec chacun des codages et options ainsi que les indices de cohérence et de rétention de ces arbres.

Codage	Option	Nb de caractères	Nb d'arbres	L	IC	IR
Position relative	AB=BA	27	1	53	0,9434	0,8235
Position relative	AB≠BA	27	3	54	0,9815	0,9375
Jonctions	AB=BA	47	2	50	0,7600	0,6129
Jonctions	AB≠BA	54	2	59	0,7797	0,6486
Jonctions signées	pas d'option	40	1	52	0,7692	0,6250

TABLEAU 3.15 – Résultats des analyses réalisées sur les génomes théoriques de la figure 3.3. Nombre d'arbres obtenus avec leurs indices de cohérence et de rétention.

Les meilleures valeurs d'indices de cohérence et de rétention sont obtenues avec le codage « Position relative », les valeurs des codages « Jonctions » et « Jonctions signées » sont équivalentes et un peu plus faibles que celles du codage « Position relative ». Ce sont des valeurs élevées qui indiquent une faible quantité d'homoplasie, les valeurs de IC sont toujours plus élevées que les valeurs de IR. La comparaison des valeurs obtenues avec les options AB=BA et AB≠BA montre un « avantage » pour l'option AB≠BA dont les valeurs sont meilleures. Les mêmes génomes, codés de plusieurs façons, indiquant des taux d'homoplasie différents, cela

suggère que le codage introduit de façon plus ou moins sensible l'homoplasie dans la matrice. Le codage « Position relative » introduit donc moins d'homoplasie que les deux autres codages. De la même manière l'option AB≠BA introduit moins d'homoplasie que l'option AB=BA.

Y a-t-il plus d'autapomorphies induites par le codage « Position relative » que par le codage « Jonctions » ? Avec le codage « Jonctions » et l'option AB=BA 13 des 32 jonctions ne sont présentes que chez un seul taxon. Avec le codage « Position Relative » sur les 82 positions différentes de gènes (12 gènes multipliés par 8 taxons moins les 14 points d'interrogations), 54 sont différentes et 35 ne sont présentes que chez un seul taxon. Cela donne un taux de 13/32 = 0,41 pour le codage « Jonctions » et un taux de 35/82 = 0,42 pour le codage « Position relative ». Grâce à cette mesure, nous pouvons conclure que le taux d'autapomorphie est équivalent pour les deux codages.

Les caractères « Position relative du gène » et « Présence Absence d'une jonction » se comportent différemment sur l'arbre. Prenons par exemple l'arbre obtenu avec le codage « Position relative » et l'option AB=BA (figure 3.4) et le consensus des arbres obtenus avec le codage « Jonctions » et l'option AB=BA (figure 3.8). Ces arbres diffèrent par les groupements (taxon 4, (taxon 5, taxon 6)) et (taxon 3, (taxon 7, taxon 8)) dans le premier cas et le groupement ((taxon 3, taxon 5), (taxon 4, taxon 6)) dans le deuxième cas. Le nœud (taxon 3, taxon 5) est soutenu par les transformations non ambiguës suivantes : perte de « A / J », gain de « C / F », gain de « D / J », perte de « E / F ». Le nœud (taxon 3, (taxon 7, taxon 8)) est soutenu par le passage, non ambiguë, pour le gène « H » de « - » à « + ». Le nœud (taxon 4, taxon 6) est soutenu par les transformations non ambiguës suivantes : gain de « C / D » alors que le nœud (taxon 4, (taxon 5, taxon 6)) n'est soutenu par aucune transformation non ambiguë. Nous pouvons remarquer qu'avec le codage « Jonctions », les branches internes sont soutenues par des changements de positions non ambigus, alors que ce n'est pas le cas avec le cas avec le codage « Position relative ». Les transformations liées à la position relative sont donc réparties sur plusieurs branches et les transformations non ambiguës n'apparaissent que sur les branches terminales. Cela explique l'écart important qu'il peut y avoir entre le nombre minimal et le nombre maximal de transformations

sur les branches de l'arbre de la figure 3.5. Ce comportement est sans doute lié au nombre élevé de positions de gènes qui ne sont présentes que chez un taxon (voir le paragraphe précédent).

Comparons maintenant la topologie des arbres obtenus avec les différents codages. Le tableau 3.16 montre le nombre de nœuds communs aux différents résultats. Il y a beaucoup plus de congruence entre les différents arbres obtenus avec le codage « Position relative » et les deux options qu'entre les arbres obtenus avec le codage « Jonctions » et les deux options. En effet, 5 nœuds sur 6 sont communs aux quatre arbres obtenus avec le codage « Position relative » alors qu'il n'y a que 2 nœuds communs aux quatre arbres obtenus avec le codage « Jonctions ». Par conséquent, la différence entre l'option AB=BA et l'option AB≠BA est plus importante avec le codage « Jonctions ». Ce n'est pas surprenant, si l'on considère qu'il n'y a que trois gènes pour lesquels il existe une position inversée alors qu'il y a 7 couples de jonctions inversées.

	PosR AB=BA	PosR AB≠BA	Jonc AB=BA	Jonc BA≠BA	JoncSig
PosR AB=BA	6				
PosR AB≠BA	5	5			
Jonc AB=BA	2	2	5		
Jonc AB≠BA	4	3	2	5	
JoncSig	3	2	5	3	6

TABLEAU 3.16 – Nombre de nœuds communs aux arbres obtenus avec les différents codages. Le nombre maximal est toujours de 6. (PosR : codage « Position relative », Jonc : codage « Jonctions », JoncSig : codage « Jonctions signées »).

Bizarrement, les arbres obtenus avec le codage « Jonctions » et l'option AB≠BA sont plus congruents avec ceux obtenus avec le codage « Position relative » et l'option AB=BA qu'avec les autres arbres. Les arbres obtenus avec le codage « Jonctions signées » sont plus congruents avec ceux obtenus avec le codage « Jonctions » qu'avec ceux obtenus avec le codage « Position relative ». Enfin, sur l'ensemble des arbres la congruence est nulle, aucun nœud n'est commun à l'ensemble des résultats. Ce résultat très fort montre à quel point le choix du codage influe sur le résultat.

3.5 Comparaison

3.5.1 Coûts

Les implications des trois codages et des options AB=BA et AB≠BA en terme de coûts des différents opérateurs sont variables. Les coûts *a priori* des différents opérateurs sont présentés dans les tableaux 3.3, 3.4, 3.6, 3.7 et 3.9 situés respectivement aux pages : 42, 42, 50, 50 et 55.

Nous pouvons constater que l'option AB≠BA impose un coût bien plus élevé à l'inversion et à la transinversion, que ce soit avec le codage « Position relative » ou le codage « Jonctions ». Ce coût supplémentaire est dû à la redondance entre les caractères « Polarité du gène » et les caractères « Position relative du gène » ou « Présence Absence d'une jonction ». En effet, lors de l'inversion d'un bloc de gènes, la prise en compte des positions opposées ou des jonctions opposées de type AB BA entraîne un nombre de pas supplémentaires qui se rajoutent à ceux du changement des caractères de polarité. L'utilisation de l'option AB≠BA introduit une information qui est déjà codée par les caractères de polarité. Le codage de la même information de deux manières différentes introduit de la redondance dans la matrice, ce qui se traduit par un coût anormalement élevé. Pour éviter ce problème nous préconisons de ne pas utiliser l'option AB≠BA.

Les coûts impliqués par les codages « Jonctions » et « Position relative » sont très proches, surtout avec l'option AB=BA. Ces codages impliquent donc à peu près le même modèle du point de vue des opérateurs : le coût de l'inversion et de la transinversion est supérieur à celui du gain–perte, lui même supérieur à celui de la transposition (qui a un coût constant). Ces coûts sont moyennement élevés.

Le codage « Jonction signées » est le seul pour lequel les coûts des différents opérateurs sont presque tous constants, faibles et équivalents entre eux, ce qui fait que ce codage implique moins de présupposés évolutifs que les deux autres.

3.5.2 Caractères liés

En parcimonie les caractères sont considérés comme indépendants les uns des autres et traités en conséquence lors de l'analyse. Ainsi, lors de la reconstruction, chaque caractère peut subir une transformation sans que cela entraîne la transformation d'un autre caractère. Des caractères non indépendants ou caractères liés, subissent des transformations conjointes, une transformation de l'un entraînant la transformation de l'autre. L'utilisation de caractères liés risque de biaiser les résultats en soutenant artificiellement certains clades.

Pour le codage de l'ordre des gènes, il est difficile d'envisager un codage où les caractères ne soient pas liés. Par nature, les données ne sont pas indépendantes. Une solution consisterait à coder le rang de chaque gène au sein de l'ordre des gènes, de cette façon la position de chaque gène serait décrite indépendamment de celle des autres. Malheureusement, cette méthode comporte de nombreux problèmes, le plus important étant celui de l'homologie. Il est très hasardeux de postuler l'homologie des rangs. Prenons par exemple deux ordres identiques, si l'un des deux perd le premier gène alors aucun gène n'a le même rang dans les deux ordres. Il est donc difficile d'éviter le problème des caractères liés dans le codage de l'ordre des gènes. Nous avons d'ailleurs illustré ces liens entre les caractères lors du calcul des coûts, par exemple une inversion entraîne le changement des caractères de positions et de polarité alors que l'on s'attendrait à ce que seule la polarité change. La question est donc de savoir dans quelle mesure le lien entre caractères influe sur le résultat dans notre approche.

Le lien entre caractères se comporte en quelque sorte comme une pondération. Par exemple, trois caractères liés subissent des transformations aux mêmes nœuds sur l'arbre, cela revient à donner un poids de trois à un seul caractère et à favoriser les regroupements soutenus par ce caractère. Qu'en est-il des caractères de position dans un ordre de gènes ? Ici, tous les caractères sont codés de la même manière et sont tous liés. La pondération des liens est donc répartie uniformément entre tous les caractères. De plus, la position d'un gène A est liée à celle d'un gène B dans une partie de l'arbre mais pas sur tout l'arbre ; mais dans une partie différente de l'arbre elle sera liée à la position d'un gène C. Ainsi, chaque événement sur l'arbre est associé

à la transformation conjointe de plusieurs caractères mais à chaque fois c'est une combinaison différente de caractères qui change, minimisant de ce fait l'influence du lien entre caractères. Le codage « Jonctions signées » qui ne comporte qu'un seul type de caractère est, dans ce contexte, moins sensible au problème des caractères liés que les deux autres codages.

3.5.3 Conclusion

Nous avons présenté et comparé, dans ce chapitre, deux codages originaux de l'ordre des gènes ainsi que le codage de Cosner *et al.* (2000b). Ces trois codages remplissent leur objectif, c'est-à-dire reconstruire une phylogénie par la méthode cladistique à partir de l'ordre des gènes sans identifier *a priori* les réarrangements. Ils présentent chacun des avantages et des inconvénients.

Les codages entièrement fondés sur la présence/absence (« Jonctions » et « Jonctions signées ») impliquent une redondance car à chaque fois qu'une jonction disparaît une autre apparaît forcément. Cependant, la redondance est la plus problématique avec les codages « Position relative » et « Jonctions » et l'option AB$\neq$BA, en effet, cette option implique une redondance importante entre les caractères de position (« Position relative du gène » ou « Présence Absence d'une jonction ») et les caractères « Polarité du gène ». Pour cette raison nous préconisons de ne pas employer l'option AB$\neq$BA.

Les temps de calcul sont beaucoup plus courts avec les codages « Jonctions » et « Jonctions signées » et conjointement le nombre d'arbres équiparcimonieux trouvés avec ces deux codages est beaucoup moins élevé qu'avec le codage « Position relative ». Ce qui représente un avantage lorsque le nombre de taxons devient élevé.

Le codage « Jonctions signées » présente deux avantages par rapport aux deux autres codages. Tout d'abord il est moins sensible aux problèmes des caractères liés et d'autre part il ne nécessite pas que l'orientation des unités fonctionnelles soit déterminée de façon absolue puisqu'il considère la position et l'orientation conjointement et de façon relative. De plus le modèle impliqué par ce codage est plus neutre, par contre, avec les codages « Position relative »

et « Jonctions » il est possible de jouer sur le modèle impliqué *a priori* en donnant des poids différents aux caractères de position et d'orientation. Néanmoins, il nous semble difficile de justifier un modèle de pondération plutôt qu'un autre.

Le codage « Position relative » présente un gros avantage sur les deux autres codages, c'est la possibilité de reconstituer les génomes ancestraux sur les arbres obtenus, ce qui n'est pas possible avec les codages « Jonctions » et « Jonctions signées ».

Ces trois codages ont été automatisés avec le programme **OrdGen**[7] qui permet d'obtenir une matrice au format NEXUS (format du logiciel PAUP) directement à partir d'un fichier contenant la description des ordres de gènes à étudier.

[7]http ://lis.snv.jussieu.fr/~gallut/These/OrdGen.html

4

Applications

4.1 Génome mitochondrial des métazoaires

4.1.1 Introduction

La mitochondrie est l'organite où a lieu la respiration cellulaire, elle est présente chez presque tous les eucaryotes. La mitochondrie mesure de 1 à 10 μm, elle possède deux membranes, une membrane externe et une membrane interne comportant de nombreux replis qui délimitent la matrice mitochondriale où se déroulent la plupart des réactions métaboliques de la respiration. Une partie des protéines intervenant dans ces réactions est codée par le génome mitochondrial. Ce génome, indépendant du génome nucléaire, a une taille, un nombre de gènes et une organisation qui varient en fonction du groupe taxonomique considéré. Il est généralement constitué d'une seule molécule d'ADN circulaire et contient au maximum une centaine de gènes. L'ordre des gènes du génome mitochondrial est variable et a conduit certains auteurs à utiliser ce type de données pour la phylogénie, notamment chez les animaux et les champignons (Sankoff *et al.*, 1992). Chez les plantes, le génome mitochondrial est constitué de populations de molécules qui recombinent entre elles formant un équilibre dynamique (Atlan et Couvet, 1993), conduisant à un taux de réarrangement élevé. Du fait de ce mode d'évolution, le génome mitochondrial des plantes est probablement mal adapté à une étude phylogénétique basée sur

l'ordre des gènes. Par contre, chez les métazoaires la conservation de l'ordre sur large échelle phylogénétique laisse envisager l'étude de la phylogénie profonde des métazoaires à partir du génome mitochondrial.

L'ordre des gènes du génome mitochondrial a déjà été utilisé pour étudier les relations de parenté entre classes d'échinodermes (Smith *et al.*, 1993) et chez les arthropodes (Boore *et al.*, 1995, 1998). Ces études sont basées sur la reconnaissance visuelle de synapomorphies impliquant un fragment d'ordre dérivé pour soutenir les groupes. Bien que ces analyses soient empiriques elles témoignent de l'intérêt du génome mitochondrial pour la phylogénie.

Les relations phylogénétiques des métazoaires ont fait l'objet de très vifs et très nombreux débats. Nous présenterons ici une vision simplifiée afin de replacer les groupes de métazoaires que nous aborderons par la suite. La phylogénie classique des métazoaires place les animaux diploblastiques à deux feuillets embryonnaires (cnidaires) à la base, en position de groupe frère des triploblastiques (animaux à trois feuillets embryonnaires). Au sein des triploblastiques, les acœlomates (plathelminthes) dont le troisième feuillet embryonnaire ne comporte pas de cœlome, sont groupe frère des pseudocœlomates (nématodes) eux-mêmes groupe frère des cœlomates (dont le troisième feuillet embryonnaire est creusé d'un cœlome). Les cœlomates sont divisés en protostomiens (annélides, arthropodes, mollusques etc.) et deutérostomiens (échinodermes, hémichordés, chordés). Ce point de vue classique est une représentation gradiste de la phylogénie des métazoaires qui structure les organismes du plus simple au plus évolué. La vision moderne, basée principalement sur l'ARN 18S (Turbeville *et al.*, 1992 ; Halanych *et al.*, 1995 ; Aguinaldo *et al.*, 1997), place les cnidaires comme groupe frère des bilatériens (animaux à symétrie bilatérale). Les bilatériens regroupent les deutérostomiens et les protostomiens (regroupant en plus les taxons placés autrefois dans les acœlomates et pseudocœlomates, les plathelminthes et les nématodes). Bien que la phylogénie des protosomiens soit encore loin d'être définitive, on reconnaît deux clades : les ecdysozoaires, animaux qui muent (nématodes et arthropodes) groupe frère des lophotrochozoaires (plathelminthes, mollusques, annélides, brachiopodes). Voir la figure 4.1.

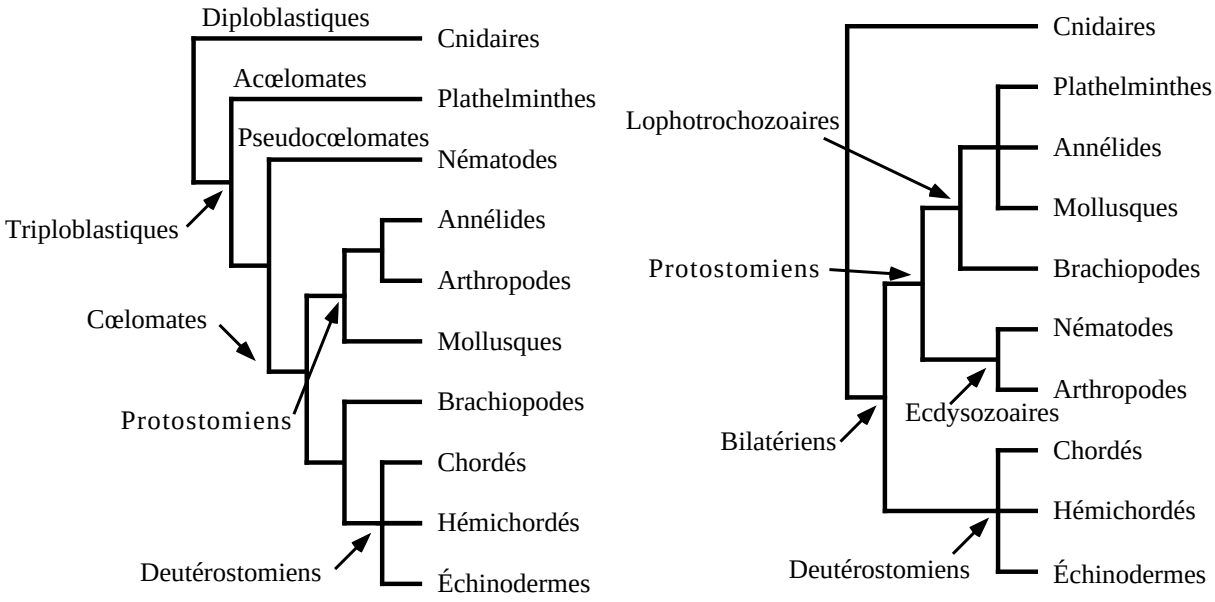


FIGURE 4.1 – Phylogénie des métazoaires : à gauche, point du vue classique, à droite, point de vue actuel.

4.1.2 Le génome mitochondrial des métazoaires

Les grandes caractéristiques du génome mitochondrial sont assez conservées chez les métazoaires. Ce génome est généralement constitué d'une seule molécule d'ADN circulaire, de petite taille. Chez les espèces dont le génome a été complètement séquencé, la taille varie de 13 Kb chez le ténia (*Taenia crassiceps*) (Le *et al.*, 2000) à 20 Kb chez la drosophile (*Drosophila melanogaster*) (Garesse, 1988). Cependant La Roche *et al.* (1990) ont montré que la taille du génome mitochondrial pouvait atteindre 40 Kb chez la coquille St Jacques (*Placopecten magellanicus*), en raison de répétitions dans des zones non codantes. Malgré quelques exceptions, la tendance générale chez les métazoaires est d'avoir un génome mitochondrial très compact, comportant peu de séquences intergéniques. Le chevauchement de gènes contigus avec un décalage du cadre de lecture est même très fréquent, ce chevauchement pouvant aller jusqu'à 30 paires de bases. Le contenu en gènes de l'ADN mitochondrial (ADNmt) est très conservé chez les animaux et se compose (figure 4.2) de :

- treize gènes codant pour des sous unités de protéines des complexes de l'oxydation phosphorylante :

  - cytochrome c oxydase sous unités I à III (cox1-3),
  - cytochrome b apoenzyme (cob),
  - NADH déshydrogénase sous unités 1 à 6 et 4L (nad1-6, 4L),
  - ATP synthase sous unités 6 et 8 (atp6 et atp8 ou encore A6 et A8).
- deux sous unités d'ARN ribosomique :
  - petite et grande sous unités (rns et rnl).
- vingt-deux ARN de transfert utilisés dans la traduction des protéines codées par l'ADNmt.

Chez certaines espèces, ces gènes peuvent être tous transcrits à partir du même brin de la molécule, ou bien à partir de l'un et de l'autre des brins. L'ADNmt comporte parfois une, voire deux (rarement plus) régions non codantes de grande taille (jusqu'à plusieurs Kb). Chez les vertébrés, cette région joue un rôle dans l'initiation de la réplication du brin principal ainsi que dans l'initiation de la transcription ; cette région est appelée région de contrôle. Chez la plupart des arthropodes, il existe aussi une région qui joue ce rôle, appelée région riche en AT (à cause de sa teneur élevée en AT) ou bien encore région de contrôle. Il est à noter que bien qu'elles aient un rôle similaire chez les vertébrés et les arthropodes ces régions n'ont pas d'identité de séquence et ne peuvent donc pas être considérées comme homologues. Enfin, la plupart des vertébrés ont une séquence capable de former une structure tige boucle à l'état de simple brin, structure qui sert de site d'initiation pour la réplication du brin opposé.

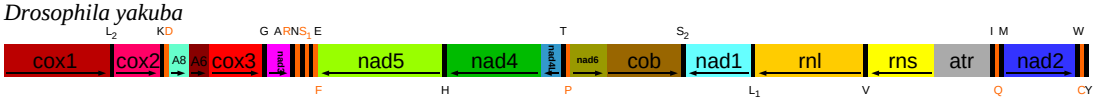


FIGURE 4.2 – Carte linéarisée du génome mitochondrial de *Drosophila yakuba*. cox1-3 (cytochrome c oxydase sous unités I à III), cob (cytochrome b apoenzyme), nad1-6, 4L (NADH déshydrogénase sous unités 1 à 6 et 4L), A6 et A8 (ATP synthase sous unités 6 et 8), rns et rnl (petite et grande sous unités d'ARN ribosomique), les ARNs de transfert sont nommés par le code à une lettre de l'acide aminé dont ils sont spécifiques, atr (région riche en AT). Les ARNt placés au-dessus de la carte sont transcrits dans le sens direct, ceux placés en dessous sont transcrits dans le sens opposé. Pour les gènes de grande taille une flèche indique le sens de transcription. En gris : région non codante.

Par rapport à ce schéma général nous pouvons remarquer quelques exceptions. Les cnidaires présentent les caractéristiques les plus divergentes. Les trois espèces de cnidaires dont le génome mitochondrial complet est connu (*Metridium senile, Sarcophyton glaucum, Renilla kolikeri*, classe Anthozoaires) n'ont qu'un ou deux ARNs de transfert sur les vingt-deux présents chez les autres espèces (Beagley *et al.*, 1998 ; Beaton *et al.*, 1998). D'autre part, l'anémone de mer (*Metridium senile*) est la seule espèce de métazoaire dont certains gènes mitochondriaux contiennent des introns (Beagley *et al.*, 1996, 1998) ; de plus cette espèce contient une séquence qui coderait pour une endonucléase. Le corail mou, *Sarcophyton glaucum*, possède une séquence qui serait l'homologue du gène bactérien MutS et Pont-Kingdon *et al.* (1998) suggèrent que ce gène serait issu d'un transfert récent du noyau vers la mitochondrie et que sa présence implique une activité de réparation des erreurs de réplication dans la mitochondrie de ce cnidaire, ce qui en fait un cas unique chez les animaux. Dernière exception importante trouvée chez les cnidaires : contrairement aux anthozoaires, les trois autres classes de cnidaires (scyphozoaires, cubozoaires et hydrozoaires) possèdent un génome mitochondrial linéaire constitué d'une ou de deux molécules (Bridge *et al.*, 1992 ; Pont-Kingdon *et al.*, 2000). Les autres phylums présentent aussi des exceptions par rapport au schéma général d'organisation du génome, néanmoins elles ne sont pas aussi importantes que celles trouvées chez les cnidaires. Les plathelminthes, les nématodes (sauf *Trichinella spiralis*), l'urochordé (*Halocynthia roretzi*) et les bivalves (*Mytilus edulis* et *Crassostrea gigas*) n'ont pas le gène atp8 codant pour l'ATP synthase sous unité 8. *Crassostrea gigas* est la seule espèce à posséder deux copies de la grande sous unité ribosomique (rnl). *Dinodon semicarinatus* possède deux copies de la région de contrôle, il est intéressant de noter que ces deux copies ont eu une évolution concertée depuis leur duplication, comme l'ont montré Kumazawa *et al.* (1998), leur séquence est identique sur plus de 1 Kb. Les annélides, les brachiopodes, *Metridium senile* (Cnidaires), *Crassostrea gigas* et *Mytilus edulis* (Mollusques), les nématodes (sauf *Trichinella spiralis*), les plathelminthes et l'urochordé ont tous leurs gènes transcrits dans le même sens. Les autres espèces ont une partie de leurs gènes transcrits dans un sens et l'autre partie dans le sens oppposé.

Phylum	**Nombre d'espèces**	**Nombre d'ordres différents**	**Nombre d'espèces total***
Annélides	2	2	15 000
Arthropodes	22	12	1 000 000
Brachiopodes	2	2	335
Chordés			
Céphalochordés	2	1	23
Crâniates	109	8	47 000
Urochordés	1	1	3 000
Cnidaires	3	2	9 000
Échinodermes	5	3	7 000
Hémichordés	1	1	85
Mollusques	8	8	50 000
Nématodes	5	4	12 000
Plathelminthes	8	5	20 000

TABLEAU 4.1 – Répartition des génomes complets parmi les différents phylums (*source : Brusca et Brusca (1990)).

4.1.3 Échantillonnage

4.1.3.1 Génomes complets

Nous disposions fin juin 2001 de 188 génomes mitochondriaux complets ce qui représente 168 espèces différentes (voir la liste complète en annexe, tableau A.1) dont 7 pour lesquelles la séquence du génome est incomplète. Il est important de noter que ces 168 génomes ne représentent que dix phylums de métazoaires parmi la trentaine de phylums connus. De plus, la répartition entre ces dix phylums est très inégale, en effet un peu plus des deux tiers des génomes complets appartiennent à des espèces de chordés. Le tableau 4.1 récapitule la répartition des génomes mitochondriaux connus entre les dix phylums. Il y a plusieurs explications à ce déséquilibre, tout d'abord les motivations pour séquencer un génome mitochondrial sont diverses : étude d'un système génétique simple, phylogénie des mammifères etc. D'autre part, l'obtention d'une nouvelle séquence est d'autant plus facile que l'on connaît des espèces proches, ainsi plus un groupe est étudié plus il est simple d'obtenir un nouveau génome au sein de ce groupe.

Ces données proviennent principalement des banques de séquences mais aussi de la littérature. Le travail d'acquisition a nécessité de nombreuses corrections des annotations de sé-

quences, en effet ces dernières sont souvent erronées et incomplètes. De plus, un même gène peut être dénommé de nombreuses façons différentes selon les auteurs des séquences. Il a donc été nécessaire de corriger les annotations et d'uniformiser les noms de gènes afin de faciliter la comparaison. Les séquences ont été corrigées à partir des articles de description ou même réannotées à partir de l'analyse de la séquence lorsque les informations des différentes sources se sont révélées contradictoires. Nous avons utilisé, pour les noms de gènes, la nomenclature de GOBASE [8] (Shimko *et al.*, 2001), base de données spécialisée dans les génomes d'organites, cela permet d'avoir une nomenclature dont l'usage est plus large que le génome mitochondrial des métazoaires. Nous avons réalisé un programme qui permet une acquisition semi-automatique de l'ordre des gènes du génome mitochondrial des métazoaires à partir des fichiers EMBL. Il utilise une base de synonymie des noms de gènes de plus de 800 entrées (que nous avons réalisée) afin d'attribuer à chaque gène le même nom. Les numéros d'accession et les références bibliographiques des 49 espèces retenues sont présentés dans le tableau A.1 en annexe. À partir de ces données nous avons réalisé les cartes d'ordres de gènes des 49 espèces présentées figures A.1 à A.12.

Dans le cadre de cette étude sur le génome mitochondrial, nous avons participé à l'analyse de la séquence complète du génome mitochondrial de la roussette *Scyliorhinus canicula* de la lamproie *Lampetra fluviatilis* et de la myxine *Eptatretus burgeri*. Le génome mitochondrial de ces trois expèces a été séquencé au laboratoire de Biologie Molèculaire du Gène de l'Institut Pasteur par l'équipe de Gabriel Gachelin. Nous avons réalisé des analyses phylogénétiques intégrant ces nouvelles séquences complètes afin de préciser les relations phylogénétiques à la base des Crâniates (Voir articles en annexes).

4.1.3.2 Ordres de gènes

Parmi les 168 génomes complets, il existe 49 ordres de gènes différents, chacun de ces 49 ordres représente un groupe d'une ou plusieurs espèces. Pour chaque groupe nous avons retenu

[8] http://megasun.bch.umontreal.ca/gobase/

comme espèce représentative la première espèce du groupe à avoir été séquencée, le tableau 4.2 regroupe les 49 espèces représentatives d'un ordre de gène distinct. Les cartes de ces espèces sont présentées dans les figures A.1 à A.12. À l'exception de trois groupes (groupes représentés respectivement par *Homo sapiens*, *Limulus polyphemus* et *Drosophila yakuba*), tous les groupes ayant le même ordre de gènes constituent des ensembles monophylétiques. *Homo sapiens* est ici le représentant d'un groupe polyphylétique, en effet parmi les crâniates, certains actinoptérygiens, les lisamphibiens, les chondrichtyens, le cœlacanthe, les dipneustes, les myxines, les monotrèmes, les euthériens, les chéloniens et certains squamates ont le même ordre de gènes. Par contre les crocodiliens, les aves, les métathériens, l'autre squamate (*Dinodon semicarinatus*), les autres actinoptérygiens (*Conger myriaster* et *Gonostoma gracile*) et les lamproies possèdent tous un ordre de gènes différent. Si l'on considère que les crâniates sont monophylétiques, ces groupes présentent des ordres de gènes autapomorphiques par rapport à l'ordre de gènes du groupe représenté par *Homo sapiens*. De la même manière *Limulus polyphemus* représente son propre taxon ainsi qu'un arachnide (la tique *Ixodes hexagonus*). Les deux autres espèces de tiques (*Boophilus microplus* et *Rhiphicephalus sanguineus*) sont représentées par *Rhiphicephalus sanguineus*. *Drosophila yakuba* regroupe à la fois des diptères des hémiptères et des crustacés. Les diptères sont divisés en deux groupes représentés par *Anopheles gambiae* et *Drosophila yakuba* (groupe polyphylétique). Les crustacés sont divisés en trois : le branchiopode *Artemia franciscana*, le branchiopode *Daphnia pulex* et le malacostracé *Penaeus monodon* ont le même ordre de gène que *Drosophila yakuba* et enfin le malacostracé *Pagurus longicarpus* a lui aussi un ordre différent.

Classification			**Genre espèce**
Annélides	Oligochètes		*Lumbricus terrestris*
Annélides	Polychètes		*Platynereis dumerilii*
Arthropodes	Arachnides		*Rhiphicephalus sanguineus*
Arthropodes	Mérostomes	**Polyphylétique**	*Limulus polyphemus*
Arthropodes	Crustacés	Branchiopodes	*Artemia franciscana*
Arthropodes	Crustacés	Malacostracés	*Pagurus longicarpus*
Arthropodes	Hexapodes	Collemboles	*Tetrodontophora bielanensis*
Arthropodes	Hexapodes	Insectes	*Anopheles gambiae*
Arthropodes	Hexapodes	Insectes	*Apis mellifera*
Arthropodes	Hexapodes	Insectes	*Bombyx mori*
Arthropodes	Hexapodes	**Polyphylétique**	*Drosophila yakuba*
Arthropodes	Hexapodes	Insectes	*Heterodoxus macropus*
Arthropodes	Hexapodes	Insectes	*Locusta migratoria*
Arthropodes	Myriapodes		*Lithobius forficatus*
Brachiopodes	Articulés		*Terebratulina retusa*
Brachiopodes	Articulés		*Laqueus rubellus*
Chordés	Céphalochordés		*Branchiostoma lanceolatum*
Chordés	Crâniates	Actinoptérygiens	*Conger myriaster*
Chordés	Crâniates	Actinoptérygiens	*Gonostoma gracile*
Chordés	Crâniates	Aves	*Gallus gallus*
Chordés	Crâniates	Crocodiliens	*Alligator mississippiensis*
Chordés	Crâniates	**Polyphylétique**	*Homo sapiens*
Chordés	Crâniates	hyperoarti	*Petromyzon marinus*
Chordés	Crâniates	Métathériens	*Didelphis virginiana*
Chordés	Crâniates	Squamates	*Dinodon semicarinatus*
Chordés	Urochordés		*Halocynthia roretzi*
Cnidaires	Anthozoaires		*Metridium senile*
Cnidaires	Anthozoaires		*Sarcophyton glaucum*
Échinodermes	Astérides		*Asterina pectinifera*
Échinodermes	Crinoïdes		*Florometra serratissima*
Échinodermes	Échinides		*Strongylocentrotus purpuratus*
Hémichordés	Entéropneustes		*Balanoglossus carnosus*
Mollusques	Bivalves		*Crassostrea gigas*
Mollusques	Bivalves		*Mytilus edulis*
Mollusques	Céphalopodes		*Loligo bleekeri*
Mollusques	Gastéropodes		*Albinaria coerulea*
Mollusques	Gastéropodes		*Cepaea nemoralis*
Mollusques	Gastéropodes		*Euhadra herklotsi*
Mollusques	Gastéropodes		*Pupa strigosa*
Mollusques	Polyplacophores		*Katharina tunicata*
Nématodes	Adenophores		*Trichinella spiralis*
Nématodes	Secernentes		*Ascaris suum*
Nématodes	Secernentes		*Meloidogyne javanica*
Nématodes	Secernentes		*Onchocerca volvulus*
Plathelminthes	Cestodes		*Echinococcus multilocularis*
Plathelminthes	Cestodes		*Hymenolepis diminuta*
Plathelminthes	Trématodes		*Fasciola hepatica*
Plathelminthes	Trématodes		*Schistosoma japonicum*
Plathelminthes	Trématodes		*Schistosoma mansoni*

TABLEAU 4.2 – Liste des espèces représentatives d'un ordre de gènes.

4.1.3.3 Groupe extérieur

Le groupe frère des eumétazoaires (cnidaires plus bilatériens) sont les porifères (ou éponges) mais malheureusement aucun génome mitochondrial complet d'éponge n'est disponible pour l'instant. Les cnidaires, comme nous l'avons vu, présentent un génome très différent de celui des bilatériens : nombre de gènes beaucoup plus faible, présence d'introns etc. Comme les cnidaires sont groupe frère des bilatériens nous les utiliserons pour enraciner les arbres obtenus mais il est important de garder à l'esprit que les caractéristiques très divergentes des deux cnidaires (*Metridium senile* et *Sarcophyton glaucum*) font que ce ne sont pas les candidats rêvés comme groupe extérieur. En effet, en qualité de groupe extérieur, il est intéressant d'avoir un taxon dont on peut faire l'hypothèse qu'il est extérieur au groupe étudié mais dont les caractéristiques ne sont pas trop éloignées, car dans ce cas, cela risque d'introduire plus d'incertitude que cela n'en lève.

4.1.4 Analyses

Les objectifs de ces analyses sont multiples :

- peut-on apporter un éclairage sur la phylogénie des métazoaires à partir de l'ordre des gènes du génome mitochondrial ?
- notre méthode de codage nécessite-t-elle des adaptations pour se conformer à ces données ?
- peut-on tirer des informations sur les avantages et les inconvénients des différents codages ?

Nous avons retenu pour l'analyse l'ensemble des gènes homologues et éliminé les régions non codantes, les régions dont l'homologie est douteuse (région de contrôle, région riche en AT), les pseudo-gènes et les gènes dont l'identification est douteuse. Pour les gènes dupliqués, nous avons éliminé la copie, selon les auteurs des séquences, qui n'est pas l'orthologue (notée par un 2 à la fin du nom sur les figures). Ce qui représente au total 37 gènes différents.

4.1.4.1 Analyses préliminaires

La simple comparaison à l'œil des cartes d'ordre de gènes du génome mitochondrial des métazoaires permet de faire quelques observations préliminaires.

- Premièrement, l'ordre semble plus conservé à l'intérieur des phylums qu'entre phylums.
- Deuxièmement, les gènes de petites tailles *i.e.* les ARNs de transfert, semblent beaucoup plus mobiles que les gènes de grande taille.
- Troisièmement, certaines espèces présentent un ordre de gènes très remanié par rapport aux espèces du groupe auxquel elles appartiennent, voir même par rapport à l'ensemble des espèces.

Ces observations nous ont conduit, dans un premier temps, à réaliser une analyse exploratoire des données afin de quantifier de façon objective ces observations. Pour cette analyse exploratoire nous avons utilisé une mesure de dissimilarité. Cette mesure doit être indépendante des résultats obtenus avec les différents codages pour ne pas fausser l'interprétation. Par conséquent, nous avons utilisé la distance de points de cassure (Sankoff et Blanchette (1998) et voir § 1.2.4, page 18).

Le nombre de points de cassure d'un ordre i vers un ordre j est égal au nombre de jonctions signées de gènes présentes dans l'ordre i qui sont absentes dans l'ordre j. De la même manière, le nombre de points de cassure de l'ordre j vers l'ordre i est égal au nombre de jonctions signées de gènes présentes dans l'ordre j qui sont absentes dans l'ordre i. Si les deux ordres i et j n'ont pas le même nombre de gènes, le nombre de points de cassure de i vers j n'est pas égal au nombre de points de cassure de j vers i. Par conséquent, si i contient plus de gènes que j, le nombre de points de cassure de i vers j est supérieur à celui de j vers i. Notons que les nombres de points de cassure de i vers et de j vers i sont identiques même si les deux ordres n'ont pas la même composition en gènes. Seul le nombre de gènes a une influence sur la symétrie des valeurs des nombres de points de cassure. Nous avons corrigé l'influence du nombre de gènes en normant le nombre de points de cassure de i vers j par le nombre de gènes de i et de j vers i par le nombre de gènes de j. De façon à ce que cette valeur soit symétrique, nous avons fait la

moyenne de i vers j et de j vers i. Ainsi, la distance de points de cassure de i vers j est égale à la distance de j vers i ce qui est beaucoup plus satisfaisant.

Laissons la place à quelques définitions :

Soit i et j deux génomes différents appartenant à $\mathcal{G}$ l'ensemble des génomes étudiés. N est le nombre de génomes de $\mathcal{G}$. $phylum(i)$ est l'ensemble des génomes appartenant au même phylum que i. n_i est le nombre de génomes de $phylum(i)$. Soit k un phylum, n_k le nombre de génomes de k et K le nombre de phylums.

b_{ij} est le nombre de points de cassure de i vers j (Notons que $b_{ii} = 0$).

Soit $\bar{b}_{int(i)}$ la moyenne des points de cassure de i vers les autres génomes de son phylum :

$$\bar{b}_{int(i)} = \frac{\sum_{j=1}^{j=n_i} B_{ij}}{n_i - 1} \; tel \; que \; i \neq j \; et \; j \in phylum(i)$$

Soit $\bar{b}_{ext}$ la moyenne des points de cassure entre génomes de phylums différents :

$$\bar{b}_{ext} = \frac{\sum_{i=1}^{i=N} \sum_{j=1}^{j=N} B_{ij}}{N^2 - \sum_{k=1}^{k=K} n_k^2} \; tel \; que \; i \neq j \; et \; j \notin phylum(i)$$

Plus $\bar{b}_{int(i)}$ est faible, plus le génome i est semblable aux taxons de son phylum, par contre plus cette valeur est élevée, plus i est différent des autres génomes de $phylum(i)$ et plus il peut être considéré comme excentrique par rapport à son phylum. Pour l'ensemble des 49 espèces $\bar{b}_{ext} = 0,89$ cette valeur est élevée. Si nous tenons compte du fait que les taxons ont une moyenne de 35,96 gènes, cela donne en moyenne 32 points de cassure pour chaque couple d'espèces qui n'appartiennent pas au même phylum. Pour identifier les taxons qui sont plus dissemblables des membres de leur phylum que des membres des autres phylums, nous retiendrons ceux pour lesquels $\bar{b}_{int(i)}$ est supérieur à $\bar{b}_{ext}$. Le tableau 4.3 présente les valeurs de $\bar{b}_{int(i)}$ pour les 49 espèces. Neuf de ces dernières ont une valeur de $\bar{b}_{int(i)}$ supérieure à $\bar{b}_{ext}$, ce qui signifie que *Heterodoxus macropus*, *Halocynthia roretzi*, *Metridium senile*, *Sarcophyton glaucum*, *Crassostrea gigas*, *Katharina tunicata*, *Mytilus edulis*, *Loligo bleekeri* et *Trichinella spiralis*

ont des ordres de gènes très différent de ceux des espèces de leurs phylums respectifs. Nous considérerons par la suite, ces neufs espèces comme « excentriques ».

$\bar{\mathbf{b}}_{\mathbf{ext}} = \mathbf{0,89}$

L. terrestris	0,32	*M. senile*	**0,94**
P. dumerilii	0,32	*S. glaucum*	**0,94**
A. gambiae	0,36	*A. pectinifera*	0,12
A. mellifera	0,54	*F. serratissima*	0,19
A. franciscana	0,31	*S. purpuratus*	0,12
B. mori	0,30	*A. coerulea*	0,62
D. yakuba	0,27	*C. nemoralis*	0,66
H. macropus	**0,94**	*C. gigas*	**0,95**
L. polyphemus	0,30	*E. herklotsi*	0,64
L. forficatus	0,34	*K. tunicata*	**0,90**
L. migratoria	0,31	*L. bleekeri*	**0,93**
P. longicarpus	0,61	*M. edulis*	**0,98**
R. sanguineus	0,40	*P. strigosa*	0,66
T. bielanensis	0,34	*A. suum*	0,82
L. rubellus	0,78	*M. javanica*	0,88
T. retusa	0,78	*O. volvulus*	0,84
A. mississippiensis	0,30	*T. spiralis*	**0,99**
B. carnosus	0,45	*E. multilocularis*	0,22
B. lanceolatum	0,44	*F. hepatica*	0,24
C. myriaster	0,29	*H. diminuta*	0,27
D. virginiana	0,32	*S. japonicum*	0,27
D. semicarinatus	0,29	*S. mansoni*	0,45
G. gallus	0,27		
G. gracile	0,31		
H. roretzi	**0,97**		
H. sapiens	0,24		
P. marinus	0,28		

TABLEAU 4.3 – Valeurs de $\bar{b}_{int(i)}$ pour chaque taxon. En gras : valeurs qui dépassent $\bar{b}_{ext}$. Note : comme *Balanoglossus carnosus* est le seul taxon du phylum des Hémichordés, il a été placé au sein des chordés pour les calculs.

Le nombre moyen de points de cassure entre deux ordres aléatoires de n gènes est de $n - \frac{1}{2}$ (Blanchette *et al.*, 1999). Il est possible de voir si un taxon a un ordre de gène qui n'est pas discernable d'un ordre aléatoire par rapport aux autres taxons, si son nombre moyen de points de cassure calculé par rapport à tous les taxons atteint la valeur de $n-\frac{1}{2}$. Nous avons calculé pour chaque espèce le nombre moyen de points de cassure (non corrigé) par rapport à toutes les autres espèces et comparé cette valeur à leur nombre de gènes (voir tableau A.2 en annexe). Cette moyenne globale par espèce n'atteint jamais le seuil de $n - \frac{1}{2}$, néanmoins certaines espèces en sont très proches : *Crassostrea gigas* : 35,1, *Halocynthia roretzi* : 35,13, *Meloidogyne javanica* : 35,15 et *Onchocerca volvulus* : 35,31 ; pour 36 gènes chacune alors que *Heterodoxus macropus* et *Laqueus rubellus* ont un nombre moyen de points de cassure de 35,48 et 35,88 pour 37 gènes.

Cela signifie que les ordres de gènes de *Crassostrea gigas*, *Halocynthia roretzi* et *Heterodoxus macropus* sont non seulement éloignés de ceux des membres de leur propre phylum mais également de ceux de tous les taxons de l'échantillon en général. Par contre, bien que les valeurs de $\bar{b}_{int(i)}$ de *Onchocerca volvulus*, *Meloidogyne javanica* et *Laqueus rubellus* ne dépassent pas $\bar{b}_{ext}$ comme nous l'avons vu précédemment (tableau 4.3), ces trois espèces ont un nombre moyen de points de cassure par rapport à tous les taxons proche de $n - \frac{1}{2}$. Cela montre premièrement, que ces trois espèces sont plus semblables aux espèces de leur phylum qu'aux autres et deuxièmement, qu'elles sont également très dissemblables de l'ensemble de l'échantillon.

Calculons maintenant le nombre moyen de points de cassure par rapport aux espèces du même phylum ($\bar{b}_{int(i)}$ non corrigé). Trois espèces se révèlent avoir une valeur très élevée : *Halocynthia roretzi* (35 pour 36 gènes), *Mytilus edulis* (35,43 pour 36 gènes) et enfin le nombre moyen de points de cassure de *Trichinella spiralis* dépasse même le seuil de $n - \frac{1}{2}$ avec un nombre moyen de 36,67 pour 37 gènes ce qui signifie que l'ordre des gènes de ce nématode (classe Adenophores) ne peut pas être distingué d'un ordre aléatoire par rapport à l'ordre de gènes des autres nématodes (classe Secernentes). Notons que le nombre moyen de points de cassure de *Trichinella spiralis* calculé par rapport à tous les taxons est de 30,73 pour 37 gènes. Ces deux résultats sont corroborés par les analyses de Lavrov et Brown (2001) qui ont montré

que le génome mitochondrial de *Trichinella spiralis* a des caractéristiques « intermédiaires » entre celui des nématodes secernentes et celui des métazoaires cœlomates, notamment celui de *Limulus polyphemus*. La classe des adenophores est considérée comme « primitive » au sein des nématodes, elle est probablement paraphylétique. Lavrov et Brown considèrent que le génome de *Trichinella spiralis* est une mosaïque de caractères dérivés propres aux nématodes et de caractères de cœlomates, ce qui placerait les adenophores à la base des nématodes. De plus, ils ont observé que l'ordre des gènes du génome de *T. spiralis* a des fragments d'ordres conservés avec *L. polyphemus* alors qu'il n'en a pas avec celui des autres nématodes.

Ces analyses exploratoires nous ont permis de montrer que certains phylums comportent des espèces dont les ordres de gènes sont très divergents :

- Arthropodes : *Heterodoxus macropus* a un ordre de gènes très différent de ceux des autres arthropodes mais aussi de ceux de tous les autres taxons,
- Brachiopodes : *Laqueus rubellus* a un ordre de gènes presque aléatoire en comparaison avec ceux des taxons non brachiopodes,
- Chordés : *Halocynthia roretzi* a un ordre de gènes presque aléatoire en comparaison avec ceux des autres chordés mais aussi en comparaison avec ceux des autres taxons,
- Cnidaires : les deux cnidaires *Metridium senile* et *Sarcophyton glaucum* ont des ordres de gènes plus dissemblables entre eux que par rapport aux autres espèces, néanmoins du point de vue de leurs génomes, ils partagent l'absence de la majorité des ARNt,
- Mollusques : *Crassostrea gigas*, *Katharina tunicata*, *Mytilus edulis* et *Loligo bleekeri* ont un ordre de gènes très dissemblables entre eux et par rapport aux autres mollusques (les gastéropodes) ; de plus *Crassostrea gigas* a un ordre de gènes presque aléatoire en comparaison avec ceux des taxons non mollusques,
- Nématodes : l'ordre des gènes du nématode adenophore *Trichinella spiralis* n'est pas discernable d'un ordre aléatoire vis-à-vis de celui des nématodes secernentes ; d'autre part *Meloidogyne javanica* et *Onchocerca volvulus* ont des ordres de gènes presque aléatoires vis-à-vis des non nématodes.

4.1.4.2 Analyses phylogénétiques

a Échantillonnage complet

Ces analyses regroupent l'ensemble des taxons comportant un ordre de gènes différent, soit 49 groupes. Nous avons analysé cet échantillonnage avec les trois codages différents. Pour le codage en positions relatives et le codage en jonctions nous avons retenu l'option AB=BA puisque nous avons vu au chapitre précédent que l'option AB≠BA impliquait des biais trop importants (voir chapitre 3.5, page 84).

a.1 Échantillonnage complet avec le codage « Position relative »

La matrice [9] comporte 49 taxons et 96 caractères, répartis de la façon suivante : 37 caractères « Position relative du gène » (un pour chaque gène), 37 caractères « Polarité du gène » (un pour chaque gène) et 22 caractères « Présence Absence du gène » (un pour le gène atp8 et un pour les gènes des ARNs de transfert à l'exception de trnM qui est toujours présent). Les caractères « Position relative du gène » ont de 11 à 28 états différents. Les caractères « Polarité du gène » des gènes cox1 et trnK sont constants, le caractère « Polarité du gène cox2 » et le caractère de « Présence Absence du gène trnW » sont non informatifs. Tous les caractères sont traités non orientés, non ordonnés et ont un poids de 1. La recherche de l'arbre le plus court a été réalisée avec un algorithme heuristique (Branch swapping TBR avec 500 agrégations aléatoires des taxons) par le logiciel PAUP* (Swofford, 1998). Le nombre d'arbres équiparcimonieux trouvés étant extrêmement élevé (>180 000) l'analyse n'a pu être menée à son terme. Néanmoins cela n'affecte pas le résultat [10]. Les arbres équiparcimonieux ont une longueur de 931 pas, un IC = 0,9076 et un IR = 0,9036. Le consensus strict des arbres équiparcimonieux est présenté figure 4.3.

[9] disponible à l'adresse http://lis.snv.jussieu.fr/~gallut/These

[10] Lorsque le nombre d'arbres trouvés est élevé, l'analyse est très longue. Il alors difficile de faire plusieurs fois l'analyse pour trouver des arbres plus courts. Nous avons donc, réalisé une première heuristique en agrégeant les taxons de façon aléatoire (500 réplications) en ne conservant qu'un seul à chaque fois. Ce qui a permi de trouver des arbres plus courts qui nous ont servi ensuite de point de départ à une nouvelle analyse où cette fois nous avons conservé tous les arbres équiparcimonieux.

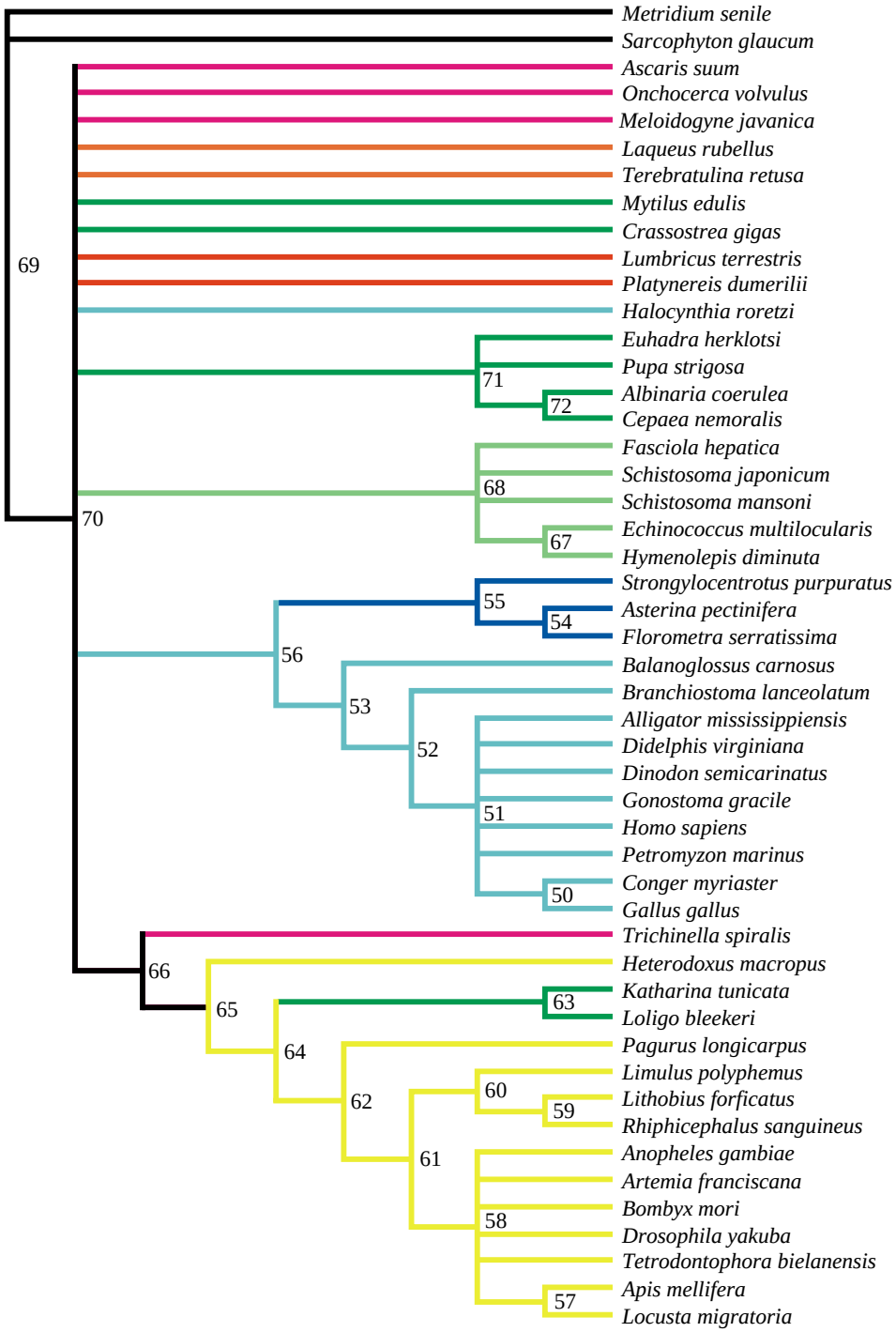


FIGURE 4.3 – Consensus strict des 180 000+ arbres équiparcimonieux obtenus avec la matrice complète et avec le codage « Position relative ». (Arbre le plus court : L = 931 pas, IC = 0,9076 IR = 0,9036). Groupe extérieur : cnidaires. Noire : cnidaires, rouge : annélides, jaune : arthropodes, orange : brachiopodes, bleu : deutérostomiens, bleu clair : chordés + hémichordés, bleu foncé : échinodermes, vert foncé : mollusques, rose : nématodes, vert clair : plathelminthes. Les chiffres correspondent aux numéros des nœuds.

a.2 Échantillonnage complet avec le codage « Jonctions »

La matrice [11] est constituée de 49 taxons et 473 caractères (414 caractères « Présence Absence d'une jonction », 37 caractères « Polarité du gène » et 22 caractères « Présence Absence du gène »). Les caractères « Présence Absence d'une jonction » des jonctions « atp6 / nad5(5') », « nad1 / nad5(5') », « nad3 / nad5(3') » et « nad5(3') / trnW » sont constants, ce qui est normal étant donné que seul *Metridium senile* a un intron dans le gène nad5. Les caractères « Polarité du gène » des gènes cox1 et trnK sont constants, et 156 caractères sont non informatifs. Tous les caractères sont traités non orientés, non ordonnés avec un poids de 1. La recherche de l'arbre le plus court a été réalisée avec un algorithme heuristique (Branch swapping TBR avec 500 agrégations aléatoires des taxons) par le logiciel PAUP* et a fourni 54 arbres équiparcimonieux. Ces derniers ont une longueur de 945 pas, un IC de 0,4942 et un IR de 0,7141. Le consensus strict des 54 arbres équiparcimonieux est présenté figure 4.4.

[11] disponible à l'adresse http://lis.snv.jussieu.fr/~gallut/These

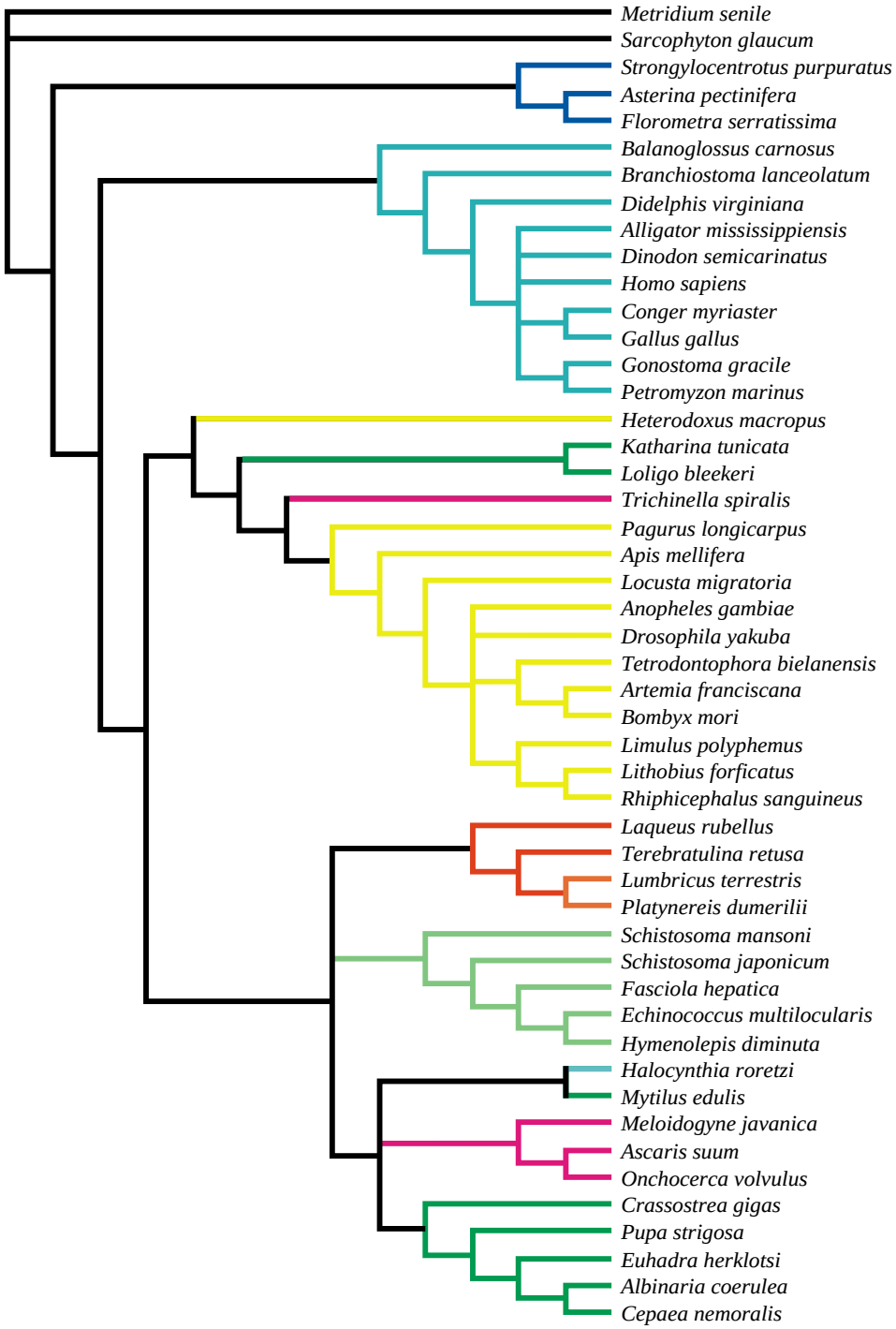


FIGURE 4.4 – Consensus strict des 54 arbres équiparcimonieux obtenus avec la matrice complète et avec le codage « Jonctions ». (Arbre le plus court : L = 945 pas, IC = 0,4942 IR = 0,7141). Groupe extérieur : cnidaires. Noire : cnidaires, rouge : annélides, jaune : arthropodes, orange : brachiopodes, bleu : deutérostomiens, bleu clair : chordés + hémichordés, bleu foncé : échinodermes, vert foncé : mollusques, rose : nématodes, vert clair : plathelminthes.

a.3 Échantillonnage complet avec le codage « Jonctions signées »

La matrice [12] est constituée de 49 taxons et 587 caractères (587 caractères « Présence Absence d'une jonction signée ») parmi lesquels 341 sont non informatifs. Tous les caractères sont traités non orientés, non ordonnés avec un poids de 1. La recherche de l'arbre le plus court a été réalisée avec un algorithme heuristique (Branch swapping TBR avec 500 agrégations aléatoires des taxons) par le logiciel PAUP* et a trouvé 16 740 arbres équiparcimonieux. Ces derniers ont une longueur de 862 pas, un IC de 0,6810 et un IR de 0,7571. Le consensus strict des arbres équiparcimonieux est présenté figure 4.5.

[12] disponible à l'adresse http://lis.snv.jussieu.fr/~gallut/These

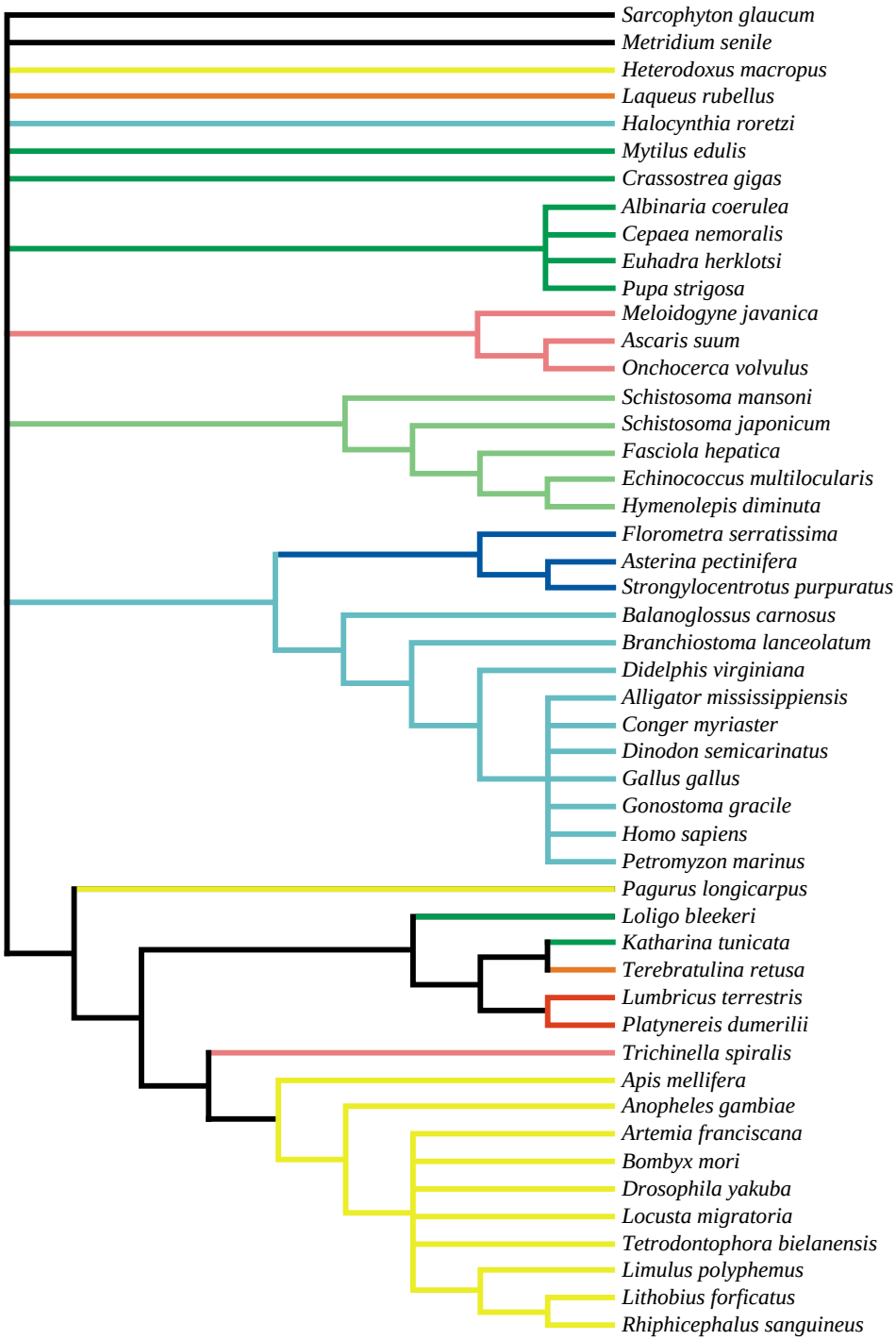


FIGURE 4.5 – Consensus strict des 16 740 arbres équiparcimonieux obtenus avec la matrice complète et avec le codage « Jonctions signées ». (Arbre le plus court : L = 862 pas, IC = 0,6810 IR = 0,7571). Groupe extérieur : cnidaires. Noire : cnidaires, rouge : annélides, jaune : arthropodes, orange : brachiopodes, bleu : deutérostomiens, bleu clair : chordés + hémichordés, bleu foncé : échinodermes, vert foncé : mollusques, rose : nématodes, vert clair : plathelminthes.

a.4 Résultats

Les résultats de ces trois analyses sont synthétisés dans le tableau 4.4, qui présente les phylums retrouvés monophylétiques dans le consensus strict des arbres obtenus avec chacun des trois codages, parmi ceux-là seuls les cnidaires, les échinodermes et les plathelminthes sont retrouvés monophylétiques dans les trois analyses. D'autre part, la résolution des arbres obtenus est relativement faible en terme de nombre de nœuds résolus. Les arbres consensus obtenus avec les codages en positions relatives, jonctions et jonctions signées ont respectivement 22, 37 et 30 nœuds contre 46 nœuds pour un arbre complètement résolu (ce qui donne les valeurs normées suivantes : 0,478, 0,652 et 0,804). Le codage en jonctions donne une bonne résolution, tandis que les deux autres codages donnent une résolution plus faible. Les trois consensus n'ont que peu de nœuds en communs : les cnidaires, les crâniates, les échinodermes, les plathelminthes et les gastéropodes apparaissent monophylétiques dans les trois arbres, mais les relations entre ces groupes sont complètement irrésolues.

Taxon	**Position relative**	**Jonctions**	**Jonctions signées**
Annélides		x	x
Arthropodes			
Brachiopodes			
Chordés + Hémichordés			
Cnidaires	x	x	x
Échinodermes	x	x	x
Mollusques			
Nématodes			
Plathelminthes	x	x	x
Deutérostomiens			
Crâniates	x	x	x
Protostomiens		x	
Gastéropodes	x	x	x

TABLEAU 4.4 – Monophylie des différents phylums en fonction du codage et avec la matrice complète. Les groupes retrouvés monophylétiques sont indiqués par un x. En bas : autres taxons.

b Échantillonnage sans les taxons excentriques

Nous avons retiré de l'analyse les 9 espèces qui se sont avérées très divergentes des autres espèces de leurs phylums (voir § 4.1.4.1, page 101). Cela nous permettra d'évaluer l'influence de ces taxons sur les résultats obtenus avec la matrice complète. En effet, les arbres obtenus avec la matrice complète sont mal résolus et peu congruents entre eux. Cela peut être dû en partie au fait que les taxons excentriques sont très différents des autres taxons et de ce fait se place de façon aléatoire sur l'arbre. Nous retenons donc au final 40 espèces.

b.1 Échantillonnage sans les taxons excentriques avec le codage « Position relative »

La matrice [13] comporte 40 taxons et 75 caractères, répartis de la façon suivante : 37 caractères « Position relative du gène » (un pour chaque gène), 37 caractères « Polarité du gène » (un pour chaque gène) et 1 caractère « Présence Absence du gène » (pour le gène atp8). Les caractères « Position relative du gène » ont de 7 à 21 états différents (7 pour le gène atp8 et 21 pour les gènes trnA, trnQ et trnW). Les caractères « Polarité du gène » des gènes cob, cox1, cox2 et trnK sont constants et un caractère est non informatif. Tous les caractères sont traités non orientés, non ordonnés avec un poids de 1.

La recherche de l'arbre le plus court a été réalisée avec un algorithme heuristique (Branch swapping TBR avec 500 agrégations aléatoires des taxons) par le logiciel PAUP* (Swofford, 1998). Le nombre d'arbres équiparcimonieux trouvés étant extrêmement élevé (>180 000) l'analyse n'a pu être menée à son terme. Les arbres équiparcimonieux ont une longueur de 628 pas, un IC = 0,9299 et un IR = 0,9417. Le consensus strict des arbres équiparcimonieux est présenté figure 4.6.

[13] disponible à l'adresse http://lis.snv.jussieu.fr/~gallut/These

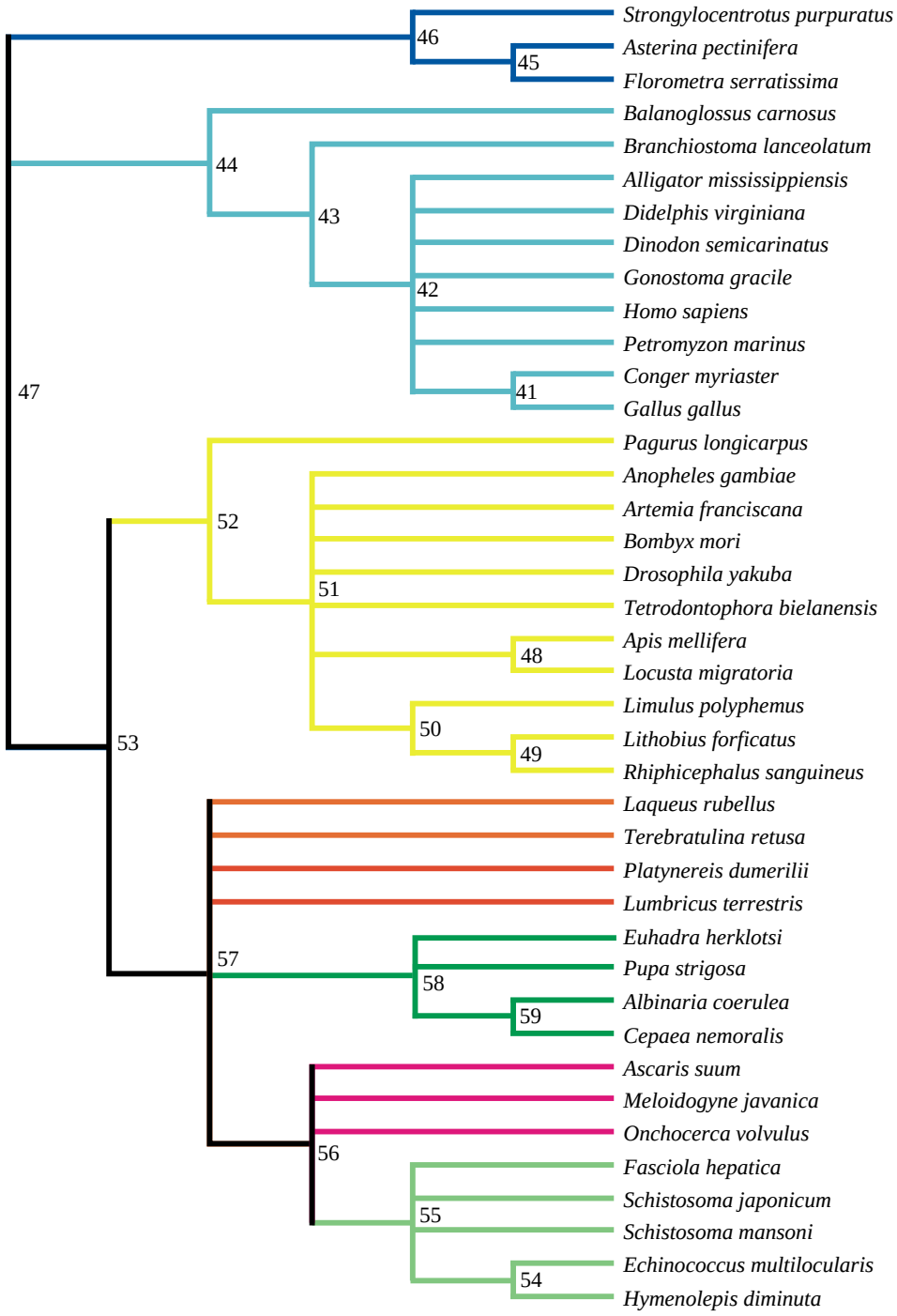


FIGURE 4.6 – Consensus strict des 180 000+ arbres équiparcimonieux obtenus avec la matrice sans les taxons excentriques et avec le codage « Position relative ». (Arbre le plus court : L = 628 pas, IC = 0,9299 IR = 0,9417). La racine a été placée entre les deutérostomiens et les protostomiens. Noire : cnidaires, rouge : annélides, jaune : arthropodes, orange : brachiopodes, bleu : deutérostomiens, bleu clair : chordés + hémichordés, bleu foncé : échinodermes, vert foncé : mollusques, rose : nématodes, vert clair : plathelminthes.

b.2 Échantillonnage sans les taxons excentriques avec le codage « Jonctions »

La matrice [14] est constituée de 40 taxons et 359 caractères (321 caractères « Présence Absence d'une jonction », 37 caractères « Polarité du gène » et 1 caractère « Présence Absence du gène »). Les caractères « Polarité du gène » des gènes cob, cox1, cox2 et trnK sont constants, et 116 caractères sont non informatifs. Tous les caractères sont traités non orientés, non ordonnés avec un poids de 1.

La recherche de l'arbre le plus court a été réalisée avec un algorithme heuristique (Branch swapping TBR avec 500 agrégations aléatoires des taxons) par le logiciel PAUP* et a fourni 63 arbres équiparcimonieux. Ces derniers ont une longueur de 641 pas, un IC de 0,5538 et un IR de 0,7880. Le consensus strict des 63 arbres équiparcimonieux est présenté figure 4.7.

[14] disponible à l'adresse http://lis.snv.jussieu.fr/~gallut/These

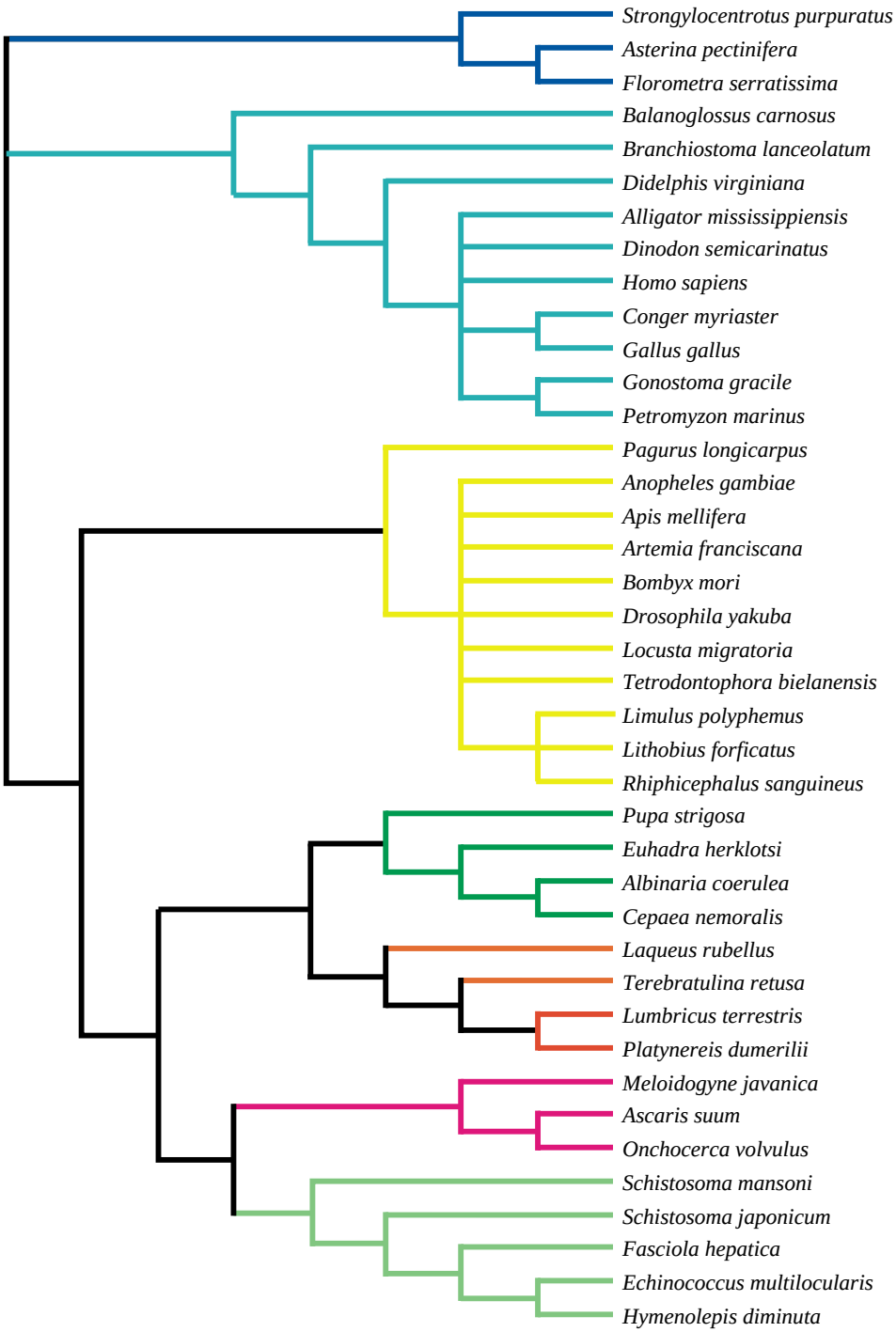


FIGURE 4.7 – Consensus strict des 63 arbres équiparcimonieux obtenus avec la matrice sans les taxons excentriques et avec le codage « Jonctions ». (Arbre le plus court : L = 641 pas, IC = 0,5538 IR = 0,7880). La racine a été placée entre les deutérostomiens et les protostomiens. Noire : cnidaires, rouge : annélides, jaune : arthropodes, orange : brachiopodes, bleu : deutérostomiens, bleu clair : chordés + hémichordés, bleu foncé : échinodermes, vert foncé : mollusques, rose : nématodes, vert clair : plathelminthes.

b.3 Échantillonnage sans les taxons excentriques avec le codage « Jonctions signées »

La matrice [15] est constituée de 40 taxons et 417 caractères « Présence Absence d'une jonction signée », pour lesquels, 201 caractères sont non informatifs et aucun n'est constant. Tous les caractères sont traités non orientés, non ordonnés avec un poids de 1. La recherche de l'arbre le plus court a été réalisée avec un algorithme heuristique (Branch swapping TBR avec 500 agrégations aléatoires des taxons) par le logiciel PAUP* et a trouvé 360 arbres équiparcimonieux. Ces derniers ont une longueur de 610 pas, un IC de 0,6836 et un IR de 0,8052. Le consensus strict des 360 arbres équiparcimonieux est présenté figure 4.8.

[15] disponible à l'adresse http://lis.snv.jussieu.fr/~gallut/These

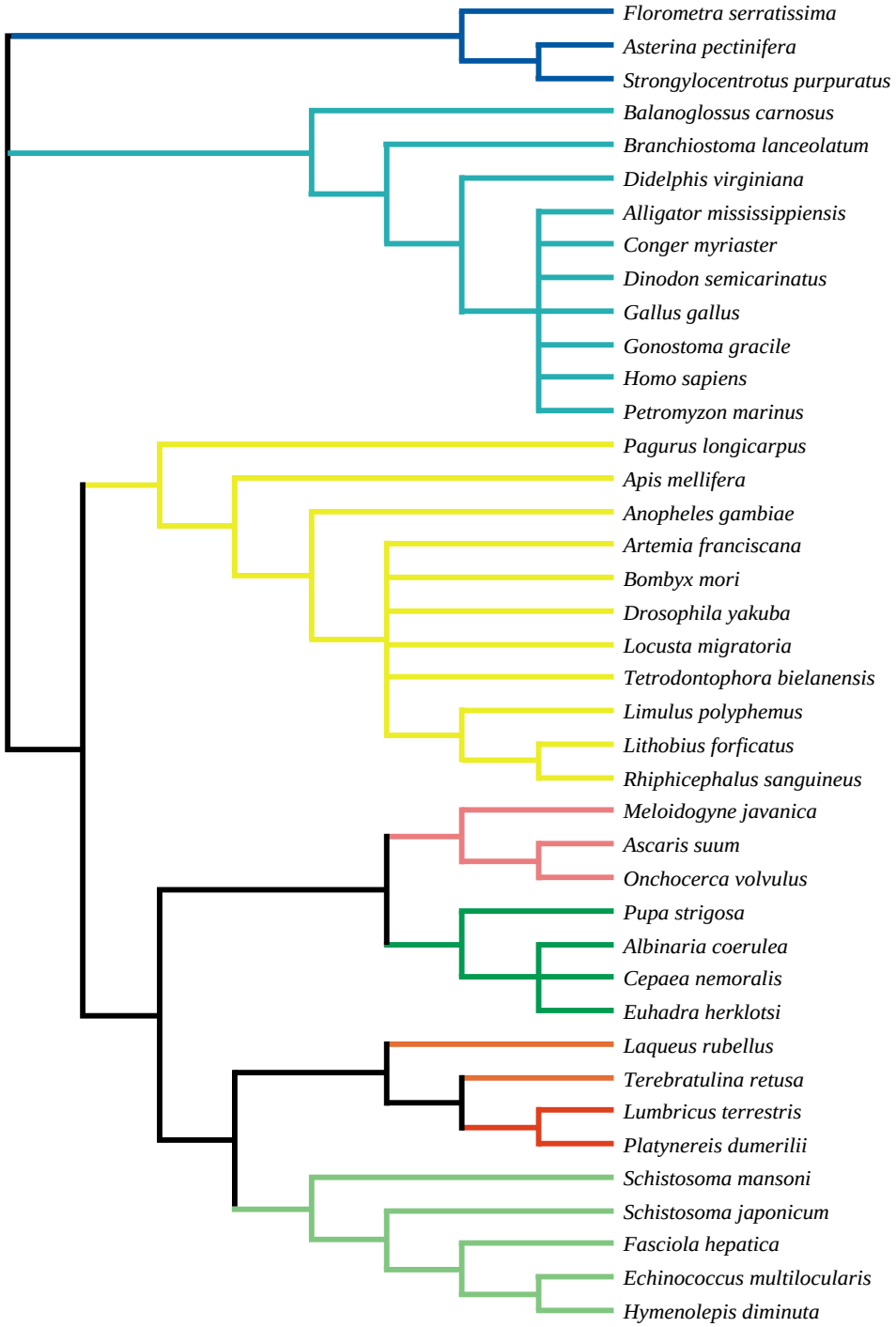


FIGURE 4.8 – Consensus strict des 360 arbres équiparcimonieux obtenus avec la matrice sans les taxons excentriques et avec le codage « Jonctions signées ». (Arbre le plus court : L = 610 pas, IC = 0,6836 IR = 0,8052). La racine a été placée entre les deutérostomiens et les protostomiens. Noire : cnidaires, rouge : annélides, jaune : arthropodes, orange : brachiopodes, bleu : deutérostomiens, bleu clair : chordés + hémichordés, bleu foncé : échinodermes, vert foncé : mollusques, rose : nématodes, vert clair : plathelminthes.

b.4 Résultats

Les résultats des trois analyses avec les trois codages de la matrice sans les taxons excentriques sont synthétisés dans le tableau 4.5. Parmi les différents phylum seuls les annélides, les brachiopodes et les nématodes ne sont pas retrouvés monophylétiques dans les trois arbres consensus. De plus les relations entre les différents phylums sont mieux résolues qu'avec la matrice complète. Les deutérostomiens et les protostomiens sont monophylétiques ; nous considérons ici les protostomiens au sens moderne du terme, c'est-à-dire en y incluant les plathelminthes et les nématodes. Les nématodes ne sont pas retrouvés en position de groupe frère des arthropodes mais inclus dans le clade (mollusques, annélides, plethelminthes, brachiopodes, nématodes) et en conséquence ni les écdizozoaires ni les lophotrochozoaires (voir figure 4.1, page 91) ne sont monophylétiques dans nos résultats.

Taxon	**Position relative**	**Jonctions**	**Jonctions signées**
Annélides		x	x
Arthropodes	x	x	x
Brachiopodes			
Chordés + Hémichordés	x	x	x
Échinodermes	x	x	x
Mollusques (Gastéropodes)	x	x	x
Nématodes		x	x
Plathelminthes	x	x	x
Deutérostomiens	x	x	x
Crâniates	x	x	x
Protostomiens	x	x	x

TABLEAU 4.5 – Monophylie des différents taxons en fonction du codage et avec la matrice sans les taxons excentriques. Les groupes retrouvés monophylétiques sont indiqués par un x. En bas : autres taxons.

Sans les taxons excentriques les trois arbres obtenus sont beaucoup mieux résolus et la congruence entre ces trois arbres est beaucoup plus importante que pour les arbres obtenus avec la matrice complètes. Ce qui confirme l'hypothèse que les taxons excentriques brouillent le signal phylogénétique et perturbent les résultats.

c Reconstitution des génomes ancestraux

Nous avons reconstitué les génomes ancestraux avec **Recons** pour l'arbre consensus des arbres obtenus avec la matrice complète (figure 4.3) ainsi que sur l'arbre consensus des arbres obtenus avec la matrice sans les excentriques (figure 4.6), codées avec le codage « Position relative ». Aucun des deux consensus n'a permis de reconstituer un génome complet à tous les nœuds. Du fait du très grand nombre d'arbres équiparcimonieux (>180 000) obtenus dans les deux cas nous n'avons pas cherché à reconstituer les génomes ancestraux sur ces arbres, ce qui aurait nécessité un temps de calcul d'environ 500 jours. Le temps de calcul de la reconstitution elle-même est faible mais le temps nécéssaire pour calculer tous les états possibles à chaque nœud est de l'ordre d'une à deux minutes (calcul effectué par le logiciel PAUP*). Il aurait été intéressant de pouvoir reconstituer les génomes ancestraux sur ces arbres afin le cas échéant d'éliminer ceux pour lesquels certains nœuds ne présentaient pas de génome complet. Il aurait peut-être été possible de diminuer le nombre d'arbres, en ne retenant que ceux pour lesquels la reconstitution était complète. Les résultats pour les deux consensus sont présentés dans le tableau 4.6.

Nœud	**Nombre**	**Nœud**	**Nombre**	**Nœud**	**Nombre**
50	0 génome	58	1 génome	66	0 génome
51	1 génome	59	0 génome	67	1 génome
52	2 génomes	60	1 génome	68	0 génome
53	2 génomes	61	0 génome	69	0 génome
54	8 génomes	62	0 génome	70	0 génome
55	6 génomes	63	0 génome	71	0 génome
56	4 génomes	64	0 génome	72	0 génome
57	1 génome	65	0 génome		
41	0 genome	48	1 génome	54	1 génome
42	1 génome	49	0 genome	55	0 genome
43	0 genome	50	1 génome	56	0 genome
44	0 genome	51	1 génome	57	0 genome
45	12 génomes	52	1 génome	58	0 genome
46	6 génomes	53	0 genome	59	0 genome
47	0 genome				

TABLEAU 4.6 – Nombre de génomes ancestraux reconstitués aux nœuds internes des arbres des figures 4.3 (en haut) et 4.6 (en bas).

Sur l'arbre consensus obtenu avec la matrice complète (figure 4.3) aucun des génomes reconstitués ne sont identiques entre eux, par contre plusieurs de ces derniers sont identiques au génome d'un taxon terminal :

- le génome du nœud 51 est égal au génome du groupe *Homo sapiens*,
- l'un des génomes du nœud 55 est égal au génome de *Strongylocentrotus purpuratus*,
- le génome du nœud 57 est égal au génome de *Locusta migratoria*,
- le génome du nœud 58 est égal au génome du groupe *Drosophila yakuba*,
- le génome du nœud 60 est égal au génome du groupe *Limulus polyphemus*,
- le génome du nœud 67 est égal au génome d'*Echinococcus multilocularis*.

De la même manière, aucun des génomes reconstitués sur l'arbre consensus obtenu avec la matrice sans les taxons excentriques (figure 4.6) n'est identique à un autre génome reconstitué mais certains sont identiques à celui d'un taxon terminal :

- le génome du nœud 42 est égal au génome du groupe *Homo sapiens*,
- le génome du nœud 48 est égal au génome de *Locusta migratoria*,
- le génome du nœud 50 est égal au génome du groupe *Limulus polyphemus*,
- le génome du nœud 51 est égal au génome du groupe *Drosophila melanogaster*,
- le génome du nœud 54 est égal au génome de *Echinococcus multilocularis*.

Nous avons choisi d'illustrer la reconstruction des événements évolutifs impliqués par les génomes reconstitués au niveau des nœuds 55 et 54 de la figure 4.3. Le nœud 55 regroupe l'ensemble des échinodermes et le 54 regroupe *Florometra serratissima* et *Asterina pectinifera*. Parmi tous les génomes reconstitués aux nœuds 54 et 55 nous avons retenu pour cet exemple ceux qui permettent de faire la reconstruction des événements évolutifs le plus simplement possible. Au nœud 55, nous avons retenu le génome qui est identique à celui de *Strongylocentrotus purpuratus*. Au nœud 54, nous avons choisi un génome de façon à ce que la combinaison des deux génomes soit compatible avec le nombre de transformations impliquées sur la branche 55–54.

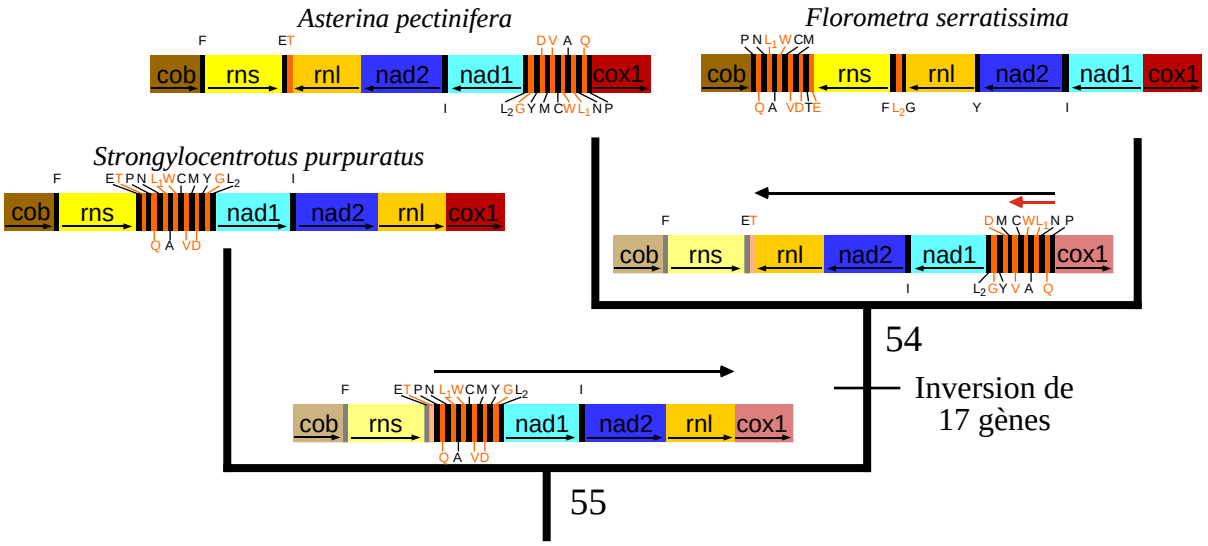


FIGURE 4.9 – Reconstruction d'événements entre les nœuds 55 et 54 de la figure 4.3. Un extrait des génomes des trois taxons terminaux est placé aux extrémités des branches. Un extrait des génomes reconstitués aux nœuds 55 et 54 sont placés au niveau des nœuds internes. Il est possible de reconstruire une inversion entre les nœuds 55 et 54 matérialisée par la flèche noire. Cette inversion n'est pas parfaite, en effet l'ordre de tous les gènes est inversé mais l'orientation de certains gènes n'a pas changé (matérialisé par la flèche rouge). Les couleurs ternes caractérisent les gènes dont la position ne change pas entre les nœuds 54 et 55.

Nous pouvons tout d'abord remarquer que les génomes de *Strongylocentrotus purpuratus* et de *Asterina pectinifera* ne diffèrent que par une seule inversion d'un bloc de 17 gènes. Avec les génomes reconstitués que nous avons retenus ici, l'inversion de 17 gènes semble avoir eu lieu sur la branche 55–54 (voir figure 4.9), c'est-à-dire entre l'ancêtre hypothétique commun des trois échinodermes et l'ancêtre hypothétique commun de *Florometra serratissima* et *Asterina pectinifera*. Cependant, l'inversion inférée à partir des génomes reconstitués n'est pas « parfaite » car bien que l'ordre des 17 gènes soit inversé, 9 de ces derniers ont la même orientation aux nœuds 55 et 54. Leur orientation ne change qu'au niveau de la branche d'*A. pectinifera*. D'autre part, plusieurs inversions et transpositions sont nécessaires pour passer du génome re-

constitué au nœud 54 au génome de *F. serratissima*. Ainsi, pour passer de l'ordre des gènes du génome de *S. purpuratus* à celui du génome d'*A. pectinifera* il faut :

- du nœud 55 à *Strongylocentrotus purpuratus* :
  - aucun événement (génomes identiques),
- du nœud 55 au nœud 54 :
  - une transinversion de 8 gènes (de -D à +rnl),
  - neuf transpositions indépendantes d'un gène (M, V, C, W, A, L1, N, Q et P).
- du nœud 54 à *Asterina pectinifera* :
  - neuf inversions indépendantes d'un gène (M, V, C, W, A, L1, N, Q et P).

Ce résultat peut paraître surprenant, surtout si l'on considère qu'une seule inversion pour passer du génome de *S. purpuratus* au génome d'*A. pectinifera* est beaucoup plus parcimonieuse que 19 événements indépendants. Cela s'explique par le fait que dans notre approche les clades ne sont pas fondés sur le partage d'événements identifiés *a priori* tels que l'inversion de 17 gènes observée entre *S. purpuratus* et *A. pectinifera* mais sur le partage de transformations impliquant des changements individuels de position et d'orientation des gènes. De plus, comme nous l'avons vu au chapitre 3, les événements identifiés *a priori* et impliquant de nombreux gènes sont généralement *a posteriori* retrouvés scindés en plusieurs événements indépendants. Ce qui est dû au fait que les caractères « Position relative du gène » et « Orientation du gène » peuvent subir des transformations sur des branches différentes. Ainsi, pour l'inversion de 17 gènes l'inversion de l'ordre des gènes a lieu sur une branche alors que l'inversion de l'orientation a lieu sur deux branches différentes.

Ces trois espèces (*A. pectinifera*, *F. serratissima* et *S. purpuratus*) appartiennent respectivement à trois des cinq classes d'échinodermes : les astérides, les crinoïdes et les échinides. Le résultat obtenu avec le génome mitochondrial de ces trois espèces et le codage « Position relative » est en contradiction avec la phylogénie couramment admise des échinodermes (Brusca et Brusca, 1990). En effet les échinides et les astérides sont considérés comme plus proches entre eux qu'ils ne le sont des crinoïdes or ici les astérides (*A. pectinifera*) et les crinoïdes

(*F. serratissima*) forment un groupe monophylétique. Nous ne pouvons pas considérer ce résultat comme définitif sans avoir au moins un représentant de chacune des deux autres classes d'échinodermes, les holoturoïdes et les ophiurides. Nous avons donc entrepris le séquençage complet du génome mitochondrial d'une holothurie et d'une ophiure, génomes qui nous permettront probablement de confirmer ou infirmer ce résultat intéressant (travail en cours).

4.1.5 Discussion

Les analyses réalisées avec la matrice complète ont permis de montrer que :

– les différents phylums ne sont pas tous retrouvés monophylétiques,

– la résolution est faible,

– il y a peu de congruence entre les résultats obtenus avec les trois de codages.

Plusieurs facteurs peuvent expliquer ces résultats.

Le fait que la plupart des différents phylums ne soient pas retrouvés monophylétiques peut tout simplement montrer qu'ils ne le sont pas, mais cela peut aussi être dû au fait que certains taxons ont un taux d'évolution beaucoup plus élevé que les autres, ce qui perturbe l'analyse. Cette deuxième hypothèse semble plus probable car, mis à part les nématodes, tous les phylums sont retrouvés monophylétiques lorsque les taxons excentriques sont retirés de l'analyse. D'autre part, l'échantillonnage des différents phylums est très variable, certains ne sont représentés que par une seule espèce alors que d'autres sont beaucoup mieux représentés. De plus l'échantillonnage dont nous disposons ne couvre pas toute la diversité des métazoaires.

La faible résolution peut être due à un taux d'homoplasie élevé qui brouille le signal phylogénétique. Cependant les indices de cohérence et de rétention des arbres obtenus sont assez élevés (respectivement, de 0,5 à 0,9 et de 0,7 à 0,9 sans les taxons excentriques) ce qui infirme cette hypothèse. Encore une fois, ce sont probablement les taxons excentriques qui sont à l'origine de ce résultat. Le phénomène de saturation peut brouiller le signal phylogénétique et donner des résultats peu résolus. Pour un caractère dont le nombre d'états possibles est fixé, les possibilités de mutations sont limitées. Ainsi, un nombre élevé de mutations peut conduire deux

taxons à partager un même état pour ce caractère sans que cela soit dû à une ascendance commune, c'est le phénomène de saturation. Le nombre d'ordres de gènes possibles avec 37 gènes est très élevé, il est donc très improbable qu'un phénomène de saturation puisse apparaître dans ces conditions.

Le manque de congruence entre les arbres obtenus avec les différents codages peut aussi être causé par la saturation et/ou un taux d'homoplasie élevé, mais pour les raisons que nous venons d'évoquer ces deux hypothèses ne peuvent pas être retenues. L'incongruence entre les arbres obtenus avec les trois codages est en partie due aux taxons excentriques, en effet la congruence est bien meilleure entre les arbres obtenus avec la matrice sans les taxons excentriques. Mais elle est aussi due en partie au fait que les différents codages ne sont pas équivalents et ne représentent pas les données de façon rigoureusement identique comme nous l'avons montré au chapitre 3. Le fait que les résultats obtenus avec la matrice sans les taxons excentriques soient congruents, est même de ce point de vue, une preuve de leur robustesse. En effet, obtenir des résultats similaires avec des approches différentes conforte ces résultats. Les différences entre les résultats obtenus avec et sans les taxons excentriques, montre de plus, que notre méthode est sensible aux taux d'évolutions variables.

Les résultats de l'ensemble de ces analyses montrent que le taux d'évolution du génome mitochondrial est très variable chez les métazoaires. Certaines espèces proches présentent des ordres de gènes complètement remaniés alors que des espèces dont les lignées sont séparées depuis des centaines de millions d'années, telles que les vertébrés et les arthropodes, présentent des ordres de gènes très similaires. Malgré ces taux d'évolution variables, le génome mitochondrial représente un jeu de données très intéressant pour reconstruire les relations de parentés entre phylums de métazoaires. L'échantillonnage dont nous disposons actuellement n'est pas suffisamment représentatif de la diversité des métazoaires, ce qui rend l'acquisition de l'ordre des gènes de nouveaux génomes mitochondriaux, absolument primordial.

4.2 Chromosomes du genre *Mastomys*

Le genre *Mastomys* (Mammalia, Rodentia, Muridae, Murinae) est un genre de rongeurs africains présents partout sauf dans les déserts et les forêts denses. Il comprend une dizaine d'espèces dont certaines n'ont pas encore été décrites. Nous avons à notre disposition les caryotypes en bandes G des autosomes de cinq espèces différentes de *Mastomys* (Volobouev V. et Aniskin V., non publié) qui ont été étudiées au laboratoire Mammifères et Oiseaux du Muséum National d'Histoire Naturelle. *Mastomys verheyeni*, *Mastomys erythroleucus* et *Mastomys species* ont 18 autosomes alors que *Mastomys natalensis* et *Mastomys awashencsis* en ont seulement 15. Chacune de ces espèces a sa propre organisation chromosomique et une espèce (*Mastomys erythroleucus*) présente deux populations d'origines différentes qui n'ont pas la même organisation. Il y a suffisamment de variation entre ces cinq espèces pour envisager de reconstruire leurs relations de parenté à partir de leur organisation chromosomique. Nous avons décidé d'utiliser ces données pour étudier les adaptations de notre approche de codage pour le traitement de chromosomes linéaires et multiples.

Une étude récente (Lecompte E., non publié) de la phylogénie moléculaire du genre *Mastomys*, à partir de la séquence du gène mitochondrial *cytochrome b*, a permis d'établir les relations de parentés entre certaines de ces espèces (figure 4.10). Nous utiliserons cette phylogénie pour évaluer les résultats obtenus avec le codage de l'organisation des chromosomes.

4.2.1 Homologies des chromosomes du genre *Mastomys*

Les homologies entre les chromosomes des cinq espèces de *Mastomys* (Volobouev V. et Aniskin V., non publié) ont été établies à partir de la comparaison des caryotypes. Nous avons utilisé ces homologies pour définir l'ensemble des segments chromosomiques homologues et ce, de façon à ce que le nombre de segments identifiés soit maximal, c'est-à-dire l'ensemble des plus petits segments chromosomiques homologues. Nous avons nommé chaque segment à partir de sa position dans le chromosome pour l'espèce *Mastomys verheyeni*. Nous avons choisi cette

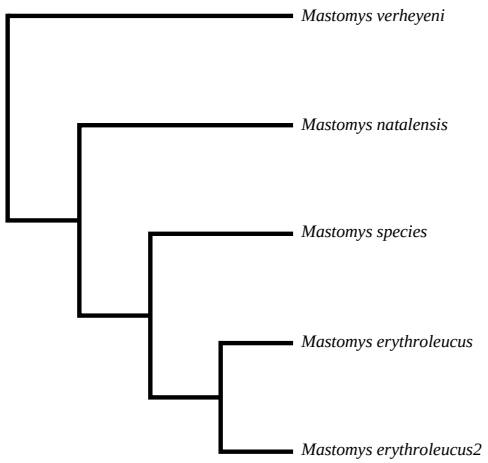


FIGURE 4.10 – Phylogénie moléculaire du genre *Mastomys*, enracinée à partir d'une espèce extérieure au genre.

espèce arbitrairement comme référence pour définir une nomenclature unifiée des segments homologues, l'objectif étant de nommer chaque segment de la même manière pour toutes les espèces. Pour chacune des espèces, la structure de chaque chromosome a ensuite été réexprimée par la succession des segments présents sur celui-ci. La comparaison de caryotypes ne permet pas d'orienter les segments de façon absolue, il est donc nécessaire d'orienter les segments de façon relative les uns par rapport aux autres. Nous avons arbitrairement attribué le signe « + » à tous les segments de *M. verheyeni*. Les segments des autres espèces dont l'orientation est la même que chez *M. verheyeni* ont reçu le signe « + » alors que ceux dont l'orientation est opposée à celle de chez *Mastomys verheyeni* ont reçu le signe « - ». Chaque chromosome est donc représenté par un ordre signé et l'organisation chromosomique de l'ensemble des espèces est représentée de façon unifiée.

Nous avons identifié 39 segments chromosomiques homologues qui sont tous présents chez chacune des cinq espèces. L'ensemble des homologies des chromosomes du genre *Mastomys* est présenté dans le tableau B.1 tandis que le tableau 4.7 présente un exemple pour un des chromosomes. Le chromosome 10 de *M. verheyeni* est homologue du chromosome 10 de *M. natalensis*, du chromosome 13 de *M. erythroleucus*, de *M. erythroleucus2* et de *M. awashencsis* et du chromosome 14 de *M. species*. La deuxième ligne du tableau 4.7 présente la description classique

des homologies entre chromosomes avec, associée au N° du chromosome, la morphologie des chromosomes telle que acrocentrique, métacentrique ou submétacentrique et avec les événements de réarrangements permettant de passer d'une morphologie à une autre. Ici, le chromosome 10 de *M. natalensis* est indiqué comme ayant subi une inversion péricentrique qui explique qu'il soit submétacentrique et non acrocentrique comme les autres. Ce point de vue orientant *a priori* les événements, nous avons donc exprimé la morphologie des chromosomes de façon à ne pas faire intervenir d'événement de réarrangements. À partir de la description classique des homologies entre chromosomes nous avons identifié les segments chromosomiques homologues et décrit l'organisation des chromosomes avec ces segments. Les segments chromosomiques homologues présents sur les chromosomes présentés dans le tableau 4.7 sont identifiés par *10prox* et *10dist* en référence à leur position sur le chromosome 10 de *M. verheyeni*. Ainsi, *10prox* correspond au segment proximal du chromosome 10. Le chromosome homologue des cinq espèces est constitué à chaque fois de ces deux segments mais l'orientation et la position de ces derniers sont variables.

M. verheyeni	*M. erythroleucus*	*M. erythroleucus2*	*M. species*	*M. natalensis*	*M. awashencsis*
10 acro	13 acro	13 acro	14 acro	10 SM (inv)	13 acro
centro	**centro**	**centro**	**centro**	**télo**	**centro**
+10prox	+10prox	+10prox	+10prox	-10prox	+10prox
+10dist	+10dist	+10dist	+10dist	**centro**	+10dist
télo	**télo**	**télo**	**télo**	+10dist	**télo**
				télo	

TABLEAU 4.7 – Segments homologues d'un même chromosome au sein du genre *Mastomys* (chromosome 10 de *M. verheyeni* et de *M. natalensis*, chromosome 13 de *M. erythroleucus*, de *M. erythroleucus2* et de *M. awashencsis* et chromosome 14 de *M. species*). La deuxième ligne présente les homologies entre chromosomes de façon classique, c'est-à-dire avec la morphologie de chacun et avec les événements de réarrangement qui permettent de passer de l'un à l'autre tel que l'inversion péricentrique (notée inv) qu'à subie le chromosome 10 de *M. natalensis*. Les lignes suivantes présentent les mêmes homologies mais exprimées avec notre formalisme. Chaque couleur correspond à un segment homologue. L'orientation relative des segments homologues est indiquée par les signes « + » et « - ». acro : acrocentrique, SM : submétacentrique, centro : centromère, télo : télomère, prox : proximal, dist : distal.

La comparaison de caryotypes ne permet pas de faire d'hypothèses sur l'homologie des centromères d'une part et sur l'homologie des télomères d'autre part. Le centromère et le télomère

ont un rôle fonctionnel dans le chromosome. Il n'est pas possible, sans analyse complémentaire, de savoir si le centromère d'un chromosome d'une espèce est homologue à celui du même chromosome d'une autre espèce, *a fortiori* si le chromosome de la première espèce est homologue à deux chromosomes de la deuxième espèce. Dans ce cas, le centromère du chromosome de la première espèce peut être l'homologue de celui de l'un ou l'autre des chromosomes de la deuxième espèce. Seule une analyse moléculaire de la séquence de ces centromères permettrait de trancher. Certains chromosomes peuvent avoir un bras constitué entièrement d'hétérochromatine (indiqué par *Hét* dans le tableau B.1), c'est le résultat d'une expansion rapide de l'hétérochromatine propre à chaque espèce. Par conséquent nous ne considérerons pas les segments d'hétérochromatine comme homologues.

4.2.2 Codage de l'organisation chromosomique du genre *Mastomys*

Les centromères, les télomères et les segments d'hétérochromatine ne pouvant pas être considérés comme homologues, nous les avons éliminés de la description des chromosomes pour le codage. Le codage « Position relative » implique que chaque unité fonctionnelle (ici les segments chromosomiques homologues) soit entourée de deux autres unités, ce qui n'est pas toujours le cas sur les chromosomes linéaires. Les unités qui se trouvent aux extrémités des chromosomes sont encadrées par une autre unité et par le « vide » ce qui pose un problème pour coder leur position. Le fait qu'une unité se trouve à l'intérieur du chromosome ou à l'extrémité n'est pas équivalent ; dans ce cas nous avons choisi d'utiliser la notion de *vide* pour la position de l'unité. Les chromosomes acrocentriques ont deux extrémités différentes : l'extrémité où se trouve le centromère et l'extrémité où se trouve le télomère alors que les chromosomes métacentriques et submétacentriques ont deux extrémités identiques, chacune constitué d'un télomère. Par conséquent, pour coder la position des segments situés aux extrémités d'un chromosome avec le codage « Position relative » nous avons représenté la notion de *vide* par le télomère et le centromère, ce dernier n'est utilisé que pour l'extrémité centromérique des chromosomes acrocentrique. Par exemple, la position du segment A situé à l'extrémité d'un chromosome mé-

tacentrique et précédé du segment B est représentée par la paire (B,télo). Si le segment A se trouve à l'extrémité d'un chromosome acrocentrique, sa position est représentée par la paire (B,centro). Les centromères et les télomères ne sont pas considérés comme des caractères mais seulement utilisés pour la position des segments situés aux extrémités, ainsi les centromères en position médiane sont complètement éliminés.

Le codage d'un chromosome linéaire nécessite une adaptation avec le codage « Position relative », ce qui n'est pas le cas avec les codages en jonctions. D'autre part, le codage de chromosomes multiples est identique au codage d'un chromosome unique, chaque chromosome est codé indépendamment et l'ensemble des caractères est utilisé conjointement. Le nombre de chromosomes et leur morphologie peuvent être retrouvés à partir du codage de l'ordre des segments, par conséquent nous n'avons pas introduit de caractères pour ces deux caractéristiques, ce qui aurait introduit de la redondance.

4.2.2.1 Codage « Position relative »

Nous avons codé l'organisation des chromosomes du genre *Mastomys* avec le codage « Position relative » et l'option AB=BA. La matrice finale est constituée de 6 taxons et de 78 caractères, 39 caractères « Position relative du segment » et 39 caractères « Orientation du segment » (la matrice est fournie en annexe, page 175).

L'analyse exhaustive de la matrice fourni un seul arbre d'une longueur de 69 pas avec un IC de 0,8696 et un IR de 0,7273 présenté sur la figure 4.11. Nous avons enraciné l'arbre obtenu avec *Mastomys verheyeni* car c'est l'espèce qui se trouve à la base de l'arbre obtenu en phylogénie moléculaire (cf. figure 4.10) et nous ne disposons pas du caryotype d'une espèce pouvant être utilisée comme groupe extérieur.

4.2.2.2 Codage « Jonctions »

Nous avons codé l'organisation des chromosomes du genre *Mastomys* avec le codage « Jonctions » et l'option AB=BA. La matrice est constituée de 6 taxons et de 69 caractères, 30 ca-

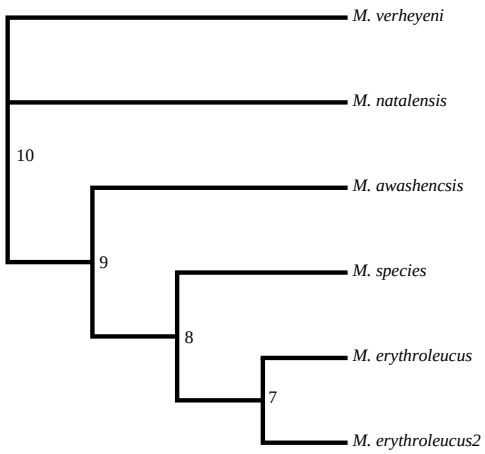


FIGURE 4.11 – Phylogénie chromosomique du genre *Mastomys* avec le codage « Position relative ». Longueur : 69 pas, IC = 0,8696 et IR = 0,7273.

ractères « Présence Absence d'une jonction » et 39 caractères « Orientation du segment » (la matrice est fournie en annexe, page 176).

L'analyse exhaustive de la matrice a fourni deux arbres équiparcimonieux d'une longueur de 42 pas avec un IC de 0,8333 et un IR de 0,7308. Le consensus strict de ces deux arbres, enraciné avec *Mastomys verheyeni*, est présenté sur la figure 4.12.

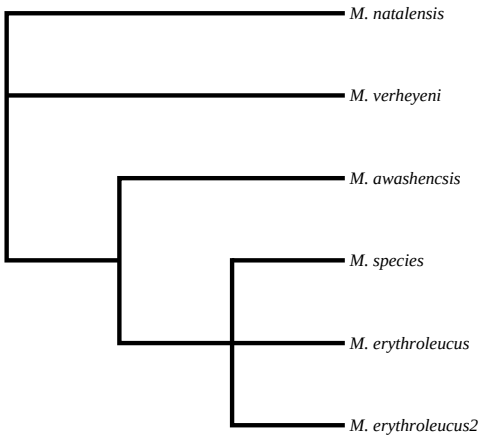


FIGURE 4.12 – Consensus des deux arbres obtenus avec le codage « Jonctions ». Longueur de l'arbre le plus court : 42 pas, IC = 0,8333 et IR = 0,7308.

4.2.2.3 Codage « Jonctions signées »

Nous avons codé l'organisation des chromosomes du genre *Mastomys* avec le codage « Jonctions signées ». La matrice est constituée de 6 taxons et de 44 caractères « Présence Absence d'une jonction signée » (la matrice est présentée en annexe, page 177).

L'analyse exhaustive de la matrice a fourni un seul arbre d'une longueur de 47 pas avec un IC de 0,8511 et un IR de 0,6818 présenté sur la figure 4.13. L'arbre obtenu est enraciné avec *Mastomys verheyeni*.

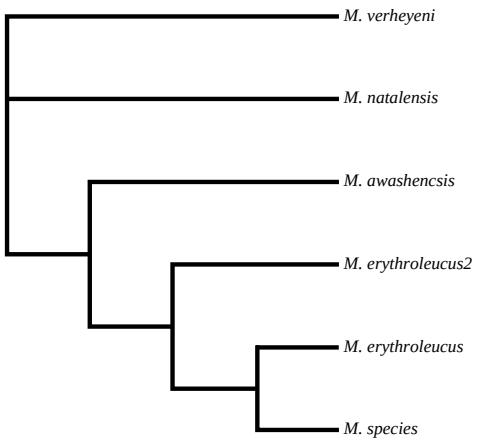


FIGURE 4.13 – Phylogénie chromosomique du genre *Mastomys* avec le codage « Jonctions signées ». Longueur : 47 pas, IC = 0,8511 et IR = 0,6818.

4.2.2.4 Résultats

Les résultats obtenus avec les trois codages sont très similaires, seule la position des deux populations de *Mastomys erythroleucus* est différentes. Elles forment un groupe monophylétique sur l'arbre obtenu avec le codage « Position relative », elles sont paraphylétiques sur l'arbre obtenu avec le codage « Jonctions signées » et elles sont retrouvées dans ces deux configurations sur les arbres obtenus avec le codage « Jonctions ». Ces différences sont dues au fait que nous avons pris en compte la notion de *vide* pour la position des segments chromosomiques situés aux extrémités des chromosomes avec le codage « Position relative ». Par rapport à ce dernier, il y a une perte d'information avec les codages « Jonctions » et « Jonctions signées », en effet,

le fait qu'un segment se trouve à l'extrémité d'un chromosome n'est pas pris en compte, il n'y a pas de jonction du type « segment / vide ».

L'arbre obtenu avec le codage « Position relative » est celui qui est le plus proche de l'arbre obtenu avec la séquence du gène *cytochrome b* (figure 4.10). Il est intéressant de constater que *M. natalensis* et *M. awashencsis*, qui contrairement aux autres n'ont que 15 chromosomes, ne sont pas retrouvés monophylétiques.

4.2.3 Retour aux caractères

Nous n'avons pas pu reconstituer les génomes des ancêtres hypothétiques communs sur les arbres obtenus, quel que soit le codage. En effet, avec les codages « Jonctions » et « Jonctions signées », il n'y a pas assez d'information pour reconstituer des génomes complets. D'autre part, l'algorithme que nous avons présenté (voir § c, page 61) pour reconstituer les génomes ancestraux sur les arbres obtenus avec le codage « Position relative » ne s'applique qu'aux chromosomes uniques et circulaires, ce qui n'est pas le cas ici.

Il est néanmoins possible d'interpréter les transformations d'états de caractères sur les branches des arbres obtenus. Prenons par exemple quelques transformations sur l'arbre obtenu avec le codage « Position relative » (figure 4.11) :

- entre le nœud 10 et le nœud 9 :
  - « Position relative de 1prox2 » : (1dist1,1prox1) → (1prox1,télo).
- entre le nœud 9 et *M. awashencsis* :
  - « Position relative de 1prox1 » : (1prox2,centro) → (1dist1,1prox2).

Sur la branche 10–9, la position relative du segment 1prox2 passe de (1dist1,1prox1) à (1prox1,télo), ce qui peut être interprété comme la fission du chromosome 1 de l'ancêtre commun hypothétique de l'ensemble des *Mastomys*. Sur la branche qui va du nœud 9 à *M. awashencsis*, la position du segment 1prox1 passe de (1prox2,centro) à (1dist1,1prox2) ce qui peut être interprété comme la fusion de deux chromosomes pour donner le chromosome 1 de *M. awashencsis*. Cette série d'événements est illustrée sur la figure 4.14.

Cela montre que les événements qui ont conduit au nombre réduit de chromosomes de *M. natalensis* et *M. awashencsis* ont eu lieu de façon indépendante ce qui explique que ces deux espèces ne soient pas retrouvées ensembles.

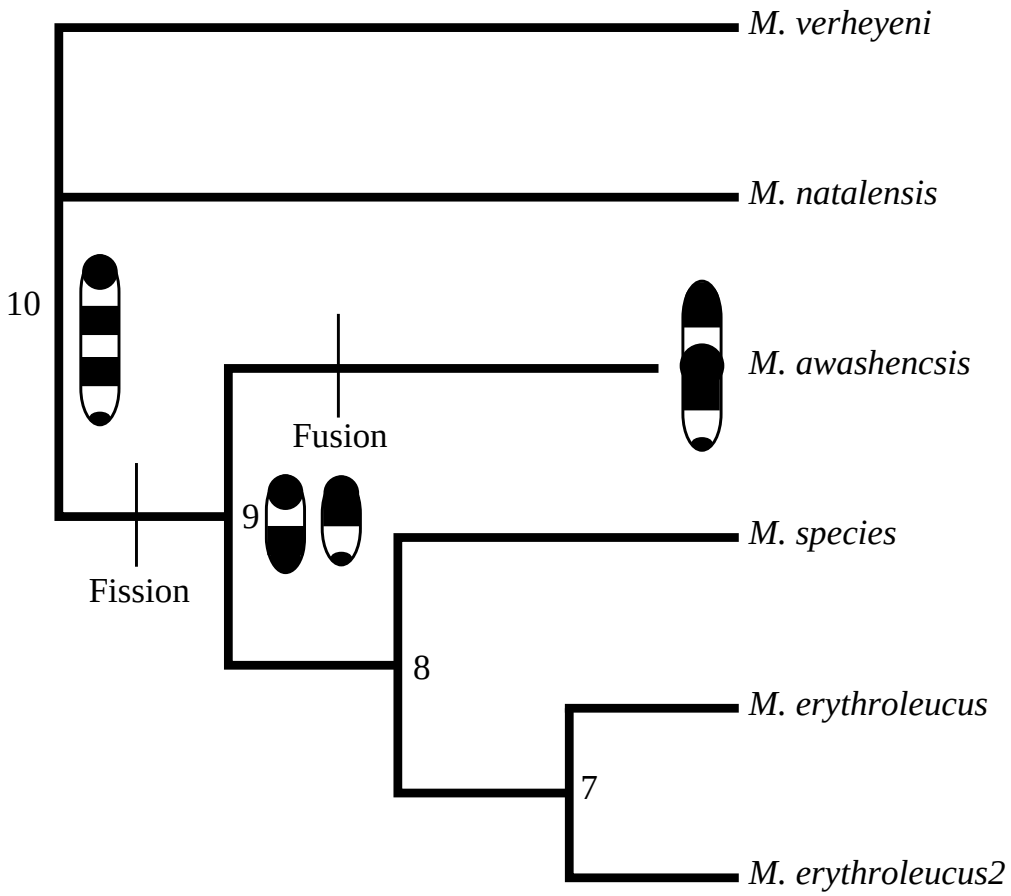


FIGURE 4.14 – Reconstitution d'événements évolutifs pour un chromosome du genre *Mastomys* sur l'arbre de la figure 4.11. Deux événements sont reconstitués : la fission du chromosome 1 de l'ancêtre commun hypothétique de l'ensemble des *Mastomys* (entre les nœuds 10 et 9) et la fusion de deux chromosomes pour donner le chromosome 1 de *M. awashencsis* (entre le nœud 9 et *M. awashencsis*).

4.2.4 Conclusion

Cette étude phylogénétique des chromosomes du genre *Mastomys* montre que notre approche de codage peut s'appliquer avec succès aux données d'organisation chromosomique. Pour coder l'organisation des chromosomes, il est nécessaire de réexprimer les homologies

entre chromosomes sous la forme d'ordres de segments chromosomiques homologues. Le fait que le génome soit constitué de plusieurs chromosomes ne pose pas de problème particulier, l'ordre des segments sur chaque chromosome étant codé indépendamment. Par contre, le fait que les chromosomes soient linéaires nécessite d'adapter le codage de la position des segments situés aux extrémités des chromosomes. Les résultats obtenus sont très intéressants car ils permettent de confirmer ceux obtenus avec le gène *cytochrome b* et permettent en plus de préciser la position de *Mastomys awashencsis*.

Conclusion

Nous proposons ici une approche originale de codage de l'organisation du génome dans le cadre de la méthode cladistique. L'organisation du génome de chaque taxon étudié est décrite sur la base d'un canevas établi à partir de la comparaison globale de l'ensemble des taxons sans introduire d'hypothèses *a priori* sur les événements évolutifs responsables de l'évolution du génome. La méthode cladistique permettant dans un deuxième temps, un retour aux caractères pour interpréter ces événements sur l'arbre obtenu. Notre approche s'articule donc en deux phases, une phase d'exploration de l'organisation du génome pour établir des hypothèses d'homologies et une phase de codage des états de caractères établis lors de la première phase. La matrice obtenue est ensuite traitée par la méthode cladistique de façon classique.

Nous proposons deux codages, « Position relative » et « Jonctions », avec deux options (AB=BA et AB≠BA) que nous avons analysé et comparé avec le codage « Jonctions signées » de Cosner *et al.* (2000b). Ces trois codages permettent de représenter l'ordre d'unités fonctionnelles homologues sur les chromosomes des taxons étudiés. Ces unités fonctionnelles peuvent être par exemple des gènes ou encore des segments chromosomiques homologues.

Nous avons montré que l'option AB≠BA introduit une redondance importante entre les caractères de position (« Position relative d'une unité » et « Présence Absence d'une jonction ») et les caractères « Orientation d'une unité ». Cette redondance a pour conséquence de surpondérer les événements d'inversion et par conséquent nous considérons qu'il est préférable de ne pas l'utiliser.

Les deux codages entièrement fondés sur la présence/absence (« Jonctions » et « Jonctions signées ») introduisent une redondance qui est intrinsèque à ce type de codage mais ils présentent l'avantage que les temps de calcul sont beaucoup moins grands qu'avec le codage à états multiples (« Position relative »). De plus, le codage « Jonctions signées » ne nécessite pas que l'orientation des unités fonctionnelles soit déterminée de façon absolue puisque l'orientation et la position sont codées conjointement et de façon relative, ce qui est avantage important sur les deux autres codages. En effet, nous avons montré que les événements évolutifs reconstitués *a posteriori* sur les arbres obtenus avec le codage « Position relative » impliquent souvent que les changements de positions et d'orientations se fassent indépendamment. Néanmoins, le codage « Position relative » présente l'avantage de fournir suffisamment d'informations au niveau des nœuds internes pour permettre la reconstitution de génomes ancestraux complets. Cette possibilité est un atout important pour pouvoir interpréter *a posteriori* l'évolution des génomes sur l'arbre obtenu.

Nous avons appliqué ces trois possibilités de codages à des données réelles : le génome mitochondrial des métazoaires et les chromosomes du genre *Mastomys*. Ces analyses ont permis de montrer que notre approche peut s'appliquer avec succès à l'organisation de génomes différents, que ce soient des génomes constitués d'un seul chromosome circulaire ou bien des génomes constitués de plusieurs chromosomes linéaires et ce, avec un minimum d'adaptations. Les résultats obtenus avec ces deux analyses sont encourageants quant à l'utilisation de ces données dans un cadre phylogénétique et mettent l'accent sur la nécessité d'élargir l'échantillonnage notamment pour le génome mitochondrial des métazoaires.

Nous espérons que la formalisation que nous proposons de l'utilisation de l'organisation du génome dans le cadre de la méthode cladistique débouchera sur de nouvelles applications. Ce domaine n'en est qu'à ses débuts et il reste de nombreuses possibilités de codage à explorer. Un codage alliant les avantages des différents codages que nous avons présentés serait très appréciable. Par exemple, un codage à états multiples prenant en compte simultanément la position et l'orientation, ce qui permettrait d'avoir suffisamment d'information pour reconstituer les gé-

nomes ancestraux tout en évitant que les événements de changement de position d'orientation soit séparés. De plus il serait très intéressant de trouver un algorithme permettant de reconstituer des génomes ancestraux constitués de chromosomes multiples et/ou linéaires. Enfin, il serait intéressant de simuler artificiellement l'évolution de génomes pour évaluer plus finement les différents codages, en simulant par exemple, des taux d'évolution variables avec différents types de topologies. Bien qu'il soit difficile dans ce genre d'études, de faire la distinction entre les implications du codage et celles de la méthode cladistique elle-même, cela permettrait de tester différents modèles évolutifs et la capacité de notre approche à les analyser.

Bibliographie

Aguinaldo, A. M., Turbeville, J. M., Linford, L., Rivera, M., Garey, J. R., Raff, R. A. et Lake, J. A. (1997). Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature*, 387(6632) : 489–493.

Albà, M. M., Das, R., Orengo, C. A. et Kellam, P. (2001). Genomewide function conservation and phylogeny in the Herpesviridae. *Genome Research*, 11(1) : 43–54.

Anderson, S., Bankier, A. T., Barrell, B. G., de Bruijn, M. H., Coulson, A., Drouin, J., Eperon, I. C., Nierlich, D. P., Roe, B. A., Sanger, F., Schreier, P. H., Smith, A. J., Staden, R. et Young, I. G. (1981). Sequence and organization of the human mitochondrial genome. *Nature*, 290(5806) : 457–465.

Anderson, S., de Bruijn, M. H., Coulson, A., Eperon, I. C., Sanger, F. et Young, I. G. (1982). Complete sequence of bovine mitochondrial DNA. Conserved features of the mammalian mitochondrial genome. *Journal of Molecular Biology*, 156(4) : 683–717.

Arnason, U. et Gullberg, A. (1993). Comparison between the complete mtDNA sequences of the blue and the fin whale, two species that can hybridize in nature. *Journal of Molecular Evolution*, 37(4) : 312–322.

Arnason, U., Gullberg, A., Gretarsdottir, S., Ursing, B. M. et Janke, A. (2000a). The mitochondrial genome of the sperm whale and a new molecular reference for estimating eutherian divergence dates. *Journal of Molecular Evolution*, 50(6) : 569–578.

Arnason, U., Gullberg, A. et Janke, A. (1997). Phylogenetic analyses of mitochondrial DNA suggest a sister group relationship between Xenarthra (Edentata) and Ferungulates. *Molecular Biology And Evolution*, 14(7) : 762–768.

Arnason, U., Gullberg, A. et Janke, A. (1998). Molecular timing of primate divergences as estimated by two nonprimate calibration points. *Journal of Molecular Evolution*, 47(6) : 718–727.

Arnason, U., Gullberg, A. et Janke, A. (1999). The mitochondrial DNA molecule of the aardvark, *Orycteropus afer*, and the position of the Tubulidentata in the eutherian tree. *Proceedings of The Royal Society of London. Series B : Biological Sciences*, 266(1417) : 339–345.

Arnason, U., Gullberg, A., Johnsson, E. et Ledje, C. (1993). The nucleotide sequence of the mitochondrial DNA molecule of the grey seal, *Halichoerus grypus*, and a comparison with

mitochondrial sequences of other true seals. *Journal of Molecular Evolution*, 37(4) : 323–330.

Arnason, U., Gullberg, A., Schweizer Burguete, A. S. et Janke, A. (2000b). Molecular estimates of primate divergences and new hypotheses for primate dispersal and the origin of modern humans. *Hereditas*, 133(3) : 217–228.

Arnason, U., Gullberg, A. et Widegren, B. (1991). The complete nucleotide sequence of the mitochondrial DNA of the fin whale, *Balaenoptera physalus*. *Journal of Molecular Evolution*, 33(6) : 556–568.

Arnason, U., Gullberg, A. et Xu, X. (1996a). A complete mitochondrial DNA molecule of the white-handed gibbon, *Hylobates lar*, and comparison among individual mitochondrial genes of all hominoid genera. *Hereditas*, 124 : 185–189.

Arnason, U. et Johnsson, E. (1992). The complete mitochondrial DNA sequence of the harbor seal, *Phoca vitulina*. *Journal of Molecular Evolution*, 34(6) : 493–505.

Arnason, U., Xu, X. et Gullberg, A. (1996b). Comparison between the complete mitochondrial DNA sequences of *Homo* and the common chimpanzee based on nonchimeric sequences. *Journal of Molecular Evolution*, 42(2) : 145–152.

Asakawa, S., Himeno, H., Miura, K. et Watanabe, K. (1995). Nucleotide sequence and gene organization of the starfish *Asterina pectinifera* mitochondrial genome. *Genetics*, 140(3) : 1047–1060.

Atlan, A. et Couvet, D. (1993). A model simulating the dynamics of plant mitochondrial genomes. *Genetics*, 135(1) : 213–222.

Bafna, V. et Pevzner, P. A. (1995). Sorting by transpositions. *In* Galil, Z. et Ukkonen, E. (rédacteurs), *Proceedings of the Sixth Symposium on Combinatorial Pattern Matching*, tome 937 de *Lecture Notes in Computer Science*, pages 614–623. Springer-Verlag, New York.

Barriel, V. et Tassy, P. (1993). Characters, observations and steps : comment on Lipscomb's "parsimony, homology and the analysis of multistate characters". *Cladistics*, 9(2) : 223–232.

Beagley, C. T., Okada, N. A. et Wolstenholme, D. R. (1996). Two mitochondrial group I introns in a metazoan, the sea anemone *Metridium senile* : one intron contains genes for subunits 1 and 3 of NADH dehydrogenase. *Proceedings of The National Academy of Sciences of The United States of America*, 93(11) : 5619–5623.

Beagley, C. T., Okimoto, R. et Wolstenholme, D. R. (1998). The mitochondrial genome of the sea anemone *Metridium senile* (Cnidaria) : introns, a paucity of tRNA genes, and a near-standard genetic code. *Genetics*, 148(3) : 1091–1108.

Beard, C. B., Hamm, D. M. et Collins, F. H. (1993). The mitochondrial genome of the mosquito *Anopheles gambiae* : DNA sequence, genome organization, and comparisons with mitochondrial sequences of other insects. *Insect Mololecular Biology*, 2(2) : 103–124.

Beaton, M. J., Roger, A. J. et Cavalier-Smith, T. (1998). Sequence analysis of the mitochondrial genome of *Sarcophyton glaucum* : conserved gene order among octocorals. *Journal of Molecular Evolution*, 47(6) : 697–708.

Bibb, M. J., Van Etten, R. A., Wright, C. T., Walberg, M. W. et Clayton, D. A. (1981). Sequence and gene organization of mouse mitochondrial DNA. *Cell*, 26(2 Pt 2) : 167–180.

Black, W. C. t. et Roehrdanz, R. L. (1998). Mitochondrial gene order is not conserved in arthropods : prostriate and metastriate tick mitochondrial genomes. *Molecular Biology And Evolution*, 15(12) : 1772–1785.

Blanchette, M., Kunisawa, T. et Sankoff, D. (1996). Parametric genome rearrangement. *Gene*, 172(1) : GC 11–17.

Blanchette, M., Kunisawa, T. et Sankoff, D. (1999). Gene Order Breakpoint Evidence in Animal Mitochondrial Phylogeny. *Journal of Molecular Evolution*, 49(2) : 193–203.

Boore, J. L. et Brown, W. M. (1994). Complete DNA sequence of the mitochondrial genome of the black chiton, *Katharina tunicata*. *Genetics*, 138(2) : 423–443.

Boore, J. L. et Brown, W. M. (1995). Complete sequence of the mitochondrial DNA of the annelid worm *Lumbricus terrestris*. *Genetics*, 141(1) : 305–319.

Boore, J. L. et Brown, W. M. (2000). Mitochondrial genomes of *Galathealinum*, *Helobdella*, and *Platynereis* : sequence and gene arrangement comparisons indicate that Pogonophora is not a phylum and Annelida and Arthropoda are not sister taxa. *Molecular Biology And Evolution*, 17(1) : 87–106.

Boore, J. L., Collins, T. M., Stanton, D. J., Daehler, L. L. et Brown, W. M. (1995). Deducing the pattern of arthropod phylogeny from mitochondrial DNA rearrangements. *Nature*, 376(6536) : 163–165.

Boore, J. L., Daehler, L. L. et Brown, W. M. (1999). Complete sequence, gene arrangement, and genetic code of mitochondrial DNA of the cephalochordate *Branchiostoma floridae* (Amphioxus). *Molecular Biology And Evolution*, 16(3) : 410–418.

Boore, J. L., Lavrov, D. V. et Brown, W. M. (1998). Gene translocation links insects and crustaceans. *Nature*, 392(6677) : 667–668.

Bridge, D., Cunningham, C. W., Schierwater, B., DeSalle, R. et Buss, L. W. (1992). Class-level relationships in the phylum Cnidaria : evidence from mitochondrial genome structure. *Proceedings of The National Academy of Sciences of The United States of America*, 89(18) : 8750–8753.

Brusca, R. C. et Brusca, G. J. (1990). *Invertebrates*. Sinauer Associates, Sunderland Massachusetts.

Campbell, N. J. et Barker, S. C. (1999). The novel mitochondrial gene arrangement of the cattle tick, *Boophilus microplus* : fivefold tandem repetition of a coding region. *Molecular Biology And Evolution*, 16(6) : 732–740.

Cantatore, P., Roberti, M., Rainaldi, G., Gadaleta, M. N. et Saccone, C. (1989). The complete nucleotide sequence, gene organization, and genetic code of the mitochondrial genome of *Paracentrotus lividus*. *Journal of Biological Chemistry*, 264(19) : 10965–10975.

Cao, Y., Waddell, P. J., Okada, N. et Hasegawa, M. (1998). The complete mitochondrial DNA sequence of the shark *Mustelus manazo* : evaluating rooting contradictions to living bony vertebrates. *Molecular Biology And Evolution*, 15(12) : 1637–1646.

Caprara, A. (1997). Sorting by reversal is difficult. *In* Istrail, S., Pevzner, P. A. et Waterman, M. S. (rédacteurs), *Proceedings of the First Annual Internationnal Conference on Computational Molecular Biology (RECOMB-97)*, pages 75–83. ACM press, Santa Fe, New Mexico.

Castresana, J., Feldmaier-Fuchs, G., Yokobori, S., Satoh, N. et Pääbo, S. (1998). The mitochondrial genome of the hemichordate *Balanoglossus carnosus* and the evolution of deuterostome mitochondria. *Genetics*, 150(3) : 1115–11123.

Chang, Y. S., Huang, F. L. et Lo, T. B. (1994). The complete nucleotide sequence and gene organization of carp (*Cyprinus carpio*) mitochondrial genome. *Journal of Molecular Evolution*, 38(2) : 138–155.

Clary, D. O. et Wolstenholme, D. R. (1985). The mitochondrial DNA molecular of *Drosophila yakuba* : nucleotide sequence, gene organization, and genetic code. *Journal of Molecular Evolution*, 22(3) : 252–271.

Cooper, A., Lalueza-Fox, C., Anderson, S., Rambaut, A., Austin, J. et Ward, R. (2001). Complete mitochondrial genome sequences of two extinct moas clarify ratite evolution. *Nature*, 409(6821) : 704–707.

Cosner, M. E., Jansen, R. K., Moret, B. M., Raubeson, L. A., Wang, L.-S., Warnow, T. et Wyman, S. (2000a). An Empirical Comparison of Phylogenetic Methods on Chloroplast Gene Order Data in Campanulaceae. *In* Sankoff, D. et Nadeau, J. H. (rédacteurs), *Comparative Genomics. Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and the Evolution of Gene Families*, tome 1 de *Computational Biology*, pages 99–121. Kluwer Academic, Boston, MA.

Cosner, M. E., Jansen, R. K., Moret, B. M., Raubeson, L. A., Wang, L.-S., Warnow, T. et Wyman, S. (2000b). A New Fast Heuristic for Computing the Breakpoint Phylogeny and Experimental Phylogenetic Analyses of Real and Synthetic Data. *In 8th International Conference on Intelligent Systems for Molecular Biology*. San Diego.

Crease, T. J. (1999). The complete sequence of the mitochondrial genome of *Daphnia pulex* (Cladocera : Crustacea). *Gene*, 233(1-2) : 89–99.

Crozier, R. H. et Crozier, Y. C. (1993). The mitochondrial genome of the honeybee *Apis mellifera* : complete sequence and genome organization. *Genetics*, 133(1) : 97–117.

Dandekar, T., Snel, B., Huynen, M. et Bork, P. (1998). Conservation of gene order : a fingerprint of proteins that physically interact. *Trends in Biochemical Sciences*, 23 : 324–328.

De Giorgi, C., Martiradonna, A., Lanave, C. et Saccone, C. (1996). Complete sequence of the mitochondrial DNA in the sea urchin *Arbacia lixula* : conserved features of the echinoid mitochondrial genome. *Molecular Phylogenetics And Evolution*, 5(2) : 323–332.

de Pinna, M. C. C. (1991). Concepts and tests of homology in the cladistic paradigm. *Cladistics*, 7 : 367–394.

Delarbre, C., Escriva, H., Gallut, C., Barriel, V., Kourilsky, P., Janvier, P., Laudet, V. et Gachelin, G. (2000). The Complete Nucleotide Sequence of the Mitochondrial DNA of the Agnathan *Lampetra fluviatilis* : Bearings on the Phylogeny of Cyclostomes. *Molecular Biology And Evolution*, 17(4) : 519–529.

Delarbre, C., Gallut, C., Barriel, V., Janvier, P. et Gachelin, G. (2002). Complete Mitochondrial DNA of the Hagfish, *Eptatretus burgeri* : The Comparative Analysis of Mitochondrial DNA Sequences Strongly Supports the Cyclostome Monophyly. *Molecular Phylogenetics And Evolution*, 22(2) : 184–192.

Delarbre, C., Spruyt, N., Delmarre, C., Gallut, C., Barriel, V., Janvier, P., Laudet, V. et Gachelin, G. (1998). The Complete Nucleotide Sequence of the Mitochondrial DNA of the Dogfish, *Scyliorhinus canicula*. *Genetics*, 150(1) : 331–344.

D'Erchia, A. M., Gissi, C., Pesole, G., Saccone, C. et Arnason, U. (1996). The guinea-pig is not a rodent. *Nature*, 381(6583) : 597–600.

Desjardins, P. et Morais, R. (1990). Sequence and gene organization of the chicken mitochondrial genome. A novel gene order in higher vertebrates. *Journal of Molecular Biology*, 212(4) : 599–634.

Dicks, J. (2000). CHROMTREE : Maximum Likelihood Estimation of Chomosmal Phylogenies. *In* Sankoff, D. et Nadeau, J. H. (rédacteurs), *Comparative Genomics. Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and the Evolution of Gene Families*, tome 1 de *Computational Biology*, pages 333–342. Kluwer Academic, Boston, MA.

Dobzhansky, T. et Sturtevant, A. H. (1938). Inversions in the chromosomes of *Drosophila pseudoobscura*. *Genetics*, 23 : 28–64.

Doiron, S., Blier, P. U. et Bernatchez, L. (1999). A comparative analysis of complete sequence of mitochondrial genome between brook char (*Salvelinus fontinalis*) and arctic char (*S. alpinus*). *Non publié*.

Doolittle, W. F. (2000). Lateral genomics. *Trends in Genetics*, 15(12) : M5–M8. Ford@is.dal.ca.

Ehrlich, J., Sankoff, D. et Nadeau, J. H. (1997). Synteny conservation and chromosome rearrangements during mammalian evolution. *Genetics*, 147(1) : 289–296.

Farris, J. S. (1972). Estimating phylogenetic trees from distance matrices. *American Naturalist*, 106 : 645–668.

Fitz-Gibbon, S. T. et House, C. H. (1999). Whole genome-based phylogenetic analysis of free-living microorganisms. *Nucleic Acids Research*, 27(21) : 4218–4222.

Flook, P. K., Rowell, C. H. et Gellissen, G. (1995). The sequence, organization, and evolution of the *Locusta migratoria* mitochondrial genome. *Journal of Molecular Evolution*, 41(6) : 928–941.

Fukunaga, M. (2000). *Echinococcus multilocularis* mitochondrial DNA sequence. *Non publié*.

Gadaleta, G., Pepe, G., De Candia, G., Quagliariello, C., Sbisa, E. et Saccone, C. (1989). The complete nucleotide sequence of the *Rattus norvegicus* mitochondrial genome : cryptic signals revealed by comparative analysis between vertebrates. *Journal of Molecular Evolution*, 28(6) : 497–516.

Gallut, C. et Barriel, V. (2001). Cladistic coding of genomic maps. *Cladistics*, Soumis.

Gallut, C., Barriel, V. et Vignes-Lebbe, R. (2000). Gene Order and Phylogenetic Information. *In* Sankoff, D. et Nadeau, J. H. (rédacteurs), *Comparative Genomics. Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and the Evolution of Gene Families*, tome 1 de *Computational Biology*, pages 123–132. Kluwer Academic, Boston, MA.

Garesse, R. (1988). *Drosophila melanogaster* mitochondrial DNA : gene organization and evolutionary considerations. *Genetics*, 118(4) : 649–663.

Gissi, C., Gullberg, A. et Arnason, U. (1998). The complete mitochondrial DNA sequence of the rabbit, *Oryctolagus cuniculus*. *Genomics*, 50(2) : 161–169.

Gu, X. (2000). A Simple Evolutionnary Model for Genome Phylogeny Based on Gene Content. *In* Sankoff, D. et Nadeau, J. H. (rédacteurs), *Comparative Genomics. Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and the Evolution of Gene Families*, tome 1 de *Computational Biology*, pages 515–523. Kluwer Academic, Boston, MA.

Haddrath, O. et Baker, A. J. (2001). Complete mitochondrial DNA genome sequences of extinct birds : ratite phylogenetics and the vicariance biogeography hypothesis. *Proceedings of The Royal Society of London. Series B : Biological Sciences*, 268(1470) : 939–945.

Halanych, K. M., Bacheller, J. D., Aguinaldo, A. M., Liva, S. M., Hillis, D. M. et Lake, J. A. (1995). Evidence from 18S ribosomal DNA that the lophophorates are protostome animals. *Science*, 267(5204) : 1641–1643.

Hannenhalli, S. et Pevzner, P. A. (1999). Transforming cabbage into turnip : polynomial algorithm for sorting signed permutations by reversals. *Journal of the ACM*, 46(1) : 1–27.

Härlid, A. et Arnason, U. (1999). Analyses of mitochondrial DNA nest ratite birds within the Neognathae : supporting a neotenous origin of ratite morphological characters. *Proceedings of The Royal Society of London. Series B : Biological Sciences*, 266(1416) : 305–309.

Härlid, A., Janke, A. et Arnason, U. (1997). The mtDNA sequence of the ostrich and the divergence between paleognathous and neognathous birds. *Molecular Biology And Evolution*, 14(7) : 754–761.

Härlid, A., Janke, A. et Arnason, U. (1998). The complete mitochondrial genome of *Rhea americana* and early avian divergences. *Journal of Molecular Evolution*, 45(6) : 669–679.

Hatzoglou, E., Rodakis, G. C. et Lecanidou, R. (1995). Complete sequence and gene organization of the mitochondrial genome of the land snail *Albinaria coerulea*. *Genetics*, 140(4) : 1353–1366.

Hauf, J., Chalwatzis, N., Joger, U. et Zimmermann, F. K. (1999). The complete mitochondrial sequence of the African Elephant (*Loxodonta africana*) and its implication on the assessment of the systematic position of the Proboscidea. *Non publié*.

Hickerson, M. J. et Cunningham, C. W. (2000). Dramatic mitochondrial gene rearrangements in the hermit crab *Pagurus longicarpus* (Crustacea, anomura). *Molecular Biology and Evolution*, 17(4) : 639–644.

Hiendleder, S., Lewalski, H., Wassmuth, R. et Janke, A. (1998). The complete mitochondrial DNA sequence of the domestic sheep (*Ovis aries*) and comparison with the other major ovine haplotype. *Journal of Molecular Evolution*, 47(4) : 441–448.

Hoffmann, R. J., Boore, J. L. et Brown, W. M. (1992). A novel mitochondrial genome organization for the blue mussel, *Mytilus edulis*. *Genetics*, 131(2) : 397–412.

Horai, S., Hayasaka, K., Kondo, R., Tsugane, K. et Takahata, N. (1995). Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. *Proceedings of The National Academy of Sciences of The United States of America*, 92(2) : 532–536.

Hurst, C. D., Bartlett, S. E., Davidson, W. S. et Bruce, I. J. (1999). The complete mitochondrial DNA sequence of the Atlantic salmon, *Salmo salar*. *Gene*, 239(2) : 237–242.

Huynen, M., Snel, B., Lathe, r., W. et Bork, P. (2000). Predicting protein function by genomic context : quantitative evaluation and qualitative inferences. *Genome Research*, 10(8) : 1204–1210.

Inoue, J. G., Miya, M., Aoyama, J., Ishikawa, S., Tsukamoto, K. et Nishida, M. (2001a). Complete Mitochondrial DNA Sequence of the Japanese Eel, *Anguilla japonica*. *Fisheries Science*, 67(1) : 118–125.

Inoue, J. G., Miya, M., Tsukamoto, K. et Nishida, M. (2000). Complete mitochondrial DNA sequence of the Japanese sardine *Sardinops melanostictus*. *Fisheries Science*, 66(5) : 924–932.

Inoue, J. G., Miya, M., Tsukamoto, K. et Nishida, M. (2001b). Complete Mitochondrial DNA Sequence of *Conger myriaster* (Teleostei : Anguilliformes) : Novel Gene Order for Vertebrate Mitochondrial Genomes and the Phylogenetic Implications for Anguilliform Families. *Journal of Molecular Evolution*, 52(4) : 311–320.

Jacobs, H. T., Elliott, D. J., Math, V. B. et Farquharson, A. (1988). Nucleotide sequence and gene organization of sea urchin mitochondrial DNA. *Journal of Molecular Biology*, 202(2) : 185–217.

Janke, A. et Arnason, U. (1997). The complete mitochondrial genome of *Alligator mississippiensis* and the separation between recent archosauria (birds and crocodiles). *Molecular Biology And Evolution*, 14(12) : 1266–1272.

Janke, A., Erpenbeck, D., Nilsson, M. et Arnason, U. (2001). The mitochondrial genomes of the iguana (*Iguana iguana*) and the caiman (*Caiman crocodylus*) : implications for amniote phylogeny. *Proceedings of The Royal Society of London. Series B : Biological Sciences*, 268(1467) : 623–631.

Janke, A., Feldmaier-Fuchs, G., Thomas, W. K., von Haeseler, A. et Pääbo, S. (1994). The marsupial mitochondrial genome and the evolution of placental mammals. *Genetics*, 137(1) : 243–256.

Janke, A., Gemmell, N. J., Feldmaier-Fuchs, G., von Haeseler, A. et Pääbo, S. (1996). The mitochondrial genome of a monotreme–the platypus (*Ornithorhynchus anatinus*). *Journal of Molecular Evolution*, 42(2) : 153–159.

Janke, A., Xu, X. et Arnason, U. (1997). The complete mitochondrial genome of the wallaroo (*Macropus robustus*) and the phylogenetic relationship among Monotremata, Marsupialia, and Eutheria. *Proceedings of The National Academy of Sciences of The United States of America*, 94(4) : 1276–1281.

Johansen, S. et Bakke, I. (1996). The complete mitochondrial DNA sequence of Atlantic cod (*Gadus morhua*) : relevance to taxonomic studies among codfishes. *Molecular Marine Biology And Biotechnology*, 5(3) : 203–214.

Kececioglu, J. et Ravi, R. (1995). Of mice and men. Evolutionary distances between genomes under translocation. *In Proceedings of the 6th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 604–613.

Keddie, E. M., Higazi, T. et Unnasch, T. R. (1998). The mitochondrial genome of *Onchocerca volvulus* : sequence, structure and phylogenetic analysis. *Molecular And Biochemical Parasitology*, 95(1) : 111–127.

Kim, K. S., Lee, S. E., Jeong, H. W. et Ha, J. H. (1998). The complete nucleotide sequence of the domestic dog (*Canis familiaris*) mitochondrial genome. *Molecular Phylogenetics And Evolution*, 10(2) : 210–220.

Kim, S. H., Je, E. Y. et Park, D. (2000). *Conger myriaster* mitochondrial DNA. *Non publié*.

Krettek, A., Gullberg, A. et Arnason, U. (1995). Sequence analysis of the complete mitochondrial DNA molecule of the hedgehog, *Erinaceus europaeus*, and the phylogenetic position of the Lipotyphla. *Journal of Molecular Evolution*, 41(6) : 952–957.

Kumazawa, Y. et Nishida, M. (1999). Complete mitochondrial DNA sequences of the green turtle and blue- tailed mole skink : statistical evidence for archosaurian affinity of turtles. *Molecular Biology And Evolution*, 16(6) : 784–792.

Kumazawa, Y., Ota, H., Nishida, M. et Ozawa, T. (1998). The complete nucleotide sequence of a snake (*Dinodon semicarinatus*) mitochondrial genome with two identical control regions. *Genetics*, 150(1) : 313–329.

Kurabayashi, A. et Ueshima, R. (2000). Complete sequence of the mitochondrial DNA of the primitive opisthobranch gastropod *Pupa strigosa* : systematic implication of the genome organization. *Molecular Biology and Evolution*, 17(2) : 266–277.

La Roche, J., Snyder, M., Cook, D. I., Fuller, K. et Zouros, E. (1990). Molecular characterization of a repeat element causing large-scale size variation in the mitochondrial DNA of the sea scallop *Placopecten magellanicus*. *Molecular Biology And Evolution*, 7(1) : 45–64.

Lavrov, D. V., Boore, J. L. et Brown, W. M. (2000a). The complete mitochondrial DNA sequence of the horseshoe crab limulus polyphemus. *Molecular Biology and Evolution*, 17(5) : 813–824.

Lavrov, D. V. et Brown, W. M. (2001). *Trichinella spiralis* mtDNA. A nematode mitochondrial genome that encodes a putative atp8 and normally structured trnas and has a gene arrangement relatable to those of coelomate metazoans. *Genetics*, 157(2) : 621–637.

Lavrov, D. V., Brown, W. M. et Boore, J. L. (2000b). A novel type of RNA editing occurs in the mitochondrial tRNAs of the centipede *Lithobius forficatus*. *Proceedings of The National Academy of Sciences of The United States of America*, 97(25) : 13738–13742.

Le, T. H., Blair, D., Agatsuma, T., Humair, P. F., Campbell, N. J., Iwagami, M., Littlewood, D. T., Peacock, B., Johnston, D. A., Bartley, J., Rollinson, D., Herniou, E. A., Zarlenga, D. S. et McManus, D. P. (2000). Phylogenies inferred from mitochondrial gene orders-a cautionary tale from the parasitic flatworms. *Molecular Biology And Evolution*, 17(7) : 1123–1125.

Lebbe, J. (1996). *Informatique et systématique*, tome 14 de *Biosystema*. SFS, Paris.

Lee, J. S., Kim, Y. S., Sung, S. H., Hwang, J. S., Lee, D. S. et Suh, D. S. (2000). The complete mitochondrial genome of *Bombyx mori*. *Non publié.*.

Lee, W. J. et Kocher, T. D. (1995). Complete sequence of a sea lamprey (*Petromyzon marinus*) mitochondrial genome : early establishment of the vertebrate genome organization. *Genetics*, 139(2) : 873–887.

Lessinger, A. C., Junqueira, A. M., Lemos, T. A., Kemper, E. L., Vettore, A. L., da Silva, F. R., Arruda, P. et Azeredo-Espin, A. M. L. (2000). The mitochondrial genome of the primary screwworm fly *Cochliomyia hominivorax* (Diptera : Calliphoridae). *Non publié.*.

Lin, Y. H. et Penny, D. (2001). Implications for bat evolution from two new complete mitochondrial genomes. *Molecular Biology And Evolution*, 18(4) : 684–688.

Lopez, J. V., Cevario, S. et O'Brien, S. J. (1996). Complete nucleotide sequences of the domestic cat (*Felis catus*) mitochondrial genome and a transposed mtDNA tandem repeat (Numt) in the nuclear genome. *Genomics*, 33(2) : 229–246.

Lyons, L. A., Laughlin, T. F., Copeland, N. G., Jenkins, N. A., Womack, J. E. et O'Brien, S. J. (1997). Comparative anchor tagged sequences (CATS) for integrative mapping of mammalian genomes. *Nature Genetics*, 15(1) : 47–56.

Macey, J. R., Larson, A., Ananjeva, N. B., Fang, Z. et Papenfuss, T. J. (1997). Two novel gene orders and the role of light-strand replication in rearrangement of the vertebrate mitochondrial genome. *Molecular Biology And Evolution*, 14(1) : 91–104.

Macey, J. R., Schulte, J. A., Larson, A. et Papenfuss, T. J. (1998). Tandem duplication via light-strand synthesis may provide a precursor for mitochondrial genomic rearrangement. *Molecular Biology And Evolution*, 15(1) : 71–75.

Milam, J. E., Broughton, R. E. et Roe, B. A. (2000). Complete Mitochondrial Genome Of *Danio rerio* (Zebrafish). *Non publié.*.

Mindell, D. P., Sorenson, M. D., Dimcheff, D. E., Hasegawa, M., Ast, J. C. et Yuri, T. (1999). Interordinal relationships of birds and other reptiles bases on whole mitochondrial genomes. *Systematic Biology*, 48(1) : 138–152.

Mitchell, S. E., Cockburn, A. F. et Seawright, J. A. (1993). The mitochondrial genome of *Anopheles quadrimaculatus* species A : complete nucleotide sequence and gene organization. *Genome*, 36(6) : 1058–1073.

Miya, M. et Nishida, M. (1999). Organization of the mitochondrial genome of a deep-sea fish *Gonostoma gracile* (Teleostei : Stomiiformes) : the first example of tRNA rearrangements in bony fish. *Marine Biotechnology*, 1(5) : 416–426.

Miya, M. et Nishida, M. (2000). Use of Mitogenomic Information in Teleostean Molecular Phylogenetics : A Tree-Based Exploration under the Maximum-Parsimony Optimality Criterion. *Molecular Phylogenetics And Evolution*, 17 : 437–455.

Mouchaty, S. K., Catzeflis, F. M., Janke, A. et Arnason, U. (2001). Molecular Evidence of an African Phiomorpha-South American Caviomorpha Clade and Support for Hystricognathi Based on the Complete Mitochondrial Genome of the Cane Rat (*Thryonomys swinderianus*). *Molecular Phylogenetics and Evolution*, 18(1) : 127–135.

Mouchaty, S. K., Gullberg, A., Janke, A. et Arnason, U. (2000a). The phylogenetic position of the Talpidae within eutheria based on analysis of complete mitochondrial sequences. *Molecular Biology and Evolution*, 17(1) : 60–67.

Mouchaty, S. K., Gullberg, A., Janke, A. et Arnason, U. (2000b). Phylogenetic position of the Tenrecs (Mammalia : Tenrecidae) of Madagascar based on analysis of the complete mitochondrial genome sequence of *Echinops telfairi*. *Zoologica Scripta*, 29(4) : 307–317.

Murakami, M., Yamashita, Y. et Fujitani, H. (1998). The complete sequence of mitochondrial genome from a gynogenetic triploid 'ginbuna' (*Carassius auratus langsdorfi*). *Zoological Science*, 15 : 335–337.

Nadeau, J. H. et Sankoff, D. (1997). Landmarks in the Rosetta Stone of mammalian comparative maps. *Nature Genetics*, 15(1) : 6–7.

Nadeau, J. H. et Taylor, B. A. (1984). Length of chromosomal segments conserved since divergence of man and mouse. *Proceedings of The National Academy of Sciences of The United States of America*, 81 : 814–818.

Nardi, F., Carapelli, A., Fanciulli, P. P., Dallai, R. et Frati, F. (2001). The Complete Mitochondrial DNA Sequence of the Basal Hexapod *Tetrodontophora bielanensis* : Evidence for Heteroplasmy and tRNA Translocations. *Molecular Biology And Evolution*, 18(7) : 1293–1304.

Nikaido, M., Harada, M., Cao, Y., Hasegawa, M. et Okada, N. (2000). Monophyletic origin of the order chiroptera and its phylogenetic position among mammalia, as inferred from the complete sequence of the mitochondrial DNA of a japanese megabat, the ryukyu flying fox (*Pteropus dasymallus*). *Journal of Molecular Evolution*, 51(4) : 318–328.

Noack, K., Zardoya, R. et Meyer, A. (1996). The complete mitochondrial DNA sequence of the bichir (*Polypterus ornatipinnis*), a basal ray-finned fish : ancient establishment of the consensus vertebrate gene order. *Genetics*, 144(3) : 1165–1180.

Noguchi, Y., Endo, K., Tajima, F. et Ueshima, R. (2000). The mitochondrial genome of the brachiopod *Laqueus rubellus*. *Genetics*, 155(1) : 245–259.

Okimoto, R., Macfarlane, J. L., Clary, D. O. et Wolstenholme, D. R. (1992). The mitochondrial genomes of two nematodes, *Caenorhabditis elegans* and *Ascaris suum*. *Genetics*, 130(3) : 471–498.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D. et Maltsev, N. (1999). The use of gene clusters to infer functional coupling. *Proceedings of The National Academy of Sciences of The United States of America*, 96(6) : 2896–2901.

Pevzner, P. A. (2000). *Computational Molecular Biology. An algorithmic Approach*. Computational Molecular Biology. MIT press.

Pleijel, F. (1995). On character coding for phylogeny reconstruction. *Cladistics*, 11(3) : 309–315.

Pont-Kingdon, G. A., Okada, N. A., Macfarlane, J. L., Beagley, C. T., Watkins-Sims, C. D., Cavalier-Smith, T., Clark-Walker, G. D. et Wolstenholme, D. R. (1998). Mitochondrial DNA of the coral *Sarcophyton glaucum* contains a gene for a homologue of bacterial MutS : a possible case of gene transfer from the nucleus to the mitochondrion. *Journal of Molecular Evolution*, 46(4) : 419–431.

Pont-Kingdon, G. A., Vassort, C. G., Warrior, R., Okimoto, R., Beagley, C. T. et Wolstenholme, D. R. (2000). Mitochondrial DNA of hydra attenuata (Cnidaria) : A sequence that includes an end of one linear molecule and the genes for l-rRNA, tRNA(f-Met), tRNA(Trp), COII, and ATPase8. *Journal of Molecular Evolution*, 51(4) : 404–415.

Pumo, D. E., Finamore, P. S., Franek, W. R., Phillips, C. J., Tarzami, S. et Balzarano, D. (1998). Complete mitochondrial genome of a neotropical fruit bat, *Artibeus jamaicensis*, and a new hypothesis of the relationships of bats to other eutherian mammals. *Journal of Molecular Evolution*, 47(6) : 709–717.

Rasmussen, A. S. et Arnason, U. (1999a). Molecular studies suggest that cartilaginous fishes have a terminal position in the piscine tree. *Proceedings of The National Academy of Sciences of The United States Of America*, 96(5) : 2177–2182.

Rasmussen, A. S. et Arnason, U. (1999b). Phylogenetic studies of complete mitochondrial DNA molecules place cartilaginous fishes within the tree of bony fishes. *Journal of Molecular Evolution*, 48(1) : 118–123.

Reyes, A., Gissi, C., Pesole, G., Catzeflis, F. M. et Saccone, C. (2000). Where do rodents fit ? Evidence from the complete mitochondrial genome of *Sciurus vulgaris*. *Molecular Biology And Evolution*, 17(6) : 979–983.

Reyes, A., Pesole, G. et Saccone, C. (1998). Complete mitochondrial DNA sequence of the fat dormouse, *Glis glis* : further evidence of rodent paraphyly. *Molecular Biology And Evolution*, 15(5) : 499–505.

Roe, B. A., Ma, D. P., Wilson, R. K. et Wong, J. F. (1985). The complete nucleotide sequence of the *Xenopus laevis* mitochondrial genome. *Journal of Biological Chemistry*, 260(17) : 9759–9774.

Saitoh, K., Hayashizaki, K., Yokoyama, Y., Asahida, T., Toyohara, H. et Yamashita, Y. (2000). The complete nucleotide sequence of Japanese flounder mitochondrial genome : structural property and cue for resolving teleostean relationships. *Non publié.*.

Sankoff, D. et Blanchette, M. (1998). Multiple genome rearrangement and breakpoint phylogeny. *Journal of Computational Biology*, 5(3) : 555–570.

Sankoff, D. et Blanchette, M. (1999). Comparative Genomics via Phylogenetic Invariants for Jukes-Cantor Semigroups. *In* Gorostiza, L. et Ivanoff, G. (rédacteurs), *Proceedings of the International Conference on Stochastic Models*, Conference Proceedings series, Canadian Mathematical Society.

Sankoff, D., Deneault, M., Bryant, D., Lemieux, C. et Turmel, M. (2000a). Chloroplast Gene Order and the Divergence of Plants and Algae, From The Normalized Number of Induced Breakpoints. *In* Sankoff, D. et Nadeau, J. H. (rédacteurs), *Comparative Genomics. Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and the Evolution of Gene Families*, tome 1 de *Computational Biology*, pages 89–98. Kluwer Academic, Boston, MA.

Sankoff, D., Leduc, G., Antoine, N., Paquin, B., Lang, B. F. et Cedergren, R. J. (1992). Gene order comparisons for phylogenetic inference : evolution of the mitochondrial genome. *Proceedings of The National Academy of Sciences of The United States of America*, 89(14) : 6575–6579.

Sankoff, D. et Nadeau, J. H. (2000). *Comparative Genomics. Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and the Evolution of Gene Families*, tome 1 de *Computational biology*. Kluwer Academic, Boston, MA.

Sankoff, D., Parent, M.-n. et Bryant, D. (2000b). Accuracy and robustness of analyses based on numbers of genes in observed segments. *In* Sankoff, D. et Nadeau, J. H. (rédacteurs), *Comparative Genomics. Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and the Evolution of Gene Families*, tome 1 de *Computational Biology*, pages 299–306. Kluwer Academic, Boston, MA.

Sasuga, J., Yokobori, S., Kaifu, M., Ueda, T., Nishikawa, K. et Watanabe, K. (1999). Gene contents and organization of a mitochondrial DNA segment of the squid *Loligo bleekeri*. *Journal of Molecular Evolution*, 48(6) : 692–702.

Schmitz, J., Ohme, M. et Zischler, H. (2000). The complete mitochondrial genome of *Tupaia belangeri* and the phylogenetic affiliation of scandentia to other eutherian orders. *Molecular Biology And Evolution*, 17(9) : 1334–1343.

Scouras, A. et Smith, M. J. (1998). The complete mitochondrial genome of the crinoid *Florometra serratissima*. *Non publié*.

Shao, R., Campbell, N. J. et Barker, S. C. (2001). Numerous Gene Rearrangements in the Mitochondrial Genome of the Wallaby Louse, *Heterodoxus macropus* (Phthiraptera). *Molecular Biology And Evolution*, 18(5) : 858–865.

Shimko, N., Liu, L., Lang, B. F. et Burger, G. (2001). GOBASE : the organelle genome database. *Nucleic Acids Research*, 29(1) : 128–132.

Smith, M. J., Arndt, A., Gorski, S. et Fajber, E. (1993). The phylogeny of echinoderm classes based on mitochondrial gene arrangements. *Journal of Molecular Evolution*, 36(6) : 545–554.

Snel, B., Bork, P. et Huynen, M. (1999). Genome phylogeny bassed on gene content. *Nature Genetics*, 21 : 108–110.

Spanos, L., Koutroumbas, G., Kotsyfakis, M. et Louis, C. (2000). The mitochondrial genome of the mediterranean fruit fly, *Ceratitis capitata*. *Insect Mololecular Biology*, 9(2) : 139–144.

Spruyt, N., Delarbre, C., Gachelin, G. et Laudet, V. (1998). Complete sequence of the amphioxus (*Branchiostoma lanceolatum*) mitochondrial genome : relations to vertebrates. *Nucleic Acids Research*, 26(13) : 3279–3285.

Staton, J. L., Daehler, L. L. et Brown, W. M. (1997). Mitochondrial gene arrangement of the horseshoe crab *Limulus polyphemus* L. : conservation of major features among arthropod classes. *Molecular Biology And Evolution*, 14(8) : 867–874.

Stechmann, A. et Schlegel, M. (1999). Analysis of the complete mitochondrial DNA sequence of the brachiopod *Terebratulina retusa* places Brachiopoda within the protostomes. *Proceedings of The Royal Society of London. Series B : Biological Sciences*, 266(1433) : 2043–2052.

Strong, E. E. et Lipscomb, D. (1999). Character Coding and Inapplicable Data. *Cladistics*, 15 : 363–371.

Swofford, D. L. (1998). PAUP* : Phylogenetic Analysis Using Parsimony (*and Other Methods).

Swofford, D. L. et Maddison, W. P. (1987). Reconstructing ancestral character states under Wagner parsimony. *Mathematical Biosciences*, 87 : 199–229.

Tekaia, F., Lazcano, A. et Dujon, B. (1999). The genomic tree as revealed from whole proteome comparisons. *Genome Research*, 9(6) : 550–557.

Terrett, J. A., Miles, S. et Thomas, R. H. (1996). Complete DNA sequence of the mitochondrial genome of *Cepaea nemoralis* (Gastropoda : Pulmonata). *Journal of Molecular Evolution*, 42(2) : 160–168.

Turbeville, J. M., Field, K. et Raff, R. A. (1992). Phylogenetic position of phylum Nemertini, inferred from 18S rRNA sequences : molecular data as a test of morphological character homology. *Molecular Biology And Evolution*, 9(2) : 235–249.

Tzeng, C. S., Hui, C. F., Shen, S. C. et Huang, P. C. (1992). The complete nucleotide sequence of the *Crossostoma lacustre* mitochondrial genome : conservation and variations among vertebrates. *Nucleic Acids Research*, 20(18) : 4853–4858.

Ursing, B. M. et Arnason, U. (1998a). Analyses of mitochondrial genomes strongly support a hippopotamus-whale clade. *Proceedings of The Royal Society of London. Series B : Biological Sciences*, 265(1412) : 2251–2255.

Ursing, B. M. et Arnason, U. (1998b). The complete mitochondrial DNA sequence of the pig (*Sus scrofa*). *Journal of Molecular Evolution*, 47(3) : 302–306.

Ursing, B. M., Slack, K. E. et Arnason, U. (2000). Subordinal artiodactyl relationships in the light of phylogenetic analysis of 12 mitochondrial protein-coding genes. *Zoologica Scripta*, 29 : 83–88.

Valverde, J. R., Batuecas, B., Moratilla, C., Marco, R. et Garesse, R. (1994). The complete mitochondrial DNA sequence of the crustacean *Artemia franciscana*. *Journal of Molecular Evolution*, 39(4) : 400–408.

von Nickisch-Rosenegk, M., Brown, W. M. et Boore, J. L. (2001). Complete Sequence of the Mitochondrial Genome of the Tapeworm *Hymenolepis diminuta* : Gene Arrangements Indicate that Platyhelminths Are Eutrochozoans. *Molecular Biology And Evolution*, 18(5) : 721–730.

Watterson, G. A., Ewens, W. J., Hall, T. E. et Morgan, A. (1982). The chromosome inversion problem. *Journal of theoretical biology*, 99 : 1–7.

Wilson, K., Cahill, V., Ballment, E. et Benzie, J. (2000). The complete sequence of the mitochondrial genome of the crustacean *Penaeus monodon* : are malacostracan crustaceans more closely related to insects than to branchiopods ? *Molecular Biology and Evolution*, 17(6) : 863–874.

Wolf, Y. I., Rogozin, I. B., Kondrashov, A. S. et Koonin, E. V. (2001). Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Research*, 11(3) : 356–372.

Wolstenholme, D. R., Macfarlane, J. L., Okimoto, R., Clary, D. O. et Wahleithner, J. A. (1987). Bizarre tRNAs inferred from DNA sequences of mitochondrial genomes of nematode worms. *Proceedings of The National Academy of Sciences of The United States of America*, 84(5) : 1324–1328.

Xu, X. et Arnason, U. (1994). The complete mitochondrial DNA sequence of the horse, *Equus caballus* : extensive heteroplasmy of the control region. *Gene*, 148(2) : 357–362.

Xu, X. et Arnason, U. (1996a). A complete sequence of the mitochondrial genome of the western lowland gorilla. *Molecular Biology and Evolution*, 13(5) : 691–698.

Xu, X. et Arnason, U. (1996b). The mitochondrial DNA molecule of Sumatran orangutan and a molecular proposal for two (Bornean and Sumatran) species of orangutan. *Journal of Molecular Evolution*, 43(5) : 431–437.

Xu, X. et Arnason, U. (1997). The complete mitochondrial DNA sequence of the white rhinoceros, *Ceratotherium simum*, and comparison with the mtDNA sequence of the Indian rhinoceros, *Rhinoceros unicornis*. *Molecular Phylogenetics And Evolution*, 7(2) : 189–194.

Xu, X., Gullberg, A. et Arnason, U. (1996a). The complete mitochondrial DNA (mtDNA) of the donkey and mtDNA comparisons among four closely related mammalian species-pairs. *Journal of Molecular Evolution*, 43(5) : 438–446.

Xu, X., Janke, A. et Arnason, U. (1996b). The complete mitochondrial DNA sequence of the greater Indian rhinoceros, *Rhinoceros unicornis*, and the Phylogenetic relationship among Carnivora, Perissodactyla, and Artiodactyla (+ Cetacea). *Molecular Biology And Evolution*, 13(9) : 1167–1173.

Yamamoto, Y., Murata, K., Matsuda, H., Hosoda, T., Tamura, K. et Furuyama, J. (2000). Determination of the complete nucleotide sequence and haplotypes in the D-loop region of the mitochondrial genome in the Oriental white stork, *Ciconia boyciana*. *Genes and Genetic Systems*, 75(1) : 25–32.

Yamazaki, N., Ueshima, R., Terrett, J. A., Yokobori, S., Kaifu, M., Segawa, R., Kobayashi, T., Numachi, K., Ueda, T., Nishikawa, K., Watanabe, K. et Thomas, R. H. (1997). Evolution of

pulmonate gastropod mitochondrial genomes : comparisons of gene organizations of Euhadra, Cepaea and Albinaria and implications of unusual tRNA secondary structures. *Genetics*, 145(3) : 749–758.

Yokobori, S., Ueda, T., Feldmaier-Fuchs, G., Pääbo, S., Ueshima, R., Kondow, A., Nishikawa, K. et Watanabe, K. (1999). Complete DNA Sequence of the Mitochondrial Genome of the Ascidian *Halocynthia roretzi* (Chordata, Urochordata). *Genetics*, 153(4) : 1851–1862.

Zardoya, R., Garrido-Pertierra, A. et Bautista, J. M. (1995). The complete nucleotide sequence of the mitochondrial DNA genome of the rainbow trout, *Oncorhynchus mykiss*. *Journal of Molecular Evolution*, 41(6) : 942–951.

Zardoya, R. et Meyer, A. (1996). The complete nucleotide sequence of the mitochondrial genome of the lungfish (*Protopterus dolloi*) supports its phylogenetic position as a close relative of land vertebrates. *Genetics*, 142(4) : 1249–1263.

Zardoya, R. et Meyer, A. (1997). The complete DNA sequence of the mitochondrial genome of a “living fossil”, the coelacanth (*Latimeria chalumnae*). *Genetics*, 146(3) : 995–1010.

Zardoya, R. et Meyer, A. (1998). Complete mitochondrial genome suggests diapsid affinities of turtles. *Proceedings of The National Academy of Sciences of The United States of America*, 95(24) : 14226–14231.

Zardoya, R. et Meyer, A. (2000). Mitochondrial evidence on the phylogenetic position of caecilians (Amphibia : Gymnophiona). *Genetics*, 155(2) : 765–775.

A

Métazoaires

A.1 Génomes mitochondriaux complets

Nº	Genre espèce	Nº d'accession	Référence
1	*Albinaria coerulea*[*]	X83390	(Hatzoglou *et al.*, 1995 ; Yamazaki *et al.*, 1997)
2	*Alligator mississippiensis*[*]	Y13113	(Janke et Arnason, 1997)
2	*Caiman crocodilus*	AJ404872	(Janke *et al.*, 2001)
3	*Anopheles gambiae*[*]	L20934	(Beard *et al.*, 1993)
3	*Anopheles quadrimaculatus*	L04272	(Mitchell *et al.*, 1993)
4	*Apis mellifera*[*]	L06178	(Crozier et Crozier, 1993)
5	*Arbacia lixula*	X80396	(De Giorgi *et al.*, 1996)
5	*Paracentrotus lividus*	J04815	(Cantatore *et al.*, 1989)
5	*Strongylocentrotus purpuratus*[*]	X12631	(Jacobs *et al.*, 1988)
6	*Artemia franciscana*[*]	X69067	(Valverde *et al.*, 1994)
7	*Anguilla japonica*	AB038556	(Inoue *et al.*, 2001a)
7	*Artibeus jamaicensis*	AF061340	(Pumo *et al.*, 1998)
7	*Aulopus japonicus*	AB047821	(Non publié)
7	*Balaenoptera musculus*	X72204	(Arnason et Gullberg, 1993)
7	*Balaenoptera physalus*	X61145	(Arnason *et al.*, 1991)
7	*Bos taurus*	J01394	(Anderson *et al.*, 1982)
7	*Canis familiaris*	U96639	(Kim *et al.*, 1998)
7	*Carassius auratus*	AB006953	(Murakami *et al.*, 1998)
7	*Cavia porcellus*	AJ222767	(D'Erchia *et al.*, 1996)
7	*Cebus albifrons*	AJ309866	(Arnason *et al.*, 2000b)
7	*Ceratotherium simum*	Y07726	(Xu et Arnason, 1997)
7	*Chalinolobus tuberculatus*	AF321051	(Lin et Penny, 2001)
7	*Chelonia mydas*	AB012104	(Kumazawa et Nishida, 1999)
7	*Chrysemys picta*	AF069423	(Mindell *et al.*, 1999)
7	*Coregonus lavaretus*	AB034824	(Miya et Nishida, 2000)
7	*Crossostoma lacustre*	M91245	(Tzeng *et al.*, 1992)
7	*Cyprinus carpio*	X61010	(Chang *et al.*, 1994)
7	*Danio rerio*	AC024175	(Milam *et al.*, 2000)
7	*Dasypus novemcinctus*	Y11832	(Arnason *et al.*, 1997)
7	*Diplophos taenia*	AB034825	(Miya et Nishida, 2000)
7	*Dogania subplana*	AF366350	(Non publié)
7	*Echinops telfairi*	AJ400734	(Mouchaty *et al.*, 2000b)
7	*Eptatretus burgeri*	AJ278504	(Delarbre *et al.*, 2002)
7	*Equus asinus*	X97337	(Xu *et al.*, 1996a)

N°	Genre espèce	N° d'accession	Référence
7	*Equus caballus*	X79547	(Xu et Arnason, 1994)
7	*Erinaceus europaeus*	X88898	(Krettek *et al.*, 1995)
7	*Eumeces egregius*	AB016606	(Kumazawa et Nishida, 1999)
7	*Felis catus*	U20753	(Lopez *et al.*, 1996)
7	*Gadus morhua*	X99772	(Johansen et Bakke, 1996)
7	*Glis glis*	AJ001562	(Reyes *et al.*, 1998)
7	*Gorilla gorilla*	D38114	(Horai *et al.*, 1995 ; Xu et Arnason, 1996a)
7	*Halichoerus grypus*	X72004	(Arnason *et al.*, 1993)
7	*Hippopotamus amphibius*	AJ010957	(Ursing et Arnason, 1998a)
7	*Homo sapiens*[*]	V00662	(Anderson *et al.*, 1981 ; Horai *et al.*, 1995)
7	*Hylobates lar*	X99256	(Arnason *et al.*, 1996a)
7	*Iguana iguana*	AJ278511	(Janke *et al.*, 2001)
7	*Lama pacos*	Y19184	(Ursing *et al.*, 2000)
7	*Latimeria chalumnae*	U82228	(Zardoya et Meyer, 1997)
7	*Loxodonta africana*	AJ224821	(Hauf *et al.*, 1999)
7	*Macaca sylvanus*	AJ309865	(Arnason *et al.*, 2000b)
7	*Mertensiella luschani*	AF154053	(Non publié)
7	*Mus musculus*	J01420	(Bibb *et al.*, 1981)
7	*Mustelus manazo*	AB015962	(Cao *et al.*, 1998)
7	*Myxine glutinosa*	AJ404477	(Non publié)
7	*Nycticebus coucang*	AJ309867	(Arnason *et al.*, 2000b)
7	*Oncorhynchus mykiss*	L29771	(Zardoya *et al.*, 1995)
7	*Ornithorhynchus anatinus*	X83427	(Janke *et al.*, 1996)
7	*Orycteropus afer*	Y18475	(Arnason *et al.*, 1999)
7	*Oryctolagus cuniculus*	AJ001588	(Gissi *et al.*, 1998)
7	*Ovis aries*	AF010406	(Hiendleder *et al.*, 1998)
7	*Pan paniscus*	D38116	(Horai *et al.*, 1995)
7	*Pan troglodytes*	D38113	(Horai *et al.*, 1995 ; Arnason *et al.*, 1996b)
7	*Papio hamadryas*	Y18001	(Arnason *et al.*, 1998)
7	*Paralichthys olivaceus*	AB028664	(Saitoh *et al.*, 2000)
7	*Pelomedusa subrufa*	AF039066	(Zardoya et Meyer, 1998)
7	*Phoca vitulina*	X63726	(Arnason et Johnsson, 1992)
7	*Physeter macrocephalus*	AJ277029	(Arnason *et al.*, 2000a)
7	*Plecoglossus altivelis*	AB047553	(Non publié)
7	*Polymixia japonica*	AB034826	(Miya et Nishida, 2000)
7	*Polypterus ornatipinnis*	U62532	(Noack *et al.*, 1996)
7	*Pongo pygmaeus*	D38115	(Horai *et al.*, 1995 ; Xu et Arnason, 1996b)
7	*Protopterus dolloi*	L42813	(Zardoya et Meyer, 1996)
7	*Pteropus dasymallus*	AB042770	(Nikaido *et al.*, 2000)
7	*Pteropus scapulatus*	AF321050	(Lin et Penny, 2001)
7	*Raja radiata*	AF106038	(Rasmussen et Arnason, 1999a)
7	*Rattus norvegicus*	X14848	(Gadaleta *et al.*, 1989)
7	*Rhinoceros unicornis*	X97336	(Xu *et al.*, 1996b ; Xu et Arnason, 1997)
7	*Salmo salar*	U12143	(Hurst *et al.*, 1999)
7	*Salvelinus alpinus*	AF154851	(Doiron *et al.*, 1999)
7	*Salvelinus fontinalis*	AF154850	(Doiron *et al.*, 1999)
7	*Sardinops melanostictus*	AB032554	(Inoue *et al.*, 2000)
7	*Sciurus vulgaris*	AJ238588	(Reyes *et al.*, 2000)
7	*Scyliorhinus canicula*	Y16067	(Delarbre *et al.*, 1998)
7	*Squalus acanthias*	Y18134	(Rasmussen et Arnason, 1999b)
7	*Sus scrofa*	AJ002189	(Ursing et Arnason, 1998b)
7	*Talpa europaea*	Y19192	(Mouchaty *et al.*, 2000a)
7	*Thryonomys swinderianus*	AJ301644	(Mouchaty *et al.*, 2001)

N°	**Genre espèce**	**N° d'accession**	**Référence**
7	*Tupaia belangeri*	AF217811	(Schmitz *et al.*, 2000)
7	*Typhlonectes natans*	AF154051	(Zardoya et Meyer, 2000)
7	*Xenopus laevis*	M10217	(Roe *et al.*, 1985)
8	*Ascaris suum**	X54253	(Wolstenholme *et al.*, 1987 ; Okimoto *et al.*, 1992)
8	*Caenorhabditis elegans*	X54252	(Wolstenholme *et al.*, 1987 ; Okimoto *et al.*, 1992)
9	*Asterina pectinifera**	D16387	(Asakawa *et al.*, 1995)
10	*Anomalopteryx didiformis*	AF338714	(Haddrath et Baker, 2001)
10	*Apteryx haastii*	AF338708	(Haddrath et Baker, 2001)
10	*Aythya americana*	AF090337	(Mindell *et al.*, 1999)
10	*Casuarius casuarius*	AF338713	(Haddrath et Baker, 2001)
10	*Ciconia boyciana*	AB026193	(Yamamoto *et al.*, 2000)
10	*Ciconia ciconia*	AB026818	(Yamamoto *et al.*, 2000)
10	*Corvus frugilegus*	Y18522	(Härlid et Arnason, 1999)
10	*Dinornis giganteus*	AY016013	(Cooper *et al.*, 2001 ; Haddrath et Baker, 2001)
10	*Dromaius novaehollandiae*	AF338711	(Haddrath et Baker, 2001)
10	*Emeus crassus*	AF338712	(Cooper *et al.*, 2001 ; Haddrath et Baker, 2001)
10	*Eudromia elegans*	AF338710	(Haddrath et Baker, 2001)
10	*Falco peregrinus*	AF090338	(Mindell *et al.*, 1999)
10	*Gallus gallus**	X52392	(Desjardins et Morais, 1990)
10	*Pterocnemia pennata*	AF338709	(Haddrath et Baker, 2001)
10	*Rhea americana*	Y16884	(Härlid *et al.*, 1998 ; Mindell *et al.*, 1999)
10	*Smithornis sharpei*	AF090340	(Mindell *et al.*, 1999)
10	*Struthio camelus*	Y12025	(Härlid *et al.*, 1997 ; Mindell *et al.*, 1999)
10	*Tinamus major*	AF338707	(Haddrath et Baker, 2001)
10	*Vidua chalybeata*	AF090341	(Mindell *et al.*, 1999)
11	*Balanoglossus carnosus**	AF051097	(Castresana *et al.*, 1998)
12	*Boophilus microplus*		(Campbell et Barker, 1999)
12	*Rhiphicephalus sanguineus**	AF081829	(Black et Roehrdanz, 1998)
13	*Branchiostoma floridae*	AF098298	(Boore *et al.*, 1999)
13	*Branchiostoma lanceolatum**	Y16474	(Spruyt *et al.*, 1998)
14	*Crassostrea gigas**	AF177226	(Kim *et al.*, 2000)
15	*Didelphis virginiana**	Z29573	(Janke *et al.*, 1994)
15	*Isoodon macrourus*	AF358864	(Non publié)
15	*Macropus robustus*	Y10524	(Janke *et al.*, 1997)
16	*Dinodon semicarinatus**	AB008539	(Kumazawa *et al.*, 1998)
17	*Cepaea nemoralis**	U23045	(Terrett *et al.*, 1996 ; Yamazaki *et al.*, 1997)
18	*Ceratitis capitata*	AJ242872	(Spanos *et al.*, 2000)
18	*Chrysomya chloropyga*	AF352790	(Non publié)
18	*Cochliomyia hominivorax*	AF260826	(Lessinger *et al.*, 2000)
18	*Daphnia pulex*	AF117817	(Crease, 1999)
18	*Drosophila melanogaster*	U37541	(Garesse, 1988)
18	*Drosophila yakuba**	X03240	(Clary et Wolstenholme, 1985)
18	*Penaeus monodon*	AF217843	(Wilson *et al.*, 2000)
18	*Triatoma dimidiata*	AF301594	(Non publié)
19	*Echinococcus multilocularis**	AB018440	(Fukunaga, 2000)
19	*Taenia crassiceps*	AF216699	(Le *et al.*, 2000)
20	*Euhadra herklotsi**		(Yamazaki *et al.*, 1997)
21	*Florometra serratissima**	AF049132	(Scouras et Smith, 1998)
22	*Gonostoma gracile**	AB016274	(Miya et Nishida, 1999)
23	*Halocynthia roretzi**	AB024528	(Yokobori *et al.*, 1999)
24	*Katharina tunicata**	U09810	(Boore et Brown, 1994)
25	*Lampetra fluviatilis*	Y18683	(Delarbre *et al.*, 2000)
25	*Petromyzon marinus**	U11880	(Lee et Kocher, 1995)

N°	Genre espèce	N° d'accession	Référence
26	*Laqueus rubellus**	AB035869	(Noguchi *et al.*, 2000)
27	*Ixodes hexagonus*	AF081828	(Staton *et al.*, 1997)
27	*Limulus polyphemus**	AF216203	(Staton *et al.*, 1997 ; Lavrov *et al.*, 2000a)
28	*Locusta migratoria**	X80245	(Flook *et al.*, 1995)
29	*Loligo bleekeri**	AB009838	(Sasuga *et al.*, 1999)
30	*Lumbricus terrestris**	U24570	(Boore et Brown, 1995)
31	*Meloidogyne javanica**		(Wolstenholme *et al.*, 1987)
32	*Metridium senile**	AF000023	(Beagley *et al.*, 1998)
33	*Mytilus edulis**		(Hoffmann *et al.*, 1992)
34	*Onchocerca volvulus**	AF015193	(Keddie *et al.*, 1998)
35	*Pagurus longicarpus**	AF150756	(Hickerson et Cunningham, 2000)
36	*Platynereis dumerilii**	AF178678	(Boore et Brown, 2000)
37	*Pupa strigosa**	AB028237	(Kurabayashi et Ueshima, 2000)
38	*Renilla kolikeri*		(Beaton *et al.*, 1998)
38	*Sarcophyton glaucum**		(Beaton *et al.*, 1998 ; Pont-Kingdon *et al.*, 1998)
39	*Terebratulina retusa**	AJ245743	(Stechmann et Schlegel, 1999)
40	*Bombyx mori**	AF149768	(Lee *et al.*, 2000)
41	*Fasciola hepatica**	AF216697	(Le *et al.*, 2000)
41	*Paragonimus westermani*	AF219379	(Le *et al.*, 2000)
42	*Schistosoma japonicum**	AF215860	(Le *et al.*, 2000)
42	*Schistosoma mekongi*	AF217449	(Le *et al.*, 2000)
43	*Schistosoma mansoni**	AF216698	(Le *et al.*, 2000)
44	*Conger myriaster**	AB038381	(Inoue *et al.*, 2001b)
45	*Heterodoxus macropus**	AF270939	(Shao *et al.*, 2001)
46	*Lithobius forficatus**	AF309492	(Lavrov *et al.*, 2000b)
47	*Trichinella spiralis**	AF293969	(Lavrov et Brown, 2001)
48	*Hymenolepis diminuta**	AF314223	(von Nickisch-Rosenegk *et al.*, 2001)
49	*Tetrodontophora bielanensis**	AF272824	(Nardi *et al.*, 2001)

TABLEAU A.1 – Liste des espèces dont le génome mitochondrial complet est connu. Les numéros correspondent aux 49 groupes au sein desquels toutes les espèces ont le même ordre de gènes. Les espèces choisies pour représenter ces groupes sont marquées d'une étoile (*). Le numéro d'accession est indiqué pour les espèces dont la séquence complète est disponible. Cette liste a été établie fin juin 2001.

A.2 Nombre moyen de points de cassure

Espèce	Nombre de gènes	Nombre moyen de points de cassure	Espèce	Nombre de gènes	Nombre moyen de points de cassure
A. coerulea	37	33,88	*H. diminuta*	36	32,46
A. mississippiensis	37	26,77	*K. tunicata*	37	29,77
A. gambiae	37	26,46	*L. rubellus*	37	**35,88**
A. mellifera	37	28,85	*L. polyphemus*	37	25,58
A. franciscana	37	26,46	*L. forficatus*	37	26,04
A. suum	36	34,96	*L. migratoria*	37	26,25
A. pectinifera	37	31,83	*L. bleekeri*	37	32,25
B. carnosus	37	27,56	*L. terrestris*	37	31,42
B. mori	37	26,44	*M. javanica*	36	**35,15**
B. lanceolatum	37	28,98	*M. senile*	18	16,90
C. nemoralis	37	33,94	*M. edulis*	36	34,67
C. myriaster	37	26,58	*O. volvulus*	36	**35,31**
C. gigas	36	**35,10**	*P. longicarpus*	37	30,85
D. virginiana	37	27,06	*P. marinus*	37	26,54
D. semicarinatus	37	26,73	*P. dumerilii*	37	32,10
D. yakuba	37	25,77	*P. strigosa*	37	34,00
E. multilocularis	36	32,38	*R. sanguineus*	37	26,85
E. herklotsi	37	33,88	*S. glaucum*	16	14,83
F. hepatica	36	32,54	*S. japonicum*	36	32,81
F. serratissima	37	32,31	*S. mansoni*	36	33,31
G. gallus	37	26,17	*S. purpuratus*	37	31,83
G. gracile	37	26,67	*T. retusa*	37	31,04
H. roretzi	36	**35,13**	*T. bielanensis*	37	26,71
H. macropus	37	**35,48**	*T. spiralis*	37	30,73
H. sapiens	37	25,92			

TABLEAU A.2 – Nombre (non corrigé) de points de cassure moyen par rapport à toutes les autres espèces, comparé au nombre de gènes (n). En gras : nombre moyen de points cassure proche de la valeur limite $n - \frac{1}{2}$.

A.3 Cartes du génome mitochondrial des métazoaires

Légendes des noms de gènes utilisés pour les cartes :

- cox1-3 (cytochrome c oxydase sous unités I à III),
- cob (cytochrome b apoenzyme),
- nad1-6, 4L (NADH déshydrogénase sous unités 1 à 6 et 4L),
- A6 et A8 (ATP synthase sous unités 6 et 8),
- rns et rnl (petite et grande sous unités d'ARN ribosomique),

- les ARNs de transfert sont nommés par le code à une lettre de l'acide aminé dont ils sont spécifiques,
- atr (région riche en AT),
- cr (région de contrôle).

Les ARNt placés au-dessus des cartes sont transcrits dans le sens direct, ceux placés en dessous sont transcrits dans le sens opposé. Pour les gènes de grande taille une flèche indique le sens de transcription. En gris : région non codante.

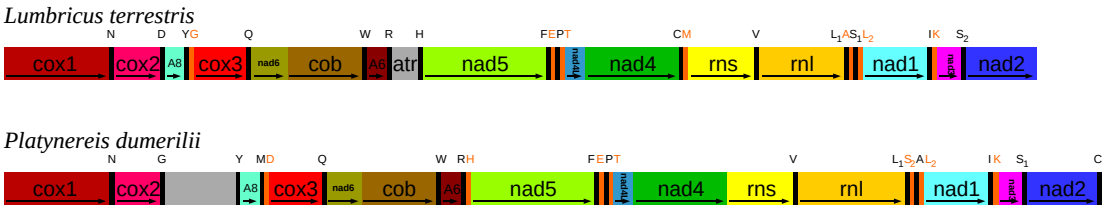


FIGURE A.1 – Cartes du génome mitochondrial des Annélides.

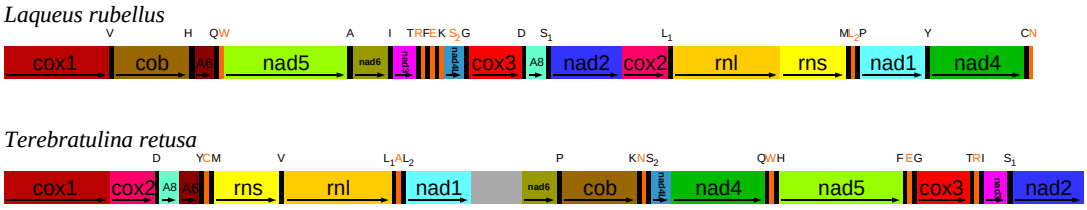


FIGURE A.2 – Cartes du génome mitochondrial des Brachiopodes.

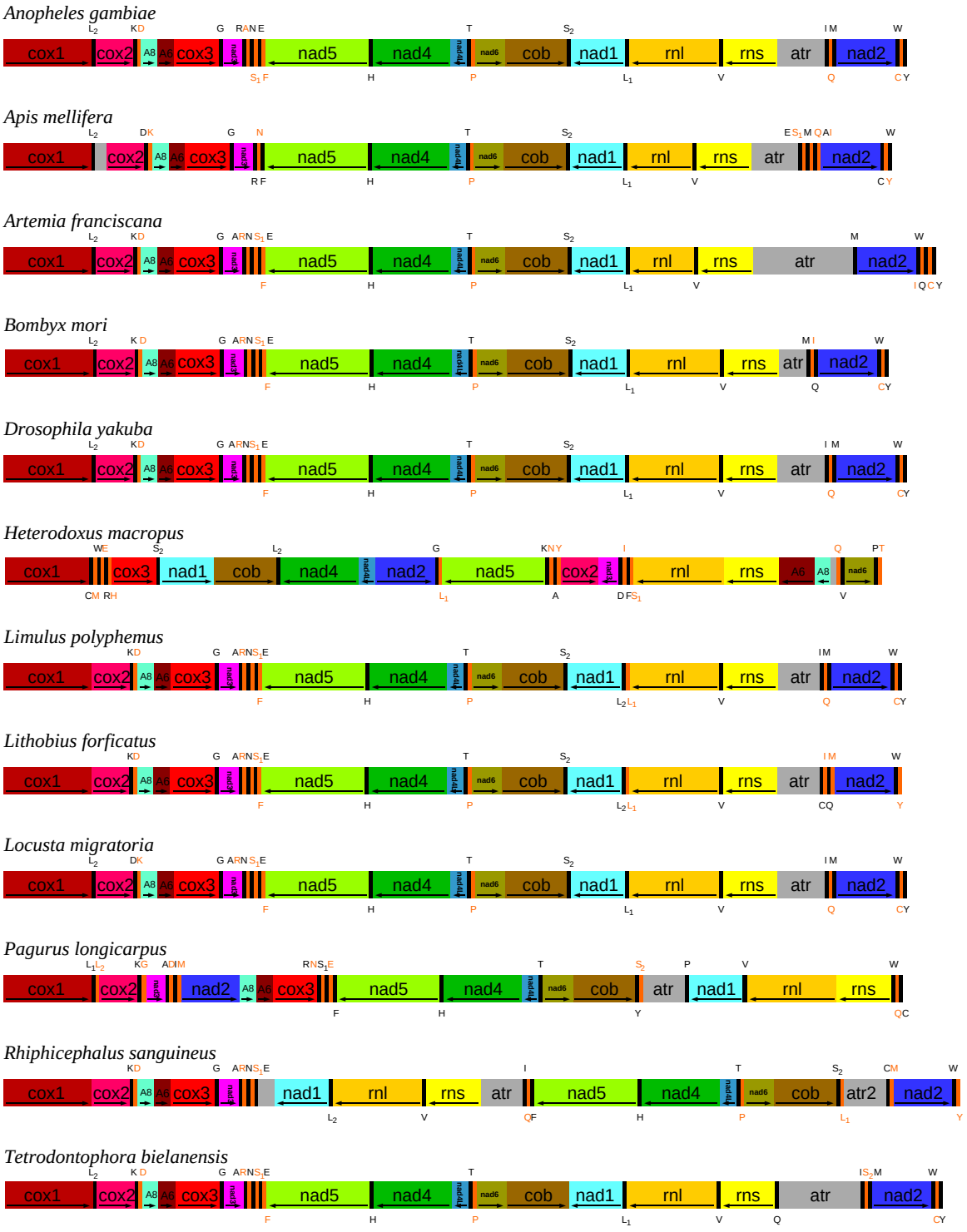


FIGURE A.3 – Cartes du génome mitochondrial des Arthropodes.

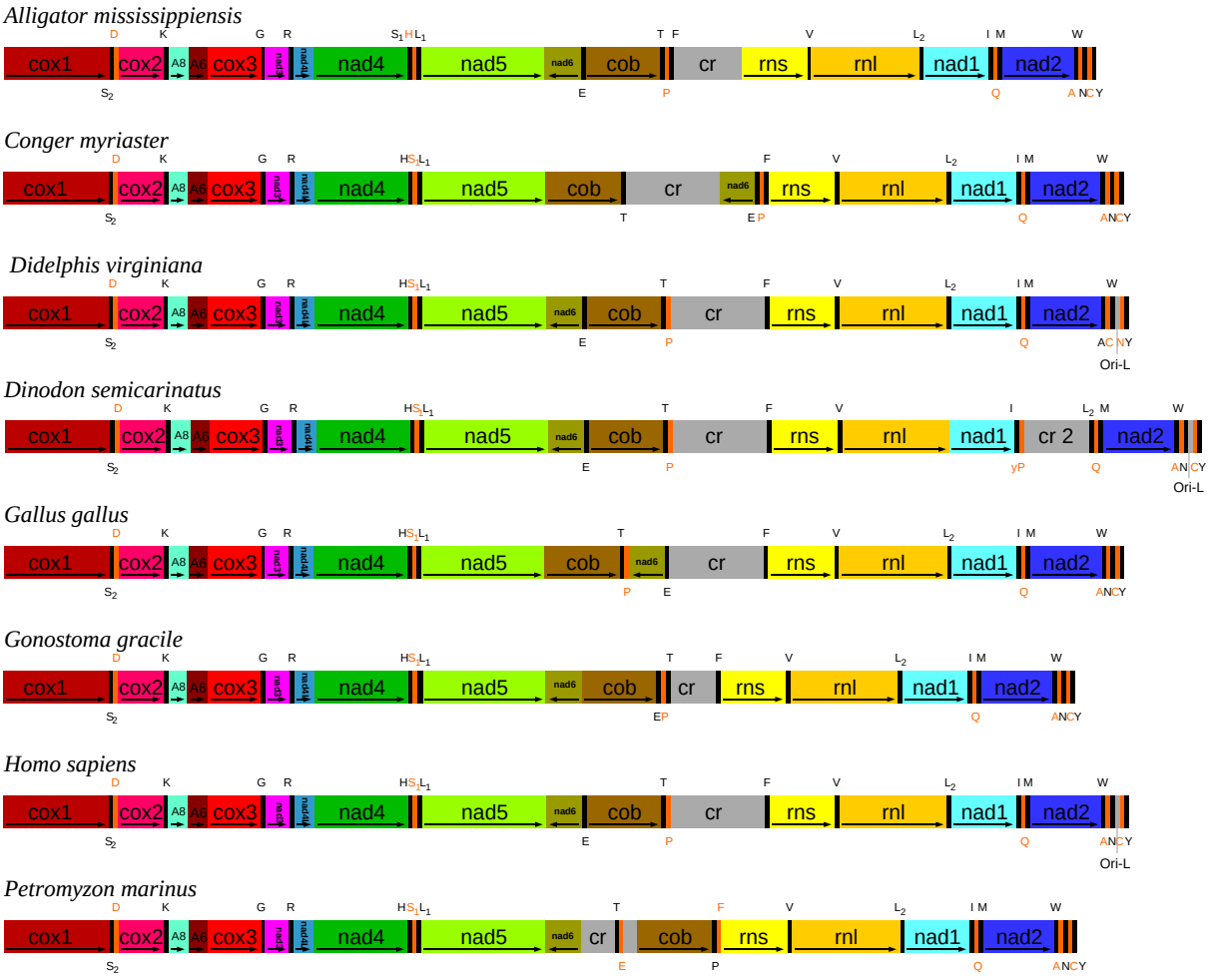


FIGURE A.4 – Cartes du génome mitochondrial des Crâniates.

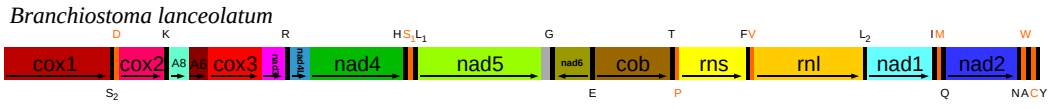


FIGURE A.5 – Cartes du génome mitochondrial des Céphalochordés.

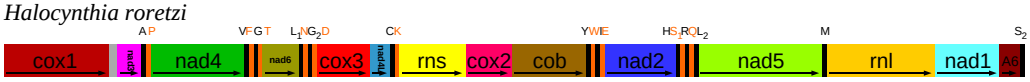


FIGURE A.6 – Cartes du génome mitochondrial des Urochordés.

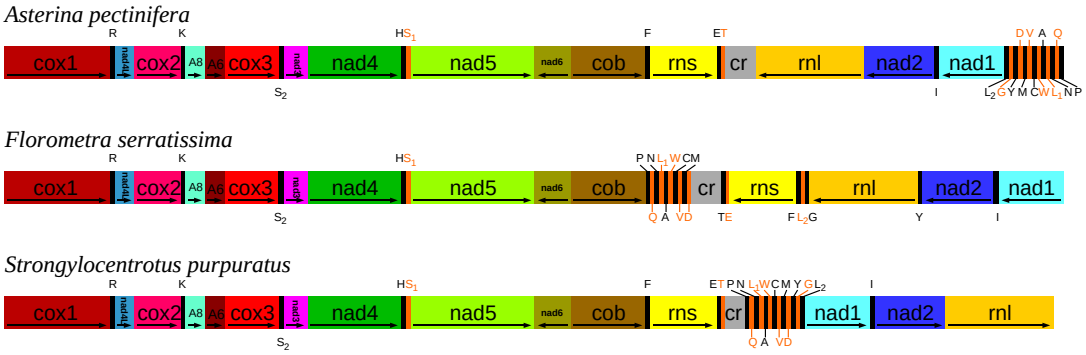


FIGURE A.7 – Cartes du génome mitochondrial des Échinodermes.

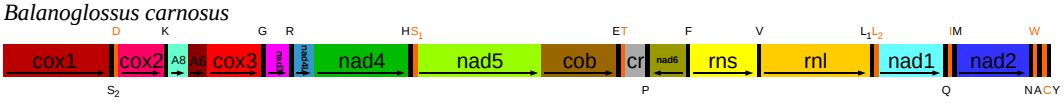


FIGURE A.8 – Cartes du génome mitochondrial des Hémichordés.

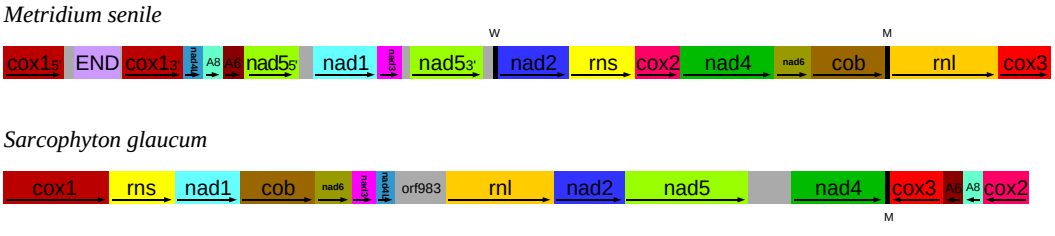
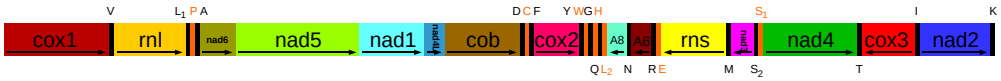
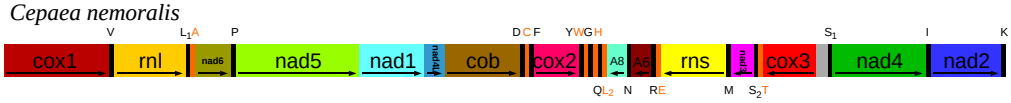


FIGURE A.9 – Cartes du génome mitochondrial des Cnidaires.

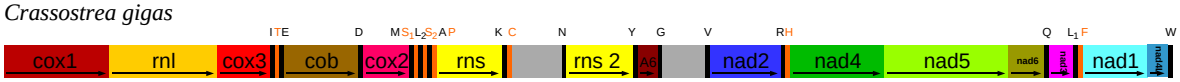
Albinaria coerulea



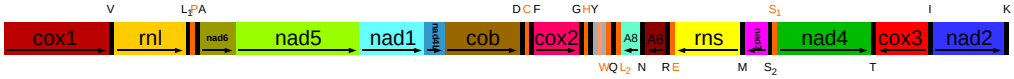
Cepaea nemoralis



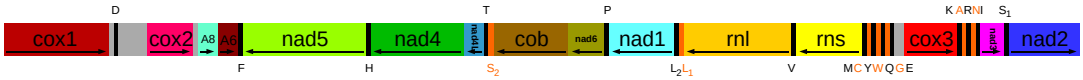
Crassostrea gigas



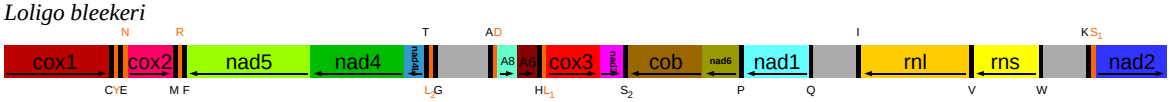
Euhadra herklotsi



Katharina tunicata



Loligo bleekeri



Mytilus edulis

Pupa strigosa

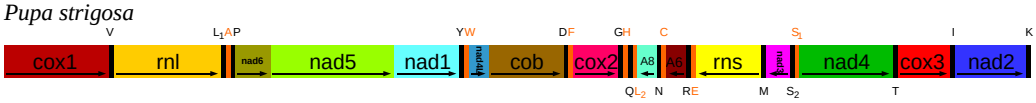
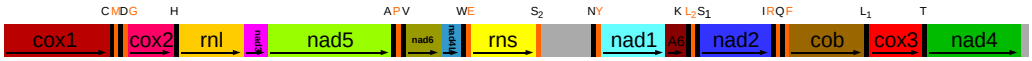


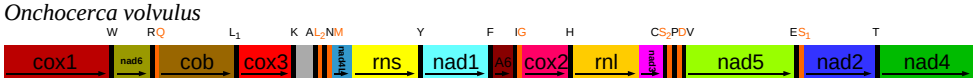
FIGURE A.10 – Cartes du génome mitochondrial des Mollusques.

Ascaris suum



Meloidogyne javanica

Onchocerca volvulus



Trichinella spiralis

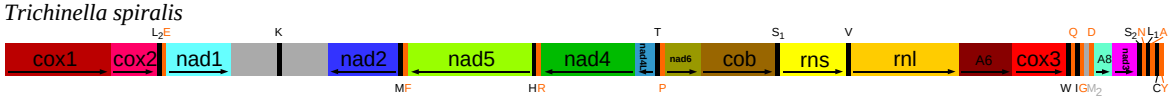


FIGURE A.11 – Cartes du génome mitochondrial des Nématodes.

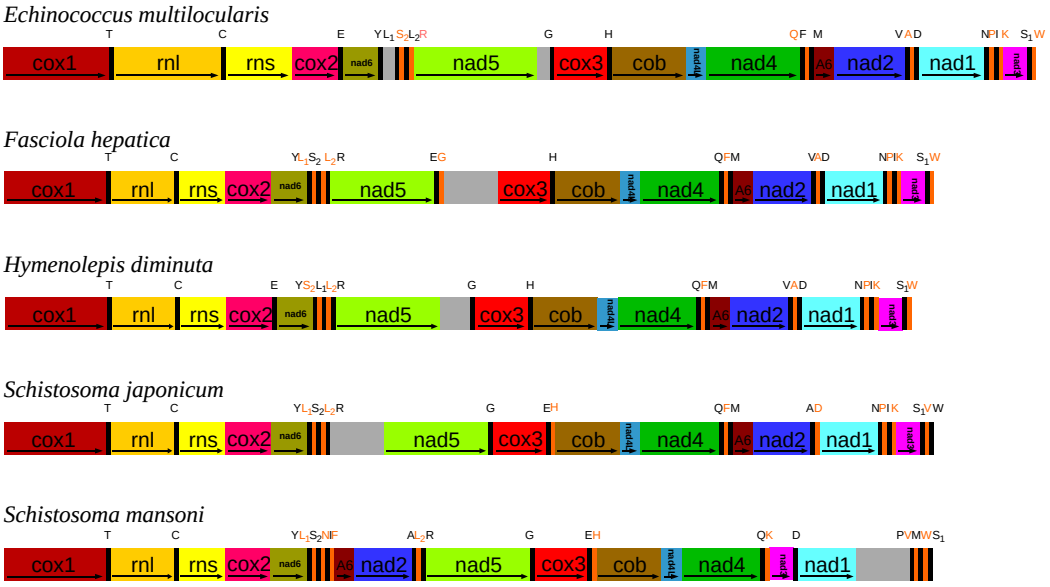


FIGURE A.12 – Cartes du génome mitochondrial des Plathelminthes.

B

Genre *Mastomys*

B.1 Homologies des chromosomes du genre *Mastomys*

M. verheyeni	*M. erythroleucus*	*M. erythroleucus2*	*M. species*	*M. natalensis*	*M. awashencsis*
1A	10A	10A	9SM	1SM	1SM
centro	**centro**	**centro**	**télo**	**télo**	**télo**
+1prox1	+1prox1	+1prox1	-1prox1	-1prox1	-1prox2
+1prox2	+1prox2	+1prox2	**centro**	**centro**	-1prox1
+1dist1	**télo**	**télo**	+1prox2	+1prox2	**centro**
+1dist2			**télo**	+1dist1	+1dist1
télo				+1dist2	+1dist2
	9A	9SM	8SM	**télo**	**télo**
	centro	**télo**	**télo**		
	+1dist1	-1dist1	-1dist1		
	+1dist2	**centro**	**centro**		
	télo	+1dist2	+1dist2		
		télo	**télo**		
2A	15A	15A	16A	2SM	2SM
centro	**centro**	**centro**	**centro**	**télo**	**télo**
+2prox1	+2prox1	+2prox1	+2prox1	-2prox1	-2prox2
+2prox2	+2prox2	+2prox2	+2prox2	**centro**	-2prox1
+2dist1	**télo**	**télo**	**télo**	+2prox2	**centro**
+2dist2				+2dist1	+2dist1
télo	8A	2SM	2SM	+2dist2	+2dist2
	centro	**télo**	**télo**	**télo**	**télo**
	+2dist1	-2dist1	-2dist1		
	+2dist2	**centro**	**centro**		
	télo	+2dist2	+2dist2		
		télo	**télo**		

M. verheyeni	*M. erythroleucus*	*M. erythroleucus2*	*M. species*	*M. natalensis*	*M. awashencsis*
16A	1SM	1SM	1SM	3SM	3SM
centro	**centro**	**centro**	**centro**	**centro**	**centro**
+16prox	+16dist	+16dist	+16dist	+16dist	+16dist
+16dist	+16prox	+16prox	+16prox	+16prox	+16prox
centro	**centro**	-3prox	**centro**	**centro**	**centro**
	+3prox	**centro**	+3prox	+3prox	+3prox
3A	+3dist	+3dist	+3dist	+3dist	+3dist
centro	**télo**	**télo**	**télo**	**télo**	**télo**
+3prox					
+3dist					
télo					
4A	2SM	3SM	3SM	5SM	5SM
centro	**télo**	**télo**	**télo**	**Hét**	**télo**
+4prox1	-4prox2	-4prox2	-4prox2	**centro**	+4prox1
+4prox2	-4prox1	-4prox1	-4prox1	+4prox1	+4prox2
+4dist	**centro**	**centro**	**centro**	+4prox2	**centro**
télo	+4dist	+4dist	+4dist	+4dist	+4dist
	télo	**télo**	**télo**	**télo**	**télo**
5A	4M	5M	5M	7ST	6SM
centro	-5prox1	-5prox1	-5prox1	**centro**	-5prox1
+5prox1	**télo**	**télo**	**télo**	+5prox1	**télo**
+5prox2	-5prox2	-5prox2	-5prox2	+5prox2	-5prox2
+5dist1	**centro**	**centro**	**centro**	-5dist2	**centro**
+5dist2	-5dist2	-5dist2	-5dist2	-5dist1	+5dist1
télo	-5dist1	-5dist1	-5dist1	**télo**	+5dist2
	télo	**télo**	**télo**		**télo**
6A	5M	6M	6M	9M	8M
centro	**télo**	**télo**	**télo**	**télo**	**télo**
+6prox	-6prox	-6prox	-6prox	-6prox	-6prox
+6dist	**centro**	**centro**	**centro**	**centro**	**centro**
télo	+6dist	+6dist	+6dist	+6dist	+6dist
	télo	**télo**	**télo**	**télo**	**télo**
7A	6M	7M	7M	6SM	9M
centro	**télo**	**télo**	**télo**	**Hét**	**télo**
+7prox	-7prox	-7prox	-7prox	**centro**	-7prox
+7dist	**centro**	**centro**	**centro**	+7prox	**centro**
télo	+7dist	+7dist	+7dist	+7dist	+7dist
	télo	**télo**	**télo**	**télo**	**télo**
9A	12A	12A	13A	13A	12A
centro	**centro**	**centro**	**centro**	**centro**	**centro**
+9	+9	+9	+9	+9	+9
télo	**télo**	**télo**	**télo**	**télo**	**télo**

M. verheyeni	*M. erythroleucus*	*M. erythroleucus2*	*M. species*	*M. natalensis*	*M. awashencsis*
10A	13A	13A	14A	10SM	13A
centro	**centro**	**centro**	**centro**	**télo**	**centro**
+10prox	+10prox	+10prox	+10prox	-10prox	+10prox
+10dist	+10dist	+10dist	+10dist	**centro**	+10dist
télo	**télo**	**télo**	**télo**	+10dist	**télo**
				télo	
11A	14A	14A	15A	14A	14A
centro	**centro**	**centro**	**centro**	**centro**	**centro**
+11	+11	+11	+11	+11	+11
télo	**télo**	**télo**	**télo**	**télo**	**télo**
15A	3SM	4SM	4SM	8SM	7SM
télo	**télo**	**télo**	**télo**	**télo**	**télo**
+15	+15	+15	+15	+15	+15
centro	**centro**	**centro**	**centro**	-12prox	**centro**
	+12prox	+12prox	+12prox	**centro**	+12prox
12A	+12dist	+12dist	+12dist	+12dist	+12dist
centro	**télo**	**télo**	**télo**	**télo**	**télo**
+12prox					
+12dist					
télo					
13A	16A	16A	17A	4SM	4SM
télo	**télo**	**télo**	**télo**	**Hét**	**télo**
+13dist	-13prox	-13prox	-13prox	**centro**	-13prox
+13prox	-13dist	-13dist	-13dist	+13dist	-13dist
centro	**centro**	**centro**	**centro**	+13prox	**centro**
				+8prox	**centro**
8A	11A	11A	10M	+8dist	+8prox
centro	**centro**	**centro**	**télo**	**télo**	+8dist
+8prox	+8prox	+8prox	-8prox		**télo**
+8dist	+8dist	+8dist	**centro**		
télo	**télo**	**télo**	+8dist		
			télo		
14A	18A	18A	18A	15A	15A
centro	**centro**	**centro**	**centro**	**centro**	**centro**
+14	+14	+14	+14	+14	+14
télo	**télo**	**télo**	**télo**	**télo**	**télo**
17M	7M	8M	11M	11M	11M
télo	**télo**	**télo**	**télo**	**télo**	**télo**
+17p	+17p	+17p	+17p	+17p	+17p
centro	**centro**	**centro**	**centro**	**centro**	**centro**
+17q	+17q	+17q	+17q	+17q	+17q
télo	**télo**	**télo**	**télo**	**télo**	**télo**

M. verheyeni	*M. erythroleucus*	*M. erythroleucus2*	*M. species*	*M. natalensis*	*M. awashencsis*
18SM	17A	17A	12SM	12A	10SM
télo	**centro**	**centro**	**télo**	**centro**	**Hét**
+18p	-18p	-18p	+18p	-18p	+18p
centro	+18q	+18q	**centro**	+18q	**centro**
+18q	**télo**	**télo**	+18q	**télo**	+18q
télo			**télo**		**télo**

TABLEAU B.1 – Homologies des chromosomes du genre *Mastomys*. Le nom d'un chromosome correspond à son numéro dans la nomenclature de l'espèce et à sa forme, par exemple 1A pour chromosome N^{o} 1 de forme acrocentrique. Les blocs de couleurs indiquent les homologies entre chromosomes et chaque segment homologue est nommé en fonction de sa position sur le chromosome de gauche, par exemple : 1prox1 signifie, premier segment proximal du chromosome 1. L'orientation relative des segments homologues est indiquée par les signes « + » et « - ». Les segments d'hétérochromatine ne sont pas homologues entre eux, ils ne sont donc pas représentés par un bloc de couleur. Termes employés pour décrire les chromosomes, A : acrocentrique, SM : submétacentrique, M : métacentrique, ST : subtélomérique. Termes employés pour décrire la morphologie des chromosomes, centro : centromère, télo : télomère, Hét : hétérochromatine, prox : proximal, dist : distal, p : petit bras, q : grand bras.

C

Matrices

C.1 Matrices de l’exemple théorique

C.1.1 Codage « Position relative »

C.1.1.1 Option AB=BA

```
#NEXUS

BEGIN DATA;
DIMENSIONS    NTAX=8 NCHAR=27;
FORMAT MISSING=? SYMBOLS= "0 1 2 3 4 5 6";

STATELABELS
1     [Pos_A]   ’C\J’    ’G\J’    ’B\D’    ’B\J’    ’B\E’    ’J\L’    ’H\L’,
2     [Pos_B]   ’C\D’    ’A\E’    ’A\C’    ’C\K’    ’C\L’,
3     [Pos_C]   ’A\B’    ’B\J’    ’E\F’    ’B\D’    ’B\F’    ’B\E’    ’B\K’,
4     [Pos_D]   ’B\E’    ’A\J’    ’C\E’    ’E\J’,
5     [Pos_E]   ’D\F’    ’B\C’    ’A\D’    ’C\F’    ’F\K’,
6     [Pos_F]   ’E\H’    ’C\I’    ’E\G’    ’C\G’    ’E\I’    ’E\J’,
7     [Pos_G]   ’I\J’    ’A\I’    ’H\J’    ’F\H’,
8     [Pos_H]   ’F\I’    ’G\I’    ’A\I’,
9     [Pos_I]   ’G\H’    ’F\H’    ’H\J’,
10    [Pos_J]   ’A\G’    ’A\C’    ’D\G’    ’A\I’    ’D\I’    ’F\G’,
11    [Pos_K]   ’B\L’    ’C\E’,
12    [Pos_L]   ’A\K’    ’A\B’,
13    [Ss_A]    ’-’ ’+’,
14    [Ss_B]    ’-’ ’+’,
15    [Ss_C]    ’-’ ’+’,
16    [Ss_D]    ’-’ ’+’,
17    [Ss_E]    ’-’ ’+’,
18    [Ss_F]    ’-’ ’+’,
19    [Ss_G]    ’-’ ’+’,
20    [Ss_H]    ’-’ ’+’,
21    [Ss_I]    ’-’ ’+’,
22    [Ss_J]    ’-’ ’+’,
23    [Ss_K]    ’-’ ’+’,
24    [Ss_L]    ’-’ ’+’,
25    [PA_D]    ’Absent’ ’Present’,
26    [PA_K]    ’Absent’ ’Present’,
27    [PA_L]    ’Absent’ ’Present’,
    ;
MATRIX
Taxon1                          0000000000?? 1100011001?? 100
Taxon2                          1010001001?? 1100011000?? 100
Taxon3                          212111211??? 1100011101?? 100
Taxon4                          3232023123?? 1100011001?? 100
Taxon5                          4243233124?? 1100011011?? 100
Taxon6                          323202312 3?? 1100011011?? 100
Taxon7                          535?34211000 110?01010111 011
Taxon8                          646?45020511 110?01110001 011
    ;
ENDBLOCK;

BEGIN ASSUMPTIONS;
    CHARSET Position    =    1-12;
    CHARSET Sens    =    13-24;
    CHARSET PreAbs    =    25-27;
ENDBLOCK;
```

C.1.1.2 Option AB≠BA

```
#NEXUS

BEGIN DATA;
DIMENSIONS  NTAX=8 NCHAR=27;
FORMAT MISSING=? SYMBOLS= "0 1 2 3 4 5 6";

STATELABELS
1     [Pos_A]    'J\C'    'G\J'    'D\B'    'J\B'    'E\B'    'J\L'    'H\L',
2     [Pos_B]    'C\D'    'A\E'    'A\C'    'K\C'    'L\C',
3     [Pos_C]    'A\B'    'J\B'    'E\F'    'B\D'    'B\F'    'B\E'    'B\K',
4     [Pos_D]    'B\E'    'J\A'    'C\E'    'J\E',
5     [Pos_E]    'D\F'    'B\C'    'D\A'    'C\F'    'K\F',
6     [Pos_F]    'E\H'    'C\I'    'E\G'    'C\G'    'E\I'    'E\J',
7     [Pos_G]    'I\J'    'I\A'    'H\J'    'F\H'    'J\I',
8     [Pos_H]    'F\I'    'I\G'    'G\I'    'I\A',
9     [Pos_I]    'H\G'    'F\H'    'H\J'    'G\H',
10    [Pos_J]    'G\A'    'A\C'    'G\D'    'I\A'    'I\D'    'F\G',
11    [Pos_K]    'L\B'    'C\E',
12    [Pos_L]    'A\K'    'A\B',
13    [Ss_A]     '-' '+',
14    [Ss_B]     '-' '+',
15    [Ss_C]     '-' '+',
16    [Ss_D]     '-' '+',
17    [Ss_E]     '-' '+',
18    [Ss_F]     '-' '+',
19    [Ss_G]     '-' '+',
20    [Ss_H]     '-' '+',
21    [Ss_I]     '-' '+',
22    [Ss_J]     '-' '+',
23    [Ss_K]     '-' '+',
24    [Ss_L]     '-' '+',
25    [PA_D]     'Absent' 'Present',
26    [PA_K]     'Absent' 'Present',
27    [PA_L]     'Absent' 'Present',
   ;
MATRIX
Taxon1                        0000000000?? 1100011001?? 100
Taxon2                        1010001001?? 1100011000?? 100
Taxon3                        2121112112?? 1100011101?? 100
Taxon4                        3232023223?? 1100011001?? 100
Taxon5                        4243233224?? 1100011011?? 100
Taxon6                        3232023223?? 1100011011?? 100
Taxon7                        535?34211000 110?01010111 011
Taxon8                        646?45433511 110?01110001 011
   ;
ENDBLOCK;

BEGIN ASSUMPTIONS;
    CHARSET Position    =    1-12;
    CHARSET Sens    =    13-24;
    CHARSET PreAbs    =    25-27;
ENDBLOCK;
```

C.1.2 Codage « Jonctions »

C.1.2.1 Option AB=BA

```
#NEXUS

BEGIN DATA;
DIMENSIONS    NTAX=8 NCHAR=47;
FORMAT MISSING=? SYMBOLS= "0 1";

STATELABELS
1     [PA_A\B]    'Absent' 'Present',
2     [PA_A\C]    'Absent' 'Present',
3     [PA_A\D]    'Absent' 'Present',
4     [PA_A\E]    'Absent' 'Present',
5     [PA_A\G]    'Absent' 'Present',
6     [PA_A\H]    'Absent' 'Present',
7     [PA_A\J]    'Absent' 'Present',
8     [PA_A\L]    'Absent' 'Present',
9     [PA_B\C]    'Absent' 'Present',
10    [PA_B\D]    'Absent' 'Present',
11    [PA_B\E]    'Absent' 'Present',
12    [PA_B\K]    'Absent' 'Present',
13    [PA_B\L]    'Absent' 'Present',
14    [PA_C\D]    'Absent' 'Present',
15    [PA_C\E]    'Absent' 'Present',
16    [PA_C\F]    'Absent' 'Present',
17    [PA_C\J]    'Absent' 'Present',
18    [PA_C\K]    'Absent' 'Present',
19    [PA_D\E]    'Absent' 'Present',
20    [PA_D\J]    'Absent' 'Present',
21    [PA_E\F]    'Absent' 'Present',
```

```
22    [PA_E\K]    'Absent' 'Present',
23    [PA_F\G]    'Absent' 'Present',
24    [PA_F\H]    'Absent' 'Present',
25    [PA_F\I]    'Absent' 'Present',
26    [PA_F\J]    'Absent' 'Present',
27    [PA_G\H]    'Absent' 'Present',
28    [PA_G\I]    'Absent' 'Present',
29    [PA_G\J]    'Absent' 'Present',
30    [PA_H\I]    'Absent' 'Present',
31    [PA_I\J]    'Absent' 'Present',
32    [PA_K\L]    'Absent' 'Present',
33    [Ss_A]      '-'      '+',
34    [Ss_B]      '-'      '+',
35    [Ss_C]      '-'      '+',
36    [Ss_D]      '-'      '+',
37    [Ss_E]      '-'      '+',
38    [Ss_F]      '-'      '+',
39    [Ss_G]      '-'      '+',
40    [Ss_H]      '-'      '+',
41    [Ss_I]      '-'      '+',
42    [Ss_J]      '-'      '+',
43    [Ss_K]      '-'      '+',
44    [Ss_L]      '-'      '+',
45    [PA_D]      'Absent' 'Present',
46    [PA_K]      'Absent' 'Present',
47    [PA_L]      'Absent' 'Present',
    ;
MATRIX
Taxon1     0100001?110??0000?101?010001110? 1100011001?? 100
Taxon2     0000101?110??0001?101?010001010? 1100011000?? 100
Taxon3     1010000?001??0110?010?001010110? 1100011101?? 100
Taxon4     1000001?100??1000?101?100010011? 1100011001?? 100
Taxon5     1001000?100??0010?110?100010011? 1100011011?? 100
Taxon6     1000001?100??1000?101?100010011? 1100011011?? 100
Taxon7     00?000111?010?1000??100010101101 110?01010111 011
Taxon8     00?001011?001?0001??110001011100 110?01110001 011
    ;
ENDBLOCK;

BEGIN ASSUMPTIONS;
    CHARSET Jonction = 1-32;
    CHARSET Sens = 33-44;
    CHARSET PreAbs = 45-47;
ENDBLOCK;
```

C.1.2.2 Option AB≠BA

```
#NEXUS

BEGIN DATA;
DIMENSIONS    NTAX=8 NCHAR=54;
FORMAT MISSING=? SYMBOLS= "0 1";

STATELABELS
1     [PA_A\B]    'Absent' 'Present',
2     [PA_A\C]    'Absent' 'Present',
3     [PA_A\J]    'Absent' 'Present',
4     [PA_A\L]    'Absent' 'Present',
5     [PA_B\C]    'Absent' 'Present',
6     [PA_B\D]    'Absent' 'Present',
7     [PA_B\E]    'Absent' 'Present',
8     [PA_C\B]    'Absent' 'Present',
9     [PA_C\D]    'Absent' 'Present',
10    [PA_C\E]    'Absent' 'Present',
11    [PA_C\F]    'Absent' 'Present',
12    [PA_C\K]    'Absent' 'Present',
13    [PA_D\A]    'Absent' 'Present',
14    [PA_D\E]    'Absent' 'Present',
15    [PA_E\A]    'Absent' 'Present',
16    [PA_E\C]    'Absent' 'Present',
17    [PA_E\F]    'Absent' 'Present',
18    [PA_F\G]    'Absent' 'Present',
19    [PA_F\H]    'Absent' 'Present',
20    [PA_F\I]    'Absent' 'Present',
21    [PA_F\J]    'Absent' 'Present',
22    [PA_G\A]    'Absent' 'Present',
23    [PA_G\H]    'Absent' 'Present',
24    [PA_G\I]    'Absent' 'Present',
25    [PA_G\J]    'Absent' 'Present',
26    [PA_H\A]    'Absent' 'Present',
27    [PA_H\G]    'Absent' 'Present',
28    [PA_H\I]    'Absent' 'Present',
29    [PA_I\G]    'Absent' 'Present',
30    [PA_I\H]    'Absent' 'Present',
31    [PA_I\J]    'Absent' 'Present',
32    [PA_J\A]    'Absent' 'Present',
33    [PA_J\C]    'Absent' 'Present',
34    [PA_J\D]    'Absent' 'Present',
35    [PA_J\G]    'Absent' 'Present',
```

```
36    [PA_K\B]    'Absent' 'Present',
37    [PA_K\E]    'Absent' 'Present',
38    [PA_L\B]    'Absent' 'Present',
39    [PA_L\K]    'Absent' 'Present',
40    [Ss_A]      '-'      '+',
41    [Ss_B]      '-'      '+',
42    [Ss_C]      '-'      '+',
43    [Ss_D]      '-'      '+',
44    [Ss_E]      '-'      '+',
45    [Ss_F]      '-'      '+',
46    [Ss_G]      '-'      '+',
47    [Ss_H]      '-'      '+',
48    [Ss_I]      '-'      '+',
49    [Ss_J]      '-'      '+',
50    [Ss_K]      '-'      '+',
51    [Ss_L]      '-'      '+',
52    [PA_D]      'Absent' 'Present',
53    [PA_K]      'Absent' 'Present',
54    [PA_L]      'Absent' 'Present',
   ;
MATRIX
Taxon1      010?0101000?0100101000000100110010000??? 11000110011?? 100
Taxon2      001?0101000?0100101001000000110001000??? 1100011000?? 100
Taxon3      100?0010001?1001000100001010010001000??? 1100011101?? 100
Taxon4      100?1000100?0100110000100001001100000??? 1100011001?? 100
Taxon5      100?1000001?0110010000100001001001000??? 1100011011?? 100
Taxon6      100?1000100?0100110000100001001100000??? 1100011011?? 100
Taxon7      00011?00?100??0010010000101001010?01001 110?01010111 011
Taxon8      00011?00?001??0010001001010001000?10110 110?01110001 011
   ;
ENDBLOCK;

BEGIN ASSUMPTIONS;
    CHARSET Jonction = 1-39;
    CHARSET Sens = 40-51;
    CHARSET PreAbs = 52-54;
ENDBLOCK;
```

C.1.3 Codage « Jonctions signées »

```
#NEXUS

BEGIN DATA;
DIMENSIONS    NTAX=8 NCHAR=40;
FORMAT MISSING=? SYMBOLS= "0 1";

STATELABELS
1     [PA_+A\+B]    'Absent' 'Present',
2     [PA_+A\+L]    'Absent' 'Present',
3     [PA_+A\-C]    'Absent' 'Present',
4     [PA_+A\-J]    'Absent' 'Present',
5     [PA_+B\-C]    'Absent' 'Present',
6     [PA_+B\-D]    'Absent' 'Present',
7     [PA_+B\-E]    'Absent' 'Present',
8     [PA_+C\+E]    'Absent' 'Present',
9     [PA_+C\+J]    'Absent' 'Present',
10    [PA_+D\-J]    'Absent' 'Present',
11    [PA_+E\+K]    'Absent' 'Present',
12    [PA_+F\+G]    'Absent' 'Present',
13    [PA_+F\-H]    'Absent' 'Present',
14    [PA_+F\-I]    'Absent' 'Present',
15    [PA_+F\-J]    'Absent' 'Present',
16    [PA_+G\+J]    'Absent' 'Present',
17    [PA_+G\-H]    'Absent' 'Present',
18    [PA_+G\-I]    'Absent' 'Present',
19    [PA_+I\+J]    'Absent' 'Present',
20    [PA_-A\+D]    'Absent' 'Present',
21    [PA_-A\+E]    'Absent' 'Present',
22    [PA_-A\-G]    'Absent' 'Present',
23    [PA_-A\-H]    'Absent' 'Present',
24    [PA_-A\-J]    'Absent' 'Present',
25    [PA_-B\+C]    'Absent' 'Present',
26    [PA_-B\-K]    'Absent' 'Present',
27    [PA_-B\-L]    'Absent' 'Present',
28    [PA_-C\+F]    'Absent' 'Present',
29    [PA_-C\-D]    'Absent' 'Present',
30    [PA_-C\-E]    'Absent' 'Present',
31    [PA_-C\-K]    'Absent' 'Present',
32    [PA_-D\-E]    'Absent' 'Present',
33    [PA_-E\+F]    'Absent' 'Present',
34    [PA_-G\+I]    'Absent' 'Present',
35    [PA_-G\+J]    'Absent' 'Present',
36    [PA_-G\-H]    'Absent' 'Present',
37    [PA_-H\+I]    'Absent' 'Present',
38    [PA_-H\-I]    'Absent' 'Present',
39    [PA_-I\+J]    'Absent' 'Present',
40    [PA_-K\-L]    'Absent' 'Present',
   ;
MATRIX
```

```
Taxon1                          00100100000010010000000110000001110001000
Taxon2                          00010100100010000000001001000000111000100
Taxon3                          10000011010001010001000000010000000110000
Taxon4                          10001000000100001000000100001001100001100
Taxon5                          10001000010100001010100000010001000010000
Taxon6                          10001000000100001010000100001001100010000
Taxon7                          01001000000001001000000101000100101010010
Taxon8                          01001000001000100100001000100010101010000
    ;
ENDBLOCK;
```

C.2 Matrices des chromosomes du genre *Mastomys*

C.2.1 Codage « Position relative »

```
#NEXUS

BEGIN DATA;
DIMENSIONS    NTAX=6 NCHAR=78;
FORMAT MISSING=? SYMBOLS= "0 1 2 3";

STATELABELS
1     [Pos_10dist]   ’10prox\télo’,
2     [Pos_10prox]   ’10dist\centro’   ’10dist\télo’,
3     [Pos_11]       ’centro\télo’,
4     [Pos_12dist]   ’12prox\télo’,
5     [Pos_12prox]   ’12dist\15’       ’12dist\centro’,
6     [Pos_13dist]   ’13prox\8prox’    ’13prox\centro’   ’13prox\télo’,
7     [Pos_13prox]   ’13dist\télo’     ’13dist\8prox’    ’13dist\centro’,
8     [Pos_14]       ’centro\télo’,
9     [Pos_15]       ’12prox\télo’     ’centro\télo’,
10    [Pos_16dist]   ’16prox\télo’,
11    [Pos_16prox]   ’16dist\3prox’    ’16dist\centro’,
12    [Pos_17p]      ’centro\télo’,
13    [Pos_17q]      ’centro\télo’,
14    [Pos_18p]      ’18q\télo’       ’18q\centro’,
15    [Pos_18q]      ’18p\télo’,
16    [Pos_1dist1]   ’1dist2\1prox1’  ’1dist2\centro’    ’1dist2\télo’     ’1dist2\1prox2’,
17    [Pos_1dist2]   ’1dist1\télo’,
18    [Pos_1prox1]   ’1dist1\1prox2’  ’1prox2\centro’    ’1prox2\télo’,
19    [Pos_1prox2]   ’1prox1\télo’    ’1dist1\1prox1’,
20    [Pos_2dist1]   ’2dist2\2prox1’  ’2dist2\centro’    ’2dist2\télo’     ’2dist2\2prox2’,
21    [Pos_2dist2]   ’2dist1\télo’,
22    [Pos_2prox1]   ’2dist1\2prox2’  ’2prox2\centro’    ’2prox2\télo’,
23    [Pos_2prox2]   ’2prox1\télo’    ’2dist1\2prox1’,
24    [Pos_3dist]    ’3prox\télo’,
25    [Pos_3prox]    ’16prox\3dist’   ’3dist\centro’,
26    [Pos_4dist]    ’4prox2\télo’    ’4prox1\télo’,
27    [Pos_4prox1]   ’4prox2\télo’    ’4dist\4prox2’     ’4prox2\centro’,
28    [Pos_4prox2]   ’4dist\4prox1’   ’4prox1\télo’,
29    [Pos_5dist1]   ’5dist2\5prox1’  ’5dist2\télo’      ’5dist2\5prox2’,
30    [Pos_5dist2]   ’5dist1\télo’    ’5dist1\5prox1’    ’5dist1\5prox2’,
31    [Pos_5prox1]   ’5dist1\5prox2’  ’5dist2\5prox2’    ’5prox2\centro’,
32    [Pos_5prox2]   ’5prox1\télo’    ’5dist2\5prox1’    ’5dist1\5prox1’,
33    [Pos_6dist]    ’6prox\télo’,
34    [Pos_6prox]    ’6dist\télo’     ’6dist\centro’,
35    [Pos_7dist]    ’7prox\télo’,
36    [Pos_7prox]    ’7dist\télo’     ’7dist\centro’,
37    [Pos_8dist]    ’8prox\télo’,
38    [Pos_8prox]    ’13dist\8dist’   ’8dist\centro’     ’13prox\8dist’     ’8dist\télo’,
39    [Pos_9]        ’centro\télo’,
40    [Ss_10dist]    ’-’ ’+’,    41    [Ss_10prox]   ’-’ ’+’,
42    [Ss_11]        ’-’ ’+’,    43    [Ss_12dist]   ’-’ ’+’,
44    [Ss_12prox]    ’-’ ’+’,    45    [Ss_13dist]   ’-’ ’+’,
46    [Ss_13prox]    ’-’ ’+’,    47    [Ss_14]       ’-’ ’+’,
48    [Ss_15]        ’-’ ’+’,    49    [Ss_16dist]   ’-’ ’+’,
50    [Ss_16prox]    ’-’ ’+’,    51    [Ss_17p]      ’-’ ’+’,
52    [Ss_17q]       ’-’ ’+’,    53    [Ss_18p]      ’-’ ’+’,
54    [Ss_18q]       ’-’ ’+’,    55    [Ss_1dist1]   ’-’ ’+’,
56    [Ss_1dist2]    ’-’ ’+’,    57    [Ss_1prox1]   ’-’ ’+’,
58    [Ss_1prox2]    ’-’ ’+’,    59    [Ss_2dist1]   ’-’ ’+’,
60    [Ss_2dist2]    ’-’ ’+’,    61    [Ss_2prox1]   ’-’ ’+’,
62    [Ss_2prox2]    ’-’ ’+’,    63    [Ss_3dist]    ’-’ ’+’,
64    [Ss_3prox]     ’-’ ’+’,    65    [Ss_4dist]    ’-’ ’+’,
66    [Ss_4prox1]    ’-’ ’+’,    67    [Ss_4prox2]   ’-’ ’+’,
68    [Ss_5dist1]    ’-’ ’+’,    69    [Ss_5dist2]   ’-’ ’+’,
70    [Ss_5prox1]    ’-’ ’+’,    71    [Ss_5prox2]   ’-’ ’+’,
72    [Ss_6dist]     ’-’ ’+’,    73    [Ss_6prox]    ’-’ ’+’,
74    [Ss_7dist]     ’-’ ’+’,    75    [Ss_7prox]    ’-’ ’+’,
76    [Ss_8dist]     ’-’ ’+’,    77    [Ss_8prox]    ’-’ ’+’,
78    [Ss_9]         ’-’ ’+’,
    ;
MATRIX
```

```
Mastomys_awashencsis     00000000000000000000000000000000000000 11111001111111111001100111111001010111
Mastomys_erythroleucus   00000100000001010101010001111110000010 11111001111110111111111111000000101011l
Mastomys_erythroleucus2  00000100000001020102010001111110000010 11111001111110101110111101000000101011l
Mastomys_natalensis      01000110000001030213021000201221000102O 10110111111111011101110111111001110111l
Mastomys_species         00000100000000020202010001111110000003O 11111001111111101010111111000000101010l
Mastomys_verheyeni       00001220101000030113011010202022010101O 11111111111111111111111111111111111111l
    ;
ENDBLOCK;

BEGIN ASSUMPTIONS;
    CHARSET Position = 1-39;
    CHARSET Sens = 40-78;
ENDBLOCK;
```

C.2.2 Codage « Jonctions »

```
#NEXUS

BEGIN DATA;
    DIMENSIONS NTAX=6 NCHAR=69;
    FORMAT DATATYPE=STANDARD SYMBOLS="01";

STATELABELS
1     [PA_10dist\10prox]   'Absent' 'Present',
2     [PA_12dist\12prox]   'Absent' 'Present',
3     [PA_12prox\15]       'Absent' 'Present',
4     [PA_13dist\13prox]   'Absent' 'Present',
5     [PA_13dist\8prox]    'Absent' 'Present',
6     [PA_13prox\8prox]    'Absent' 'Present',
7     [PA_16dist\16prox]   'Absent' 'Present',
8     [PA_16prox\3prox]    'Absent' 'Present',
9     [PA_18p\18q]         'Absent' 'Present',
10    [PA_1dist1\1dist2]   'Absent' 'Present',
11    [PA_1dist1\1prox1]   'Absent' 'Present',
12    [PA_1dist1\1prox2]   'Absent' 'Present',
13    [PA_1prox1\1prox2]   'Absent' 'Present',
14    [PA_2dist1\2dist2]   'Absent' 'Present',
15    [PA_2dist1\2prox1]   'Absent' 'Present',
16    [PA_2dist1\2prox2]   'Absent' 'Present',
17    [PA_2prox1\2prox2]   'Absent' 'Present',
18    [PA_3dist\3prox]     'Absent' 'Present',
19    [PA_4dist\4prox1]    'Absent' 'Present',
20    [PA_4dist\4prox2]    'Absent' 'Present',
21    [PA_4prox1\4prox2]   'Absent' 'Present',
22    [PA_5dist1\5dist2]   'Absent' 'Present',
23    [PA_5dist1\5prox1]   'Absent' 'Present',
24    [PA_5dist1\5prox2]   'Absent' 'Present',
25    [PA_5dist2\5prox1]   'Absent' 'Present',
26    [PA_5dist2\5prox2]   'Absent' 'Present',
27    [PA_5prox1\5prox2]   'Absent' 'Present',
28    [PA_6dist\6prox]     'Absent' 'Present',
29    [PA_7dist\7prox]     'Absent' 'Present',
30    [PA_8dist\8prox]     'Absent' 'Present',
31    [Ss_10dist]          '-'      '+',
32    [Ss_10prox]          '-'      '+',
33    [Ss_11]              '-'      '+',
34    [Ss_12dist]          '-'      '+',
35    [Ss_12prox]          '-'      '+',
36    [Ss_13dist]          '-'      '+',
37    [Ss_13prox]          '-'      '+',
38    [Ss_14]              '-'      '+',
39    [Ss_15]              '-'      '+',
40    [Ss_16dist]          '-'      '+',
41    [Ss_16prox]          '-'      '+',
42    [Ss_17p]             '-'      '+',
43    [Ss_17q]             '-'      '+',
44    [Ss_18p]             '-'      '+',
45    [Ss_18q]             '-'      '+',
46    [Ss_1dist1]          '-'      '+',
47    [Ss_1dist2]          '-'      '+',
48    [Ss_1prox1]          '-'      '+',
49    [Ss_1prox2]          '-'      '+',
50    [Ss_2dist1]          '-'      '+',
51    [Ss_2dist2]          '-'      '+',
52    [Ss_2prox1]          '-'      '+',
53    [Ss_2prox2]          '-'      '+',
54    [Ss_3dist]           '-'      '+',
55    [Ss_3prox]           '-'      '+',
56    [Ss_4dist]           '-'      '+',
57    [Ss_4prox1]          '-'      '+',
58    [Ss_4prox2]          '-'      '+',
59    [Ss_5dist1]          '-'      '+',
60    [Ss_5dist2]          '-'      '+',
61    [Ss_5prox1]          '-'      '+',
62    [Ss_5prox2]          '-'      '+',
63    [Ss_6dist]           '-'      '+',
64    [Ss_6prox]           '-'      '+',
65    [Ss_7dist]           '-'      '+',
66    [Ss_7prox]           '-'      '+',
```

```
67    [Ss_8dist]           '-'      '+',
68    [Ss_8prox]           '-'      '+',
69    [Ss_9]               '-'      '+',
    ;
MATRIX
Mastomys_awashencsis      11111011111011101011110001111 11111001111111111001100111111001010111
Mastomys_erythroleucus    11110011110011001110110010111 11111001111110111111111111100000001010111
Mastomys_erythroleucus2   11110011110011001110110010111 11111001111110101110111110100000001010111
Mastomys_natalensis       11110111110111011101110001111 10110111111110111011101111110011101111
Mastomys_species          11110011110011001110110010111 11111001111111101010111111100000001010101
Mastomys_verheyeni        11010010110111011101110010111 11111111111111111111111111111111111111111
    ;
END;

BEGIN ASSUMPTIONS;
    CHARSET Jonction = 1-30;
    CHARSET Sens = 31-69;
ENDBLOCK;
```

C.2.3 Codage « Jonctions signées »

```
#NEXUS

BEGIN DATA;
    DIMENSIONS NTAX=6 NCHAR=44;
    FORMAT DATATYPE=STANDARD SYMBOLS="01";

STATELABELS
1     [PA_+12prox\-15]       'Absent' 'Present',
2     [PA_+13dist\+13prox]   'Absent' 'Present',
3     [PA_+13prox\+8prox]    'Absent' 'Present',
4     [PA_+16dist\+16prox]   'Absent' 'Present',
5     [PA_+16prox\+3prox]    'Absent' 'Present',
6     [PA_+16prox\-3prox]    'Absent' 'Present',
7     [PA_+18p\+18q]         'Absent' 'Present',
8     [PA_+1dist1\+1dist2]   'Absent' 'Present',
9     [PA_+1prox1\+1prox2]   'Absent' 'Present',
10    [PA_+2dist1\+2dist2]   'Absent' 'Present',
11    [PA_+2prox1\+2prox2]   'Absent' 'Present',
12    [PA_+4prox1\+4prox2]   'Absent' 'Present',
13    [PA_+5dist1\+5dist2]   'Absent' 'Present',
14    [PA_+5dist2\+5prox1]   'Absent' 'Present',
15    [PA_+5dist2\-5prox2]   'Absent' 'Present',
16    [PA_+5prox1\+5prox2]   'Absent' 'Present',
17    [PA_-10dist\+10prox]   'Absent' 'Present',
18    [PA_-10dist\-10prox]   'Absent' 'Present',
19    [PA_-12dist\+12prox]   'Absent' 'Present',
20    [PA_-12dist\-12prox]   'Absent' 'Present',
21    [PA_-12prox\-15]       'Absent' 'Present',
22    [PA_-13dist\+8prox]    'Absent' 'Present',
23    [PA_-16dist\-16prox]   'Absent' 'Present',
24    [PA_-18p\+18q]         'Absent' 'Present',
25    [PA_-1dist1\+1dist2]   'Absent' 'Present',
26    [PA_-1dist1\+1prox1]   'Absent' 'Present',
27    [PA_-1dist1\-1prox2]   'Absent' 'Present',
28    [PA_-1prox1\+1prox2]   'Absent' 'Present',
29    [PA_-2dist1\+2dist2]   'Absent' 'Present',
30    [PA_-2dist1\+2prox1]   'Absent' 'Present',
31    [PA_-2dist1\-2prox2]   'Absent' 'Present',
32    [PA_-2prox1\+2prox2]   'Absent' 'Present',
33    [PA_-3dist\+3prox]     'Absent' 'Present',
34    [PA_-3dist\-3prox]     'Absent' 'Present',
35    [PA_-4dist\+4prox1]    'Absent' 'Present',
36    [PA_-4dist\-4prox2]    'Absent' 'Present',
37    [PA_-5dist1\+5prox1]   'Absent' 'Present',
38    [PA_-5dist1\-5prox2]   'Absent' 'Present',
39    [PA_-6dist\+6prox]     'Absent' 'Present',
40    [PA_-6dist\-6prox]     'Absent' 'Present',
41    [PA_-7dist\+7prox]     'Absent' 'Present',
42    [PA_-7dist\-7prox]     'Absent' 'Present',
43    [PA_-8dist\+8prox]     'Absent' 'Present',
44    [PA_-8dist\-8prox]     'Absent' 'Present',
    ;
MATRIX
Mastomys_verheyeni        01000011111110010101001000100010010101010101
Mastomys_awashencsis      01011011111110010101110001000100010110101001
Mastomys_erythroleucus    01011001111111010101100100000000011000101001
Mastomys_erythroleucus2   01010100101111010101100110001000101000101001
Mastomys_natalensis       11111001010110111010000100110011010100100101
Mastomys_species          01011010001111010101100010011000011000101010
    ;
ENDBLOCK;
```

D

Articles

GENE ORDER AND PHYLOGENETIC INFORMATION

Cyril Gallut

Véronique Barriel

Régine Vignes

We present a new gene order cladistic coding approach. This approach is based on the physical description of the genome's organization. It takes into account the relative positions of genes in the genome, their transcription sense and their presence-absence. We compare it to the junction coding approach in a phylogenetic frame. Different coding options are considered and discussed.

1 Introduction

The ever increasing number of metazoan mitochondrial genomes available provides a useful data set for phylogenetic reconstruction. Molecular phylogenetic inference is usually based on comparison of homologous sequences. The mitochondrial gene sequences are widely used for this purpose, but it is difficult to distinguish homoplasy from homology when comparing animals from different phyla. The important accumulation of mutations in the mitochondrion covers the evolutionary signal. It has been suggested to infer mitochondrial phylogeny from the gene order rather than from gene sequences (Sankoff et al., 1990, 1992; Boore and Brown, 1998). Unlike gene sequences, organellar gene orders are conserved among metazoan, see Saccone et al. (1999); Boore (1999) for recent reviews.

The evolution of gene rearrangements has been addressed with different methods; Sankoff et al. (1992) proposed an edit distance based on transposition, inversion, insertion and deletion of genes; Blanchette et al. (1999); Sankoff and Blanchette (1998, 1999) proposed a breakpoint distance analysis. Bridge et al. (1992) inferred relationships among cnidarian classes using the number and the shape of the molecule(s) composing the mitochondrial genome. Smith et al. (1993) undertook echinoderm phylogeny and Boore et al. (1995, 1998) the arthropods phylogeny, grouping taxa on the share of rearrangements. This implies the recognition of the rearrangements that occurred between taxa.

We aim to free our selves from the *a priori* interpretation of rearrangements that happened during the evolution of taxa. We propose and compare two cladis-

tic coding of the mitochondrial genome organization. The purpose of a cladistic approach is to compare taxa globally instead of pairwisely. The comparison is not based on observation of rearrangements between two taxa but on a physical description of every genome. Only homologous features are retained in this description.

2 Genome morphology

The metazoan mitochondrial genome is a closed-circular DNA molecule (except in some Cnidarian (Bridge et al., 1992)), ranging in size from 13 kb to 42 kb. The gene content is highly conserved and is made up of 13 protein subunits of the phosphorylative oxidation complexes, 2 ribosomal RNA subunits and 22 tRNAs used in the translation of the mitochondrial encoded proteins (reviewed in Wolstenholme (1992)). There is one (or more) non coding sequence(s) of very variable size. Some of these non coding sequences contain structures involved in the initiation of the replication and translation processes. Genes can be encoded exclusively on the same strand or on both strands.

As for usual morphological descriptions where organisms are divided in different levels of integration, the genome morphology can be divided in two levels. First, the broad organization level and second, the genome constituents level. The general level encompasses the number, the shape (*i.e.* linear or circular) and the size of the DNA molecule(s) so as the presence of particular structures. By structure we mean any distinctive feature part of the genome like genes, binding sites, stem-loop structure, repeated sequences etc. A genome is depicted by the distribution of these structures. Each one of these has a set of characteristics including: 1) its function (if any), 2) its length, 3) its nucleotide sequence, 4) its transcription sense (if appropriate), 5) its position in the genome and so on.

In order to perform a parsimony analysis in this morphological frame, we retain as "characters" only the genome features postulated as homologous. The shape of the DNA molecule can be considered as a character, whereas the size cannot. Indeed many sets of non-homologous molecular events can lead to the same size, making the homology of the genome size very questionable. We consider each genome structure as possible characters as far as the structure itself is homologous between the involved taxa. For example the non-coding sequences can hardly be considered as homologous on a large phylogenetic scale but the Cytochrome b gene is more likely to be homologous. For each homologous structure we can draw several characters from their characteristics, some of these are the position and the transcription sense. We do not retain the structure's function because it is used to identify the structure, besides it is generally unchanging, except in tRNAs shifts (this could bring valuable information). The structure's size is neither a character just like the genome's size. The set of retained characters depends on the taxonomic sampling, for example the character "shape of the genome" is only interesting if the sampling involves taxa with linear and circular mtDNAs.

3 Coding

Among the characters extracted from the genome morphology we focus on the characters involved in the representation of the mitochondrion's gene order. The gene order problem can be addressed at both levels of organization of the genome. One can consider, on the one hand, the global distribution of genes in the genome, and on the other hand the position of each gene by itself. This led us to consider two coding approaches, the "junctions" and the "relative positions". In both cases we must take into account the transcription sense and the presence-absence of each individual gene.

As the molecule is circular, to define the transcription sense of a gene we must distinguish the two strands. For the vertebrates the two strands are identified by their molecular weight and the replication origins. When such elements are unidentified or missing, the main strand is, by convention, the one that encodes more genes. Genes encoded on the main strand have the "main sense", the others have the reverse sense. We can then define an overall main transcription sense of the genome used to identify every individual gene transcription sense. Even if the gene contents is very conserved among metazoan, a few taxa can have some genes lacking. Nematodes for example (Okimoto et al., 1992) miss the gene coding for the ATPase subunit 8. We introduce a binary character for the transcription sense of each gene and a character of presence-absence for each of the missing genes in our two coding approaches. (e.g. see Table 1).

In order to empirically evaluate these coding approaches we designed eight hypothetical taxa, showing different gene arrangements, see Figure 1. They consist of ten or eleven genes, labeled from A to L. The genes transcription sense is whether clockwise or anticlockwise, which is represented by the inner arrows. We coded this set of taxa with our two approaches.

Sense

Characters	A	B	C	D	E	F	G	H	I	J	K	L
Taxon 7	+	+	-	Abs	-	+	-	+	-	+	+	+
Coding	1	1	0	?	0	1	0	1	0	1	1	1

Presence-absence

Characters	D	K	L
Taxon 7	Absent	Present	Present
Coding	0	1	1

Table 1: Transcription sense and gene presence-absence coding of taxon 7 from Figure 1.

3.1 Junctions

This approach is adapted from Sankoff and Blanchette (1999) who compare it to their breakpoint distance. It relies on genes junction, which is the contact of two contiguous genes. Every junction encountered in the taxa is viewed as a

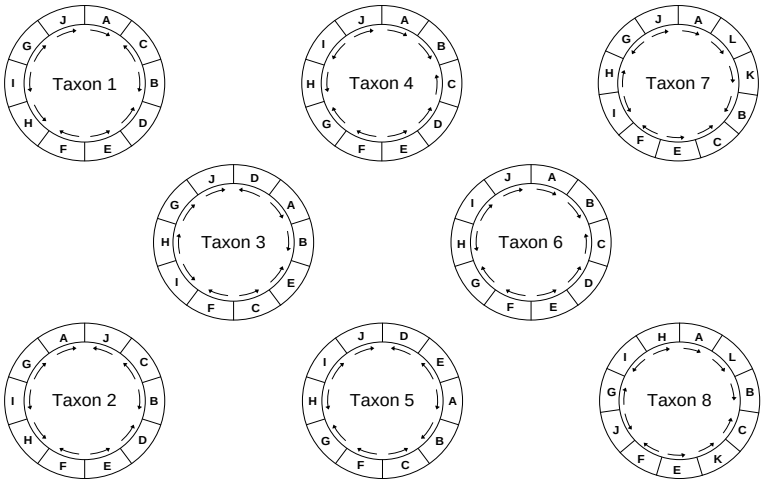


Figure 1: Theoretical genomes. Genes are labeled from A to L, the transcription sense is symbolized by the inner arrows.

binary character in terms of presence-absence. The complete set of junctions, for a particular analysis, is extracted from the whole taxa. Each taxon is then coded by the presence or absence of every junction from this complete set. As said above we add to these "junctions" characters the "sense" and "gene presence-absence" characters (Table 1). It allows a complete description of the genome organization as illustrated in Table 2.

Junctions													
Characters	AB	AC	AJ	AL	BC	BD	BE	CB	CD	CE	CF	CK	DA
Taxon 7				yes	yes					yes			
Coding	0	0	0	1	1	0	0	0	0	1	0	0	0
...	DE	EA	EC	EF	FG	FH	FI	FJ	GA	GH	GI	GJ	HA
...				yes			yes					yes	
...	0	0	0	1	0	0	1	0	0	0	0	1	0
...	HG	HI	IG	IH	IJ	JA	JC	JD	JG	KB	KE	LB	LK
...	yes			yes		yes				yes			yes
...	1	0	0	1	0	1	0	0	0	1	0	0	1

Table 2: Junction coding of taxon 7 from Figure 1.

We distinguish two possible junction coding: junctions are viewed as pairs or as couples. In the first case the junction of genes A and B is a different character from the junction of genes B and A. In the second case it is the same character, it means that junctions AB and BA are identical. This point will be discussed later.

3.2 Relative positions

Here, the gene order problem is addressed at the gene level, unlike in the preceding approach. We deem as a character the relative positions of each gene present at least once in the sampled taxa. The relative position of a gene B, in a genome, is characterized by the two genes surrounding B. The groups of genes surrounding gene B in the involved genomes constitutes the different states of the "relative position of the gene B" character; which is treated as multi-states. (e.g. see Table 3). When a gene is missing in a taxon its position is coded with a question mark instead of adding an "absent" state to character (Barriel and Tassy, 1993) which would be redundant with the presence-absence characters.

Taxon	1	2	3	4	5	6	7	8
Relative position	CD	CD	AE	AC	AC	AC	KC	LC
Coding	0	0	1	2	2	2	3	4

Table 3: Relative positions of gene B in taxa from Figure 1.

Alike junctions, the character states can be treated as pairs or as couples, the state AB is or is not equivalent to the state BA. We also add the "sense" and "gene presence-absence" characters (Table 1). See Gallut and Barriel (2000) for a complete description of this coding.

Relative positions

Character	A	B	C	D	E	F	G	H	I	J	K	L
Taxon 7	JL	KC	BE	Abs	CF	EI	HJ	IG	FH	GA	LB	AK
Coding	5	3	5	?	3	4	2	1	1	0	1	1
Taxon 8	HL	LC	BK	Abs	KF	EJ	JI	IA	GH	FG	CE	AB
Coding	6	4	6	?	4	5	4	3	3	5	2	2

Table 4: Relative positions coding of taxon 7 and taxon 8 from Figure 1.

The Table 4 shows the coding of the taxon 7 and 8 (considering states AB and BA as different).

4 Theoretical example

We coded the eight hypothetical taxa with our two coding approaches, see Figure 1, in both cases we made a distinction between the two possibilities AB≠BA and AB=BA, yielding four matrices. We ran these matrices with PAUP* (Swofford, 1998) using an exhaustive search.

4.1 Junctions

The matrix using AB≠BA is composed of 54 binary characters, among which 12 characters of transcription sense, 3 of presence-absence of a gene and 39 of presence-absence of a junction. Among the characters presence-absence of a junction, 21 are parsimony informative and no character is constant, which means that there is no junction shared by all the taxa. This matrix produced one most parsimonious tree with a length of 60 steps, the consistency index is 0.783 and the retention index is 0.675.

The matrix using AB=BA is composed of 47 binary characters, among which 12 characters of transcription sense, 3 of presence-absence of a gene and 32 of presence-absence of a junction. Among the characters presence-absence of a junction, 17 are parsimony informative and one character is constant, the junction HI (or IH) is shared by all the taxa. This matrix produced three equally parsimonious trees of 52 steps with a consistency index of 0.750 and a retention index of 0.617.

4.2 Relative positions

The matrix using AB≠BA is composed of 27 multistates characters, among which 12 characters of transcription sense, 3 of presence-absence of a gene and 12 of relative position of gene. Among these, 7 are parsimony informative, no character is constant, which means that there is no gene with the same position in all the taxa. This matrix produced three equally parsimonious trees with 54 steps, the consistency index is 0.981 and the retention index is 0.937.

The matrix using AB=BA is composed of 27 multistates characters, among which 12 characters of transcription sense, 3 of presence-absence of a gene and 12 of relative position of gene, It is the same number of characters as before, only the number of states changes. Among these, 7 are parsimony informative, no character is constant. This matrix produced one most parsimonious tree with 53 steps; the consistency index and the retention index are 0.943 and 0.823 respectively.

Upon the tree obtained in the last case we reconstructed the genome of every common ancestor. At each internal node many characters can take different states, character transformations can then be placed at several nodes. Choosing the placement of transformations on the tree is called optimization. There are many equally parsimonious optimizations, among these we only retained the ones yielding "possible" genomes at each internal node. It means that the position of each gene must be compatible with the position of the others (e.g. if A is localized between B and C, the position of B must be something like XA and the position of C something like AY and so on). This gave rise to two equally parsimonious reconstructions differing only on two internal genomes see Figure 2.

5 Discussion

A good coding has to represent the observed data as fairly as possible, but it must avoid redundancy, in other words the same information must not be coded by

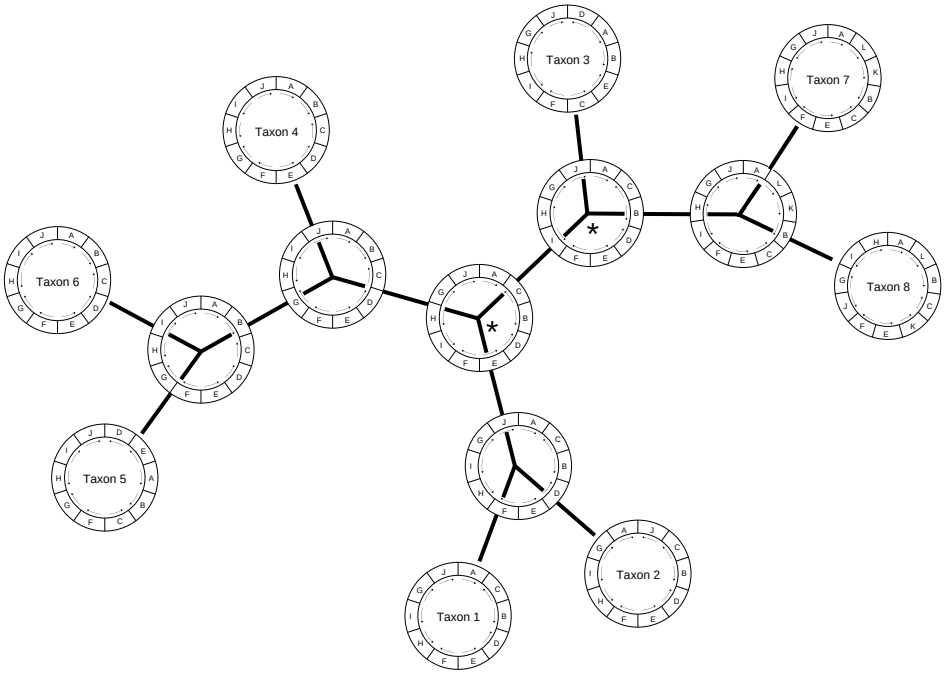


Figure 2: One of the two equally parsimonious reconstruction of ancestral states upon the tree obtained with relative positions coding (with AB=BA option). In the other reconstruction, genomes labelled with a * only differ on the order of A C B which is A B C.

the mean of different characters. The use of three kinds of characters, positions (junctions or relative) transcription sense and presence-absence of genes, allows the complete description of genome topology. It appeared that the distinction between the states AB and BA (in both junction and relative approaches) brings redundant information with the transcription sense characters. When, for example, a block of genes undergoes an inversion, the both kinds of type characters change simultaneously. This is unsatisfying, we thus tried to avoid this by the non distinction of AB and BA. These two options have advantages and drawbacks. In the extreme case of the two seemingly similar genomes of the Figure 3, the use of AB≠BA allows to distinguish them, the other option would code the genomes exactly alike. Nevertheless we must remark that it is rather improbable to encounter this situation with real genomes. The preceding theoretical example shows that with AB≠BA the indexes are higher and there are more informative characters. On the other hand, the AB≠BA option in junction coding can yield clades supported by redundant characters. For example the clade (Taxon3, Taxon7, Taxon8) is supported by the presence of junction IH and by the absence of junction HI, the presence of one out the two opposite junctions implies the absence of the other. Therefore we prefer to employed the AB=BA option rather than the other.

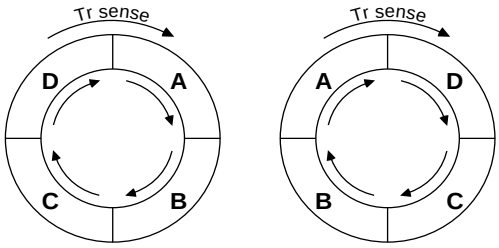


Figure 3: Genomes with opposite gene orders. The upper arrows show the overall transcription sense. The inner arrows show the transcription sense of each individual gene.

The results obtained from the theoretical example show that there is not much difference between the junction and the relative position coding. The indexes of relative positions trees are higher than those from junctions trees but this is due to the higher number of autapomorphies included in relative position matrices. These autapomorphies are character states present only in one taxon. The main difference is that it is not possible to reconstruct the genomes of internal nodes upon junctions trees unlike with relative positions. On trees obtained with junction coding, internal nodes are characterized by the presence of few junctions and by many junction absences. There are not enough present junctions to reconstruct a complete genome. A genome made of ten genes has ten junctions, if for example a node is sustained only by the presence of seven junctions it is not possible to reconstruct the corresponding genome. This is a strong argument to use the relative positions coding.

The relative position coding were applied to a sampling of twenty five metazoan mitochondrial genomes (Gallut, 1998). This analysis yielded interesting results and most of them are congruent with the actual metazoan phylogeny. This approach could be employed to address the karyotype arrangements.

Acknowledgments

CG receives a Phd grant from the Ministère de l'Éducation Nationale, de la Recherche et de la Technologie.

References

Barriel, V. and Tassy, P. (1993). Characters, observations and steps: comment on lipscombs "parsimony, homology and the analysis of multistate characters". *Cladistics*, 9(2):223–232.

Blanchette, M., Kunisawa, T., and Sankoff, D. (1999). Gene order breakpoint evidence in animal mitochondrial phylogeny. *Journal of Molecular Evolution*, 49(2):193–203.

Boore, J. L. (1999). Animal mitochondrial genomes. *Nucleic Acids Research*, 27(8):1767–1780.

Boore, J. L. and Brown, W. M. (1998). Big trees from little genomes: mitochondrial gene order as a phylogenetic tool. *Current Opinion in Genetics And Development*, 8(6):668–674.

Boore, J. L., Collins, T. M., Stanton, D., Daehler, L. L., and Brown, W. M. (1995). Deducing the pattern of arthropod phylogeny from mitochondrial dna rearrangements. *Nature*, 376(6536):163–165.

Boore, J. L., Lavrov, D., and Brown, W. M. (1998). Gene translocation links insects and crustaceans. *Nature*, 392(6677):667–668.

Bridge, D., Cunningham, C. W., Schierwater, B., DeSalle, R., and Buss, L. W. (1992). Class-level relationships in the phylum cnidaria: evidence from mitochondrial genome structure. *Proceedings of The National Academy of Sciences of The United States of America*, 89(18):8750–8753.

Gallut, C. (1998). *Codage de l'ordre des gènes du génome mitochondrial animal en vue d'une analyse phylogénétique*. Mémoire de dea, Paris VI.

Gallut, C. and Barriel, V. (2000). Mitochondrial gene order coding. *Cladistics*, Submitted.

Okimoto, R., Macfarlane, J. L., Clary, D. O., and Wolstenholme, D. R. (1992). The mitochondrial genomes of two nematodes, caenorhabditis elegans and ascaris suum. *Genetics*, 130(3):471–498.

Saccone, C., De Giorgi, C., Gissi, C., Pesole, G., and Reyes, A. (1999). Evolutionary genomics in metazoa: the mitochondrial dna as a model system. *Gene*, 238(1):195–209.

Sankoff, D. and Blanchette, M. (1998). Multiple genome rearrangement and breakpoint phylogeny. *Journal Of Computational Biology*, 5(3):555–570.

Sankoff, D. and Blanchette, M. (1999). Comparative genomics via phylogenetic invariants for jukes-cantor semigroups. In Gorostiza, L. and Ivanoff, G., editors, *Proceedings of the International Conference on Stochastic Models*, Conference Proceedings series, Canadian Mathematical Society.

Sankoff, D., Cedergren, R. J., and Abel, Y. (1990). Genomic divergence through gene rearrangement. *Methods in Enzymology*, 183:428–438.

Sankoff, D., Leduc, G., Antoine, N., Paquin, B., Lang, B. F., and Cedergren, R. J. (1992). Gene order comparisons for phylogenetic inference: evolution of the mitochondrial genome. *Proceedings of The National Academy of Sciences of The United States of America*, 89(14):6575–6579.

Smith, M. J., Arndt, A., Gorski, S., and Fajber, E. (1993). The phylogeny of echinoderm classes based on mitochondrial gene arrangements. *Journal of Molecular Evolution*, 36(6):545–554.

Swofford, D. L. (1998). Paup*: Phylogenetic analysis using parsimony (*and other methods).

Wolstenholme, D. R. (1992). Animal mitochondrial dna: structure and evolution. *International Review of Cytology*, 141:173–216.

Laboratoire Informatique et Systématique, Université Pierre et Marie Curie. 12, rue Cuvier 75005 Paris FRANCE
E-mail: gallut@ccr.jussieu.fr

Service de Systématique Moléculaire, Muséum National d'Histoire Naturelle. 43, rue Cuvier 75005 Paris FRANCE
E-mail: barriel@mnhn.fr

Laboratoire Informatique et Systématique, Université Pierre et Marie Curie. 12, rue Cuvier 75005 Paris FRANCE
E-mail: vignes@ccr.jussieu.fr

Soumis le 9 juillet 2001 à Cladistics

Cladistic coding of genomic maps

Cyril Gallut [1,2] and Véronique Barriel [1,3]

1. Institut de Systématique (CNRS FR 1541).
2. Laboratoire Informatique & Systématique, jeune équipe Classification Évolution et Biosystématique (JE 2160), Université Pierre et Marie Curie. 12 rue Cuvier 75005 PARIS FRANCE

   Tel: 33 (0) 1 44 27 65 21; Fax: 33 (0) 1 44 27 65 60; email: gallut@ccr.jussieu.fr
3. Service de Systématique Moléculaire, Muséum National d'Histoire Naturelle, 43 rue Cuvier F-75231 PARIS CEDEX 05 FRANCE

   Tel: 33 (0) 1 40 79 31 71; Fax : 33 (0) 1 40 79 38 44; email: barriel@mnhn.fr

Abstract

A new method of genomic maps analysis in a cladistic framework is described. The purpose is reconstructing phylogenetic relationships from genomic organization of taxa. Our approach is based on gene order coding. This coding allows the description of genome topology without a priori hypothesis about evolutionary events and phylogenetic relationships. Different characters are used for each gene: 1) presence/absence, 2) orientation and 3) relative position. The relative position of a particular gene inside genome is the pair of genes surrounding it. The relative position character represents all the positions of a gene in the sampled genomes. It is coded as a multistate character. Our coding method has a priori variable costs implications on operators like inversion, transposition, gene loss/gain; these implications are discussed. The overall approach best fits the “duplication, random loss” evolutionary model. The coding method allows the reconstitution of a possible hypothetical common ancestor genome at each node of the tree. This reconstitution is based on the character states optimisation; it comes down to choose among all possible optimisations, the optimisation compatible with a complete genome topology at each internal node. This reconstitution is permitted by the multistate coding of gene relative position, which is an undeniable advantage of this method.

Table of contents:

List of figures:

List of tables:

Introduction

The development of automated DNA sequencing strategies have generated so large amounts of molecular data to phylogeneticists that computation of phylogenetic trees from sequences data has become exceedingly difficult owing to the size of the data sets which now include extremely long and numerous DNA sequences. In an anticipation of the problem, ten years ago, (Sankoff *et al.,* 1992) proposed to use for phylogenetic purposes not only the comparison of linear nucleotide or amino-acid sequences, but also molecular maps of segments of the genome. More recently Boore and Brown (1998) pointed out for the need of methods adapted to this new family of molecular data.

Indeed, the use of gene order in phylogeny may provide several advantages over that of nucleotide or amino-acid sequences:

- gene order is likely to evolve more slowly than sequences, and consequently genomic maps contrarily to nucleotide sequences, may provide better information concerning ancient cladogenetic events;
- as they are far less frequent than substitutions, rearrangements of genes are likely to be less deeply hidden than mutations affecting nucleotides, by later rearrangements, when analysing fast radiations;
- structural data likely provide data sets smaller in size and thus more manageable than nucleotide data, a feature which may become an operational necessity as complete genomes data are accumulating.

However, and as for other methods, a method for coding genomic structural data should fulfil several general conditions:

- it should allow a complete description of gene topology without any *a priori* hypothesis about the underlying clades and phylogenetic events;
- whenever possible, and since no model of genome structural evolution exists so far, the coding method should not be alienated to a predetermined model. This requirement implies that coding must authorize the independent move of each gene, keeping in mind that this condition is not totally freed of evolutionary assumptions (by making no assumption on rearrangement of large blocks of genes, the trend is to conform to the model of duplication/random loss (Macey *et al.*, 1997; Macey *et al.*, 1998)).

Finally it is worth noting that a method permitting the coding of gene order should be applicable to all kinds of structures consisting in a linear succession of elements the relative position of which varies among clades, such as cytogenetic data currently used empirically to infer phylogenetic relationships or the gene order of homologous clusters of genes.

However, the coding of relative positions is a delicate procedure the consequences of which have to be evaluated before the method is generalized. In order to explore the direct and indirect effects of coding attempts, we have explored the relationships between coding, results and actual patterns of evolution using the defined example of synthetic circular genomes (similar to mitochondrial or chloroplastic genomes).

First attempts to use gene maps to inferring relationships among taxa were based on edit distances (Sankoff *et al.*, 1992; Blanchette *et al.*, 1996). Edit distances are pairwise comparisons based on different operators, like inversion, transposition and weighted combination of inversion and transposition (see references in Blanchette *et al.*, (1999) for additional details). Recently methods implementing global comparison with breakpoint distances have come out (Sankoff and Blanchette, 1998; Blanchette *et al.*, 1999; Sankoff *et al.*, 2000). Cosner *et al.*, (2000a) used "Maximum Parsimony on Binary Encodings" as heuristic for breakpoint distance phylogeny reconstruction using chloroplasts gene order (Cosner *et al.*, 2000b).

Mitochondrial gene order has already been successfully used to inferring echinoderms (Smith *et al.*, 1993) and arthropods relationships (Boore *et al.*, 1995; Boore *et al.*, 1998). However, these works reconstructed phylogenetic patterns "by hand", with no computation method that could be implemented. Such an approach can be made on small data sets associated with minor variations. The limits of the empirical approach are easily reached as the size of the data set increases, thus strengthening the need for a more general method. We propose here a formal method implementing a coding of gene order to reconstruct phylogenetic relationships in a cladistic frame.

Genomic morphology

We will consider here genomic morphology as structural organization, excluding functional organization like chromatin condensation. Indeed chromatin organization depends on the physiological state of the cell and thus cannot be considered as a phylogenetic character. On the other hand, structural organization encompasses genetic maps *sensu lato*, like caryotypes, synthenic gene clusters, mitochondrial or chloroplastic gene orders. This kind of genomic morphology carries valuable phylogenetic characters.

Organization

In phylogenetic studies, hypotheses concerning homology define characters, and the way characters are coded in addition to the definition of characters, determine the final results. In molecular phylogeny, the positions within aligned sequences are the characters, and the nucleotides or amino acids are character states. In morphological analysis, the different parts (organs or parts of organs) composing the organisms are usually taken for characters, and the various forms observed for each character are character states. In molecular as well as in morphological analyses, the definition of a character is based upon its position in the plan of the organism or in the sequence of the gene, in other words its connexions (“principe des connexions”, Geoffroy Saint-Hilaire, *in* Philosophie anatomique (1818-1822)).

When genome organizations or maps are to be compared, the homology of the genes is *a priori* defined through the comparison of their nucleotide sequence, which is not correlated to connections. One of the questions to answer deals with the structural properties that may be considered as characters, *i.e.* which can be the object of a (primary) homology hypothesis. Character states are considered in a second time. For that purpose we need to identify the levels of organization of the genome, what the components of the genomes at each level are, and finally the features of the components and of the organization.

Genome organization may be analysed at the level of the DNA molecule or at the component and sequence levels. The nucleotide sequences being widely used in phylogeny will not consequently be addressed here. At the genome level, for example that of mtDNA, potential characters are the number, size and shape (linear or circular) of the DNA molecule(s), the presence or absence of some functional regions (e.g. control region in mt genomes). However, most of these features are labile (number of molecules, length…), or seemingly unique and then not allowing comparison (two linear DNA molecules of some hydrozoan mt genomes (Bridge *et al.*, 1992)). Conversely, presence/absence of potentially homologous functional region(s) may obviously be retained as character.

The component level corresponds to the functional units of the genome, such as genes, repeated units, protein binding sites, spatial structures such as stems and loops etc. These components may be described through: 1) their function; 2) their length; 3) the succession of their nucleotides; 4) the strand on which they are located; 5) their position relatively to other genome components. Once more, length is too labile to be of any use, whereas function and

sequence, which are obviously inter-related, are used to hypothesize homology of components : they thus cannot be used to compare genome organizations. Conversely, the relative position of the genes and, once strand homology is hypothesized, their orientation, can be compared in terms of primary homology (de Pinna, 1991).

As a matter of consequence our approach is based on considering genes (or functional regions) as characters and minimizing, for each of them, the number of changes in their relative position and transcription polarity, in addition to their presence or absence. Each step thus represents a change in the position, or a change of polarity, or a gain/loss event (indel). This approach is intended to avoiding, as much as possible, evolutionary assumptions due to any genomic evolutionary model.

Strand homology

In order to code for the transcription polarity of a gene, one has to define the strand of DNA from which it is transcribed. In the particular case of circular DNA molecules such as mtDNA, this requires the identification of both strands and the establishment of the homology between strands of the different species. As an example, the two vertebrate mitochondrial DNA strands possesses different molecular weights, and are thus commonly named “heavy” and “light” strand. In the usual absence of such a criterion, the strands are usually named “main” and “opposite” or “leading” and “lagging”. When a nearly- objective criterion, such as the origin of replication or a base composition bias, is missing, thus not allowing a straightforward identification of strand homology between species, it becomes necessary to test all the strands

homology hypothesizes. This can be done by inverting the main and opposite strands of ambiguous species and analysing all the possible combinations of inversion.

Characters and coding

The purpose of this approach is to generate an independent description of each genome with no *a priori* concerning the phylogenetic relationships of the taxa, and thus in the absence of any information on the rearrangements involved in the phylogeny of the taxa, an approach which is not is feasible by using a pairwise comparison.

The characters observed can be coded in different ways, from complete binary coding to complete multistate coding (Pleijel, 1995). In general we used multistate coding every time multiple observations suggested homology. However, absence was not considered a character state (*i.e.* not as a positive observation), and was consequently coded as a question mark for absence (Barriel and Tassy, 1993). Presence or absence of a character was coded as a distinct character, in addition to multistate.

"Gene Presence/absence"

If a gene is not present in all taxa studied we used a "presence/absence " character for this gene. It is a binary character. When the gene was absent from a given genome, its position and transcription strand were coded by using a question mark. For example (Figure 1) since taxa 7

and 8 miss gene D, we added a “presence/absence ” character for this gene, its position and strand being coded as “?” in taxa 7 and 8.

“Gene polarity”

“Gene polarity” character depicts the way the gene is transcribed and was coded as binary character. When a gene was transcribed from the main strand the character takes the state “+” and when transcribed from the opposite strand the character state is “-”. For example (Figure 1) the gene J is “+” in taxon 1 and “-” in taxon 2.

“Gene Relative position”

In order to depict the relative position of genes (or functional regions) on the DNA molecule we used a character defined as the “relative position of the gene”. This character is coded as unordered multistate. In a given genome the surrounding two neighbours define the relative position of the considered gene. It then becomes possible to describe the gene order of a genome by using for each gene the couple of genes, which surround it. The entire couples of genes surrounding a given gene in different genomes represent the states of the “relative position” character. For example (Figure 1) the gene A is surrounded by: J\C in taxon 1, G\J in taxon 2, D\B in taxon 3, J\B in taxon 4, E\B in taxon 5, J\B in taxon 6, J\L in taxon 7, H\L in taxon 8. The states of the character “relative position of the gene A” are then: {J\C, G\J, D\B, J\B, E\B, J\L, H\L}.

The combined use of “relative position” and “polarity” characters can lead to some redundancy in the data matrix. As an example, if gene X is surrounded by the couple of genes (A, B) in a taxon and by the couple (B, A) in an other taxon, the inversion of gene X is represented jointly by both characters “relative position” and “polarity”. Indeed, the character “relative position of gene X” passes from state A/B to state B/A and the character “polarity of gene X” also changes its state. In order to eliminate this redundancy we have decided to consider the couple (A, B) and (B, A) as a pair, which means that A/B and B/A are a single state, as discussed elsewhere (Gallut *et al.*, 2000).

As an example of relative position coding, the Table 1 shows the coding of taxa 6, 7 and 8 from Figure 1. See appendix 1 for the complete matrix.

	Position of											
	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**
Taxon 6	J\B	A\C	B\D	C\E	D\F	E\G	F\H	G\I	H\J	I\A	/	/
Coding	0	0	0	0	0	0	0	0	0	0	?	?
Taxon 7	J\L	K\C	B\E	/	C\F	E\I	H\J	G\I	F\H	G\A	L\B	A\K
Coding	1	1	1	?	1	1	1	0	1	1	0	0
Taxon 8	H\L	C\L	B\K	/	F\K	E\J	I\J	A\I	G\H	F\G	C\E	A\B
Coding	2	2	2	?	2	2	2	1	2	2	1	1

	Sense of												Presence/absence of		
	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**	**D**	**K**	**L**
Taxon 6	+	+	-	-	-	+	+	-	+	+	/	/	Present	Absent	Absent
Coding	1	1	0	0	0	1	1	0	1	1	?	?	1	0	0
Taxon 7	+	+	-	/	-	+	-	+	-	+	+	+	Absent	Present	Present
Coding	1	1	0	?	0	1	0	1	0	1	1	1	0	1	1
Taxon 8	+	+	-	/	-	+	+	+	-	-	-	+	Absent	Present	Present
Coding	1	1	0	?	0	1	1	1	0	0	0	1	0	1	1

Table 1

Coding the position of a gene inside a genome can be done in several ways, “relative position” is one of them. We have compared it to another coding (Gallut *et al.*, 2000), presence/absence of genes junctions, which is a complete binary coding.

In our approach, each step represents either a change of position or a change of strand, in order to avoid, as far as possible, any genomic evolutionary model. Despite this intend, our coding used along with the cladistic method, inescapably implies some kind of underlying evolutionary assumptions: in particular, we can expect that decomposing processes into such elementary steps may induce some artificial combinations of characters when the data are processed by the parsimony algorithms.

Theoretical example

In order to outline the evolutionary implications of the proposed way of coding, we have defined complex evolutionary operators, corresponding to evolutionary changes implying one or more steps as defined above in a single event. This model does not imply that natural genomes evolve through these operators: those were designed uniquely to get an estimation of the implications of our approach prior to any analysis of natural data.

The operators, which range from displacement of a single gene to complex translocations, such as inversion of four genes and deletion of two genes in a single evolutionary step, are described in Figure 1: eight different artificial genomes are only connected by unoriented rearrangements implementing each operator and applied onto clusters of genes, supposedly representing single step evolutionary changes. We retained four usual operators: inversion, transposition, inverted transposition (transinversion) and gain/loss (indel). Inversion shifts altogether the transcription strand of a cluster of genes within the same relative position: all the inverted genes still share the same relative position with one another. Transposition alters the position of a cluster of genes, but the relative positions remain the same within the transposed block of genes. Transinversion represents a transposition along with an inversion. The gain/loss inserts or removes a block of genes from the genome. These operators are sufficient to explain any genomic rearrangement apart from translocations. Sankoff *et al.,* (1992) used inversion and transposition to calculate edit distances between mitochondrial genomes.

We have addressed the implications of our method at two distinct levels, first before performing an analysis *i.e. a priori* implications and later by using *a posteriori* implications.

A priori Implications

Due to our coding characteristics, rearrangements implementing our operators do not have the same cost in terms of number of steps. The number of steps varies with the operator and with the number of genes involved. The costs also vary, for transposition and transinversion, with the distance, *i.e.* with the number of genes located between the points of departure and of arrival. We thus varied the number of genes involved in rearrangements and distances.

Inversion

Let us first consider an inversion applied to a single gene. Taxon 4 and taxon 6 are connected by an inversion of gene I. The relative position of gene I is the same in both taxa, gene I is surrounded by H and J. The inversion of a single gene does not change its position whereas its polarity would do. Thus the cost of the inversion of a single gene is equivalent to a single step.

On the other hand, taxon 6 and taxon 7 are connected by an inversion of several genes: genes G H and I. The relative position of gene H is the same in both taxa, gene H is surrounded by I and G. But the relative positions of genes F G I and J are not the same. The relative position of gene F in taxon 6 is E/G and E/I in taxon 7. In the same way, the relative position of gene G

in taxon 6 is F/H and H/J in taxon 7. The transcription strand of G H and I genes changes. Thus the inversion of three genes costs 7 steps (4 of position and 3 of strand).

The cost of an inversion can be computed by using the following equation:

Number of steps = n + [(0, if n = 1) or (4, if n ≥ 2)]. n = number of genes inverted.

Transposition

In the case of a transposition event, the transcriptional orientation does not change, only the relative of the genes do. The costs are different whether the transposed element moves on a "one gene" distance or on a longer distance.

Distance of one gene:

- Taxon 1 and taxon 4 are connected by a transposition of gene C. It costs 4 steps for such a transposition, because the relative positions of genes A B C and D have shifted.
- A transposition of genes H and I links taxon 1 and taxon 4, this rearrangement costs 5 steps, because the relative positions of genes F G H I and J have shifted.

Distance of several genes:

- Taxon 3 and taxon 5 are connected by a transposition of gene E, which costs 5 steps, because the relative positions of genes D A B E and C have shifted.

- A transposition of genes A B and C connects taxon 5 and taxon 6. This transposition costs 6 steps, because the relative positions of genes J D E A C and F have shifted but the relative position of gene B remains the same.

The cost of a transposition can be found by using the following formula:
Number of steps = 2 + min{2, n} + min{2, distance of transposition}. n = number of genes transposed.

Transinversion

The transinversion flips the polarity of the involved genes and moves them through the genome.

Distance of one gene:

- Taxon 1 and taxon 2 are connected by a transinversion of gene J. This transinversion costs 5 steps: one for the flip of gene J and four for the shift of the relative positions of genes G J A and C.
- Taxon 3 and taxon 5 are connected by the transinversion of genes H and I. The cost of this rearrangement is 6 steps: two for the flip of genes H and J and four for the shift of the relative positions of genes F I G and J. Despite gene H is involved in the transinversion, its relative position is identical in both taxa.

Distance of several genes:

- Taxon 7 and taxon 8 are connected by a transinversion of gene K. It costs 6 steps: one for the flip of gene K and five for the shift of the relative positions of genes L B K C and E.
- Taxon 7 and taxon 8 are connected by a transinversion of genes G and J. This rearrangement costs 8 steps: two for the flip of genes G and J and six for the shift of the relative positions of genes F I H G J and A.

The cost of a transinversion can be found by using the following equation:

Number of steps = n + 2 + [(2, if distance of transposition = 1) or (min{2,n} + 2, if distance of transposition ≥ 2)]. n = number of transinverted genes.

Gain/Loss

- Taxon 6 and taxon 7 are connected by an indel of gene D which costs 3 steps because the relative position of genes C and E has changed and the presence/absence of gene D also has been altered.
- Taxon 6 and taxon 7 are connected by an indel of genes K and L. It costs 4 steps because the relative position of genes A and B has changed along with the presence/absence of genes K and L.

The cost of a gain/loss can be calculated by using the following equation:

Number of steps = n + 2. n = number of genes implicated.

Table 2 summarizes the costs of every operator applied on different numbers of genes.

	Inversion	Transposition /1	Transposition /P	Transinversion /1	Transinversion /P	Indel
1 gene	**1**	**4**	**5**	**5**	**6**	**3**
2 genes	**6**	**5**	**6**	**6**	**8**	**4**
3 genes	**7**	**5**	**6**	**7**	**9**	**5**
4 genes	**8**	**5**	**6**	**8**	**10**	**6**
5 genes	**9**	**5**	**6**	**9**	**11**	**7**

Table 2

The *a priori* cost of the twelve rearrangement events, each taken independently, is 60 steps.

A posteriori Implications

Analysis

We have then encoded the eight artificial taxa shown in Figure 1 with each gene being coded using the characters defined above. This resulted in a matrix of 8 taxa and 27 characters (12 for position, 12 for strand and 3 for presence/absence, as shown in Appendix 1), (7 characters are constant, 7 uninformative and 13 informative). We then computed the matrix in PAUP (Swofford, 1993) with an exhaustive search, yielding a single parsimonious tree, Figure 2 (unrooted). This tree is 53 steps long, CI = 0,943, RI = 0,823.

This topology is not relevant to any historical relationship among taxa but it was of interest to explore the characters behaviour, and to estimate the consequences of the way of coding upon the result. For that purpose we tried to reconstitute hypothetical common ancestors genomes out of which the resulting tree has been generated.

Hypothetical Common Ancestor

It is of common acceptation that hypothetical common ancestors need not to be viable since they are merely a set of character states sampled out all of the characters composing an organism. We are interested however in getting viable genomes at internal nodes in order to reconstruct rearrangements along the tree. Each internal node is composed of a set of unambiguous and ambiguous synapomorphies. Ambiguous synapomorphies cannot be assigned specifically to a node but to a series of nodes. Unambiguous synapomorphies are assigned to a precise node (not meaning that they are unique on the tree, but only that they are fixed on that node). The distribution of ambiguous synapomorphies on the tree depends on the chosen optimisation procedure with many possible ways of optimisation (Farris, 1972; Swofford and Maddison, 1987). At each node, we retained among the possible optimisations, only those that allow the reconstitution of complete genomes. In order to reconstruct a genome we need to assign a position to every present gene, the genes with unambiguous position allowing the choice of the position of the others. Because the position of a gene is a couple of two other genes, the position involves the two adjacent genes of the gene under study; we then have to find, among the different states available for genes with an ambiguous

position, the state compatible with the positions of fixed genes. The genome is reconstructed step-by-step setting the position to a gene and then to its neighbours and so on. There can be states, of different characters, incompatible with one another driving to the inability of reconstructing an entire genome.

In the frame of that theoretical example, we have been able to reconstitute at least one genome per node and even two genomes at three different nodes giving four different optimisations. Considering that there are three nodes with two possible genomes reconstituted each, it yielded eight combinations but only four were compatible with the length of the most parsimonious tree (53 steps), the other four had a length higher than 53 steps and therefore were not retained. The Figure 3 presents the first optimisation, showing at each internal node a hypothetical common ancestor (HCA) with a completely reconstituted genome.

It appears thus feasible to reconstruct rearrangements along this tree; out of the twelve rearrangements inferred *a priori* (see Figure 1) five of them can be inferred as single events whereas the other seven are split into several events. Furthermore among these twelve rearrangements, eight cost the same number of steps as implicated by the coding (Table 1) but four rearrangements cost more on the tree (Table 3).

- Rearrangements retrieved as a single event:

Rearrangement	**Cost *a priori***	**Cost on the tree**
taxon1 <-> taxon2 (transinversion of gene J).	5 steps	same
taxon1 <-> taxon4 (transposition of gene C).	4 steps	same

taxon4 <-> taxon6 (inversion of gene I).	1 step	same
taxon5 <-> taxon6 (transposition of genes A B C).	6 steps	same (see Figure 3)
taxon7 <-> taxon8 (transposition of gene K).	6 steps	same

• Rearrangements retrieved as several events:

Rearrangement	**Cost *a priori***	**Cost on the tree**	**Number of events *a posteriori***
taxon1 <-> taxon4 (transposition of genes I H).	5 steps	8 steps	3
taxon3 <-> taxon5 (transinversion of genes I H).	6 steps	same	4 (see text and Figure 3)
taxon3 <-> taxon5 (transposition of gene E).	5 steps	17 steps	4
taxon6 <-> taxon7 (inversion of genes I H G).	7 steps	same	5
taxon7 <-> taxon8 (transinversion of genes G J).	8 steps	same	3
taxon6 <-> taxon7 (indel of gene D).	3 steps	5 steps	2 (see text)
taxon6 <-> taxon7 (indel of genes K L).	4 steps	6 steps	2

Table 3

More than half the rearrangements inferred *a priori* between pairs of taxa, are not retrieved *a posteriori* on the resulting tree. For example the transinversion of genes I H inferred *a priori* as a single rearrangement between taxon3 and taxon5, is divided into four events and implies 6 steps:

- inversion of I between HCA (Hypothetical Common Ancestor) 11 and 12
- transposition of H between HCA 12 and 13
- transposition of I between HCA 12 and 13
- inversion of H between HCA 13 and 10

These rearrangements are not retrieved because some of the pairs of taxa they link are not closely related on the tree and the path linking the two taxa entails character transformations which are spread over branches, thus implying various events. Since these events are involved into several of the rearrangements inferred *a priori*, the sum of *a posteriori* rearrangements costs (Table 3) is larger than the length of the tree. The indels are not retrieved as single events since the genes surrounding the gene involved in the indels are implicated in other events along the branches linking taxa 6 and 7. For example gene C that is close to gene D in taxon 6 is transposed twice, once between HCA 12 and 13 and once between HCA 10 and 9. It is then, not only the presence/absence of gene D that changes but also its relative position since one of its neighbours is shifted (gene C is replaced by gene B). Therefore, the indel of gene D between taxa 6 and 7 costs *a posterori* more than the indel of a single gene.

Discussion

The results of analyses conducted with “relative position” method are free of phylogenetic beforehand assumptions but not free of an evolutionary model. The relative position coding implicates that each gene can independently move inside the genome as long as it can switch its polarity or can be inserted or deleted from the genome. This model is not really compliant with a “rearrangement upon large blocks of genes” model, specially for inner branches, indeed as each gene can move separately it is rather unlikely to get a rearrangement involving several genes upon the resulting tree. In other respects, this model of independent changes conforms to the model of duplication/random loss (Macey *et al.*, 1997; Macey *et al.*, 1998). The model of duplication/random loss suggests that a duplication implying several genes can be followed by a series of random loss, deleting without distinction the original gene or the copy, leading to a novel gene order. This kind of rearrangement can be easily retrieved upon a tree obtained with the relative position coding.

The example of gene D indel outlines a character linkage between “gene relative position” characters. Indeed, as the position of a gene is represented by its two neighbours, this position is linked to the position of the other two genes. As discussed above, the consequence of this linkage is the variable costs of the different evolutionary events. Character linkage becomes problematic when it induces a bias in the phylogenetic results. In the case of gene order, character linkage is inescapable; as a matter of fact localizing a gene inside an order without referring to its neighbourhood can only be done using ranks. Thus, gene positions could be

represented by gene ranks in the genomes. But this method as severe drawbacks: first, an origin to measure ranks is usually missing, second, primary homology based on ranks is not acceptable (the movement of a single gene can change all the ranks). Linked characters tend to group taxa even if they are not closely related, acting as an overly weighted character. In the present example, this problem is avoided, since all the genes are coded in the very same way, giving to all “gene relative position” characters the same weight and thus neglecting the risk of wrongly group unrelated taxa. Furthermore the relative position coded as multistate character allows reconstituting hypothetical common ancestors genome at each node, permitting an *a posteriori* character reinterpretation. This reinterpretation of characters allows inferring evolutionary events consistent with parsimony criteria used to reconstruct the tree.

References

Barriel, V. and Tassy, P. 1993. Characters, observations and steps: comment on Lipscomb's "parsimony, homology and the analysis of multistate characters". *Cladistics* **9,** 223-232.

Blanchette, M., Kunisawa, T. and Sankoff, D. 1996. Parametric genome rearrangement. *Gene* **172,** GC 11-17.

Blanchette, M., Kunisawa, T. and Sankoff, D. 1999. Gene Order Breakpoint Evidence in Animal Mitochondrial Phylogeny. *J. Mol. Evol.* **49,** 193-203.

Boore, J. L., Collins, T. M., Stanton, D. J., Daehler, L. L. and Brown, W. M. 1995. Deducing the pattern of arthropod phylogeny from mitochondrial DNA rearrangements. *Nature* **376,** 163-165.

Boore, J. L. and Brown, W. M. 1998. Big trees from little genomes: mitochondrial gene order as a phylogenetic tool. *Curr. Opin. Genet. Dev.* **8,** 668-674.

Boore, J. L., Lavrov, D. V. and Brown, W. M. 1998. Gene translocation links insects and crustaceans. *Nature* **392,** 667-668.

Bridge, D., Cunningham, C. W., Schierwater, B., DeSalle, R. and Buss, L. W. 1992. Class-level relationships in the phylum Cnidaria: evidence from mitochondrial genome structure. *Proc. Natl. Acad. Sci. U S A* **89,** 8750-8753.

Cosner, M. E., Jansen, R. K., Moret, B. M. E., Raubeson, L. A., Wang, L.-S., Warnow, T. and Wyman, S. 2000a. An Empirical Comparison of Phylogenetic Methods on Chloroplast Gene Order Data in Campanulaceae. *In* D. Sankoff and J. H. Nadeau (eds), Comparative Genomics. Empirical and Analytical Approaches to Gene Order

Dynamics, Map Alignment and the Evolution of Gene Families. Kluwer Academic, Boston, MA, pp. 99-121.

Cosner, M. E., Jansen, R. K., Moret, B. M. E., Raubeson, L. A., Wang, L.-S., Warnow, T. and Wyman, S. 2000b. A New Fast Heuristic for Computing the Breakpoint Phylogeny and Experimental Phylogenetic Analyses of Real and Synthetic Data *In* 8th International Conference on Intelligent Systems for Molecular Biology, San Diego.

de Pinna, M. C. C. 1991. Concepts and tests of homology in the cladistic paradigm. *Cladistics* **7,** 367-394.

Farris, J. S. 1972. Estimating phylogenetic trees from distance matrices. *Am. Nat.* **106,** 645-668.

Gallut, C., Barriel, V. and Vignes-Lebbe, R. 2000. Gene Order and Phylogenetic Information. *In* D. Sankoff and J. H. Nadeau (eds), Comparative Genomics. Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and the Evolution of Gene Families. Kluwer Academic, Boston, MA, pp. 123-132.

Macey, J. R., Larson, A., Ananjeva, N. B., Fang, Z. and Papenfuss, T. J. 1997. Two novel gene orders and the role of light-strand replication in rearrangement of the vertebrate mitochondrial genome. *Mol. Biol. Evol.* **14,** 91-104.

Macey, J. R., Schulte, J. A., Larson, A. and Papenfuss, T. J. 1998. Tandem duplication via light-strand synthesis may provide a precursor for mitochondrial genomic rearrangement. *Mol. Biol. Evol.* **15,** 71-75.

Pleijel, F. 1995. On character coding for phylogeny reconstruction. *Cladistics* **11,** 309-315.

Sankoff, D., Leduc, G., Antoine, N., Paquin, B., Lang, B. F. and Cedergren, R. J. 1992. Gene order comparisons for phylogenetic inference: evolution of the mitochondrial genome. *Proc. Natl. Acad. Sci. U S A* **89,** 6575-6579.

Sankoff, D. and Blanchette, M. 1998. Multiple genome rearrangement and breakpoint phylogeny. *J. Comput. Biol.* **5,** 555-570.

Sankoff, D., Deneault, M., Bryant, D., Lemieux, C. and Turmel, M. 2000. Chloroplast Gene Order and the Divergence of Plants and Algae, From The Normalized Number of Induced Breakpoints. *In* D. Sankoff and J. H. Nadeau (eds), Comparative Genomics. Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and the Evolution of Gene Families. Kluwer Academic, Boston, MA, pp. 89-98.

Smith, M. J., Arndt, A., Gorski, S. and Fajber, E. 1993. The phylogeny of echinoderm classes based on mitochondrial gene arrangements. *J. Mol. Evol.* **36,** 545-554.

Swofford, D. L. and Maddison, W. P. 1987. Reconstructing ancestral character states under Wagner parsimony. *Math. Biosci.* **87,** 199-229.

Swofford, D. L. 1993. PAUP: Phylogenetic analysis Using Parsimony. Ver. 3.1.1. Computer program distributed by the Illinois Natural History Survey, Champaign.

Appendix 1

```
#NEXUS

BEGIN DATA;
DIMENSIONS  NTAX=8 NCHAR=27;
FORMAT MISSING=? SYMBOLS= "0 1 2 3 4 5 6";

STATELABELS
1     [Pos_A]     'C\J' 'G\J' 'B\D' 'B\J' 'B\E' 'J\L' 'H\L',
2     [Pos_B]     'C\D' 'A\E' 'A\C' 'C\K' 'C\L',
3     [Pos_C]     'A\B' 'B\J' 'E\F' 'B\D' 'B\F' 'B\E' 'B\K',
4     [Pos_D]     'B\E' 'A\J' 'C\E' 'E\J',
5     [Pos_E]     'D\F' 'B\C' 'A\D' 'C\F' 'F\K',
6     [Pos_F]     'E\H' 'C\I' 'E\G' 'C\G' 'E\I' 'E\J',
7     [Pos_G]     'I\J' 'A\I' 'H\J' 'F\H',
8     [Pos_H]     'F\I' 'G\I' 'A\I',
9     [Pos_I]     'G\H' 'F\H' 'H\J',
10    [Pos_J]     'A\G' 'A\C' 'D\G' 'A\I' 'D\I' 'F\G',
11    [Pos_K]     'B\L' 'C\E',
12    [Pos_L]     'A\K' 'A\B',
13    [Ss_A]      '-' '+',
14    [Ss_B]      '-' '+',
15    [Ss_C]      '-' '+',
16    [Ss_D]      '-' '+',
17    [Ss_E]      '-' '+',
18    [Ss_F]      '-' '+',
19    [Ss_G]      '-' '+',
20    [Ss_H]      '-' '+',
21    [Ss_I]      '-' '+',
22    [Ss_J]      '-' '+',
23    [Ss_K]      '-' '+',
24    [Ss_L]      '-' '+',
25    [PA_D]      'Absent' 'Present',
26    [PA_K]      'Absent' 'Present',
27    [PA_L]      'Absent' 'Present',
      ;
MATRIX

Taxon1      00000000000?? 1100011001?? 100
Taxon2      10100010010?? 1100011000?? 100
Taxon3      21211121120?? 1100011101?? 100
Taxon4      32320231230?? 1100011001?? 100
Taxon5      42432331240?? 1100011011?? 100
Taxon6      32320231230?? 1100011011?? 100
Taxon7      5350?34211000 1100?01010111 011
Taxon8      6460?45020511 1100?01110001 011
      ;

ENDBLOCK;

BEGIN ASSUMPTIONS;
      CHARSET Position  =     1-12;
      CHARSET Orientation     =     13-24;
      CHARSET PreAbs    =     25-27;
ENDBLOCK;
```

Figure 1

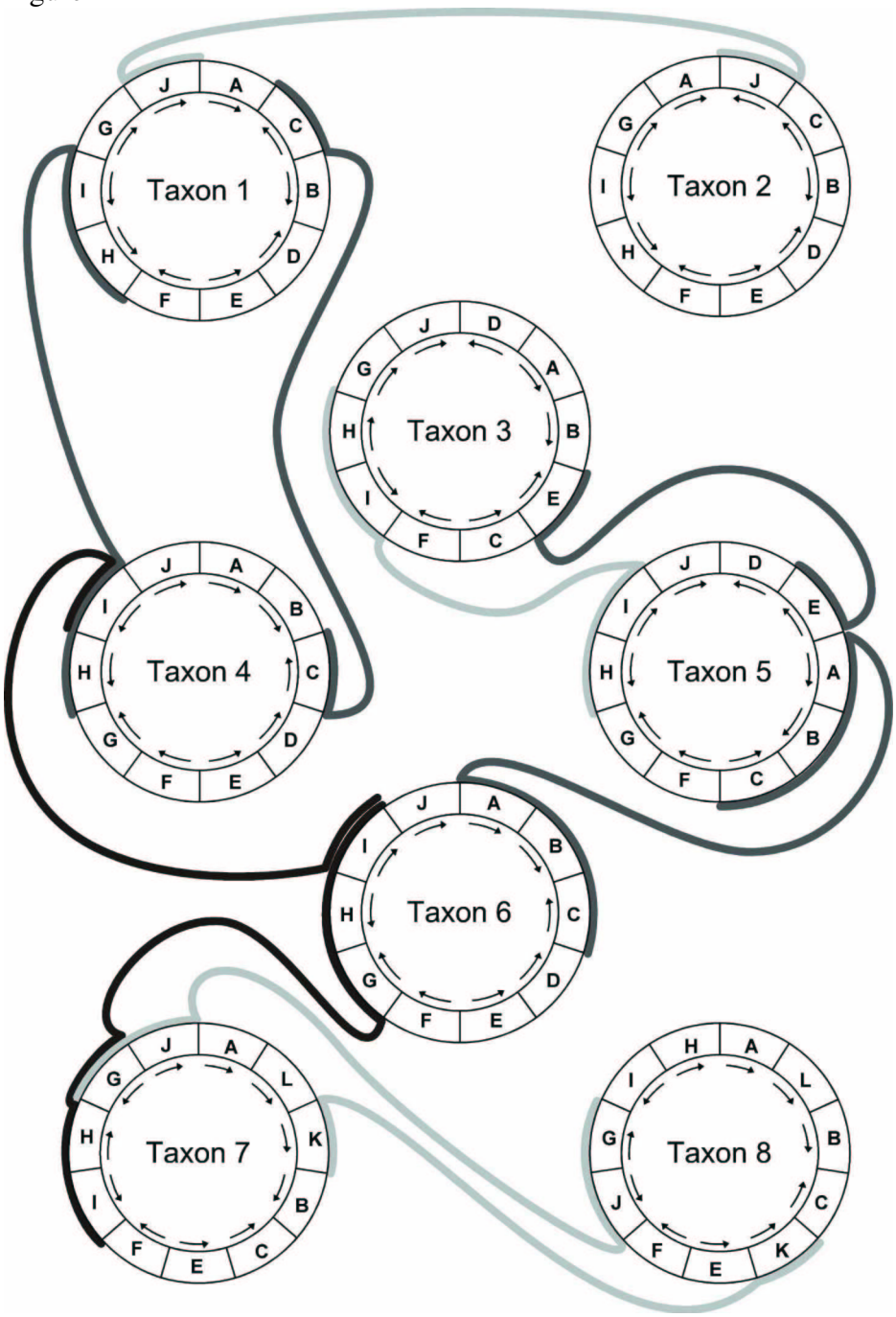


Figure 2

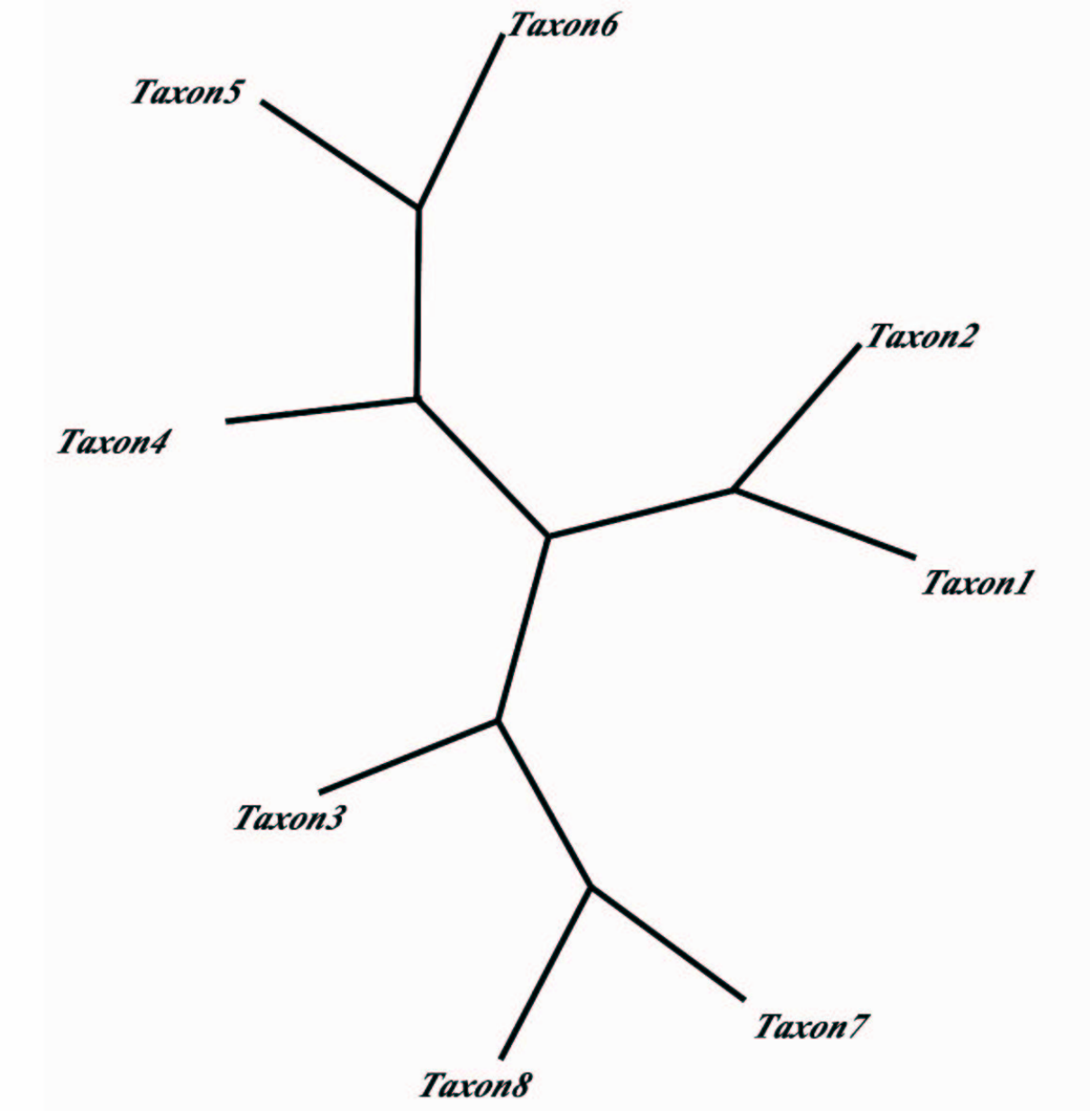
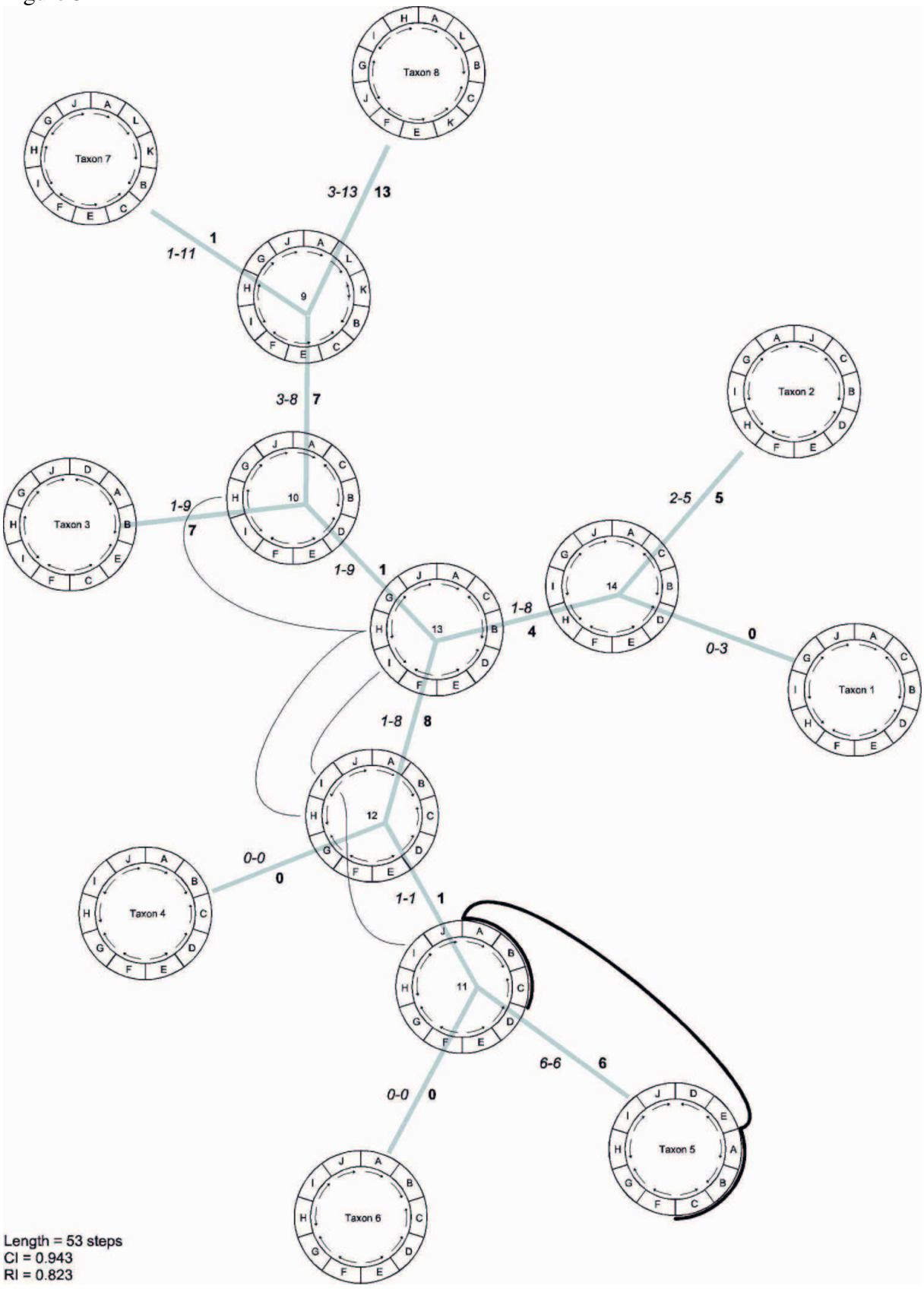


Figure 3


Taxon 8
Taxon 7
3-13
13
1
1-11
9
3-8
7
Taxon 2
Taxon 3
1-9
7
10
2-5
5
1-9
1
14
13
1-8
4
0
0-3
Taxon 1
1-8
8
12
0-0
0
Taxon 4
1-1
1
11
6-6
6
0-0
0
Taxon 5
Taxon 6
Length = 53 steps
CI = 0.943
RI = 0.823

The Complete Nucleotide Sequence of the Mitochondrial DNA of the Dogfish, *Scyliorhinus canicula*

Christiane Delarbre,*[,1] Nathalie Spruyt,[†,1] Cateline Delmarre,[†] Cyril Gallut,[‡] Véronique Barriel,[‡] Philippe Janvier,[§] Vincent Laudet[†] and Gabriel Gachelin*

**Département d'Immunologie, Institut Pasteur, 75015 Paris, France, [†]Institut de Biologie de Lille, Institut Pasteur de Lille, 59021 Lille, France, [‡]Laboratoire de Systématique Moléculaire, Muséum National d'Histoire Naturelle, 75005 Paris, France and [§]Laboratoire de Paléontologie, Muséum National d'Histoire Naturelle, 75005 Paris, France*

ABSTRACT

We have determined the complete nucleotide sequence of the mitochondrial DNA (mtDNA) of the dogfish, *Scyliorhinus canicula*. The 16,697-bp-long mtDNA possesses a gene organization identical to that of the Osteichthyes, but different from that of the sea lamprey *Petromyzon marinus*. The main features of the mtDNA of osteichthyans were thus established in the common ancestor to chondrichthyans and osteichthyans. The phylogenetic analysis confirms that the Chondrichthyes are the sister group of the Osteichthyes.

THE mitochondrial DNA (mtDNA) of most animals is a self-replicating, 15- to 17-kb-long, circular DNA molecule. Animal mtDNA codes for 13 mitochondrial proteins, 22 mitochondrial tRNAs, and two mitochondrial-specific ribosomal RNAs, the 12S and 16S rRNAs. It also contains DNA regions aimed at controlling its replication and transcription. From the evolutionary viewpoint, mtDNAs are "small genomes" that coevolve at their own rate with the organism in which they are lodged. Thus, mtDNA sequences are widely used to construct phylogenetic trees. The genomic organization of the mtDNA (*i.e.*, the relative location of the genes and the structure of the control regions of the mitochondrial genome) has also evolved, and the comparison of the genetic maps of mtDNA has been used to describe some key points in the evolution of animals [*e.g.*, the myxine/lamprey/dogfish relationships (Delarbre *et al.* 1997) and the complex evolution of snakes (Kumazawa and Nishida 1995)]. However, although a number of complete nucleotide sequences of mtDNA have been published, the relative lack of data hinders efficient statistical analysis. Concerning the evolution of fish, the sequence of mtDNA of the sea lamprey [*Petromyzon marinus* (Lee and Kocher 1995)] is known and so are several sequences of mtDNA of the Osteichthyes (Tzeng *et al.* 1992; Chang *et al.* 1994; Zardoya *et al.* 1995; Noack *et al.* 1996; Zardoya and Meyer 1996, 1997). No complete sequence of the mtDNA of any chondrichthyan has been published yet, however, although the mtDNA sequences of several members of this group are needed to examine their relationships with osteichthyans. In this article, we describe the complete nucleotide sequence of a chondrichthyan fish, the dogfish *Scyliorhinus canicula*. The sequence extending from the *tRNA-Leu* gene to the beginning of the *COI* gene of the *S. canicula* mtDNA has already been reported (Delarbre *et al.* 1997). From the comparison of the mtDNA genomic map with those of other animals, as well as a study of the main features of the DNA sequence, the dogfish appears very similar to osteichthyans and significantly different from the sea lamprey.

Corresponding author: Gabriel Gachelin, Unité de Biologie Moléculaire du Gène, Unité INSERM U277, Institut Pasteur, 25 rue du Dr. Roux, 75015 Paris, France. E-mail: ggachel@pasteur.fr

[1] These authors contributed equally to this work.

MATERIALS AND METHODS

Animals: A specimen of *S. canicula* was caught offshore Roscoff (Brittany, France). The animal was anesthesized, killed, and dissected. The organs were immediately frozen in liquid nitrogen and stored at −78°.

Preparation of DNA: Total DNA was prepared from dogfish muscles by proteinase K digestion according to conventional procedures (Hogan *et al.* 1994).

Isolation and sequencing of mtDNA: Overlapping fragments of mtDNA were obtained by PCR that was run on total genomic DNA using degenerate primers. The sequences of the primers used for PCR amplification are given in Table 1. Two strategies for obtaining PCR-amplified DNA were used.

For PCR using the 1–18 primers (*i.e.*, from the *16S rRNA* to *ND4*), the conditions were the following: 300 ng total DNA; the enzyme was the Pfu Exo+ polymerase (Stratagene, La Jolla, CA); and the thermal cycles were 3 min at 94°, followed by 50 cycles 1 min at 94°, 1 min at 48–55°, depending upon the primers, and 5 min at 72°. The last cycle was ended by incubating for 10 min at 72°. These PCR products were purified by electroelution, phosphorylated using the polynucleotide kinase (Pharmacia Fine Chemicals, Piscataway, NJ), and ligated at the *Eco*RV site of dephosphorylated KS Bluescript vector (Stratagene) using the Rapid DNA Ligation kit (Boehringer Mannheim, Indianapolis, IN). XL1Blue competent bacteria (Stratagene) were transformed with the ligation products. Several recombinant clones were selected for each PCR product, and plasmid DNA was recovered using the Clearcut Miniprep kit (Stratagene). The cloned PCR products were sequenced first using the M13 (−40) and KS reverse primers and subsequently using primers derived from the sequence

TABLE 1

PCR and sequencing primers used in the determination of the nucleotide sequence of the dogfish mitochondrial genome

PCR primers			
	Name	Sequence (5′–3′)	Location[a]
1.	LEU	TGCRMAAGRYYTAAGCYCTT	16,646 to 922
2.	NAD1SCY	GAATAATTGCTAAGGTTAATGGTAG	
3.	NAD1	CCACGATTCCGATATGATCAACT	849 to 2650
4.	COI	CCRATRTCYTTGTGWTTAGTAGA	
5.	COI5S	CTCTCAGCCATCTTACCTGTGGC	2608 to 3963
6.	COI-3	GCATCTGGGTARTCKGAGTADCGKCG	
7.	R9	CTTCCCCCAACATTTCTTAGGTC	3901 to 4991
8.	R10	CGAAGTGTTCAAGTGGGAC	
9.	COII	TGTGGCGCAAAYCAYAGYTTTATRCC	4931 to 6553
10.	COIII-3	CGTGAAAGCCKGTGGCKACAAARAAKG	
11.	R11	CGAAGCACCATTCACCATCGC	6486 to 7127
12.	ND33	CCTTGTRKTCATTCATARATTAGKCC	
13.	R18	CCATTCTCTACACTACTTTGAAC	7051 to 8182
14.	NAD43	GGGRGCYTCTACRTGRGCTTTDGG	
15.	16S5	ATYAAAYCTYGTACCTTTTGCATCAT	15,141 to 16,535
16.	16S3	CTAYAYTTTATYTKTCCTTTCGTACTA	
17.	16S	CGTGATCTGAGTTCAGACCGG	16,448 to 57
18.	R3	GATAGGGATAATGTAGGCGAGAG	
VL1		GAACGCCTCAATGCAGGTACTTATTTTTTA	7931 to 15,259
VL2		CGGGTTAGCTCTATAATGCTGCTTCGGAGT	

Sequencing primers		
Location	Name	Sequence (5′–3′)
ND1	NAD5S	AGCCCTTTTAATATGAATACCAC
ND2	NAD4 (reverse)	GTAATTATTCAGCCTAGGTTAGCG
id	NAD5	TGGGGGGGCCTAAATCAAACTCA
id	NAD6 (reverse)	CAATACTTCCGGGAGTCAAAAGTG
id	NAD7	GGAGGCCTACCTCCCCTATCAG
COI	R5	CACCAGATATAGCCTTCCCTC
id	R6 (reverse)	CAATATCTAAAGAAGAGTTGG
id	R14	CCCTCCCTGTCCTTGCAGCC
id	R15	GAAGAACCAGCATTTGTACAAG
ATP8	R7	CACCTATTTAATAATGAACCTG
ATP6	R13	TTTAGGAGGACATAAATGAGC
COIII	R8 (reverse)	GCCCATGGTGGGCTCAAGTC
id	R12	GAATTATGGTTAGGAGAAGTAGG
ND3	R16 (reverse)	GCGATTAGGAAAAATCGTATGG
ND4	R17 (reverse)	GGAGGGATTAGGAAGGTGGG
id	R19 (reverse)	GCTGGCTAGAATTATTAGTGGG
id	M3673 (reverse)	TACATCTTTGACTACCAAAAGC
id	M3672	AAAATCCATAATCGCTTACT
id	M3671 (reverse)	AATCCTAATCCAAACAC
id	M3726	TCACCCTCCATGTCCTACCT
tRNA-His	M3725 (reverse)	ATTCTGTATTTATAGTTTAA
tRNA-Leu	M3841 (reverse)	ATCCTTGGTGCAACTCCAAG
ND5	M3801	TAGCCTGATTCCACTATTTA
id	M3799	ATTGTGACTAATTGAAATTG
id	M3840 (reverse)	CGAATCGGCGATATCGGGTT
id	M3822	TGATTGCCATAAACCTCAA
id	M3843 (reverse)	AACCCTCTTCACAGCAACAT
id	M3842	ATTGGCTTAAATCAACCCCA
id	M3901 (reverse)	ACGCCTGAGCCCTAATCTTA
id	M3887	AACGAAAATAACCCCATAG
id	M3940 (reverse)	CCTAACACACTCACAATTCA

(*continued*)

TABLE 1

(Continued)

Sequencing primers		
Location	Name	Sequence (5′–3′)
id	M4001 (reverse)	ACTTGGATATTTTCCCCCAA
id	M3917	ATTAATTAAATTATCCACCC
id	M3939	AACACTTACATTAATTATCC
ND6	M3951	CTGCCGTATAAGCAAAAACT
id	M3999 (reverse)	AACTAATAAACCACATCCCA
CYTB	M4000	TCTCCCAGCCCCATCAAATA
id	M3998 (reverse)	TTAATTATACAAATTATTAC
id	M4100 (reverse)	CAAAGAAGCATGAAACATCG
id	M4085 (reverse)	TTATTCTTACTTATAGCTAC
id	M4045	CTACAGCCTTCGTAGGCTAT
id	M4083 (reverse)	TAAAGACTTATTTGGCTTCT
id	M4084	ACTTACTAGGGGATGCTGAA
id	M4146	TACACATTCAACCTGAATGA
id	M4099 (reverse)	TAGTGCCACTCCTTCACACC
id	M4103 (reverse)	AGGGCAACCAGTAGAACAAC
id	M4145	CTCCTTATTCCTTTTCGTGA
tRNA-Pro	M4088	TAAACTCCTGCCTTTGGCTC
C.R.	M4098 (reverse)	TCTGAAGGCTGTTACTGAGC
id	M4102	ATAACACTCTGGTTTTTAGC
id	M4089 (reverse)	TATATGACATGGCCCACATA
id	M4144	TGTATACATAATACATTCAT
id	M4101	ATTTATGCGGGCTGGATAGA
id	M3930 (reverse)	AACCGGTTTATCCCTATTAA
id	M3991 (reverse)	TTATTTTCCTCCCAGTTTTT
id	M4082	TTATTTTCCTCCCAGTTTTT
id	M3916 (reverse)	GGGATGAAGCAAATCGCTAT
id	M3997 (reverse)	AATCTCCTCATTACTTTTCA
id	M4046 (reverse)	TAGACGTATCAGTATAGTAA
id	M4002	GAGACAAATCATTAATCC
id	M3886 (reverse)	TATCTGGCATAGGTTTAATT
id	M3900	ATTAGATATACCCGGTCTTG
12S rRNA	M3839 (reverse)	ATGCAAGTTTCAGCCCCCCT
id	M3838	AGGGGCGGGCATCAGGCACA
id	M3798	CGAAAGTGACTCTAAATTAC
id	M3800 (reverse)	CACGACAGTTGGGCCCCAAA
id	M3821 (reverse)	ATACCCTACTATGCCCTACC
id	M3837	TAAACGTCAGGTCGAGGTGT
id	M3724 (reverse)	CGCACACACCGCCCGTCACT
tRNA-Val	M3836	CTTACACTGAGGAAATATCC
16S rRNA	M3722	TTATGTACCGCAAGGGAAAG
id	M3737 (reverse)	GCAGCATTATAGAGCTAACC
id	R1	AGTTTAGTGATAGCTGGTTAC
id	R2 (reverse)	TGAATACTCGGTCGATAAGG
id	R4 (reverse)	CAAGTGATTACGC-TACCTTTGC

[a] The numbering of the primers is that of Figure 1 of the manuscript.
C.R., control region; IUB code: R, A/G; Y, C/T; M, A/C; W, A/T; D, G/A/T; K, G/T.

determined (Table 1). Thus, primer walking was used throughout, using primers located on both strands. Three hundred base pairs were read using each primer. The entire sequence was determined on both strands. Finally, the 16S rRNA was sequenced twice on both strands starting from two independent clones. All overlapping sequences were found to be identical.

For PCR using the VL1 and VL2 primers, namely for the isolation of the *ND4-16S rRNA* segment of the mtDNA, the conditions were the following: the Gene Amp XL PCR kit from Perkin Elmer-Cetus (Norwalk, CT) was used with the rTth DNA polymerase XL and the Amplivax system. To determine the frequency of errors, three independent amplifications were carried out, and all three products were sequenced. The PCR conditions were as follows. Oligonucleotides (200 ng) were mixed with 100–200 ng DNA/tube. A complex PCR cycle with autoextension was used: 1 min at 93°, followed by 1 min at 93° and 12 min at 68° for 16 cycles; this was followed by 20 cycles, 1 min at 93° and 12 min at 68°, with an autoextension of 15 sec/cycle after the sixth cycle. The last cycle was followed by 15 sec at 72°. The final PCR product was not cloned, but was directly sequenced on both strands by primer walking (300 bp in each walk) using the primers indicated in Table 1. Three independent PCR products were sequenced.

Phylogenetic analyses: All protein-coding sequences of the mtDNA of *S. canicula*, *Branchiostoma lanceolatum* (N. Spruyt, C. Delarbre, G. Gachelin and V. Laudet, unpublished results), *Crossostoma lacustre* (Tzeng *et al.* 1992), *Cyprinus carpio* (Chang *et al.* 1994), *Gallus gallus* (Desjardins and Morais 1990), *Latimeria chalumnae* (Zardoya and Meyer 1997), *Oncorhynchus mykiss* (Zardoya *et al.* 1995), and *P. marinus* (Lee and Kocher 1995) were aligned manually. The alignment was based on the translation of the nucleotide sequences into amino acid sequences using the vertebrates' mitochondrial genetic code, except for the AGA codon in *B. lanceolatum*, which is read as a glycine (Delarbre *et al.* 1997). The alignment was refined using the physical amino acid similarity criterion. All inserted gaps were triplets and were treated as both missing data and a fifth nucleotide (additional state). Overlapping positions exist in the *ATPase8*/*ATPase6*, *ND4L*/*ND4*, and *ND5*/*ND6* genes. They were used either as duplicates (giving these positions double the weight of the others) or as singles. All protein-coding genes of each sample were combined to create the data set.

The maximum parsimony (MP) method was applied to the data set using an exhaustive search implemented in PAUP (version 3.1.1; Swofford 1993) to find the most parsimonious tree. Different ponderation schemes were used: transitions and transversions were given equal weight, or transversions were given double the weight of transitions. The neighbor-joining (NJ) method (Saitou and Nei 1987) was used to reconstruct a tree based on the Kimura two-parameter distance matrix (Kimura 1980; transversions were given double the weight of transitions) implemented in PHYLIP (version 3.5c; Felsenstein 1993). The robustness of the MP and NJ trees were tested using the bootstrap method, with 200 replications each (Felsenstein 1985), implemented in PHYLIP and using the decay index, *i.e.*, Bremer support (Bremer 1988).

A Dayhoff distance matrix was computed using the Pam matrix (Dayhoff 1978) implemented in PHYLIP. The same amino acid alignment described above was used.

RESULTS

Genomic organization of the mtDNA of *S. canicula*: The mtDNA of *S. canicula* (GenBank accession number Y16067) is 16,697 bp long. The 13 protein-coding genes, the 22 tRNA genes, and the two rRNA genes were identified through the zones of sequence similarity with the corresponding genes and encoded proteins of the carp, *C. carpio* (Chang *et al.* 1994).

The gene order of the *S. canicula* mtDNA is depicted in Table 2. The complete nucleotide and deduced amino acid sequences are given in Figure 1.

Irrespective of minor differences in the size of several genes, the organization of *S. canicula* mtDNA was found to be identical to that of the bony fish, the complete mtDNA sequence of which is known [namely *C. carpio* (Chang *et al.* 1994), *Protopterus dolloi* (Zardoya and Meyer 1996), *Polypterus ornatipinnis* (Noack *et al.* 1996), *L. chalumnae* (Zardoya and Meyer 1997), *Gadus morhua* (Johansen and Bakke 1996), *O. mykiss* (Zardoya *et al.* 1995), and *C. lacustre* (Tzeng *et al.* 1992)]. The orientation of the genes also is the same as in the Osteichthyes: all coding sequences are located on the H strand, with the exception of the *ND6* gene and the eight tRNA genes, which are located on the L strand (Table 2).

TABLE 2

Localization of the genes in the mitochondrial genome of the dogfish

Gene	From	To	Size bp	Size aa	Strand
ND1	1	975	975	324	
tRNA-Ile	979	1,048	70		
tRNA-Gln	1,050	1,122	73		L
tRNA-Met	1,123	1,192	70		
ND2	1,193	2,239	1,047	348	
tRNA-Trp	2,239	2,307	69		
tRNA-Ala	2,309	2,377	69		L
tRNA-Asn	2,378	2,450	73		L
L-strand ori	2,448	2,498	51		
tRNA-Cys	2,488	2,554	67		L
tRNA-Tyr	2,556	2,625	70		L
COI	2,627	4,180	1,554	517	
tRNA-Ser(TCN)	4,181	4,251	71		L
tRNA-Asp	4,256	4,325	70		
COII	4,334	5,024	691	230	
tRNA-Lys	5,025	5,098	74		
ATPase 8	5,100	5,267	168	55	
ATPase 6	5,258	5,941	684	227	
COIII	5,941	6,726	786	261	
tRNA-Gly	6,729	6,798	70		
ND3	6,799	7,149	351	116	
tRNA-Arg	7,148	7,217	70		
ND4L	7,218	7,514	297	98	
ND4	7,508	8,888	1,381	460	
tRNA-His	8,889	8,957	69		
tRNA-Ser(AGY)	8,958	9,024	67		
tRNA-Leu(CTN)	9,025	9,096	72		
ND5	9,097	10,926	1,830	609	
ND6	10,923	11,444	522	173	L
tRNA-Glu	11,445	11,514	70		L
CYTB	11,517	12,660	1,144	381	
tRNA-Thr	12,661	12,732	72		
tRNA-Pro	12,733	12,801	69		L
Control Region	12,802	13,851	1,050		
tRNA-Phe	13,852	13,920	69		
12S rRNA	13,921	14,877	957		
tRNA-Val	14,878	14,949	72		
16S rRNA	14,950	16,622	1,673		
tRNA-Leu(TTR)	16,623	16,697	73		

ATPase 6 and 8, ATP synthase subunits 6 and 8; COI, II, and III, cytochrome C oxidase subunits I, II, and III; CYTB, cytochrome B; ND1–6, NADH dehydrogenase subunits 1–6.

By contrast, three main features make the gene organization of *S. canicula* different from that of the agnathan *P. marinus* (Lee and Kocher 1995):

1. The 1050-bp-long control region of *S. canicula* (about as long as in the carp, 928 bp; the lungfish, 1184 bp; and the trout, 1000 bp) is located between *tRNA-Pro* and *-Phe* genes, whereas the structurally different 675-bp-long control region of *P. marinus* (Lee and Kocher 1995) is split into two regions, NC1 and NC2, respectively, by the insertion of two tRNA genes

(*Thr* and *Glu*). It is located between the *ND6* and *CYTb* genes.

2. As already described (Delarbre *et al.* 1997), the origin of replication of the L strand of the mtDNA is localized between the *tRNA-Cys* and *tRNA-Asn* genes of *S. canicula*, but could not be identified in *P. marinus*.
3. The *tRNA-Thr* gene is located 3′ to the *CYTb* gene in *S. canicula*, whereas in *P. marinus*, it is located in the middle of the control region, adjacent to the *tRNA-Glu* gene. The *tRNA-Glu* of *S. canicula* lies between the *ND6* and the *CYTb* genes.

Salient features of the control region: Several structural elements were evidenced in the 1050-bp-long control region; they are identified in Figure 1. These structural elements were two hairpin-like regions (hairpins 1 and 2), a 10-bp-long perfect palindrome (12893–12902), two 23-bp-long repeats (repeats 1 and 2), and three putative conserved sequence blocks [CSB I, II, and III, (Walberg and Clayton 1981)]. Several putative termination-associated sequences (TAS; Doda *et al.* 1981) were identified. Among them, the sequence ACATTAAACTAAT (13433–13446) is nearly identical to mouse consensus TAS and TAS-4 of the bichir, and the sequence ACATACTATG (12999–13008) is nearly identical to TAS-3 of the bichir. The AT-rich sequence of *P. marinus* [non-coding region 2 (Lee and Kocher 1995)] located between *tRNA-Glu* and *CYTb* was not identifiable in either *S. canicula* or the Osteichthyes. The presence of the origin of replication of the L strand in *S. canicula* had already been reported (Delarbre *et al.* 1997). The presence / absence of these control elements makes the organization of the control region of *S. canicula* very similar to that of the Osteichthyes (Johansen and Bakke 1996).

Lack of distinctive traits of the *S. canicula* tRNAs: The 22 tRNA genes were identified on the basis of their location and sequence. They were from 67 to 75 bp long. All of them, except for *tRNA-Ser* (*AGY*), were able to fold into a cloverleaf secondary structure. There is no feature that would make them different from the tRNAs of osteichthyans. As shown in Table 2, most of the tRNA genes are located on the H strand, with the exception of the *tRNA-Glu*, *-Pro*, *-Gln*, *-Ala*, *-Asn*, *-Ser (TCN)*, *-Cys*, and *-Tyr*, which are on the L strand.

Distinctive features of the protein-coding sequences and of the encoded proteins: The complete nucleotide sequence of the mtDNA of *S. canicula* is depicted in Figure 1, as well as the translation of the open reading frames using the vertebrate mitochondrial code and the limits of the genes.

The codon usage was determined (Table 3). We have compared the codon usage of *S. canicula* with that of *P. marinus* (Lee and Kocher 1995) and *C. carpio* (Chang *et al.* 1994). There are no major differences between the three codon usages. In all three animals, the TTG and GTG codons were predominantly used in ND6. Some codons preferentially used by *S. canicula* are also preferentially used by *P. marinus*: for example, the codon for Phe is predominantly TTT in *P. marinus* and *S. canicula*, whereas it is TTC for *C. carpio*. Conversely, among the codons coding for Ser, AGC is preferentially used in *C. carpio* and *S. canicula*, whereas AGT is predominant for Petromyzon. The codon TCG for Ser is very rarely used in *S. canicula*. Overall, the minor preferences shown by the codon usage of *S. canicula* are rather on the "Petromyzon side." The initiation codons were ATG for all genes, with the exception of GTG for the *COI* gene, as in most chordates studied so far. The stop codon for the *ND6* gene was TAG. All other stop codons were TAA. The TAA stop codons of COII, ND4, and cytochrome b mRNAs should have been created by direct polyadenylation on a terminal T (Ojala *et al.* 1981). The same stop codons are used by the carp and rainbow trout. AGA is not used as a stop codon in *S. canicula*, as it is in the carp, trout, and cod. In some other fish, the AGA or AGG are used as stop codons, once in *L. chalumnae* and *P. dolloi* and twice in *P. ornatipinnis*. In *P. marinus*, the AGA is used five times as a stop codon.

Concerning the nature of the codons, we noted the comparatively rare usage of G at the third position in the fourfold degenerate codons (5.8%), as in most vertebrates, a low value that contrasts with the 12 and 24% observed in the hagfish and lancelet, respectively [based on the nucleotide sequences of *ND1*, *ND2*, *ATPase 8*, and *ATPase 6* genes (Delarbre and Gachelin 1997; Delarbre *et al.* 1997)]. The A nucleotide is the most used nucleotide at the third position (39.7%). The figures for T and C were 27.8 and 26.7%, respectively.

The coding sequences of several proteins were found to partly overlap, as in most species studied. The sequence coding for ATPase 8 overlaps by 10 nucleotides with the coding for ATPase 6 as in the sea lamprey, lungfish, bichir, coelacanth, and chicken, whereas the overlap is seven nucleotides in the lancelet (Delarbre and Gachelin 1997) and carp. The sequences coding for the ND4L and ND4 proteins overlap by seven nucleotides as in all chordates studied so far. Finally, the ND5 / ND6 coding genes overlap by four nucleotides as in the lungfish, carp, and sea urchin, whereas the overlap is as long as 16 nucleotides in the sea lamprey.

The overall amino acid usage has also been determined. It shows no preferential usage of any given amino acid, compared to the sea lamprey and several teleosts. This is in contrast with preferential usages already reported in the partial sequences of the hagfish and lancelet (Delarbre *et al.* 1997).

Intragroup and intergroup similarities, as defined by the sequences of the cytochrome b protein: No complete mtDNA sequence of other chondrichthyans is available so far. However, the sequence of the *cytochrome b* gene has been determined in numerous species, including other elasmobranchs. This allowed a more pre-

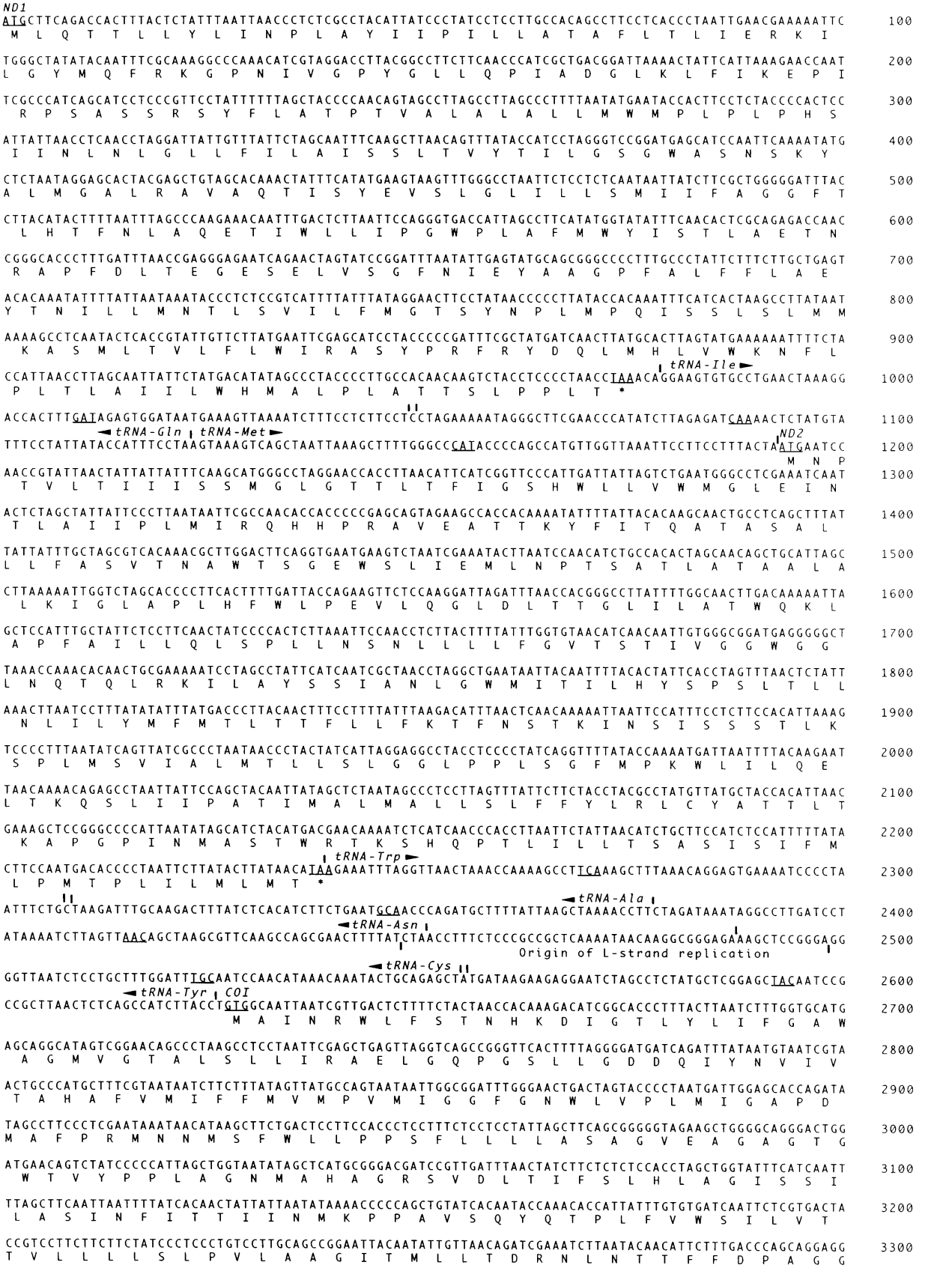


Figure 1.—Complete nucleotide and encoded amino acid sequences of the mtDNA of *S. canicula.* As in Table 2, the nucleotides are numbered starting from the first nucleotide of the gene coding for the ND1 protein. The name of each gene is placed at the 5′ of the coding sequence and above it. The limits of the genes are indicated by bars. The initiation codons and the tRNA anticodons are underlined. The stop codons are indicated by an asterisk. The putative regulatory sequences of the control are indicated above the corresponding nucleotide sequences, which are also underlined. The arrowheads indicate the orientation of the tRNAs.

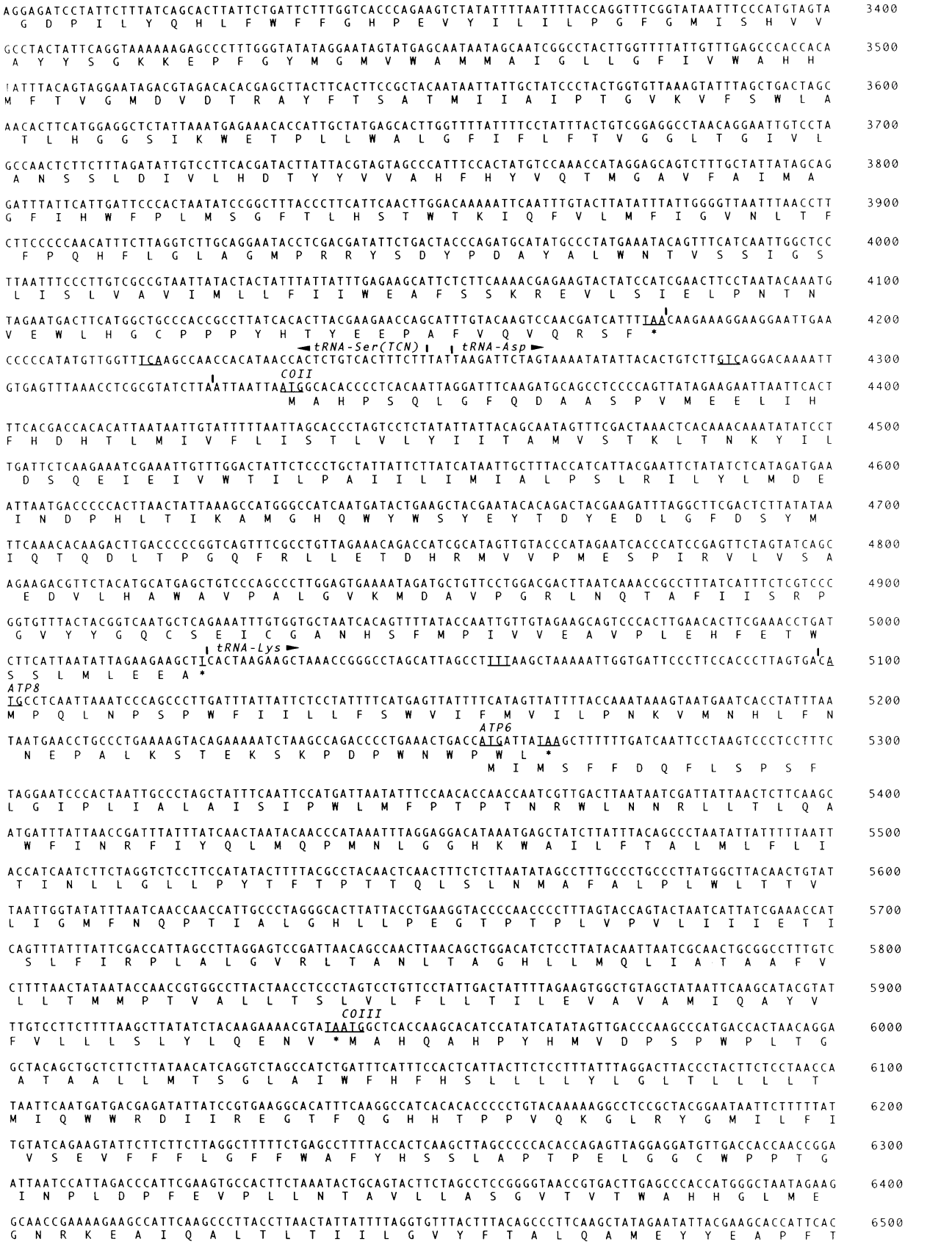


Figure 1.—*Continued.*

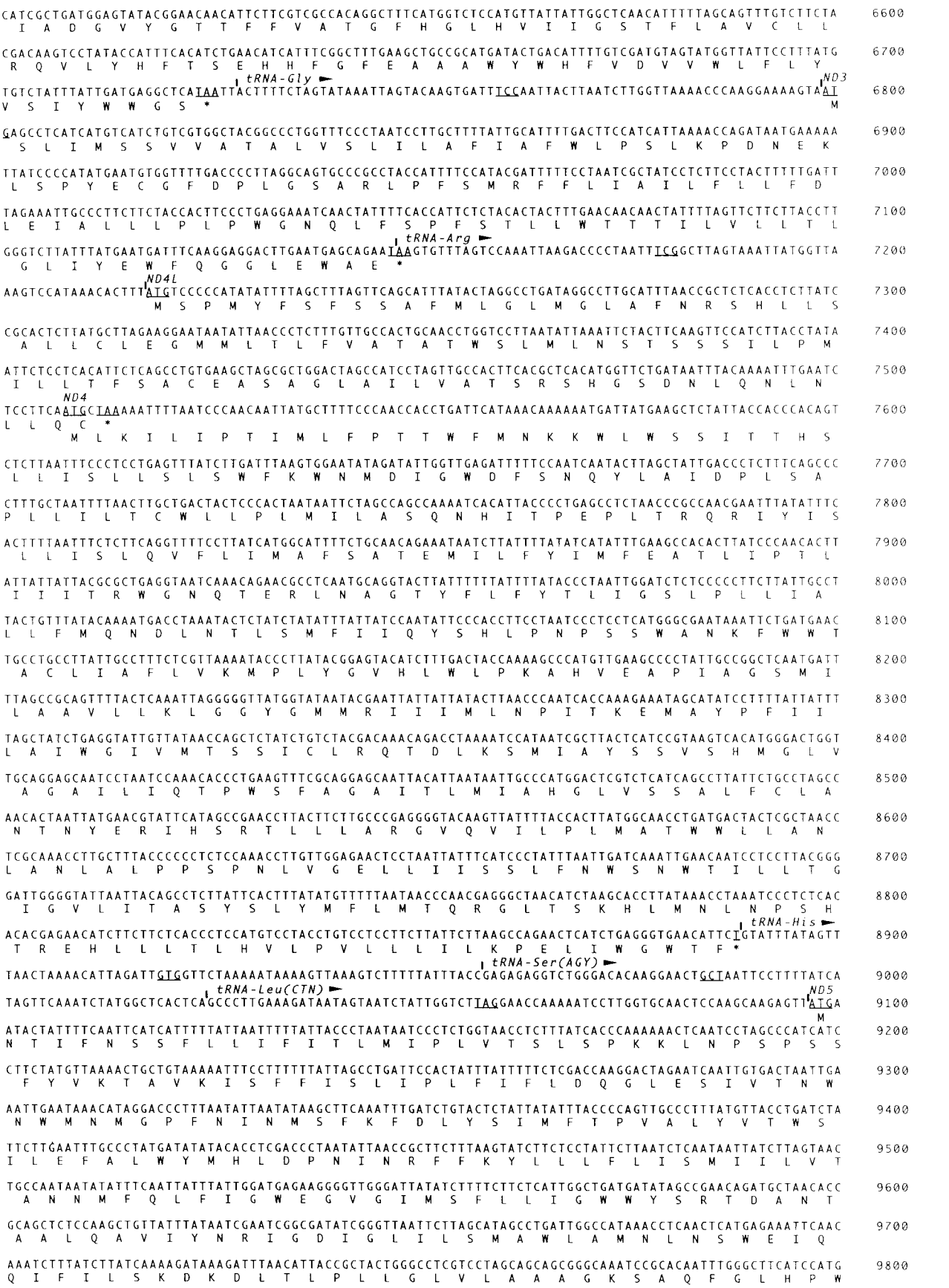


Figure 1.—*Continued.*

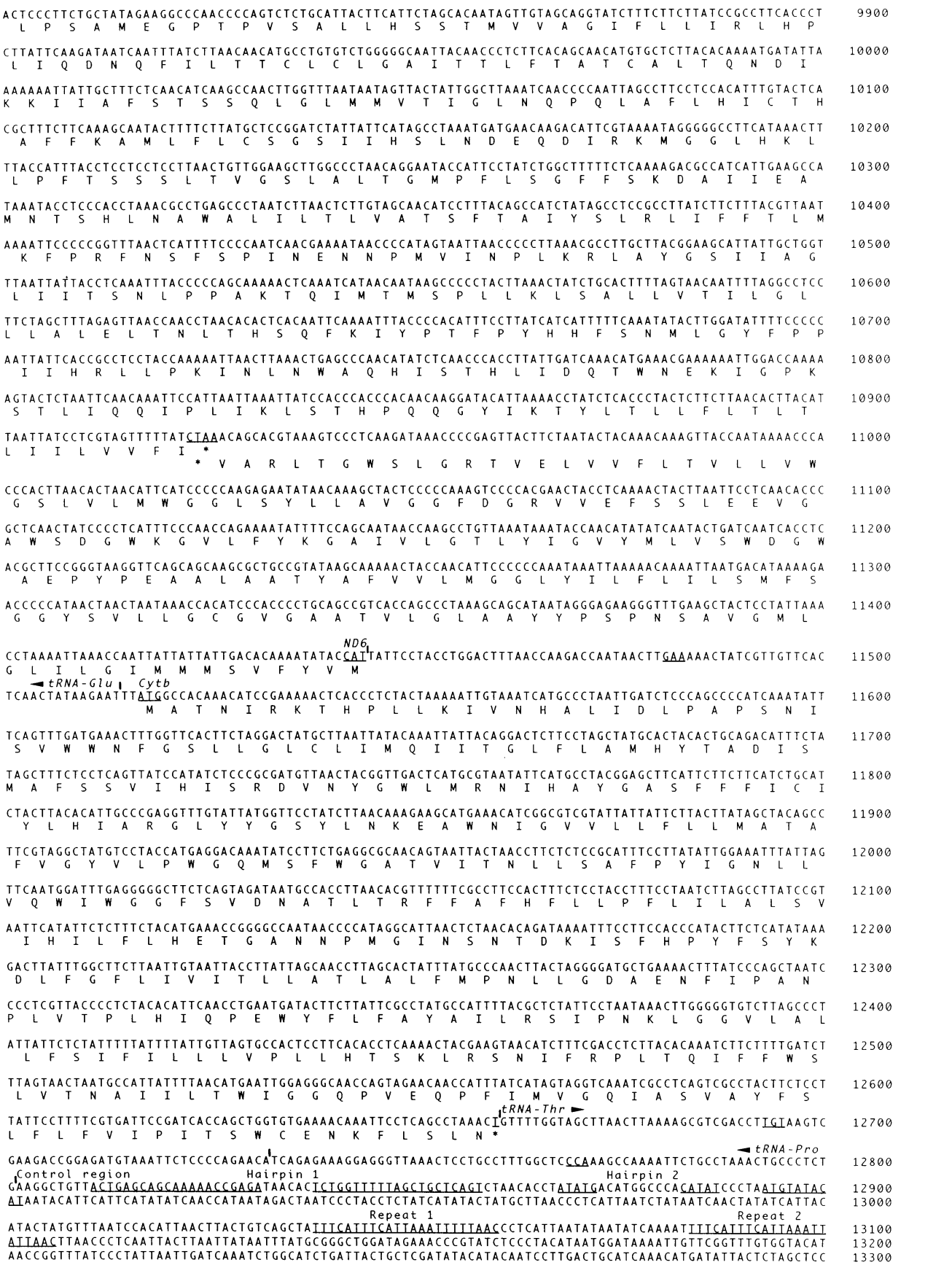


Figure 1.—*Continued.*

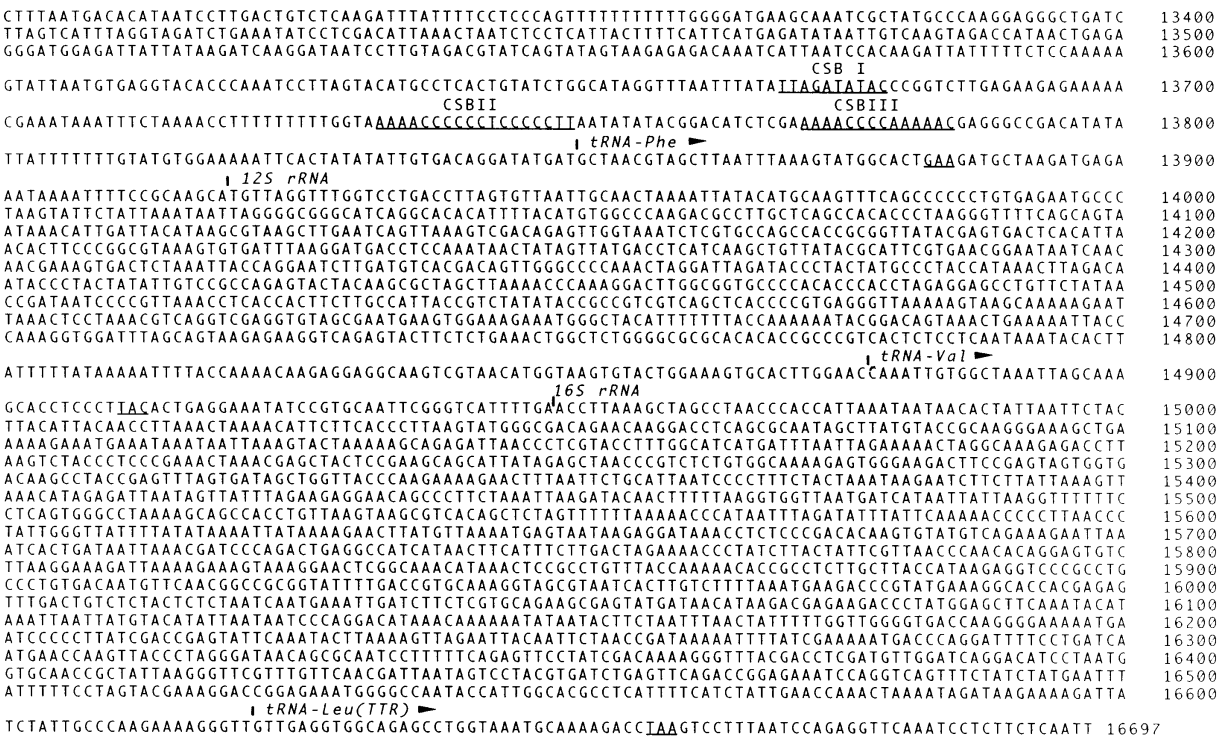


Figure 1.—*Continued.*

cise analysis of the relation of *S. canicula* to different groups of fish. Thus, the deduced amino acid sequence of the cytochrome b of *S. canicula* was aligned with that of several chondrichthyans, to check for the homogeneity of the Chondrichthyes group, then with that of the agnathan sea lamprey and a subset of bony fishes. The percentage of identity of the amino acid sequences was 83, 87, 87, and 83% with *Heterodontus francisci*, *Sphyrna tiburo vespertina*, *Carcharinus plumbeus*, and *Carcharodon carcharias* (Martin and Palumbi 1993), respectively. It was 67% with *P. marinus* (Lee and Kocher 1995) and 78, 77, 77, 75, 74, and 76% for *C. carpio* (Chang *et al.* 1994), *O. mykiss* (Zardoya *et al.* 1995), *C. lacustre* (Tzeng *et al.* 1992), *P. dolloi* (Zardoya and Meyer 1996), *P. ornatipinnis* (Noack *et al.* 1996), and *L. chalumnae* (Zardoya and Meyer 1997), respectively. The percentage of sequence similarity among cytochrome b amino acid sequences between chondrichthyans was 87%. It was 92% between the carp and loach and 89% between the carp and rainbow trout. On the basis of the cytochrome b amino acid sequences, the mtDNA of the dogfish thus behaves as a "standard" chondrichthyan mtDNA, but once more appears to be much closer to the Osteichthyes than to the Agnatha.

Intergroup similarities of all mitochondrial proteins: Considering that *S. canicula* is representative of the Chondrichthyes, we have aligned, with minimal insertions/deletions, the deduced amino acid sequences of all proteins of S. *canicula* with the corresponding proteins of several animals ranging from cephalochordates to birds (alignments are available upon request). The rate of evolution of the amino acid sequences differed for each protein. Considering the two extremes, the COI sequence was found to be extremely well conserved throughout evolution, whereas the ATPase 8 was the least conserved (data not shown). A distance Dayhoff matrix was computed (Table 4). It showed a low similarity of *S. canicula* mitochondrial proteins with those of a cephalochordate (*B. lanceolatum*) and an agnathan (*P. marinus*). The sequence similarity of the *S. canicula* mitochondrial proteins was higher with those of actinopterygians (*C. carpio*, *C. lacustre*, and *O. mykiss*) and an actinistian (*L. chalumnae*), than with those of a tetrapod (*G. gallus*).

Phylogenetic position of the dogfish: The final data set was composed of eight taxa and 11,532 sites when overlapping positions are duplicated, with 7465 variable sites, including 4888 informative sites. The lancelet was used as an outgroup. Using the MP method without weighting (equal weight given to transitions and transversions) and gaps treated as missing data, a single most-parsimonious tree was obtained (Figure 2A, length = 18,171 steps, C.I. = 0.685, R.I. = 0.311). If gaps were treated as a fifth state, the same resulting tree was computed (length = 18,665 steps, C.I. = 0.690, R.I. = 0.311). When transversions were given double the weight of transitions, the same result was obtained, and any gaps were treated as missing data (length = 27,746 steps) or as a fifth nucleotide (length = 27,943 steps). Thus, the dogfish appeared as the sister group of the Osteichthyes (sarcopterygians and actinopterygians). The node was supported by a 100% bootstrap value and a 202 Bremer

TABLE 3

Codon usage for mtDNA-encoded proteins of *S. canicula* (S), *P. marinus* (P), and *C. carpio* (C)

		S	P	C			S	P	C			S	P	C			S	P	C
Phe	TTT	152 (8)	137 (15)	76 (7)	Ser	TCT	55 (4)	73 (6)	29 (6)	Tyr	TAT	78 (10)	68 (6)	50 (6)	Cys	TGT	12 (1)	27 (7)	5 (2)
	TTC	95 (0)	98 (2)	146 (2)		TCC	65 (0)	44 (0)	57 (0)		TAC	41 (1)	39 (2)	65 (2)		TGC	11 (0)	12 (0)	20 (0)
Leu	TTA	225 (17)	199 (16)	100 (15)		TCA	95 (4)	87 (1)	82 (1)	End	TAA	9 (0)	5 (0)	7 (1)	Trp	TGA	108 (6)	99 (1)	112 (2)
	TTG	23 (8)	12 (7)	11 (9)		TCG	1 (0)	3 (1)	9 (1)		TAG	1 (1)	2 (0)	3 (0)		TGG	12 (1)	10 (2)	8 (3)
Leu	CTT	145 (2)	137 (2)	82 (7)	Pro	CCT	48 (2)	53 (4)	33 (3)	His	CAT	48 (0)	47 (1)	23 (1)	Arg	CGT	11 (2)	21 (1)	9 (4)
	CTC	86 (0)	53 (0)	97 (0)		CCC	57 (1)	32 (0)	52 (0)		CAC	52 (0)	58 (0)	79 (0)		CGC	21 (0)	8 (0)	5 (0)
	CTA	148 (0)	194 (2)	297 (0)		CCA	98 (0)	110 (0)	105 (0)	Gln	CAA	90 (0)	92 (0)	98 (0)		CGA	40 (0)	37 (2)	45 (0)
	CTG	19 (2)	15 (1)	37 (1)		CCG	6 (1)	4 (0)	12 (1)		CAG	7 (0)	6 (0)	1 (0)		CGG	4 (1)	2 (1)	10 (1)
Ile	ATT	244 (6)	232 (3)	157 (4)	Thr	ACT	62 (3)	106 (3)	41 (1)	Asn	AAT	88 (0)	78 (2)	45 (1)	Ser	AGT	23 (5)	33 (2)	9 (3)
	ATC	90 (0)	102 (3)	137 (0)		ACC	81 (0)	73 (0)	104 (2)		AAC	59 (1)	67 (0)	76 (0)		AGC	36 (0)	21 (0)	46 (0)
Met	ATA	139 (5)	187 (0)	126 (1)		ACA	119 (1)	125 (0)	146 (0)	Lys	AAA	79 (2)	91 (0)	73 (0)	End	AGA	0	5 (1)	0
	ATG	34 (4)	29 (5)	49 (5)		ACG	9 (2)	2 (1)	8 (1)		AAG	7 (0)	10 (0)	4 (1)		AGG	0	0	0
Val	GTT	60 (9)	79 (14)	43 (16)	Ala	GCT	87 (10)	96 (13)	55 (12)	Asp	GAT	42 (2)	32 (3)	11 (2)	Gly	GGT	54 (8)	51 (3)	32 (8)
	GTC	35 (0)	31 (2)	30 (0)		GCC	104 (0)	102 (0)	144 (2)		GAC	27 (1)	35 (1)	64 (0)		GGC	53 (1)	43 (2)	44 (3)
	GTA	81 (10)	73 (11)	118 (5)		GCA	85 (2)	99 (2)	129 (3)	Glu	GAA	88 (4)	82 (3)	93 (3)		GGA	80 (6)	85 (3)	137 (4)
	GTG	18 (4)	8 (4)	18 (7)		GCG	9 (3)	0	10 (1)		GAG	11 (2)	11 (1)	8 (2)		GGG	42 (11)	33 (11)	35 (11)

The values given in the table correspond to the total number of codons actually used in all genes of the animals studied. Numbers in parentheses are the values for the *NAD6* gene alone.

TABLE 4

Dayhoff distance matrix

	Scyliorhinus	Branchiostoma	Petromyzon	Cyprinus	Crossostoma	Oncorhynchus	Latimeria	Gallus
Scyliorhinus	0.00000							
Branchiostoma	0.71087	0.00000						
Petromyzon	0.53280	0.77742	0.00000					
Cyprinus	0.27102	0.69949	0.52162	0.00000				
Crossostoma	0.27833	0.71335	0.53175	0.12720	0.00000			
Oncorhynchus	0.26584	0.70567	0.53335	0.16626	0.16905	0.00000		
Latimeria	0.29705	0.72973	0.52837	0.26916	0.28440	0.26824	0.00000	
Gallus	0.44857	0.77822	0.58467	0.41917	0.42229	0.42236	0.43922	0.00000

The values correspond to the distance between the considered two taxa, which was calculated using the replacement Dayhoff's Pam matrix.

index. Using the NJ method, the NJ tree topology was found to be identical to the MP tree topology (Figure 2B). The position of the dogfish was also supported by a 100% bootstrap value. All internal branches appeared short in comparison to terminal branches. This holds particularly true for the chicken; in the same way, the minimal and maximal number of autapomorphies of the chicken were high compared to other taxa.

DISCUSSION

The primary conclusion reached from the study of the mtDNA of *S. canicula* is its high similarity with the mtDNA of the Osteichthyes: identical genomic organization, same loop of replication of the L strand, same codon usage, and high similarity of the amino acid sequences. Conversely, the genome organization and the DNA and amino acid sequences make the mtDNA of *P. marinus* very different from that of *S. canicula*. When studying the ND1-COI sequence, we had already suggested that some traits of the archetypal organization of the mtDNA of the Osteichthyes were present in the Chondrichthyes. The present data demonstrate that the overall organization and, presumably, the ways of the replication of the mtDNA of osteichthyans, were established in the common ancestor to chondrichthyans and osteichthyans.

The phylogenetic position of the dogfish we have found corresponds to the admitted consensus concerning the phylogeny of Recent chordates, with the chondrichthyans as the sister group of osteichthyans: a unique and identical tree was obtained using either MP or NJ. The result obtained using these eight taxa, which are relevant to the problem of the phylogenetic position of the dogfish, was reached using nearly all codons, in contrast to the need for a selection of a subset of codons pointed out by others (Naylor and Brown 1998).

However, the search for identity at the amino acid sequence level reveals a much higher similarity of the dogfish with teleosts and actinistians. This result does not, however, conflict with the phylogenetic position of the dogfish. The overall identity of amino acid sequences does not necessarily indicate close relationships. The different rates of evolution of mtDNA can account for the observed similarities between the amino acid sequences of the dogfish and those of teleosts. Indeed, the substitution rate of mtDNA is slow in sharks and teleosts and faster in homeothermic species (Thomas and Beckenbach 1989; Martin *et al.* 1992; Adachi *et al.* 1993).

The results of the phylogenetic analysis and of the amino acid sequences, however, must be considered in the light of the paleontological record. The current consensus about extant jawed vertebrate (gnathostome) interrelationships is chondrichthyans (including elasmobranchs and holocephalans) are the sister group of the osteichthyans (including actinopterygians and sarcopterygians). The monophyly of fossils and Recent chondrichthyans is supported by six morphological characters (Maisey 1986), but only one of them (the presence of prismatic calcified cartilage) is unambiguous or nonhomoplastic. The monophyly of fossil and recent osteichthyans is supported by at least 18 unambiguous morphological characters (Maisey 1986). The fact that the amino acid sequence of chondrichthyans (at any rate an elasmobranch) is closer to that of actinopterygians than to that of lampreys is no surprise, and it provides further support for the monophyly of the gnathostomes, which is already supported by at least 37 morphological characters (Janvier 1996a,b; Maisey 1986). The minimum age provided by paleontological data for the divergence between lampreys and gnathostomes is ~470 million years, but the minimum age for the divergence of the gnathostome taxa is practically identical because the earliest known chondrichthyans, actinopterygians, and sarcopterygians are all late Silurian in age (~410 million years).

In the framework of the current gnathostome phylogeny, the similarity between the dogfish and teleost amino acid sequences results from either homoplasy or symplesiomorphy. In the latter case, the close similarity between the dogfish and teleost sequences reflects the general (ancestral) condition for Recent gnathostomes.

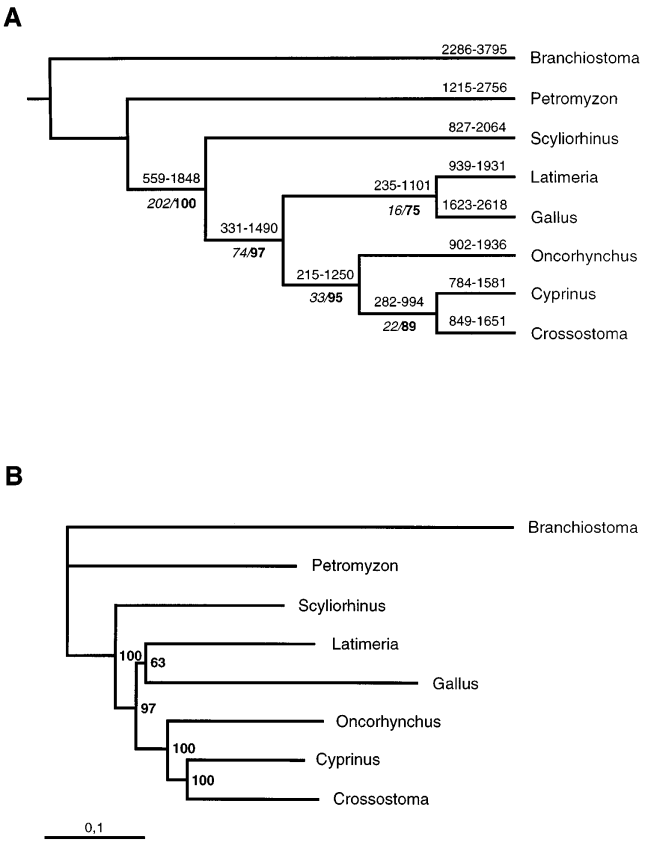


Figure 2.—Phylogenetic position of *S. canicula.* (A) The most parsimonious tree (PAUP 3.1.1, exhaustive search, unweighted parsimony, gaps = missing data) computed out of the 11,532 aligned positions. Tree length = 18,171 steps, C.I. = 0.685, R.I. = 0.311. The values above the branches indicate either synapomorphies or autapomorphies, according to the optimization of informative sites (table of linkage); numbers below the branches indicate the Bremer support (italics) and the bootstrap proportion (bold). (B) Phenogram constructed using the NJ method (Kimura's two-parameter distance, PHYLIP 3.5c). The numbers placed at the nodes indicate bootstrap proportion. The length of the branches is proportional to divergence between taxa (ranging from 0 to 1). A divergence of 0.1 corresponds to the length of the bar below the figure.

The work carried out in Paris has been supported in part by grants from the Institut National de la Santé et de la Recherche Médicale (INSERM). The work carried out in Lille has been supported by grants from the Institut Pasteur de Lille and the Centre National de la Recherche Scientifique (CNRS).

LITERATURE CITED

Adachi, J., Y. Cao and M. Hasegawa, 1993 Tempo and mode of mitochondrial DNA evolution in vertebrates at the amino acid sequence level: rapid evolution in warm-blooded vertebrates. J. Mol. Evol. **36:** 270–281.

Bremer, K., 1988 The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. Evolution **42:** 795–803.

Chang, Y.-c., F.-l. Huang and T.-b. Lo, 1994 The complete nucleotide sequence and gene organization of carp (*Cyprinus carpio*) mitochondrial genome. J. Mol. Evol. **38:** 138–155.

Dayhoff, M. O., R. M. Schwartz and B. C. Orcutt, 1978 *Atlas of Protein Sequence and Structure*, Vol. 5, pp. 342–352. National Biomedical Research Foundation, Washington, DC.

Delarbre, C., and G. Gachelin, 1997 A unique cDNA coding for subunits 8 and 6 of mitochondrial adenosine triphosphatase of the lancelet *Branchiostoma lanceolatum*, an ancestor of vertebrates. Biochem. Gen. **35:** 181–187.

Delarbre, C., V. Barriel, S. Tillier, P. Janvier and G. Gachelin, 1997 The main features of the craniate mitochondrial DNA between the ND1 and the COI genes were established in the common ancestor to the lancelet. Mol. Biol. Evol. **14:** 807–813.

Desjardins, P., and R. Morais, 1990 Sequence and gene organization of the chicken mitochondrial genome. A novel gene order in higher vertebrates. J. Mol. Biol. **212:** 599–634.

Doda, J. N., C. T. Wright and D. A. Clayton, 1981 Elongation of displacement loop strands in human and mouse mitochondrial DNA is arrested near specific template sequences. Proc. Nat. Acad. Sci. USA **78:** 6116–6120.

Felsenstein, J., 1985 Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39:** 783–791.

Felsenstein, J., 1993 PHYLIP (Phylogeny Inference Package), version 3.5c. Computer program distributed by the author. Department of Genetics, University of Washington, Seattle.

Hogan, B., R. Beddington, F. Costantini and E. Lacy, 1994 Isolating high molecular weight DNA from mouse tails, in *Manipulating the Mouse Embryo: A Laboratory Manual.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Janvier, P., 1996a Early vertebrates, in *Oxford Monographs in Geology and Geophysics*, No. 33. Oxford University Press, Oxford.

Janvier, P., 1996b The dawn of vertebrates: characters versus common ascent in the rise of current vertebrate phylogenies. Paleontology **39:** 259–287.

Johansen, S., and I. Bakke, 1996 The complete mitochondrial DNA sequence of Atlantic cod (*Gadus morhua*): relevance to taxonomic studies among codfishes. Mol. Marine. Biol. Biotechnol. **5:** 203–214.

Kimura, M., 1980 A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16:** 111–120.

Kumazawa, Y., and M. Nishida, 1995 Variations in mitochondrial tRNA gene organization of reptiles as phylogenetic markers. Mol. Biol. Evol. **12:** 759–772.

Lee, W.-J., and T. D. Kocher, 1995 Complete sequence of a sea lamprey (*Petromyzon marinus*) mitochondrial genome: early establishment of the vertebrate genome organization. Genetics **139:** 873–887.

Maisey, J. G., 1986 Heads and tails: a chordate phylogeny. Cladistics **2:** 201–256.

Martin, A. P., and S. R. Palumbi, 1993 Protein evolution in different cellular environments: cytochrome b in sharks and mammals. Mol. Biol. Evol. **10:** 873–891.

Martin, A. P., G. J. P. Naylor and S. R. Palumbi, 1992 Rates of mitochondrial DNA evolution in sharks are slow compared with mammals. Nature **357:** 153–155.

Naylor, G. J. P., and W. M. Brown, 1998 Amphioxus mitochondrial DNA, chordate phylogeny, and the limits of inference based on comparisons of sequences. Syst. Biol. **47:** 61–76.

Noack, K., R. Zardoya and A. Meyer, 1996 The complete mitochondrial DNA sequence of the Bichir (*Polypterus ornatipinnis*), a basal ray-finned fish: ancient establisment of the consensus vetebrate gene order. Genetics **144:** 1165–1180.

Ojala, D., J. Montoya and G. Attardi, 1981 tRNA punctuation model of RNA processing in human mitochondria. Nature **290:** 470–474.

Saitou, N., and M. Nei, 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol. Evol. **4:** 406–425.

Swofford, D. L., 1993 PAUP: Phylogenetic Analysis Using Parsimony, version 3.1.1. Computer program distributed by the Illinois Natural History Survey, Champaign, IL.

Thomas, W. K., and A. T. Beckenbach, 1989 Variation in salmonid mitochondrial DNA: evolutionary constraints and mechanisms of substitution. J. Mol. Evol. **29:** 233–245.

Tzeng, C.S., C.F. Hui, S.C. Shen and P.C. Huang, 1992 The complete nucleotide sequence of the *Crossostoma lacustre* mitochondrial genome, conservation and variation among vertebrates. Nucleic Acids Res. **20:** 4853–4858.

Walberg, M. W., and D. A. Clayton, 1981 Sequence and properties of the human KB cell and mouse L cell D-loop regions of mitochondrial DNA. Nucleic Acids Res. **9:** 5411–5421.

Zardoya, R., and A. Meyer, 1996 The complete nucleotide sequence of the mitochondrial geome of the lungfish (*Protopterus dolloi*) supports its phylogenetic position as a close relative of land vertebrates. Genetics **142:** 1249–1263.

Zardoya, R., and A. Meyer, 1997 The complete DNA sequence of the mitochondrial genome of a "living fossil," the coelacanth (*Latimeria chalumnae*). Genetics **146:** 995–1010.

Zardoya, R., A. Garrido-Pertierra and J. M. Bautista, 1995 The complete nucleotide sequence of the mitochondrial DNA genome of the rainbow trout, *Oncorhynchus mykiss*. J. Mol. Evol. **41:** 942–951.

Communicating editor: N. Takahata

The Complete Nucleotide Sequence of the Mitochondrial DNA of the Agnathan *Lampetra fluviatilis*: Bearings on the Phylogeny of Cyclostomes

Christiane Delarbre, Hector Escriva,† Cyril Gallut,*‡ Véronique Barriel,‡ Philippe Kourilsky,* Philippe Janvier,§ Vincent Laudet,† and Gabriel Gachelin**

*Département d'Immunologie, Unité de Biologie Moléculaire du Gène, Institut Pasteur, Paris, France; †Laboratoire de Biologie Moléculaire et Cellulaire, Ecole Normale Supérieure Lyon, France; ‡Service de Systématique Moléculaire, Institut de Systématique Centre du National de la Recherche Scientifique, Muséum National d'Histoire Naturelle, Paris, France; and §Laboratoire de Paléontologie, Muséum National d'Histoire Naturelle, Paris, France

There are two competing theories about the interrelationships of craniates: the cyclostome theory assumes that lampreys and hagfishes are a clade, the cyclostomes, whose sister group is the jawed vertebrates (gnathostomes); the vertebrate theory assumes that lampreys and gnathostomes are a clade, the vertebrates, whose sister group is hagfishes. The vertebrate theory is best supported by a number of unique anatomical and physiological characters. Molecular sequence data from *18S* and *28S* rRNA genes rather support the cyclostome theory, but mtDNA sequence of *Myxine glutinosa* rather supports the vertebrate theory. Additional molecular data are thus needed to elucidate this three-taxon problem. We determined the complete nucleotide sequence of the mtDNA of the lamprey *Lampetra fluviatilis.* The mtDNA of *L. fluviatilis* possesses the same genomic organization as *Petromyzon marinus,* which validates this gene order as a synapomorphy of lampreys. The mtDNA sequence of *L. fluviatilis* was used in combination with relevant mtDNA sequences for an approach to the hagfish/lamprey relationships using the maximum-parsimony, neighbor-joining, and maximum-likelihood methods. Although trees compatible with our present knowledge of the phylogeny of craniates can be reconstructed by using the three methods, the data collected do not support the vertebrate or the cyclostome hypothesis. The present data set does not allow the resolution of this three-taxon problem, and new kinds of data, such as nuclear DNA sequences, need to be collected.

Introduction

The relationships between hagfishes, lampreys, and jawed vertebrates (Gnathostoma) are one of the still-unresolved three-taxon problems in craniate phylogeny. Since Dumeril (1806) classified hagfishes and lampreys in the taxon Cyclostomi, characterized by horny teeth, a large notochord and pouch-shaped gills, the monophyly of this group has rarely been questioned, despite the subsequent discovery of apparently unique characters shared by the lampreys and the gnathostomes only. The apparent primitiveness of hagfishes has long been regarded as a consequence of ''degeneracy'' due to borrowing habits, a theory which still has many adherents (Fernholm 1985; Yalden 1985). Lovtrup (1977) was the first to propose that, according to character distribution, lampreys should be regarded as the sister group of the gnathostomes and, thus, that lampreys and hagfishes are paraphyletic. Janvier (1978) coined the name Myopterygii (''muscularized fins'') for the group including lampreys and the gnathostomes, but later considered it a synonym of Vertebrata, since only lampreys and gnathostomes have vertebral elements (Janvier 1981). This theory implies that the horny teeth and complex ''rasping tongues'' of hagfishes and lampreys are either homoplastic or, more likely, a general craniate character, lost in jawed vertebrates. Thus, there exist two competing theories about the interrelationships of craniates, i.e., animals with a skull. The ''cyclostome theory'' assumes that lampreys and hagfishes are a clade, the cyclostomes, whose sister group is the gnathostomes. The ''vertebrate theory'' assumes that lampreys and gnathostomes are a clade, the vertebrates, whose sister group is hagfishes.

The vertebrate theory is supported by a large number (about 50) of unique anatomical and physiological characters (Lovtrup 1977; Hardisty 1982; Janvier 1996). Although only few DNA sequences related to the phylogeny of cyclostomes have been determined, sequence data from rRNA support the cyclostome, rather than the vertebrate, theory (Stock and Whitt 1992; Mallat and Sullivan 1998). Sequence data from the mtDNA of some species including the hagfish *Myxine glutinosa* tend to support the vertebrate theory (Rasmussen, Janke, and Arnason 1998) although the data set was limited in size. Several phylogenetically important mtDNA sequences have been determined since the sequence of the mtDNA of the hagfish was published. They include two lancelets, *Branchiostoma lanceolatum* (Spruyt et al. 1998) and *Branchiostoma floridae* (Boore, Daehler, and Brown 1999); two chondrichthyans; *Scyliorhinus canicula* (Delarbre et al. 1998) and *Mustelus manazo* (Cao et al. 1998); and a hemichordate; *Balanoglossus carnosus* (Castresana, Feldmaier-Fuchs, and Pääbo 1998). Also, since the genomic organization of the mtDNA of the lamprey *Petromyzon marinus* (Lee and Kocher 1995) was found to be different from the consensus organization of the mtDNA of vertebrates, it was of interest to determine the nucleotide sequence and genomic organization of the mtDNA of another lamprey so as to ascertain the homogeneity of the zoological group and make available an extended data set with which to approach the hagfish-lamprey-gnathostomes relationships.

Key words: *Lampetra fluviatilis,* mitochondrial DNA, phylogeny, cyclostomes.

Address for correspondence and reprints: Gabriel Gachelin, Unité de Biologie Moléculaire du Gène, Département d'Immunologie, Institut Pasteur, 25 rue du Dr. Roux, 75724 Paris cedex 15, France. E-mail: ggachel@pasteur.fr.

Mol. Biol. Evol. 17(4):519–529. 2000

Table 1
PCR Primers Used in the Determination of the Nucleotide Sequence of the *Lampetra fluviatilis* Mitochondrial Genome

PCR Primers	Name	Sequence (5′–3′)	Location
1	LEU	TGCRMAAGRYYTAAGCYCTT	
	NAD1L	GAGTAAAAAGGGCTAGGTTGAGG	16106–912
2	ND1	CCACGATTCCGATATGATCAACT	
	L23	TTGGCGAGTTCAACTGATGCTAA	838–2021
3	85	TATTATCTCTAGGTGGCCTC	
	COI	CCRATRTCYTTGTGWTTAGTAGA	1950–2650
4	TYRL	GCCCACCGCCTAAACATTCGG	
	COI3	GCATCTGGGTARTCKGAGTADCGKCG	2572–3940
5	L7	ACATTCTTCCCCCAACATTTCC	
	L11	CCCTAGAAGGTTGATTGTTAG	3872–5479
6	R13	TTTAGGAGGACATAAATGAGC	
	ND33	CCTTGTRKTCATTCATARATTAGKCC	5407–7088
7	ND35	GAAAAACTMTCWCCATACGAATGYGG	
	ND43	GGGRGCYTCTACRTGRGCTTTDGG	6858–8155
8	NAD4	TCAGATCTCTCGCTACTCTCAAACACACCA	
	S16	TTATAGGTAATCCCAGGGTAGCTCGTCTAG	8021–14744
9	16S5	ATYAAAYCTYGTACCTTTTGCATCAT	
	16S3	CTAYAYTTTATYTKTCCTTTCGTACTAA	14648–15982
10	16S	CGTGATCTGAGTTCAGACCGG	
	NAD1L	GAGTAAAAAGGGCTAGGTTGAGG	15903–912

NOTE.—IUB code: R = A/G; Y = C/T; M = A/C; W = A/T; D = G/A/T; K = G/T.

We thus determined the entire mtDNA sequence of another lamprey, *Lampetra fluviatilis.* The mtDNA of *L. fluviatilis* possesses the same genomic organization as *P. marinus.* This finding validates this distinctive gene order as a synapomorphy of lampreys. The mtDNA sequence was used, in combination with relevant mtDNA sequences, to approach the present three-taxon problem using three different computational methods. In contrast with previous conclusions (Naylor and Brown 1998), the present analysis shows that complete coding regions of the mtDNA may be used to study the phylogenetic relationships between recent but anciently rooted animals but does not, however, settle the problem of the monophyly/paraphyly of agnathans. Finally, the present study points to the need for additional mtDNA and nuclear DNA sequences, such as those of urochordates and cyclostomes.

Materials and Methods

Animals

A specimen of *L. fluviatilis* was caught in the Atlantic Ocean on the shores of the estuary of the Garonne river. The animal was anesthesized, killed, and dissected. Organs were immediately frozen in liquid nitrogen and stored at −78°C.

Preparation of DNA

Total DNA was prepared from lamprey muscles by Proteinase K digestion according to conventional procedures (Hogan et al. 1994).

Isolation and Sequencing of mtDNA

Overlapping fragments of mtDNA were obtained by PCR run on total genomic DNA and using degenerate primers. The sequences of the primers used for PCR amplification are described in table 1. Two strategies for obtaining PCR-amplified DNA were used.

For PCR aimed at isolating the 16S rRNA to ND4 sequence, the conditions were as follows: 300 ng total DNA, 200 μM dNTP, and 500 nM primers; the enzyme was the *Pfu* Exo + polymerase (Stratagene); the thermal cycles were 3 min at 94°C followed by 50 cycles of 1 min at 94°C, 1 min at 48°–55°C depending upon the primers, and 5 min at 72°C. The last cycle was ended by incubating for 10 min at 72°C. These PCR products were purified by electroelution, phosphorylated using the polynucleotide kinase (Pharmacia), and ligated at the *Eco*RV site of dephosphorylated KS Bluescript vector (Stratagene) using the Rapid DNA Ligation kit (Boehringer). XL1Blue-competent bacteria (Stratagene) were transformed with the ligation products. Several recombinant clones were selected for each PCR product, and plasmid DNA was recovered using the ClearCut Miniprep kit (Stratagene). The cloned PCR products were sequenced first using the M13 (−40) and KS reverse primers, and subsequently using primers derived from the sequence determined. Thus, primer walking was used throughout with primers located on both strands. Three hundred base pairs were read using each primer. The entire sequence was determined on both strands. All overlapping sequences were found to be identical.

For PCR aimed at the isolation of the ND4-16S rRNA segment of the mtDNA, the Expand Long Template PCR system (Boehringer) was used. The PCR conditions were as follows: 500 ng DNA, 350 μM dNTP, and 300 nM primers. Denaturation of the template for 2 min at 92°C was followed by 10 cycles of 10 s at 92°C, 30 s at 60°C, and 8 min at 68°C, followed by 20 cycles with an elongation step increased by 20 s/cycle. The last cycle was ended by incubating for 7 min at 68°C. The 8-kb-long fragment was cloned in the TOPO vector (Invitrogen), digested with *Eco*RI, and subcloned

in pBCKS (Stratagene). The subcloned fragments were sequenced using forward and reverse primers. The two strands were sequenced.

Phylogenetic Analyses

The 13 protein-coding sequences of the mtDNA of 12 taxa; *Paracentrotus lividus* (Cantatore et al. 1989), *Asterina pectinifera* (Asakawa et al. 1995), *B. carnosus, B. lanceolatum, B. floridae* (we used the nucleotide sequence corrected by Boore, Daehler, and Brown [1999]), *M. glutinosa, L. fluviatilis, P. marinus, S. canicula, M. manazo, Cyprinus carpio* (Chang, Huang, and Lo 1994), and *Oncorhynchus mykiss* (Zardoya, Garrido-Pertierra, and Bautista 1995), were first aligned manually following translation of the nucleotide sequences using the mitochondrial genetic code specific for each taxon. The alignment was refined using the physical amino acid similarity criterion. All inserted gaps were triplets. The 5′ parts of the *ND5* genes (about 200 bp) were not amenable to alignment and were thus omitted. Whenever alignment problems due to deletion or insertion events were locally encountered, a parsimony analysis search using PAUP was carried out on that zone in order to define the ''optimal'' alignment that would (1) minimize the number of inferred mutations (number of steps), (2) test the number of weighted mutations (one transition [Ts] preferred to one transversion [Tv]), (3) minimize the number of variable sites, and (4) minimize the *a priori* phylogenetic implications of the alignments (Barriel 1994*a*). Thus, each alignemnt was tested with Tv = 1 or Tv = 2 for three different codings of the gaps (gap = missing data ?, gap = new state, and gap = ID). Indeed, standard procedures for coding gaps offer alternative strategies which suffer from several weaknesses: either the different sites are analyzed independently (gap = new state) so that each gap is artificially weighted relatively to the number of sites, or each site is coded ''?'' (gap = missing data) and the optimization procedure makes the whole zone devoid of phylogenetic information. Our new coding strategy (gap = ID) is aimed at expressing the potential phylogenetic information contained in complex zones with internested insertions/deletions (indels) and substitutions (Barriel 1994*b*). It precludes loss or distortion of information in all of those cases in which gaps are present in the aligned sequences. Gaps are coded as multiple-state characters; following an indel, shared subsequent mutations (substitutions as well as indels) should be interpreted in terms of common descent. According to the hierarchy of internested states of characters, this strategy introduces question marks in the data matrix, which are optimized *in fine* in cladograms based on all data. They are thus not missing data, but rather methodological codes, neutral to any *a priori* phylogenetic analysis (Barriel 1994*a*, 1994*b*). The accession numbers of the aligned sequences are 39776, 39786, 39787, 39789, and 39795–39803.

Three different analysis were carried out. The maximum-parsimony (MP) method was applied, using the branch-and-bound search implemented in PAUP, version 3.1.1 to determine the most parsimonious tree. The robustness of the MP trees was tested using the bootstrap method with 500 replications each implemented in PHYLIP (Felsenstein 1985, 1993) and using the Bremer (1988) support. The analysis of nucleotide and amino acid sequences was also performed using the neighbor-joining (NJ) algorithm (BioNJ; Gascuel 1997) with Kimura's (1980) two-parameter model for nucleotides and Kimura's (1983) model for amino acids. Finally, the fastDNAml program was used for maximum-likelihood (ML) analysis of DNA sequences (Felsenstein 1981; Olsen et al. 1994). The protml program was used for ML analysis of amino acid sequences using the Jones, Taylor and Thornton model (jtt; Adachi and Hasegawa 1996).

Results

Comparison of the mtDNA of *L. fluviatilis* with the mtDNA of *P. marinus*

Overall Genomic Organization

The mtDNA of *L. fluviatilis* (accession number Y18683) is 16,159 bp long, thus slightly smaller (42 bp) than the mtDNA of *P. marinus,* a size difference mostly due to a shorter noncoding region II (see below). The mtDNA of *L. fluviatilis* harbors 13 protein-coding genes, 22 tRNAs genes and 2 rRNAs, and a noncoding region (including the control region). The molecular map is identical to that of *P. marinus* (table 2) and is thus characteristic of the lampreys and different from the map of the mtDNA of all other vertebrates.

Control Regions

As in *P. marinus,* the control region of the mtDNA of *L. fluviatilis* is located between the *ND6* and *CYTb* genes, instead of being located between the *CYTb* and *12S rRNA* genes. It is split into two parts (noncoding regions I and II) which are 491 and 151 bp long, respectively, in *L. fluviatilis,* separated by the *tRNA-Thr* and *tRNA-Glu* genes. The percentage of similarity between the nucleotide sequences of the noncoding regions of the two lampreys reaches 90.6%. The noncoding region II of *Lampetra* is only 151 bp long, whereas that of *Petromyzon* is 199 bp long. This difference is due to the absence of two repeats in *Lampetra.* The percentage of identity is 86.7% in the part of this region common to the two species.

We have previously shown that the origin of replication of the light chain of the mtDNA of *L. fluviatilis* is not located between the *ND1* and *ND2* genes (Delarbre et al. 1997). No other obvious origin of replication of the light chain can be evidenced elsewhere in the noncoding and control regions of the mtDNA of *L. fluviatilis* or *P. marinus.*

RNAs

The percentages of identity of the *12S* and *16S rRNA* genes of *Petromyzon* and *Lampetra* are 96.3% and 93.9%, respectively.

The locations of all *tRNA* genes are found to be identical in both species. However, the sizes of the

Table 2
Localization of the Genes in the Mitochondrial Genome of *Lampetra fluviatilis*

Gene	Beginning Position	Ending Position	Size		Codons	
			bp	Amino Acids	Start	Stop
ND1	1	966	966	321	ATG	TAA
tRNA-Ile	997	1065	69			
tRNA-Gln	1068	1138	71 (L)			
tRNA-Met	1139	1206	68			
ND2	1208	2251	1,044	347	ATG	TAG
tRNA-Trp	2250	2316	67			
tRNA-Ala	2321	2388	68 (L)			
tRNA-Asn	2391	2459	69 (L)			
tRNA-Cys	2461	2527	67 (L)			
tRNA-Tyr	2532	2602	71 (L)			
COI	2604	4157	1,554	517	GTG	AGA
tRNA-Ser (*TCN*)	4148	4219	72 (L)			
tRNA-Asp	4220	4287	68			
COII	4291	4980	690	229	ATG	TAA
tRNA-Lys	4993	5059	67			
ATPase 8	5061	5228	168	55	ATG	TAG
ATPase 6	5219	5897	679	226	ATG	T– –
COIII	5898	6683	786	261	ATG	TAA
tRNA-Gly	6693	6760	68			
ND4	7475	8851	1,377	458	ATG	AGA
tRNA-His	8853	8921	69			
tRNA-Ser (*AGY*)	8922	8989	68			
tRNA-Leu (*CTN*)	8991	9062	72			
ND5	9064	10860	1,797	598	ATG	AGG
ND6	10845	11363	519 (L)	172	ATG	AGA
Noncoding I	11364	11854	491			
tRNA-Thr	11855	11926	72			
tRNA-Glu	11928	11998	71 (L)			
Noncoding II	11999	12149	151			
CYTb	12150	13340	1,191	396	ATG	AGA
tRNA-Pro	13345	13415	71 (L)			
tRNA-Phe	13427	13491	65			
12S rRNA	13492	14394	903			
tRNA-Val	14395	14465	71			
16S rRNA	14466	16081	1,616			
tRNA-Leu (*TTR*)	16082	16155	74			

Note.—ATPase 6 and 8 = ATP synthase subunits 6 and 8; COI, II, and III = cytochrome C oxidase subunits I, II, and III; CYTb = cytochrome b; ND1–6 = NADH dehydrogenase subunits 1–6. "L" in parentheses means that the gene is located on the L strand of the mtDNA.

tRNA-Trp, *-Cys*, *-Asp*, *-Phe*, and *-Ser* (*AGY*) genes differ by a single (either missing or additional) nucleotide. The sequences of the tRNAs are slightly divergent, with an average 94.7% similarity, with *tRNA-Arg* and *tRNA-Gln* genes being strictly identical in the two species. The overall 80 nucleotide differences are mainly located in the DHU (13) or in the TΨC (34) loops, although 22 are located in the stems (table 3). The more divergent tRNA genes are the *tRNA-Cys*, *-Pro*, and *-Trp* genes. The absence of loop in the T arm of the *tRNA-Phe* gene of *Lampetra* is worthy of note: all five nucleotides of the T arm can pair in *Lampetra*, whereas only three nucleotides pair in *Petromyzon*, and an additionnal nucleotide yields a loop at the end of the T arm.

Protein-Coding Genes

The sizes of the genes coding for the 13 proteins are the same in the two species. The size of the cytochrome *b* protein is the same in both species but is 12 amino acids longer than in fishes. The initiation codons have no distinctive features. Two termination codons are found to be different between *Lampetra* and *Petromyzon*. They are TAG for the *ND3* gene of *Lampetra* (TAA for *Petromyzon*) and AGG for the *ND5* gene of *Lampetra* (AGA for *Petromyzon*). We have controlled the identity of the stop codons used by *ATPase 6* and *COI* mRNAs by cloning the corresponding cDNAs and sequencing the 3′ part. In *Petromyzon*, Lee and Kocher (1995) found the ATPase 6 amino acid sequence to be 11 amino acids longer (with an overlap of 35 nt with the *COIII* gene and a stop codon AGA) than in other animals. In *L. fluviatilis*, the mRNA coded by the *ATPase 6* gene is polyadenylated after the T located immediately before the ATG of the *COIII* gene, giving a stop codon TAA and an ATPase 6 protein the same size as those in other animals. We assume that the same process is used to terminate the ATPase 6 coding sequence in *Petromyzon*. The mRNA coding for the COI protein (using the stop codon AGA) shows that the *tRNA-Ser* (*TCN*) gene which follows the coding sequence of COI is transcribed on the same RNA, with the polyadenylation occuring after the *tRNA-Ser* (*TCN*) gene sequence. The *COI* gene overlaps the *tRNA-Ser*

Table 3
Distribution of Substitutions in the tRNAs of *Lampetra* and *Petromyzon*

	AA Stem	DHU Stem	DHU Loop	DHU Stem	AC Stem	AC Loop	AC Stem	VAR Loop	TC Stem	TC Loop	TC Stem	AA Stem	Total
Ala			1							3			4
Arg													0
Asn					1					2			3
Asp			1							2			3
Cys	2			1	1				1	5		1	11
Gln													0
Glu								1				2	3
Gly			1			2				1			4
His			2					1		1			4
Ile			1										1
Leu (CTN)			1							1			2
Leu (TTR)	1		1										2
Lys									1	2			3
Met	1									5			6
Phe	1					1		1		1	2		6
Pro			2		1	1		2	1	1			8
Ser (AGY)	1							1				1	3
Ser (TCN)	1		1										2
Thr								1		2			3
Trp			1							6			7
Tyr			1	1						1		1	4
Val										1			1
Total	7	0	13	2	3	4	0	7	3	34	2	5	80

gene by 10 nt. Incidentally, a similar situation is found for *M. glutinosa*: AGG is used as the stop codon, and the *tRNA-Ser* (*TCN*) gene which immediately follows the *COI* gene is transcribed on the same RNA, with the polyadenylation occuring after the *tRNA-Ser* (*TCN*) gene. In *M. glutinosa*, the overlap between the *COI* gene and the *tRNA-Ser* (*TCN*) gene is 13 nt long. A similar finding has been reported for humans except for the absence of overlap between the *COI* and *tRNA-Ser* genes (Anderson et al. 1981).

Lampetra fluviatilis employs essentially the same codon usage as *P. marinus.* However, the codons ended by G are more frequently used. In particular, the GCG codon is used eight times by *Lampetra* and is not used at all in the mitochondrial genes of *Petromyzon.* In the *ND6* gene, which is located on the L strand of the mitochondrial DNA, there is not the same usage of C and A in *Petromyzon* and *Lampetra* in the third nucleotides of fourfold-degenerate codons (table 4).

Table 4
Third-Nucleotide Usage in the Fourfold-Degenerate Codons

	T	C	A	G
All genes but ND6				
Lampetra	28.7	23.3	42.9	5.1
Petromyzon	31.9	21.3	44.1	2.7
Myxine	28.9	27.8	35.7	7.6
ND6				
Lampetra	52.3	12.8	12.8	22.1
Petromyzon	51.1	4.5	23.3	21.1
Myxine	40.5	12.2	16.2	31.1

NOTE.—The values given correspond to the percentages of the nucleotide used at the third positions on the fourfold-degenerate codons.

The amino acid compositions of the proteins coded by the mitochondrial DNA are identical in *P. marinus* and *L. fluviatilis.* The deduced amino acid sequences were easy to align at both the amino acid and the nucleotide levels. The percentage of similarity at the nucleotide level was 82.5%–87.9% (average 85.8%) depending on the genes (table 5). The sequence of the *COI* gene is the most conserved as usual, with the least conserved being the gene coding for ATPase 6. At the amino acid level, a high percentage of similarity is noted; it ranges from 97.9% for ND4L to 83.6% for ATPase 8, with an average of 92.5%.

Substitutions

When the nucleotide sequence of *Petromyzon* was compared with that of *Lampetra*, the C-T or T-C substitutions were found to be the most frequent, with an average of 55.7% of total substitutions in the 13 genes. The average of the percentages of A-G or G-A substitutions was 16.8%. The frequency of G-C or C-G substitutions was only 1.2% of the total number of substitutions. The ratio of nonsilent to total substitutions varies with the genes, ranging from 8.1% to 42.9% with peaks above 30% for the *ATPase 6, ATPase 8,* and *ND6* genes. The passage of C to T or T to C is usually synonymous (particularly in the *COI* gene, where only 0.9% of the mutations are nonsynonymous), in contrast to the *ND6* gene, for which 44% of the C-T or T-C substitutions are nonsynonymous. In the *ATP8* and *ATP6* genes, the percentage of nonsilent substitutions is high for nearly all kinds of substitutions.

Table 5
Comparison Between Mitochondrial Protein-Coding Sequences of *Lampetra* and *Petromyzon*

	% Similarity		No. of Substitutions		
	Nucleotides	Amino Acids	Total	Nonsilent	% Nonsilent
ND1	85.6	95.9	139	15	10.8
ND2	85.7	90.5	149	33	22.1
COI	87.9	96.7	188	19	10.1
COII	87.1	93	89	17	19.1
ATP8	87.5	83.6	21	9	42.9
ATP6	82.5	85.8	119	42	35.3
COIII	87.3	95.8	100	12	12
ND3	84.9	96.6	53	6	11.3
ND4L	87.3	97.9	37	3	8.1
ND4	85	94.5	206	25	12.1
ND5	85.3	93.8	264	39	14.8
ND6	84	84.9	83	27	32.5
CYTb	85.3	94.2	175	25	14.3

Comparison of the mtDNA of the Lampreys to that of the Hagfish

At this stage, it can be concluded that few differences exist between the mtDNA and the deduced amino acid sequences of *Petromyzon* and *Lampetra*, making them a homogeneous group. In contrast, the data available so far concerning the hagfish (i.e., the protein-coding sequences) point to a remarkable divergence between the lampreys and a third representative of cyclostomes, the hagfish *M. glutinosa.* First, the gene map of the mtDNA of *M. glutinosa* is identical to that of the gnathostomes in general (Rasmussen, Janke, and Arnason 1998) and thus differs from that of the lampreys. Second, the nucleotide sequences of the protein-coding genes of the hagfish were found to be difficult to align for phylogenetic analysis (see below) with the corresponding genes of the lampreys, and the sizes of the genes were also not always identical. More importantly, the average percentages of similarity of the 13 mitochondrial proteins were found to be only 52.6% (*Petromyzon* compared with *Myxine*) and 53% (*Lampetra* vs. *Myxine*), whereas the percentage of similarity between *Petromyzon* and *Lampetra* was 92.5% and that between *Petromyzon* and the chondrichthyan *Scyliorhinus canicula* was 62.4%. Some additional differences between the mtDNA of the lampreys and the hagfish could also be noted: the codon usage of *M. glutinosa* was markedly different, and serine and phenylalanine were more frequently used by the lampreys than by the hagfish, in contrast with threonine and alanine, which exhibited the opposite behavior.

Analysis of the Lamprey/Hagfish Relationships

The high divergence between hagfishes and lampreys does not tell much about their phylogenetic relationships. Thus, a phylogenetic analysis was carried out on the mtDNA of 12 taxa ''flanking'' the hagfish/lamprey node and including *L. fluviatilis,* including *P. lividus, A. pectinifera, B. carnosus, B. lanceolatum, B. floridae, M. glutinosa, L. fluviatilis, P. marinus, S. canicula, M. manazo, C. carpio,* and *O. mykiss.* The analysis was carried out on the complete data set and on the subset of it as defined below.

Analysis of Nucleotide Sequences

Use of All Protein-Coding Mitochondrial Genes.— The complete data set (13 genes) was composed of 12 taxa, and 11,583 nt were used as sites (3,424 invariant and 7,266 parsimony-informative sites). Three taxa were used as outgroups: *B. carnosus, A. pectinifera,* and *P. lividus.* Using the MP method without weighting (equal weight given to transitions and transversions) and gaps treated as missing data, a single most-parsimonious tree was obtained in which cyclostomes are paraphyletic (tree A) (fig. 1; length = 26,979 steps, consistency index [CI] = 0.577, and retention index [RI] = 0.498). All the nodes were supported by a 100% bootstrap value, except the craniate node (69% bootstrap value) and the vertebrate node (64% bootstrap value). If gaps were recoded as in Barriel (1994*a*, 1994*b*) (Gap = ID), the same unique resulting tree was computed (length = 27,114 steps, CI = 0.578, and RI = 0.500). The same result was obtained when transversions were given double the weight of transitions and when gaps were treated as missing data (length = 40,918 steps).

We used two additional methods to approach the hagfish/lamprey relationships. The same data set was used in a neighbor-joining analysis in the Kimura's distance analysis using dnadist implemented in PHYLIP. All coding nucleotides were used, with a weighting of 2 for transversions. A unique tree denominated ''B'' was obtained (fig. 2) in which cyclostomes appeared to be monophyletic, with bootstrap values of 64.5% for the cyclostome node and 100% for all other nodes. Upon ML analysis using dnaml, the same data set also resulted in a unique tree of the B type in which cyclostomes were monophyletic. All nodes were at $P < 0.01$. Bootstraps values were 100% for all nodes except for the node supporting the monophyly of cyclostones (98.5%).

Use of a Selected Subset of Data.—Naylor and Brown (1998) suggested that subsets of genes should be selected prior to computational analysis and preferably used in place of the complete mtDNA-coding sequences. In order to select the proper set of genes to be used, every individual gene in each taxon was first studied using MP and NJ strategies and using the same weighting and coding as for the complete nucleotide sequence. Several different topologies were generated. Among them, trees A and B were the most commonly observed and were also in agreement with the generally admitted phylogenetic relationships, as well as with the cladograms obtained using the complete mtDNA-coding sequences. The nucleotide sequences of the genes which generated mostly A or B patterns (the subset comprised COI, COII, COIII, ND1, ND4, and ND5, with a total of 7,239 nucleotides, of which 2,376 were invariant and 4,321 were parsimony-informative) were then assembled into a unique nucleotide sequence and analyzed as above.

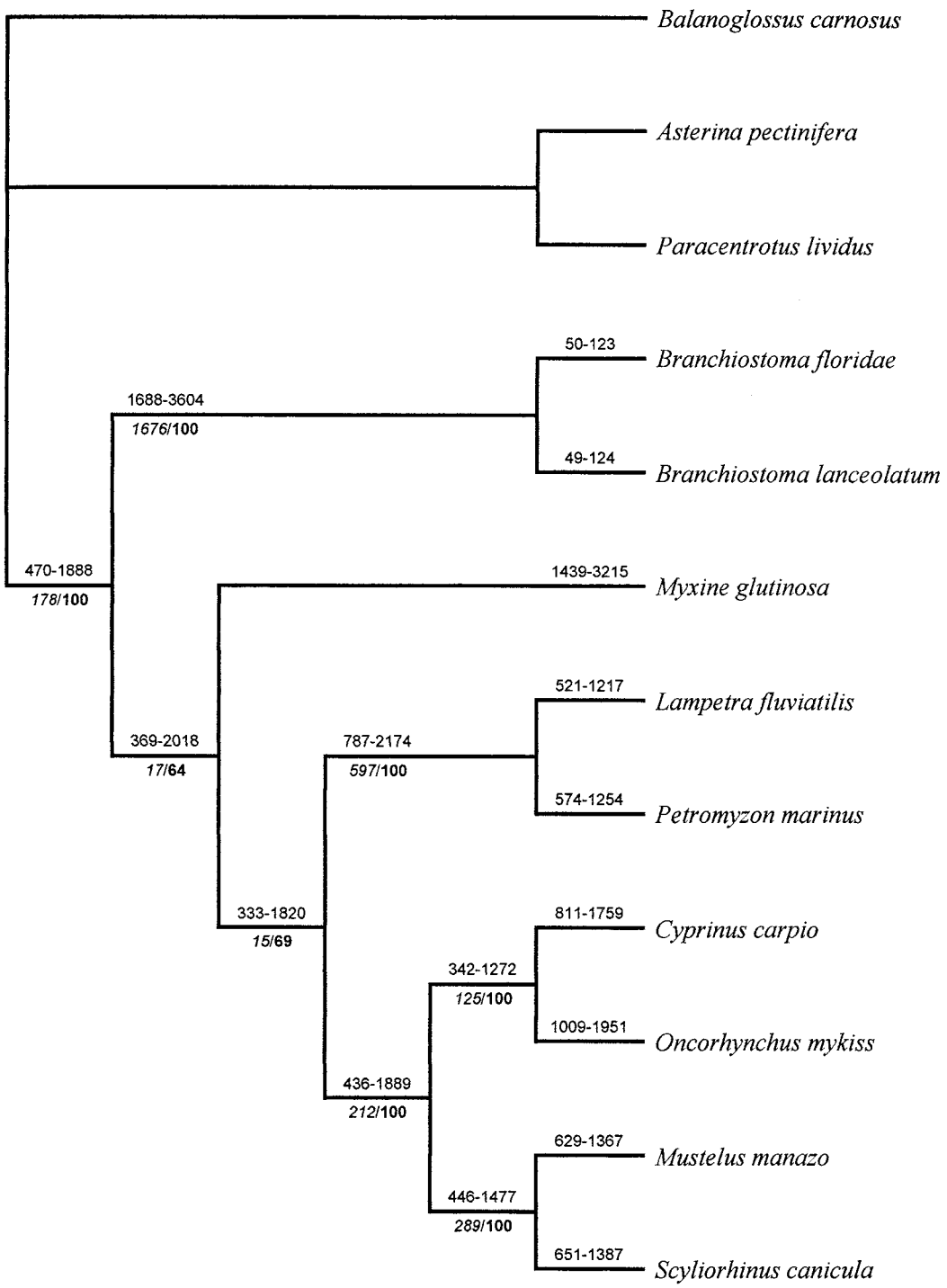


FIG. 1.—Phylogenetic position of agnathans determined by using the maximum-parsimony strategy. The most parsimonious tree (pattern A) was computed out of the 11,583 aligned positions using PAUP, version 3.1.1 (Swofford 1993), with branch-and-bound search, unweighted parsimony, and gaps treated as missing data. The values above the branches indicate both the minimum and the maximum possible lengths (synapomorphies or autapomorphies) of the branch according to the optimization of informative sites according to the table of linkage present in the ''Describe Trees'' dialog box of PAUP. Numbers below the branches indicate the Bremer (1994) support, i.e., the number of steps needed for a node to disappear in a most-parsimonious tree (italics), and the bootstrap proportions (bold type).

Upon parsimony analysis, the subset of data yielded a unique parsimonious tree (length = 15,685 steps, CI = 0.577, and RI = 0.504) identical to that obtained using unselected mtDNA sequences and in which cyclostomes appeared paraphyletic. The cyclostomes also appeared paraphyletic when the NJ method was used but became monophyletic when the ML strategy was used with all nodes at $P < 0.01$.

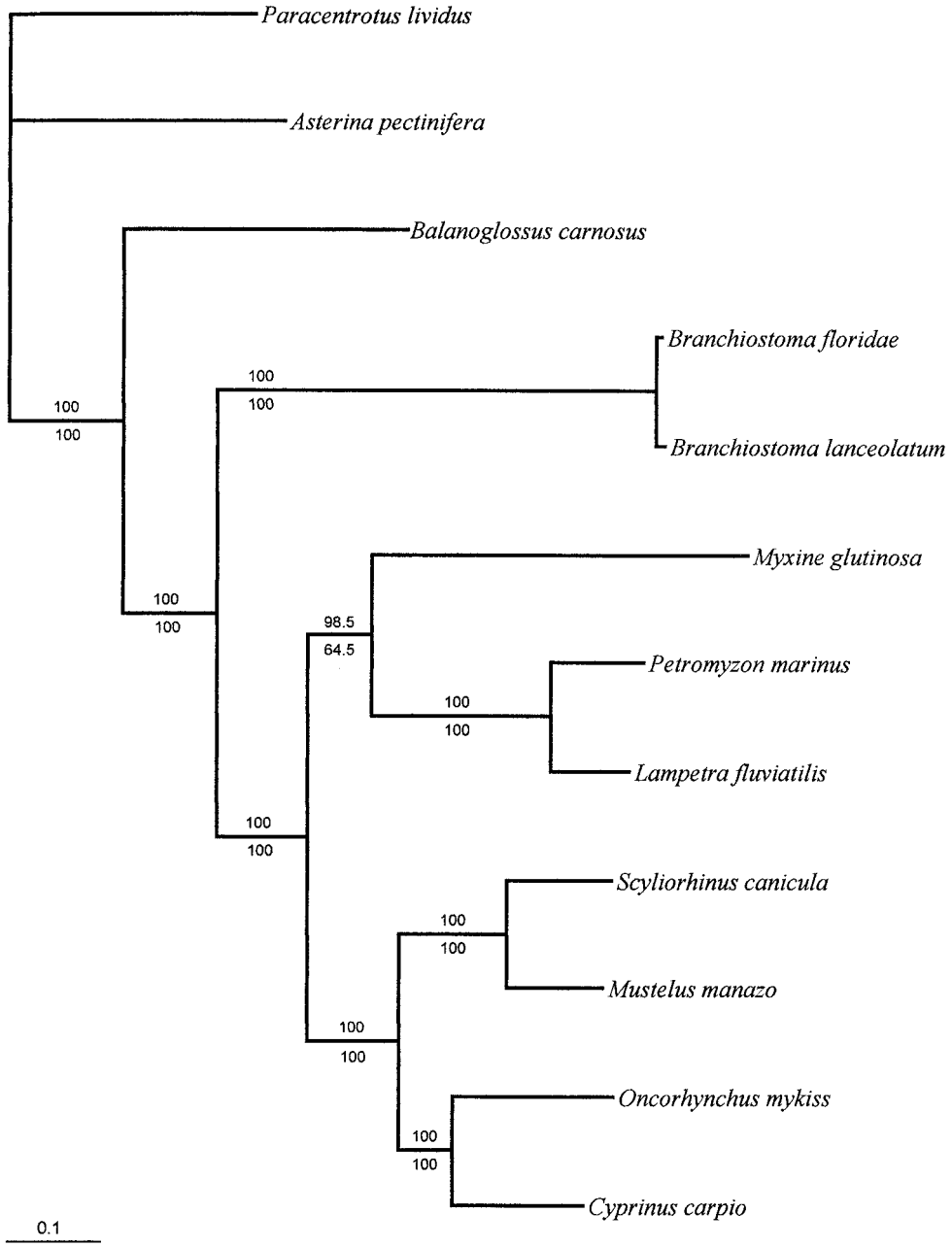


FIG. 2.—Phylogenetic position of agnathans determined using the neighbor-joining (NJ) and maximum-likelihood (ML) analyses. The same data set as in figure 1 was used, generating the unique B tree. The values above the lines are bootstrap values obtained using ML analysis. The values below the lines are bootstrap values obtained using NJ analysis.

Naylor and Brown (1998) also suggested usage of only the first and second codon positions of sites modally coding for proline, cysteine, methionine, glutamine, and asparagine. Using the first and second codon positions of these sites in the above defined subset of genes (a total of 776 sites, 359 invariant and 338 parsimony-informative), MP analysis yielded a unique A-type tree, with cyclostomes being paraphyletic. Interestingly enough, the consistency and retention indices were higher (length = 996 steps, CI = 0.719, and RI = 0.685). Bootstrap values were 100%/97%. The same result (tree A) was obtained by using the three nucleotides of each codon encoding modal amino acids.

Analysis of Amino Acid Sequences

Analysis of All Protein Sequences.—The amino acid data set is less prone to saturation effects than the nucleotide data set. Thus, the amino acid sequences already aligned as a requisite for the alignment of the nucleotide sequences were also analyzed using the same three methods. MP analysis (3,848 sites, 1,276 constant

and 2,183 parsimony-informative) yielded a single B-type parsimonious tree (cyclostomes are monophyletic) (length = 8,729 steps, CI = 0.812, and RI = 0.720). Bootstrap values were 98% for the cyclostome node and 100% for the other nodes. NJ analysis yielded a single tree of the A type (*M. glutinosa* appears to be paraphyletic). Bootstrap values were 63% for the vertebrate node and 100% for the others. ML analysis yielded two equally possible trees (A and B, with Δ1n −54,568.0 and −54,567.3, respectively).

Analysis of Protein Subsets.—The amino acid sequences of the individual genes were analyzed using the MP analysis (2,407 sites, 910 constant and 1,256 parsimony-informative). Again, several topologies were obtained, with the A and B patterns predominating in the proteins whose genes had yielded A and B patterns in the above section, thus defining the same subset of proteins as the nucleotide sequences. MP analysis of the subset yielded a single parsimonious tree belonging to the B type in which cyclostomes are monophyletic (length = 4,734 steps, CI = 0.815, and RI = 0.734). Bootstrap values were 85% for the cyclostome node and 100% for the other nodes. NJ analysis of the same data set yielded a single tree of the A type in which hagfish appears to be paraphyletic. Bootstrap values were 100%/62%. ML analysis yielded the A and B patterns with equal probability.

Table 6
Summary of the Results Obtained Using Different Methods and Data Sets

	MP	NJ	ML
Nucleotides			
Entire DNA sequence			
All genes	A	B	B
Gene subset	A	A	B
Codons encoding modal amino acids			
All genes	A		
Gene subset	A		
Amino acids			
All genes	B	A	A/B
Gene subset	B	A	A/B

NOTE.—MP = maximum parsimony; NJ = neighbor joining, ML = maximum likelihood. "A" refers to the tree shown in figure 1. "B" refers to the tree shown in figure 2. The "gene subset" has been defined for the present data set according to Naylor and Brown (1998).

Discussion

The mtDNA of *L. fluviatilis* is highly similar in all respects to that of *P. marinus.* It displays the same genomic organization as *P. marinus,* a finding which validates this particular gene order as a synapomorphy of lampreys.

In contrast, the mtDNA of the lampreys profoundly differs from that of *M. glutinosa.* Thus, the mtDNA sequence of *L. fluviatilis* was used in combination with the mtDNAs of taxa which, to the best of our present knowledge, flank the rooting under study to construct cladograms aimed at better defining the lamprey/hagfish relationships. The use of complete mitochondrial DNA to approach the phylogeny of chordates has recently been questioned by Naylor and Brown (1998). These authors concluded that only a small fraction of the DNA sequence could be retained for such studies and excluded the use of the entire coding sequence. They retained as informative the first and second positions of the triplets coding for proline, cysteine, methionine, glutamine, and asparagine in a subset of mitochondrial genes to be defined for each data set studied. In contrast to these conclusions, the present analysis shows that, at least within the range of taxa we have selected, the complete protein-coding mitochondrial DNA sequences can generate cladograms compatible with our present knowledge of the phylogeny of craniates, whether using the MP, NJ, or ML method. Moreover, the subsets of data selected according to Naylor and Brown's (1998) criteria resulted in cladograms identical to those obtained using complete mtDNA. Thus, there appears to exist no need to select subsets of data; complete coding mtDNA sequences can be used to construct cladograms in which the accepted phylogenetic relationships are retrieved. Although difficult to ascertain on a statistical basis, the influence of different evolution rates of the genes coding for different proteins is most probably minimized if a sum of coding genes is used, provided a maximum of phylogenetic information is extracted from the sequences. The use of all positions in the triplets is justified by different codon usages in different species and also by the preferential nucleotide usage at the third position, observations which both reflect distinctive properties of the translation machinery in the different taxa.

However, concerning the precise problem of the lamprey/hagfish relationships, the cladograms generated from the nucleotides and amino acids using the MP, NJ, and ML analyses support either the cyclostome or the vertebrate hypothesis (table 6); the question remains unanswered, since the conclusions produced out of the same data set by using the same outgroups are markedly dependent on the kind of mathematical analysis used.

Using a limited number and a less focused choice of taxa, Rasmussen, Janke, and Arnason (1998) deduced the paraphyly of agnathans. However, earlier analyses (Stock and Whitt 1992) based on 1,631 nt of the 18S rDNA with tunicates and cephalochordates as outgroups, as well as a more recent analysis of the 28S and 18S rDNA sequences (Mallatt and Sullivan 1998), supported the cyclostome theory, but weakly and with some ambiguity. These different conclusions most probably reflect the number of informative sites used and intrinsic differences in the uses of rRNA-coding and protein-coding DNA sequences in phylogenetic studies: rDNA consists of highly conserved sequences which are likely to be poorly informative, surrounded by highly variable stretches of DNA which are not amenable to alignment. The choice and number of taxa, the quality of the alignments, the coding strategies, and the choice of the outgroups also obviously all influence the conclusions reached.

Part of these difficulties can be due to the too small number of sequences which are directly relevant to the

problem under study. Additional mitochondrial and nuclear DNA sequences, such as those of other agnathans and urochordates, must be determined. They may still be unsufficient to definitely elucidate the present three-taxon problem, and additional morphological data, including information on the embryonic development of hagfishes (Wicht and Tusch 1998), will be needed.

Acknowledgments

C.G. acknowledges receipt of a fellowship of the Ministère de l'Education Nationale de la Recherche et de la Technologie. The work carried out in the Unité de Biologie Moléculaire du Gène has been supported by the EEC, the Institut National de la Santé et de la Recherche Médicale, and the Institut Pasteur. The work carried out in the Unité Mixte de Recherche CNRS/ENS 49 was supported in part by the Centre National de la Recherche Scientifique and the Ecole Normale Supérieure de Lyon. C.D. and H.E. contributed equally to the sequencing of the mtDNA of *L. fluviatilis.*

LITERATURE CITED

ADACHI, J., and M. HASEGAWA. 1996. MOLPHY version 2.3: programs for molecular phylogenetics. I. PROTML: maximum likelihood inference of protein phylogeny. Comput. Sci. Monogr. **28**:1–150.

ANDERSON, S., A. T. BANKIER, B. G. BARRELL et al. (11 coauthors). 1981. Sequence and organization of the human mitochondrial genome. Nature **290**:457–465.

ASAKAWA, S., H. HIMENO, K.-I. MIURA, and K. WATANABE. 1995. Nucleotide sequence and gene organization of the starfish *Asterina pectinifera* mitochondrial genome. Genetics **140**:1047–1060.

BARRIEL, V. 1994*a*. Phylogénies moléculaires et insertion délétion de nucléotides. C. R. Acad. Sci. III **317**:693–701.

———. 1994*b*. Relations de parenté au sein des Hominoidea et la place de *Pan paniscus.* Comparaison et analyse méthodologique des phylogénies morphologique et moléculaire. Ph.D. thesis, University of Paris VI, Paris.

BOORE, J. L., L. L. DAEHLER, and W. M. BROWN. 1999. Complete sequence, gene arrangements and genetic code of the mitochondrial DNA of the cephalochordate *Branchiostoma floridae* (Amphioxus). Mol. Biol. Evol. **16**:410–418.

BREMER, K. 1988. The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. Evolution **42**:795–803.

———. 1994. Branch support and tree stability. Cladistics **10**: 295–304.

CAO, Y., P. J. WADDELL, N. OKADA, and M. HASEGAWA. 1998. The complete mitochondrial DNA sequence of the shark *Mustelus manazo*: evaluating rooting contradictions to living bony vertebrates. Mol. Biol. Evol. **15**:1637–1646.

CANTATORE, P., M. ROBERTI, G. RAINALDI, M. N. GADALETA, and C. SACCONE. 1989. The complete nucleotide sequence, gene organization and genetic code of the mitochondrial genome of *Paracentrotus lividus.* J. Biol. Chem. **264**: 10695–10975.

CASTRESANA, J., G. FELDMAIER-FUCHS, and S. PÄÄBO. 1998. Codon reassignment and amino acid composition in hemichordate mitochondria. Proc. Natl. Acad. Sci. USA **95**: 3703–3707.

CHANG, Y.-C., F.-L. HUANG, and T.-B. LO. 1994. The complete nucleotide sequence and gene organization of carp (*Cyprinus carpio*) mitochondrial genome. J. Mol. Evol. **38**:138–155.

DELARBRE, C., V. BARRIEL, S. TILLIER, P. JANVIER, and G. GACHELIN. 1997. The main features of the craniate mitochondrial DNA between the ND1 and the COI genes were established in the common ancestor to the lancelet. Mol. Biol. Evol. **14**:807–813.

DELARBRE, C., N. SPRUYT, C. DELMARRE, C. GALLUT, V. BARRIEL, P. JANVIER, V. LAUDET, and G. GACHELIN. 1998. The complete nucleotide sequence of the mitochondrial DNA of the dogfish, *Scyliorhinus canicula.* Genetics **150**:331–344.

DUMERIL, A. M. C. 1806. Zoologie analytique ou méthode naturelle de classification des animaux. Didot, Paris.

FELSENSTEIN, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **17**:368–376.

———. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39**:783–791.

———. 1993. PHYLIP (phylogeny inference package). Version 3.5c. Distributed by the author, Department of Genetics, University of Washington, Seattle.

FERNHOLM, B. 1985. Evolutionary biology of primitive fishes. Pp. 113–192. *in* R. E. FOREMAN, A. GORBMANN, J. M. DODD, and R. OLSSON, eds. Plenum Press, New York.

GASCUEL, O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol. Biol. Evol. **14**:685–695.

HARDISTY, M. W. 1982. Lampreys and hagfishes: analysis of cyclostome relationships. Pp. 166–260 *in* M. W. HARDISTY and I. C. POTTER, eds. The biology of lampreys. Vol. 4B. Academic Press, London.

HOGAN, B., R. BEDDINGTON, F. R. COSTANTINI, and E. LACY. 1994. Isolating high molecular weight DNA from mouse tails. Manipulating the mouse embryo, a laboratory manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

JANVIER, P. 1978. Les nageoires paires des ostéostracés et la position systématique des céphalaspidomorphes. Ann. Paleontol. **64**:113–142.

———. 1981. The phylogeny of the Craniata, with particular reference to the significance of the fossil agnathans. J. Vertebr. Paleontol. **1**:121–159.

———. 1996. The dawn of the vertebrates: characters versus common ascent in the rise of current vertebrate phylogenies. Paleontology **39**:259–287.

KIMURA, M. 1980. A simple model for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111–120.

———. 1983. The neutral theory of evolution. Cambridge University Press, Cambridge, England.

LEE, W.-J., and T. D. KOCHER. Complete sequence of a sea lamprey (*Petromyzon marinus*) mitochondrial genome: early establishment of the vertebrate genome organization. Genetics **139**:873–887.

LOVTRUP, S. 1997. The phylogeny of Vertebrata. Wiley, New York.

MALLATT, J., and J. SULLIVAN. 1998. 28S and 18S rDNA sequences support the monophyly of lampreys and hagfishes. Mol. Biol. Evol. **15**:1706–1718.

NAYLOR, G. J. P., and W. M. BROWN. 1998. Amphioxus mitochondrial DNA, chordate phylogeny and the limits of inference based on comparisons of sequences. Syst. Biol. **47**: 61–76.

OLSEN, G. J., H. MATSUDA, R. HAGSTROM, and R. OVERBEEK. 1994. FastDNAml version 1.2: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. Comput. Appl. Biosci. **10**:41–48.

RASMUSSEN, A.-S., A. JANKE, and U. ARNASON. 1998. The mitochondrial DNA molecule of the hagfish (*Myxine glutinosa*) and vertebrate phylogeny. J. Mol. Evol. **46**:382–388.

SPRUYT, N., C. DELARBRE, G. GACHELIN, and V. LAUDET. 1998. Complete sequence of the amphioxus (*Branchiostoma lanceolatum*) mitochondrial genome: relations to vertebrates. Nucleic Acids Res. **26**:3279–3285.

STOCK, D. W., and G. S. WHITT. 1992. Evidence from 18S ribosomal RNA sequences that lampreys and hagfishes form a natural group. Science **257**:787–789.

SWOFFORD, D. L. 1993. PAUP. Version 3.1. Illinois Natural History Survey, Champaign.

WICHT, H., and U. TUSCH. 1998. Ontogeny of the head and nervous system of myxinoids. Pp. 431–451 *in* J. M. JIRGENSEN, J. P. LOMHOLT, R. E. WEBER, and H. MALTE, eds. The biology of hagfishes. Chapman & Hall, London.

YALDEN, D. W. 1985. Feedings mechanisms as evidence for cyclostome monophyly. Zool. J. Linn. Soc. **84**:291–300.

ZARDOYA, R., A. GARRIDO-PERTIERRA, and J. M. BAUTISTA. 1995. The complete nucleotide sequence of the mitochondrial DNA genome of the rainbow trout, *Oncorhynchus mykiss.* J. Mol. Evol. **41**:942–951.

AXEL MEYER, reviewing editor

Accepted December 10, 1999

Molecular Phylogenetics and Evolution
Vol. 22, No. 2, February, pp. 184–192, 2002
doi:10.1006/mpev.2001.1045, available online at http://www.idealibrary.com on IDEAL®

Complete Mitochondrial DNA of the Hagfish, *Eptatretus burgeri*: The Comparative Analysis of Mitochondrial DNA Sequences Strongly Supports the Cyclostome Monophyly

Christiane Delarbre,* Cyril Gallut,† Veronique Barriel,† Philippe Janvier,‡ and Gabriel Gachelin,*[1]

*Unité de Biologie Moléculaire du Gène, Institut Pasteur, 75724 Paris Cedex 15, France; †Service de Systématique Moléculaire, Muséum National d'Histoire Naturelle, 75005 Paris, France; ‡UMR 2569, Laboratoire de Paléontologie, Muséum National d'Histoire Naturelle, 75005 Paris, France

Received January 29, 2001; revised June 27, 2001

The phylogenetic position of cyclostomes, i.e., the relationships between hagfishes, lampreys, and jawed vertebrates is an unresolved problem. Anatomical data support the paraphyly of cyclostomes, whereas nuclear genes data support monophyly of cyclostomes. Previous results obtained using mitochondrial DNA are ambiguous, presumably due to a lack of informative sequences. By adding the complete mtDNA of a hagfish, *Eptatretus burgeri*, we have generated a novel data set for sequences of hagfishes and of lampreys. The addition of this mtDNA sequence to the 12 taxa we have already used becomes sufficient to obtain unambiguous results. This data set, which includes sequences of mtDNA of animals closely related to the lamprey/hagfish node, was used in a phylogenetic analysis with two independent statistical approaches and unequivocally supported the monophyly of cyclostomes. Thus molecular data, i.e., our results and those obtained using nuclear genes, conclude that hagfishes and lampreys form a clade.

Key Words: **mitochondrial DNA; hagfish; *Eptatretus*; phylogeny.**

INTRODUCTION

Extant vertebrates are classified into gnathostomes (jawed vertebrates) and cyclostomes. Cyclostomes (Cyclostomi), or jawless fishes, are ancient and the earliest undoubted fossils of hagfishes and lampreys date from the late Carboniferous. The possible affinities of conodonts with hagfishes as well as recent reinterpretation of fossils considerably enlarge the geological range of the group (Donoghue *et al.*, 2000), which thus appears among the earliest known vertebrates. Extant cyclostomes are subdivided into lampreys (or Hyperoartia) and hagfishes (or Hyperotreta). The phylogenetic relationships between hagfishes, lampreys, and gnathostomes are controversial. The problem of the monophyly vs paraphyly of the Cyclostomi has been vigorously debated since Løvtrup (Lovtrup, 1977) first advocated cyclostome paraphyly on the basis of an impressive number of morphological and physiological characters shared uniquely by lampreys and gnathostomes. This hypothesis of phylogenetic relationships has been widely accepted by most morphologists and palaeontologists (Janvier, 1978, 1981, 1996; Forey, 1984; Forey and Janvier, 1993; Gagnier, 1993; Rovainen, 1996). It is currently referred to as the "vertebrate hypothesis," as it implies that the presence of vertebrae is unique to lampreys and gnathostomes and thus that only these two taxa should be included in the taxon Vertebrata. There is now overwhelming evidence for at least 48 unique anatomical and physiological characters shared by lampreys and gnathostomes. Conversely, the number of unique characters shared by hagfishes and lampreys remains low, but there are possibly up to 22 and there is practically no character shared uniquely by hagfishes and gnathostomes.

Since then, only a few morphologists have opposed the "vertebrate hypothesis" (Schaeffer and Thomson, 1980; Yalden, 1985). However, most nucleotide sequence data have supported the "cyclostome hypothesis," i.e., the monophyly of cyclostomes (Mallatt and Sullivan, 1998; Kuraku *et al.*, 1999), except for one analysis of mitochondrial sequence data (from some species including two cyclostomes), which tended to favor the "vertebrate hypothesis" (Rasmussen, *et al.*, 1998). In a previous paper (Delarbre *et al.*, 2000), we attempted to solve the problem of the phylogeny of cyclostomes using mtDNA sequences, including the sequence of an additional cyclostome, *Lampetra flu-*

[1] To whom correspondence and reprint requests should be addressed at Unité de Biologie Moléculaire du Gène, Unité Inserm 277, Institut Pasteur, 25 rue du Dr. Roux, 75015 Paris, France. Fax: +33 1 45 68 85 48. E-mail: ggachel@pasteur.fr.

1055-7903/02 $35.00

TABLE 1

PCR Primers Used in the Determination of the Nucleotide Sequence of *Eptatretus burgeri* Mitochondrial Genome

	Sequences (5′–3′) of PCR primers	Location (bp)	Length of PCR (bp)
1:	CGTGATCTGAGTTCAGACCGG	16,939	
	CCRATRTCYTTGTGWTTAGTAGA	2594	2823
2:	ATCACTATGTATCTTTCCCGATG	2564	
	GCATCTGGGTARTCKGAGTADCGKCG	3906	1342
3:	GATTTCTGAGTAATATTCACAGGAG	3806	
	TCAAATGTTTTAAGAGGAACTAC	4916	1110
4:	TGTGGCGCAAAYCAYAGYTTTATRCC	4855	
	CGTGAAAGCCKGTGGCKACAAARAAKG	6462	1607
5:	ACTTCACTTCAGGCAATAGAATAC	6369	
	TTAGTCATTCGTAGATTAAACC	7044	675
6:	GAAAAACTMTCWCCATACGAATGYGG	6816	
	GGGRGCYTCTACRTGRGCTTTDGG	8090	1274
7:	GATTAGGTTGTTTTTCTGCCCTAC	8005	
	GTGCTGARACAGGAGTTGGGCCYTCTAT	9719	1714
8:	TCCGCCCAATTTCTTCTCCATC	9653	
	AARTAYCAYTCTGGYTTRATRTGRGGNGG	12,184	2531
9:	TTCCCTGCTTCTACCATACTTAACCACAGATCCTG	12,089	
	GTTAATTACTGCTGAATACCCTTGGGGGTGTGAAG	14,781	2692
10:	CCTCATTACAGTGAGTTCCTTG	14,673	
	CGGGATTCAGGCTCAGAAAGTCTG	15,763	1090
11:	ATYAAAYCTYGTACCTTTTGCATCAT	15,697	
	CTAYAYTTTATYTKTCCTTTCGTACTAA	17,017	1320

Note. IUB code: R, A/G; Y, C/T; M, A/C; W, A/T; D, G/A/T; K, G/T.

viatilis. However, our data set did not allow an unambiguous conclusion, presumably because of too few cyclostome sequences. Since then, the nucleotide sequence of the mtDNA of another specimen of the hagfish, *Myxine glutinosa* (Delarbre *et al.*, 2001), and of another hagfish, *Eptatretus burgeri* (present paper), have been determined. We have used this extended data set, enriched in mtDNA sequences obtained from cyclostomes, to approach the issue of basal relationships among the vertebrates. We show, by using two statistical methods, that the monophyly of cyclostomes is strongly supported. Thus, all presently available molecular data lead to the conclusion that cyclostomes form a natural group, a conclusion that cannot easily be reconciled with the distribution of anatomical data.

MATERIALS AND METHODS

Isolation and Sequencing of mtDNA

The specimens of *Eptatretus burgeri* were collected in Kingo bay, Kagoshima prefecture, Japan. A fragment of skeletal muscle kept in 80% ethanol was a gift of the Muséum National d'Histoire Naturelle, Paris.

Total DNA was prepared from *E. burgeri* muscle using the proteinase K digestion procedure, as described elsewhere (Hogan *et al.*, 1994). Overlapping fragments of mtDNA were obtained by PCR amplification carried out on total DNA using degenerate primers. The sequences of the primers used for PCR amplification are indicated in Table 1. The general conditions for PCR were the following: 300 ng total DNA, 200 μM dNTP, 500 nM primers, and *Pfu* Exo+ polymerase (Stratagene, La Jolla, CA); the thermal cycles were 3 min at 94°C, followed by 50 cycles 1 min at 94°C, 1 min at 48 to 55°C depending upon the primers, 5 min at 72°C. The last cycle was ended by incubating for 10 min at 72°C. For the long or problematic DNA fragments of the control region, we used the Expand Long Template PCR System (Roche Molecular Biochemicals) with 300 ng total DNA, 500 μM dNTP, 500 nM primers, and buffer with 22.5 mM $MgCl_2$ and detergents. PCR products were purified by the QIAquick Gel Extraction Kit (Qiagen, Chatsworth, CA), phosphorylated using the polynucleotide kinase (Pharmacia, Piscataway, NJ), and ligated at the *Eco*RV site of dephosphorylated KS Bluescript vector (Stratagene) using the Rapid DNA Ligation kit (Roche Molecular Biochemicals). XL1Blue competent bacteria (Stratagene) were transformed with the ligation products and plasmid DNA of several recombinant clones was recovered using the QIAprep Spin Miniprep Kit (Qiagen). Cloned PCR products were sequenced with the Sequenase Version 2.0 DNA Sequencing Kit (USB), first using M13 (−40) and KS reverse primers, and subsequently using primer walking on both strands.

Phylogenetic Analyses

The data set used for phylogenetic analyses is composed of 13 taxa: *Paracentrotus lividus* (Cantatore *et al.,* 1989), *Asterina pectinifera* (Asakawa *et al.,* 1995), *Balanoglossus carnosus* (Castresana *et al.,* 1998), *Branchiostoma lanceolatum* (Spruyt *et al.,* 1998), *Branchiostoma floridae* (Boore *et al.,* 1999), *Myxine glutinosa* (Delarbre *et al.,* 2001), *Lampetra fluviatilis* (Delarbre *et al.,* 2000), *Petromyzon marinus* (Lee and Kocher, 1995), *Scyliorhinus canicula* (Delarbre *et al.,* 1998), *Mustelus manazo* (Cao *et al.,* 1998), *Cyprinus carpio* (Chang *et al.,* 1994), *Oncorhynchus mykiss* (Zardoya *et al.,* 1995), and *Eptatretus burgeri* (present paper). The 13 protein-coding sequences of the mtDNAs were first aligned manually following translation of the nucleotide sequences using the mitochondrial genetic code specific for each taxon. The alignment was refined using the physical amino acid similarity criterion. All inserted gaps were triplets. Standard procedures for coding gaps offer alternative strategies which suffer of several weaknesses; either the different sites are analyzed independently (gap = new state) so that each gap is artificially weighted relatively to the number of sites, or each site is coded "?" (gap = missing data) and the optimization procedure makes the whole zone devoid of phylogenetic information. Our coding strategy (gap = ID) is aimed at expressing the potential phylogenetic information contained in complex zones with internested insertions/deletions (indels) and substitutions (Barriel, 1994b and discussions in Lutzoni *et al.*, 2000, Simmons and Ochoterena, 2000). It precludes loss or distortion of information in all those cases in which gaps are present in the aligned sequences. Gaps are coded as multiple state characters: following an indel, shared subsequent mutations (substitutions as well as indels) should be interpreted in terms of common descent. According to the hierarchy of internested states of characters, this strategy introduces question marks "?" in the data matrix, which are optimized in cladograms based on all data. They are thus not missing data, but rather methodological codes, neutral to any *a priori* phylogenetic analysis (Barriel, 1994a, 1994b). Whenever alignment problems due to deletion or insertion events were locally encountered, a parsimony analysis search using PAUP was carried out on that zone in order to define the "optimal" alignment. The alignment was constructed by hand using the following criteria. (1) Minimize the number of inferred mutations (number of steps); (2) test the number of weighted mutations (one transition Ts will be preferred to one transversion Tv); (3) minimize the number of variable sites; and (4) minimize the *a priori* phylogenetic implications of the alignments (Barriel, 1994a). Thus, each alignment was tested giving transversions an equal weight (Tv = 1) or a double weight (Tv = 2) to transitions for three different codings of the gaps (gap = missing data ("?"), gap = new state, and gap = ID).

Two different analyses were carried out on nucleotide and amino acid sequences alignments. A chi-square test of homogeneity of base frequencies across taxa was run using PAUP (version 4.0b3) (Swofford, 1999 PAUP*) for both datasets, and indicated a stationarity of base frequencies across taxa.

a. Maximum-parsimony (MP) method was applied, using branch-and-bound (13 taxa) searches implemented in PAUP (version 4.0b3) (Swofford, 1999 PAUP*) to determine the most parsimonious tree. The robustness of the MP trees was tested using the bootstrap method with 200 replications each and using the Bremer support (Bremer, 1988).

b. PAUP (version 4.0b3) was used for maximum-likelihood (ML) analysis of DNA sequences and boostrap values were computed using fastDNAml program (Felsenstein, 1981; Olsen *et al.,* 1994) using the F84 model (TS/TV = 2) (Felsenstein, 1984). The robustness of the DNA ML tree was tested using the bootstrap method and 200 replications. The protml program (version 2.3b3) (Adachi and Hasegawa, 1995) was used for ML analysis of amino acid sequences using the mtREV24-F model (Adachi and Hasegawa, 1996) which has been adapted to the analyses of mtDNA. The robustness of the amino acids ML tree was tested using the bootstrap method and 1000 replications. The comparison of competing topologies was carried out using the Shimodaira-Hasegawa test (Shimodaira and Hasegawa, 1999), using bootstrap with full optimization (1000 bootstrap replicates).

RESULTS

Genomic Organization of the mtDNA of Eptatretus burgeri

The mtDNA of *Eptatretus burgeri* (accession number: AJ278504) was 17168 bp long and included the 13 usual protein-coding genes, 22 tRNAs, 12S and 16S rRNAs, and a noncoding sequence corresponding to the control region (Table 2). The overall gene map of the mtDNA was identical to the "common" vertebrate mtDNA gene arrangement. As in all other cyclostomes studied (Delarbre *et al.,* 1997), the origin of replication of the light chain of the mtDNA could not be identified between the *ND1* and *ND2* genes. The most remarkable feature of the mtDNA was the control region. The control region, 1858-bp long, located between *tRNA-Pro* and *tRNA-Phe,* was longer than in most other vertebrates (around 1000 bp). The 5′ part of the control region showed two identical 257-bp long repeats, a third 256-bp long repeat with a missing T, and a fourth partial 134-bp long, repeat. The first three repeats were found to contain a TACATTTTAT sequence very similar to a termination associated sequence (TAS). A

TABLE 2

Localization of the Genes in the Mitochondrial Genome of *Eptatretus burgeri*

Genes	From	To	Size bp	Size aa	Codons Start	Codons Stop
NDI	1	957	957	318	ATG	TAG
tRNA-Ile	957	1022	66			
tRNA-Gln	1028	1098	71 (L)			
tRNA-Met	1103	1170	68			
ND2	1183	2226	1044	347	ATA	TAA
tRNA-Trp	2225	2289	65			
tRNA-Ala	2289	2355	67 (L)			
tRNA-Asn	2356	2421	66 (L)			
tRNA-Cys	2422	2486	65 (L)			
tRNA-Tyr	2495	2560	66 (L)			
COI	2570	4123	1554	517	ATG	AGG
tRNA-Ser (TCN)	4111	4181	71 (L)			
tRNA-Asp	4187	4252	66			
COII	4255	4944	690	229	ATG	AGA
tRNA-Lys	4937	5000	64			
ATPase 8	5003	5167	165	54	ATG	TAA
ATPase 6	5161	5847	687	228	ATG	TAA
COIII	5850	6635	786	261	ATG	TAA
tRNA-Gly	6635	6701	67			
ND3	6720	7069	350	116	ATA	TA-
tRNA-Arg	7070	7134	65			
ND4L	7135	7425	291	96	ATA	TAA
ND4	7419	8795	1377	458	ATG	TAA
tRNA-His	8791	8855	65			
tRNA-Ser (AGY)	8856	8916	61			
tRNA-Leu (CTN)	8919	8989	71			
ND5	8990	10792	1803	600	ATT	TAA
ND6	10788	11291	504 (L)	167	ATG	AGG
tRNA-Glu	11293	11359	67 (L)			
CYTb	11361	12518	1158	385	ATG	TAA
tRNA-Thr	12542	12611	70			
tRNA-Pro	12617	12687	71 (L)			
Noncoding	12688	14545	1858			
tRNA-Phe	14546	14615	70			
12S rRNA	14616	15474	859			
tRNA-Val	15475	15545	71			
16S rRNA	15546	17091	1546			
tRNA-Leu (TTR)	17092	17166	75			

Note. ATPase 6 and 8: ATP synthase subunits 6 and 8; COI, II, and III: cytochrome *c* oxidase subunits I, II, and III; CYTb: cytochrome b; ND1-6. NADH dehydrogenase subunits 1 to 6. (L) means that the gene is located on the L strand of the mtDNA.

fourth, slightly different TAS sequence (TA-CATATTTA) was identified in the fourth repeat. The potentially regulatory sequence conserved sequence block CSBI, AGGACATA, was located between nucleotides 14243 and 14250. By contrast, no conserved sequence blocks CSBII and CSBIII could be identified. Finally, a total of 17 possible hairpins and six palindromes were evidenced within the control region sequence of *Eptatretus,* suggesting a densely packed local structure of the mtDNA. The protein and tRNA coding genes did not show features that would be unique to *Eptatretus*. For protein-coding genes, the initiation codons were ATG or ATA in all genes with the exception of *ND5* gene where it was ATT. TAA was the most frequently used stop codon, but that of the *ND3* gene was incomplete (TA-) and must be completed by polyadenylation. However, the termination codons were AGG or AGA for the *COI, COII,* and *ND6* genes. Finally, as in all vertebrate mtDNAs studied so far, some genes were found to partially overlap: the *tRNA-Ser (TCN)* gene overlapped the *COI* gene by 13 nucleotides, the *tRNA-Lys* gene overlapped by 8 nt the *COII* gene, and the *tRNA-His* gene overlapped by 5 nt the *ND4* gene; *ATPase 8* and *ATPase 6* genes overlapped by 7 nt, *ND4L* and *ND4* genes by 7 nt, and *ND5* and *ND6* genes by 5 nt.

Comparison of the mtDNAs of the Two Hagfishes, Eptatretus burgeri and Myxine glutinosa

The mtDNA of *Eptatretus burgeri* was shorter than that of the other hagfish already determined, *Myxine glutinosa* (18909 bp) (Delarbre *et al.,* 2001; accession number: AJ404477) due to the size of the control region of the latter (3628 bp). Excluding the control region, the lengths of the two mtDNAs were quite similar: 15310 and 15281 bp for *Eptatretus* and *Myxine,* respectively. The detailed gene maps were identical, and thus both exhibit the common vertebrate arrangement. The overlaps found in *Eptatretus* between the tRNAs and the protein-coding genes were precisely the same as those observed in the mtDNA of *Myxine,* and, otherwise, sizes and features of the genes of the two hagfishes differed only by details. This observation extends to initiation and termination codons. The initiation codons were ATG or ATA for all genes of the two hagfishes with the exception of *ND5* in *Eptatretus* (ATT) and *ND4L* in *Myxine* (ATC) and the AGG or AGA stop codons were used by the same genes (*COI, COII,* and *ND6*) in the two animals. The codon usage by the two hagfishes was slightly different concerning the fourfold degenerate codons with A being more frequently used than C and G by *Eptatretus*. On the whole (Table 3), the uncorrected percent sequence divergence between the two hagfishes at the nucleotide level was 18.8 to 35.8% (with an average of 24.2%). At the amino acid level, the average divergence was 17.5% with a minimum value of 6.3% for COI and a maximum value of 40.7% for ATPase 8. Nonsilent substitutions between *Eptatretus* and *Myxine* were more frequent in *ND6* and *ATPase 8* genes with 50% of the total number of substitutions and with only 13% for the most conserved *COI* gene.

The tRNA-coding genes were 84% identical. The *tRNAs-Tyr* and *-Ser (TCA)* showed the highest similarity (93.9 and 92.9% identity, respectively) in *Eptatretus* and *Myxine,* whereas *tRNA-Ser (AGY)* gene was the most divergent with a lower 67.2% identity. The rRNA genes were 859 and 862 bp long for the 12S rRNA and 1546 bp and 1541 bp long for the 16S rRNA for *Ep-*

TABLE 3

Comparison between Mitochondrial Protein-Coding Sequences of *Myxine* and *Eptatretus*

	% of divergence		Number of substitutions		
	Nucleotides	Amino acids	Total	Non silent	% of nonsilent
ND1	21.7	12.2	207	49	23.7
ND2	26.7	20.1	279	90	32.2
COI	18.8	6.3	292	38	13
COII	19.9	13.4	137	31	22.6
ATP8	35.8	40.7	58	29	50
ATP6	27.8	21.5	189	71	37.5
COIII	20.8	14.1	163	41	25.1
ND3	27.3	20	95	38	40
ND4L	26	16.7	75	18	24
ND4	23.5	14.6	323	80	24.7
ND5	25.4	21.4	458	155	33.8
ND6	29.2	30.3	148	75	50.7
CYTb	28	24.4	324	125	38.6

tatretus and *Myxine*, respectively. The percentages of similarity of the 12S and 16S rRNAs of *Eptatretus* and *Myxine* were 85.5 and 84.3%, respectively.

The control region of the mtDNA of *Eptatretus* was much shorter than that of *Myxine* (1858 bp vs 3628 bp) due to a shortened 3′ region of the control region of *Eptatretus*, which lacks the short repeats found in *Myxine*. With the exception of interspersed short DNA stretches, the sequences of the control regions of the two hagfishes could not be aligned. However, they have retained the same general organization, as an impressive array of repeated sequences and regulatory sequences able to fold into a highly compact structure. Indeed, long repeated sequences were found in the 5′ part of both control regions, including four TAS sequences. Three of the latter (TACATTTTAT) were found identical in the two animals and only the fourth TAS sequence of *Eptatretus* was found slightly different (TACATATTTA). A part of the CSBI sequence (AGGACATATATATTAA) was identical in *Eptatretus* and *Myxine*. In both hagfishes these CSBI sequences were preceded by a T-rich sequence. C-rich sequences (22 bp for *Eptatretus* and 25 bp for *Myxine*) and T-rich sequences followed by several G were observed in the two hagfishes upstream CSBI. On the whole, the control region of *Eptatretus* was found less rich in secondary structures (17 hairpins and six palindromes) than that of *Myxine* (37 hairpins and two palindromes), and weakly similar in terms of nucleotide sequence, but able to fold into a quite similar secondary structure.

We have previously shown the high similarity of the mtDNAs of the two lampreys *Petromyzon* and *Lampetra*. The percentages of identity between *Lampetra* et *Petromyzon* were 96.3% for the 12S rRNA, 93.9% for the 16S rRNA, 94.7% for the tRNAs, 85.8% for the protein-coding genes, and about 90% for the control regions (Delarbre *et al.*, 2000). By contrast, the above features of *Myxine* and *Eptatretus* were suggestive of a greater heterogeneity of the hagfishes than of the lampreys.

Genome Organization

We then compared the mtDNAs of the two representative groups of the cyclostomes. First, the molecular maps of the mtDNAs differed by the location of the control region between *ND6* and *CYTb* genes in lampreys and between *CYTb* and *12S rRNA* genes in hagfishes. Hagfishes demonstrate a common vertebrate pattern. Second, the control regions of the two groups of cyclostomes were profoundly different in size and fine organization. The control regions of lampreys were shorter and divided into two parts separated by *tRNA-Glu* and -*Thr* genes. They contained no TAS and CSBI conserved sequences and did not show stable secondary structures. Third, the nucleotide sequences of 12S and 16S rRNAs and the protein-coding genes of the hagfishes were difficult to align with the corresponding genes of the lampreys. Fourth and more importantly, whereas the percentage of similarity of the amino acid sequences of the 13 mitochondrial proteins taken together of *Petromyzon* vs *Lampetra* was 92.5%, and that of *Myxine* vs *Eptatretus* was 82.5%, the percentage of similarity for all the mitochondrial amino acid sequences between *Lampetra* and *Myxine* was found to be as low as 53%. Some additional differences between the mtDNA of the lampreys and the hagfishes could also be noted: the codon usage and the amino acid composition of the protein-coding genes of *M. glutinosa* and *E. burgeri* was found to be markedly different from that of *Lampetra* and *Petromyzon*. Codons CTA, GTT, ACT, ACA, GCT, GCC, GCA, and CAA were more frequent in lampreys, whereas TCT, TCC, TCA, TAC, and TTT were more often used in hagfishes. Serine and phenylalanine were more frequently used by hagfishes than by lampreys, in contrast with threonine, alanine, and glutamine with an opposite usage.

On the whole, despite a higher heterogeneity of the hagfishes, the mtDNAs of lampreys and hagfishes have very different genome organizations.

Phylogenetic Analysis

The divergence noted between hagfishes and lampreys is not indicative of their respective phylogenetic relationships with other craniates. Thus, a phylogenetic study was carried out using the mtDNA of 13 taxa “flanking” the hagfish/lamprey node. Because we had shown in an earlier paper (Delarbre *et al.*, 2000) that the use of the subset of data defined following Naylor and Brown criteria (Naylor and Brown, 1998) was misleading for phylogenetic analysis, the analysis was carried out on the complete data set of protein-coding genes.

The 13 taxa retained for analysis have been deliberately chosen for the following reasons. (1) They surrounded the node of interest. (2) Only two species for each high rank taxa (i.e. Actinopterygii, Elasmobranchii, Hyperoartia, Hyperotreta, Cephalochordata, Hemichordata (only one taxon is available), and Echinodermata as outgroups) were retained to maintain a balanced amount of informative sites among high rank taxa. (3) They constituted a "continuous" series of taxa with no sampling gap among them. (4) They were all cold-blooded animals living in an aquatic environment.

Nucleotide sequences analyses. The complete data set (13 protein-coding genes) was composed of 13 taxa, and 11787 nucleotides were used as sites (3405 invariant and 7632 parsimony informative sites). Two echinoderm taxa were used as outgroups: *Asterina pectinifera* and *Paracentrotus lividus*. The results are summarized in Fig. 1.

Using the maximum-parsimony (MP) method, giving equal weight to transitions and transversions (TV1) and gaps treated as missing data, a single most parsimonious tree was obtained in which cyclostomes were monophyletic (length = 29213 steps, CI = 0.555, and RI = 0.509). All the nodes were supported by a 100% bootstrap value, except for the cyclostome node (64% bootstrap value). If gaps were recoded (Barriel, 1994a, 1994b) (gap = ID, 11891 sites, 3405 invariant, 7701 parsimony informative sites), the same unique tree was computed (length = 29337 steps, CI = 0.556, RI = 0.510) with bootstrap values of 100% for all nodes except for the cyclostome node for which it was 55%. The same tree was obtained when transversions were given double the weight of transitions (TV2) and gaps treated as missing data (length = 44119 steps with bootstrap values 100% for all nodes except for the cyclostome node: 66%). Finally, retaining only the first and second codon positions, MP analysis yielded the same unique tree using TV1 (with bootstrap values of 100% for all nodes except for the cyclostome node for which it was 70 and 86% for craniate node) or TV2 (with bootstrap values of 100% for all nodes except for the cyclostome node for which it was 67 and 98% for craniate node).

Upon maximum-likelihood analysis using PAUP□, the same data set resulted in the same unique tree (−ln L = 125217.32) in which cyclostomes were monophyletic. Bootstrap values were 100% for all nodes except 96% for the node supporting the cyclostomes. Using the Shimodaira-Hasegawa test, the likelihood of the tree supporting the monophyly of cyclostomes is significantly higher than that supporting the paraphyly of cyclostomes.

Amino acid sequences analyses. The amino acid sequences, previously aligned as a requisite in the alignment of the nucleotide sequences, were also analyzed using the same two phylogenetic methods. Maximum-parsimony analysis (3930 sites, 1225 constant and 2397 parsimony informative sites) yielded a single parsimonious tree in which cyclostomes were monophy-

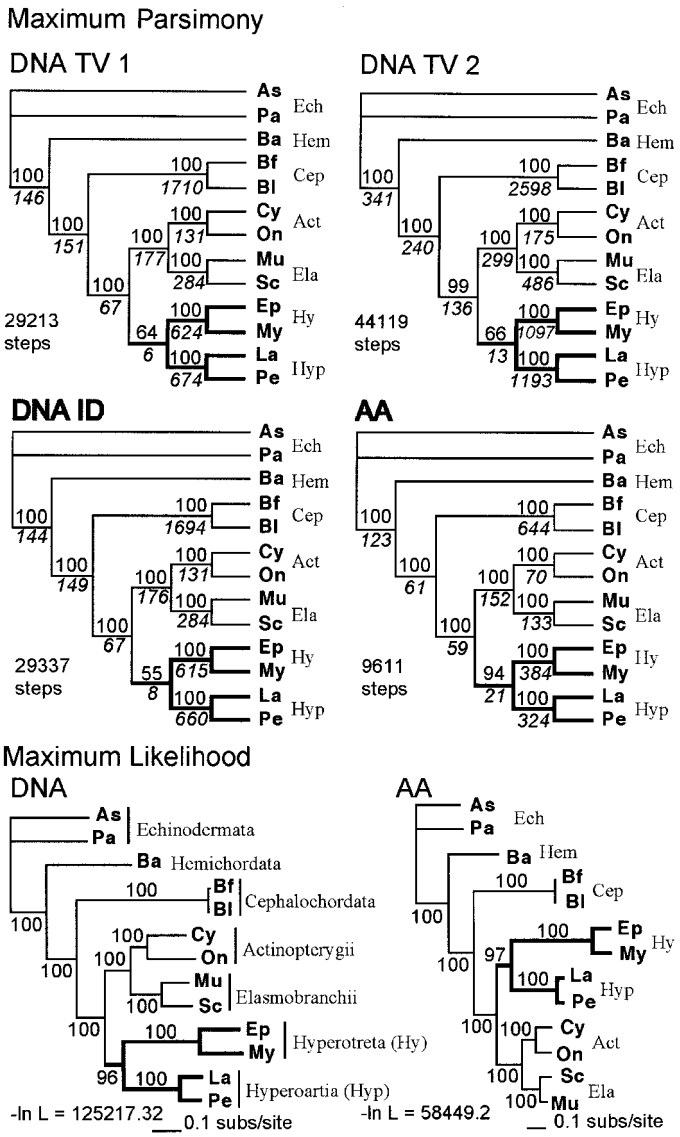


FIG. 1. Phylogenetic position of cyclostomes was determined by using maximum-parsimony and maximum-likelihood strategies. Details of the computations are given in Materials and Methods. "DNA" refers to 11787 aligned positions of the protein coding genes. "AA" refers to 3930 aligned amino acids. "TV1" and "TV2" correspond to the weight attributed to transversions. "ID" refers to gaps treated as indels. The abbreviated names of the animals are (Pa) *Paracentrotus lividus,* (As) *Asterina pectinifera,* (Ba) *Balanoglossus carnosus,* (Bl) *Branchiostoma lanceolatum,* (Bf) *Branchiostoma floridae,* (My) *Myxine glutinosa,* (Ep) *Eptatretus burgeri,* (La) *Lampetra fluviatilis,* (Pe) *Petromyzon marinus,* (Sc) *Scyliorhinus canicula,* (Mu) *Mustelus manazo,* (Cy) *Cyprinus carpio,* and (On) *Oncorhynchus mykiss.* In the frame corresponding to trees computed using the maximum-parsimony strategy, the values above the lines are boostrap values; numbers below the lines (italics) indicate the Bremer support, i.e., the number of steps needed for a node to disappear. In the trees computed using the maximum-likelihood strategy, only boostrap values are given

letic (9519 steps, CI = 0.8125, and RI = 0.7513). Bootstrap values were 96% for the cyclostome node and 100% for all other nodes. Maximum-likelihood analysis also yielded the same unique tree (-In L = 58449.2), in which cyclostomes were monophyletic with bootstrap values of 97% for the cyclostome node.

DISCUSSION

The choice of the taxa we have used may be questioned. As mentioned in the text, we have kept to a balanced set of data surrounding the node of interest. This choice was justified by the results of similar analyses carried out on a 16 taxa data set, including three representative taxa of tetrapods. The use of the same 13 taxa as above, combined with the use of an amphibian *Typhlonectes natans* (Zardoya and Meyer, 2000), the chicken *Gallus gallus* (Desjardins and Morais, 1990), and the bonobo *Pan paniscus* (Horai *et al.*, 1995), led to novel nucleotide and amino acid alignments. We performed the same analyses as before (MP and ML) on these new alignments except that we used heuristic search in place of branch-and-bound search to determine the most parsimonious tree. Whatever the methods, models and weighting were used, phylogenetic trees (data available upon request) were diverse with, however, several shared traits. In addition to the constant monophyly of craniates, osteichthyes and chondrichthyes always unexpectedly appeared as sister groups. Concerning the cyclostome node, the results sustained either monophyly and paraphyly with equal weakness. Finally, the relationships among tetrapods were also found unstable. The generation of phylogenetic trees which were not compatible with accepted phylogeny of vertebrates justifies the choice of our data set.

In a previous study aimed at the phylogenetic resolution of the cyclostomes, we concluded that additional mtDNA sequences of hagfishes are required to reach unambiguous conclusions (Delarbre *et al.*, 2000). We have determined the complete mtDNA sequence of a different hagfish, *Eptatretus burgeri*, to generate the same number of sequences for hagfishes (*Myxine* and *Eptatretus*) and lampreys (*Petromyzon* and *Lampetra*). Using these new data, we reached the following conclusions: lampreys and hagfishes have very different mtDNA gene organization and sequence similarity. The phylogenetic analysis of the thus newly defined set of data including four cyclostomes, using two different methods and analyzing nucleotide or amino acid alignments, supports unequivocally the monophyly of cyclostomes, i.e., that lampreys and hagfishes form a clade.

Such a conclusion strengthens earlier findings which supported the monophyly of cyclostomes. These included the study of 28S and 18S rRNA sequences (Mallatt and Sullivan, 1998; Stock and Whitt, 1992), nuclear DNA-coded single copy genes (Kuraku *et al.*, 1999), and sets of duplicated proteins (Suga *et al.*, 1999). An earlier study (Rasmussen *et al.*, 1998) based on the mtDNA of *Myxine glutinosa* supported the paraphyly of cyclostomes, but the choice of the data set and the way the data were exploited, were questionable (Mallatt and Sullivan, 1998). Thus, our present data combined with earlier molecular studies of nuclear genes point to the probable monophyly of cyclostomes.

This conclusion obviously conflicts with the hypothesis of cyclostome paraphyly, although the "vertebrate hypothesis" is supported by the largest number of morphological and physiological characters, even when certain cyclostome characters, such as the "lingual" apparatus, are split into several discrete characters (Yalden, 1985). Such a basic conflict between opposite results generated by using either morphological or molecular data, is not unusual, and has recently been discussed in detail. In the present case, the unique anatomical and physiological characters shared by lampreys and gnathostomes should be reexamined. Beyond the number of the characters involved in the analysis, their functional significance may be regarded as a weighting criterion, in particular when there is no other instance of loss or homoplasy on these characters among all other vertebrates. For example, the absence of cardiac innervation in hagfishes could be reconciled with the "cyclostome hypothesis" by considering that cardiac innervation has either been lost in hagfishes or developed independently, in both lampreys and gnathostomes. In both cases, the process involved is problematic. Cardiac innervation is per se a remarkable character and involves a profound modification of the autonomic nerve system patterning during ontogeny. However, its loss (and replacement by a hormonal control) may be even less likely, although Nilsson and Hølmgren (1998) favor this explanation (without justifying their opinion). In this particular case, one may note that outgroup comparison with cephalochordates supports the plesiomorphous state of the aneural heart, as the ventral artery of amphioxus that plays the same role as the craniate heart is under hormonal control exclusively. In addition, the embryonic heart of all other craniates is also aneural. This ontogenetic argument of polarity may, however, be used to invoke reversion through paedomorphy, as it could also be the case for several other character states of hagfishes (e.g., absence of eye lens, disseminate pancreas, lack of neuromasts, etc.). This would lead us back to the old concept of "degeneracy" of hagfishes (and lampreys) that has pervaded the literature on cyclostomes since Linnaeus. If all the apparently plesiomorphous character states of hagfishes are actually the result of losses, possibly through paedomorphy, then hagfishes represent the most extraordinary case of overall anatomical and physiological reversion among craniates.

Conversely, if all the characters shared uniquely by lampreys and gnathostomes were homoplastic, then this would be the most impressive case of homoplasy in vertebrates.

ACKNOWLEDGMENTS

C. Gallut acknowledges receipt of a fellowship of the Ministère de l'Education Nationale, de la Recherche et de la Technologie. The work was carried out in the "Unité de Biologie Moléculaire du Gène" and has been supported by the EEC, the Institut National de la Santé et de la Recherche Médicale, the College de France, and the Institut Pasteur.

REFERENCES

Adachi, J., and Hasegawa, M. (1995). MOLPHY: Programs for Molecular Phylogenetics Ver. 2.3, Institute of Statistical Mathematics, Tokyo.

Adachi, J., and Hasegawa, M. (1996). Model of amino acid substitution in proteins encoded by mitochondrial DNA. *J. Mol. Evol.* **42:** 459–468.

Asakawa, S., Himeno, H., Miura, K.-I., and Watanabe, K. (1995). Nucleotide sequence and gene organization of the starfish *Asterina pectinifera* mitochondrial genome. *Genetics* **140:** 1047–1060.

Barriel, V. (1994a). Relations de parenté au sein des Hominoidea et la place de *Pan paniscus*. Comparison et analyse méthodologique des phylogénies morphologique et moléculaire. Ph.D. thesis, University of Paris VI.

Barriel, V. (1994b). Phylogénies moléculaires et insertion délétion de nucléotides. *C.R. Acad. Sci. Paris, Ser.III (Sci. Vie)* **317:** 693–701.

Boore, J. L., Daehler, L. L., and Brown, W. M. (1999). Complete sequence, gene arrangements and genetic code of the mitochondrial DNA of the cephalochordate *Branchiostoma floridae* (Amphioxus). *Mol. Biol. Evol.* **16:** 410–418.

Bremer, K., (1988). The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. *Evolution* **42:** 795–803.

Cantatore, P., Roberti, M., Rainaldi, G., Gadaleta, M. N., and Saccone, C. (1989). The complete nucleotide sequence, gene organization and genetic code of the mitochondrial genome of *Paracentrotus lividus*. *J. Biol. Chem*. **264:** 10695–10975.

Cao, Y., Waddell, P. J., Okada, N., and Hasegawa, M. (1998). The complete mitochondrial DNA sequence of the shark *Mustelus manazo*: Evaluating rooting contradictions to living bony vertebrates. *Mol. Biol. Evol.* **15:** 1637–1646.

Castresana, J., Feldmaier-Fuchs, G., and Pääbo, S. (1998). Codon reassignment and amino acid composition in hemichordate mitochondria. *Proc. Natl. Acad.Sci. USA* **95:** 3703–3707.

Chang, Y.-C., Huang, F.-L., and Lo, T.-B. (1994). The complete nucleotide sequence and gene organization of carp (*Cyprinus carpio*) mitochondrial genome. *J. Mol. Evol.* **38:** 138–155.

Delarbre, C., Barriel, V., Tillier, S., Janvier, P., and Gachelin G. (1997). The main features of the craniate mitochondrial DNA between the ND1 and the COI genes were established in the common ancestor to the lancelet. *Mol. Biol. Evol.* **14:** 807–813

Delarbre, C., Spruyt, N., Delmarre, C., Gallut, C., Barriel, V., Janvier, P., Laudet, V., and Gachelin, G. (1998). The complete nucleotide sequence of the mitochondrial DNA of the dogfish, *Scyliorhinus canicula*. *Genetics* **150:** 331–344.

Delarbre, C., Escriva, H., Gallut, C., Barriel, V., Kourilsky, P., Laudet V., and Gachelin G. (2000). The complete nucleotide sequence of the mitochondrial DNA of the agnathan *Lampetra fluviatilis*: Bearings on the phylogeny of cyclostomes. *Mol. Biol. Evol.* **17:** 519–529.

Delarbre, C., Rasmussen, A.-S., Arnason, U., and Gachelin, G. (2001). The complete mitochondrial genome of the hagfish *Myxine glutinosa:* Unique features of the control region. *J. Mol. Evol.*, in the press.

Desjardins, P., and Morais, R. (1990). Sequence and gene organization of the chicken mitochondrial genome. A novel gene order in higher vertebrates. *J. Mol. Biol.* **212:** 599–634.

Donoghue, P. C. J., Forey, P. L., and Aldridge, R. J. (2000). Conodont affinity and chordate phylogeny. *Biol. Rev.* **75:** 191–251.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* **17:** 368–376.

Felsenstein, J. (1984). Distance methods for inferring phylogenies: A justification. *Evolution* **38:** 16–24.

Forey, P. L. (1984). Yet more reflections on agnathan–gnathostome relationships. *J. Vertebr. Paleontol.* **4:** 330–343.

Forey, P. L., and Janvier, P. (1993). Agnathans and the origin of jawed vertebrates. *Nature (London)* **361:** 129–134.

Gagnier, P. Y. (1993). *Sacabambaspis janvieri,* Vertébré ordovicien de Bolivie. 2. Analyse phylogénétique. *Ann. Paléontol.* **79:** 119–166.

Hogan, B., Beddington, R., Costantini, F. R., and Lacy, E. (1994). "Isolating High Molecular Weight DNA from Mouse Tails. Manipulating the Mouse Embryo, A Laboratory Manual," Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Horai, S., Hayasaka, K., Kondo, R., Tsugane, K., and Takahata, N. (1995). Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. *Proc. Natl. Acad. Sci. USA* **92:** 532–536.

Janvier, P. (1978). Les nageoires paires des ostéostracés et la position systématique des céphalaspidomorphes. *Ann. Paléontol.* **64:** 113–142.

Janvier, P. (1981). The phylogeny of the Craniata, with particular reference to the significance of the fossil agnathans. *J. Vertebr. Paleontol.* **1:** 121–159.

Janvier, P. (1996). "Early Vertebrates," Oxford University Press, Oxford.

Kuraku, S., Hoshiyama, D., Katoh, K., Suga, H., and Miyata, T. (1999). Monophyly of lampreys and hagfishes supported by nuclear DNA-coded genes. *J. Mol. Evol.* **49:** 729–735.

Lee, W.-J., and Kocher, T. D. (1995). Complete sequence of a sea lamprey (*Petromyzon marinus*) mitochondrial genome: Early establishment of the vertebrate genome organization. *Genetics* **139:** 873–887.

Lovtrup, S. 1977. "The Phylogeny of Vertebrata," Wiley, New York.

Lutzoni, F., Wagner, P., Reeb, V., and Zoller, S. (2000). Integrating ambigously aligned regions of DNA sequences in phylogenetic analyses without violating positional homology. *Syst. Biol.* **49:** 628–651.

Mallatt, J., and Sullivan, J. (1998). 28S and 18S rDNA sequences support the monophyly of lampreys and hagfishes. *Mol. Biol. Evol.* **15:** 1706–1718.

Naylor, G. J. P., and Brown, W. M. (1998). Amphioxus mitochondrial DNA, chordate phylogeny and the limits of inference based on comparisons of sequences. *Syst. Biol.* **47:** 61–76.

Nilsson, S., and Holmgren, S. (1998). "The Biology of Hagfishes" (J. M. Jorgensen, J. P. Lomholt, R. E. Weber, and H. Malte, Eds.), pp. 478–493. Chapman and Hall, London.

Olsen, G. J., Matsuda, H., Hagstrom, R., and Overbeek, R. (1994). FastDNAml version 1.2: A tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. *Comput. Appl. Biosci.* **10:** 41–48.

Rasmussen, A.-S., Janke, A., and Arnason, U. (1998). The mitochon-

drial DNA molecule of the hagfish (*Myxine glutinosa*) and vertebrate phylogeny. *J. Mol. Evol.* **46:** 382–388.

Rovainen, C. M. (1996). Neurobiology of jawless fishes. *Brain Behav. Evol.* **48:** 235–236.

Schaeffer, B., and Thomson, K. S. (1980). "Aspects of Vertebrate Life" (L. L. Jacobs, Ed.), pp. 787–789. Museum of Northern Arizona Press, Flagstaff, AE.

Shimodaira, H., and Hasegawa, M. (1999). Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* **16:** 1114–1116.

Simmons, M. P., and Ochoterena, H. (2000). Gaps as characters in sequence-based phylogenetic analyses. *Syst. Biol.* **49:** 369–381.

Spruyt, N., Delarbre, C., Gachelin, G., and Laudet, V. (1998). Complete sequence of the amphioxus (*Branchiostoma lanceolatum*) mitochondrial genome: Relations to vertebrates. *Nucleic Acids Res.* **26:** 3279–3285.

Stock, D. W., and Whitt, G. S. (1992). Evidence from 18S ribosomal RNA sequences that lampreys and hagfishes form a natural group. *Science* **257:** 787–789.

Suga, H., Hoshiyama, D., Kuraku, S., Katoh, K., Kubokawa, K., and Miyata, T. (1999). Protein tyrosine kinase cDNAs from amphioxus, hagfish, and lamprey: Isoform duplications around the divergence of cyclostomes and gnathostomes. *J. Mol. Evol.* **49:** 601–608.

Swofford, D. L. (1999). PAUP*. Phylogenetic analysis using parsimony *. Version 4.0. Sinauer Associates, Sunderland, MA.

Yalden, D. W. (1985). Feedings mecanisms as evidence for cyclostome monophyly. *Zool. J. Linn. Soc.* **84:** 291–300.

Zardoya, R., Garrido-Pertierra, A., and Bautista, J. M. (1995). The complete nucleotide sequence of the mitochondrial DNA genome of the rainbow trout, *Oncorhynchus mykiss. J. Mol. Evol.* **41:** 942–951.

Zardoya, R., and Meyer, A. (2000). Mitochondrial evidence on the phylogenetic position of caecilians. *Genetics* **155:** 765–775.

Résumé

Une approche originale de reconstruction phylogénétique à partir de l'organisation de génomes entiers dans un contexte cladistique est proposée. Cette approche se fonde sur la comparaison globale de l'organisation des génomes étudiés, sans hypothèses *a priori* sur les remaniements. Deux codages sont proposés : « Position relative » et « Jonctions », avec deux options différentes. Ces possibilités de codage sont analysées et comparées avec le codage « Jonctions signées » de Cosner *et al.* ; ils permettent de représenter l'ordre d'unités fonctionnelles le long des chromosomes, les unités fonctionnelles pouvant être des gènes ou des segments chromosomiques homologues par exemple. Le codage « Position relative » regroupe trois types de caractères : des caractères de position, d'orientation et de présence/absence des unités. Le codage jonction est basé sur des caractères binaires : polarité d'unité, présence/absence d'unités et de jonctions d'unités. Le codage « Jonctions signées » est un codage entièrement fondé sur la présence/absence de jonctions signées. Ce dernier présente l'avantage de prendre en compte l'orientation et la position simultanément. Par contre, le codage « Position relative » permet de reconstituer les génomes ancestraux *a posteriori*, ce qui est d'un grand intérêt pour l'interprétation de l'évolution du génome. Le génome mitochondrial des métazoaires ainsi que les chromosomes du genre *Mastomys* ont été analysés avec succès au moyen des trois codages.

Mots-clés : Cladistique, phylogénie, codage, génomique, chromosome, évolution, ordre de gènes, génome mitochondrial.

Abstract

An original approach of phylogenetic reconstruction from whole genomes' organization in a cladistic framework is proposed. This approach is based on global comparison of genomes' organization with no a priori assumptions about rearrangements. Two codings are proposed: "Relative position" and "Junctions", with two different options. These coding possibilities are analyzed and compared along with the "Signed junctions" coding of Cosner et al.; they allow to represent the functional units order along the chromosome, as an example, functional units can be gene or homologous chromosomal segments. The "Relative position" coding is composed of three different kinds of characters: position characters, polarity and units presence/absence characters. The "Junctions" coding is based on binary characters: units' polarity, units and units junctions presence/absence. The "Signed junctions" coding is only composed of signed junctions presence/absence characters. This coding has the advantage of considering polarity and position simultaneously. On the other hand, the "Relative position" coding allows reconstituting ancestral genomes a posteriori, which is of great interest to interpreting genome evolution. The metazoan mitochondrial genome and *Mastomys* genus chromosomes have been successfully analyzed with the three coding approaches.

Keywords: Cladistic, phylogeny, coding, genomic, chromosome, evolution, gene order, mitochondrial genome.