Université de Montréal

Phylogénétique basée sur les cassures du génome

par

Mathieu Blanchette

Département d'informatique et de recherche opérationnelle

Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en informatique

Juin, 1998

Université de Montréal

Faculté des études supérieures

Ce mémoire intitulé:

Phylogénétique basée sur les cassures du génome

présenté par:

Mathieu Blanchette

a été évalué par un jury composé des personnes suivantes:

Jacques Ferland, président-rapporteur

David Sankoff, directeur de recherche

Dana Hahn, membre du jury

François Major co-directeur

Mémoire accepté le: 30.10.1998

Sommaire

La phylogénétique est la science qui tente de reconstituer le processus d'évolution ayant mené aux espèces contemporaines. Cette étude se fait aujourd'hui majoritairement à partir d'informations génétiques: séquences d'ADN, ordre des gènes sur les chromosomes, etc. L'ordre des gènes diffère d'une espèce à l'autre à cause de mutations appelées réarrangements du génome. Cet ordre est à la base de la méthode présentée dans ce mémoire.

Toutes les méthodes de construction d'arbres phylogénétiques reposent sur la définition d'une métrique donnant la distance évolutive entre deux espèces. Dans le cas des réarrangements du génome, la métrique la plus utilisée est le nombre minimal de réarrangements permettant de transformer un génome en un autre. Cette métrique est habituellement difficile à calculer. Ce mémoire propose et discute de l'utilisation d'une autre métrique: le nombre de cassures du génome. Cette métrique est très facile à calculer, et garde une signification biologique intéressante. Cette simplicité de calcul permet d'espérer des solutions à des problèmes plus difficiles: celui de la médiane de k génomes (où le problème est de trouver le génome à distance minimale de k génomes donnés), et sa version plus générale, celui des arbres de Steiner dans l'espace des génomes (où le problème est de trouver les génomes associés aux points de divergence d'un arbre de façon à minimiser la taille totale de l'arbre). Ces deux problèmes semblent extrêmement complexes sous la métrique du nombre minimal de réarrangements, mais légèrement plus simple avec le nombre de cassures du génome.

Ce mémoire présente un algorithme permettant de trouver la solution exacte au problème de la médiane de k génomes sous la métrique du nombre de cassures du génome, en le transformant en un problème bien connu, celui du commis-voyageur. Le cas des génomes orientés et non-orientés est traité, de même que celui de génomes ne contenant pas le même ensemble de gènes. Des méthodes heuristiques sont proposées pour traiter le problème des arbres de Steiner. La

performance de ceux-ci est évaluée, et les solutions obtenues sont comparées entre elles.

Les algorithmes développés sont ensuite appliqués à un cas réel, celui du génome mitochondrial des animaux. La méthode proposée permet de résoudre certains problèmes phylogénétiques liés à l'évolution des animaux rencontrés lors de l'utilisation des méthodes classiques dans ce domaine. Par ailleurs, elle permet de faire de fortes hypothèses quant à l'ordre des gènes dans les génomes ancestraux.

Mots clés : Phylogénétique, réarrangements du génome, cassures du génome, génomes ancestraux, arbres de Steiner, problème du commis-voyageur.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

À ma famille

CHAPITRE 1

Introduction

Ce mémoire présente une application de l'informatique à la biologie de l'évolution. Il est présenté sous la forme de quatre articles publiés ([57], [8], [58]) ou soumis ([9]). Ceux-ci sont précédés d'une courte introduction aux éléments de biologie nécéssaires à la compréhension et à la mise en contexte des problèmes traités par la suite, de même que d'une revue de l'état actuel des connaissances dans ce domaine.

1.1 L'évolution des espèces

Selon la théorie de Darwin ([17]), l'évolution des espèces, il existe des processus naturels qui engendrent et maintiennent une certaine variabilité des caractères à l'intérieur des espèces. Cette variabilité est aléatoire, c'est-à-dire qu'elle n'est pas orientée dans le but d'améliorer l'espèce. Certains de ces caractères donnent néanmoins à leurs propriétaires des avantages qui leur permettront de mieux survivre dans un environnement spécifique et donc d'avoir une progéniture plus nombreuse. C'est ce qu'on appelle la sélection naturelle. Comme les enfants héritent de certains des traits de leurs parents, les enfants d'individus bien adaptés auront de bonnes chances d'être, eux aussi, bien adaptés. Ainsi, ils prendront de plus en plus de place dans la population, jusqu'à devenir la norme. La population sera alors généralement mieux adaptée aux conditions qui l'ont forgée.

A l'époque de Darwin, la science de la génétique était très peu connue, mais

les découvertes subséquentes dans ce domaine confirmèrent ce scénario. On sait maintenant que les caractères génétiques d'un individu lui sont transmis (parfois imparfaitement) de ses parents par leur ADN, ce qui entraine la variabilité de la population. On peut donc voir la sélection comme un gigantesque processus parallèle d'optimisation d'une certaine fonction d'adaptation au milieu. Ce principe est d'ailleurs au coeur des algorithmes génétiques ([33]), très utilisés en optimisation.

Cela donne une explication satisfaisante aux légères variations entre populations voisines d'une même espèce, mais permet difficilement d'expliquer l'immense diversité des formes vivantes. En effet, dans une population homogène, le processus d'optimisation défini plus haut devrait converger vers une solution unique (donc une espèce unique) plutôt que vers autant de solutions si différentes. La variabilité observée s'explique par le fait que la fonction d'adaptation optimisée varie selon l'espèce et le temps. Elle dépend du milieu, de la niche écologique, et elle change au fur et à mesure de l'évolution de l'espèce.

Cela nous porte à nous demander ce qui fait qu'une espèce se divise en deux sous-espèces qui évolueront vers des formes différentes, qu'on appelle le phénomène de *spéciation*. Plusieurs évènements peuvent causer cette différenciation ([36]). Par exemple, lorsque les individus d'une espèce sont répartis sur de grandes étendues géographiques et que les contraintes environnementales des différentes régions varient, les individus de chaque région s'adapteront aux contraintes de leur milieu, ce qui les différenciera des autres. Après plusieurs générations, les génomes des différentes sous-espèces seront tellement différents qu'elles ne pourront plus se reproduire entre elles, et elles formeront donc deux espèces différentes. La diversité actuelle du monde vivant reflète donc ce processus d'adaptation et de spéciation.

1.2 La phylogénétique

La phylogénétique est la science qui étudie ce processus de spéciation et tente d'établir les liens de parenté entre différentes espèces. Elle tente donc de retrouver le processus de filiation ayant mené aux espèces contemporaines. D'un point de vue mathématique, on peut considérer l'évolution comme un processus de branchement. Le but est donc de retrouver l'arbre phylogénétique représentant le processus de branchement. Les feuilles (les noeuds terminaux) de cet arbre représentent les espèces actuelles (sur lesquelles on peut recueillir de l'information), alors que les noeuds internes représentent des ancêtres communs aux feuilles qui descendent de ce noeud (sur lesquelles on ne possède habituellement que très peu d'information).

1.3 L'information disponible

Comment peut-on affirmer que telles espèces sont plus proches parentes que telles autres? Tout dépend du type d'information dont on dispose. Il existe une très grande variété d'informations permettant la construction d'arbres phylogénétiques. Pour être utilisable, cette information doit satisfaire à un critère simple : elle doit être préservée *partiellement* (ou au moins être parfois imparfaitement transmise) d'une génération à l'autre. Évidemment, ce critère est très large, et un très grand nombre de caractères y satisfont. Selon le niveau de conservation du caractère entre les générations, les arbres pouvant être déduits de ce caractère s'étaleront sur quelques générations ou plusieurs millions de générations.

Voici un bref résumé des sources d'information utilisables en phylogénétique. L'accent sera mis sur les réarrangements du génome, qui font l'objet de ce mémoire.

1.3.1 Les caractères phénotypiques

Le phénotype est l'expression des gènes (génotype) d'un individu. Il est formé de données morphologiques, physiologiques, embryologiques, comportementales et autres. Comme ces informations sont souvent directement observables sans outil très spécialisé, ce sont elles qui ont servi aux premiers chercheurs du domaine. Ce genre de critère était déja utilisé par Linné ([42]) pour sa classification des être vivants (mais sans le principe d'évolution).

1.3.2 Les séquences génétiques

La plupart des études phylogénétiques modernes sont basées sur l'étude de la séquence d'ADN. Cette séquence, en fait un très long "mot" sur un alphabet de quatre symboles (les nucléotides A, C, G et T), contient toute l'information qui détermine le génotype d'un individu. La très grande majorité de l'ADN d'un individu eukaryote (comme le sont toutes les espèces évoluées: animaux, plantes, levures, etc.) se trouve dans le noyau de ses cellules, mais on en retrouve aussi dans les mitochondries (responsables de la respiration cellulaire) et, chez les plantes, dans les chloroplastes (responsables de la photosynthèse).

L'essentiel de la séquence d'ADN d'un individu lui a été transmis de ses parents. Cependant, le mécanisme de recopie des brins d'ADN ne se fait pas toujours parfaitement et il en résulte ce qu'on appelle des *mutations*. Il en existe au moins trois types : la suppression (en anglais, *deletion*) (où un nucléotide est tout simplement effacé de la séquence), l'insertion (où un nouveau nucléotide est introduit dans la séquence) et la substitution (où un nucléotide est remplacé par un autre).

Ces mutations sont relativement fréquentes à l'intérieur d'une population, ce qui fait des séquences d'ADN un bon outil pour distinguer et classifier des

individus relativement proches parents. Par ailleurs, vu l'immensité de l'information disponible (on estime que le génome humain compte 3.4 milliards de paires de résidus), il peut permettre d'établir des liens de parenté entre des espèces très éloignées, à l'aide de certaines régions mieux conservées du génome.

1.3.3 Les méthodes basées sur l'ordre des gènes

La séquence d'ADN est séparée en plusieurs régions codantes appellées *gènes*. Chaque gène est responsable de la synthèse d'un type de protéine. Les gènes sont disposés linéairement sur la séquence d'ADN, comme les mots dans une phrase. Certaines régions de la séquence n'encodent pas de gènes (chez les eukaryotes, c'est le cas de la grande majorité de la séquence), et il y a généralement pas de superposition de gènes. L'ordre dans lequel apparaissent les gènes constitue donc une information bien définie et propre à une espèce.

Le phénomène de réarrangement du génome a été observé par Dobzhansky et Sturtevant en 1936 ([66], [20], [19]), lors de leur étude du génome de différentes espèces et sous-espèces de mouches drosophiles de l'ouest américain. Ceux-ci constatèrent que **l'ordre des gènes** sur leurs chromosomes était très différent, et utilisèrent cette information pour construire un arbre phylogénétique de ces drosophiles.

On connaît au moins trois types de réarrangements du génome: l'inversion, la transposition et la translocation (voir figure 1). L'inversion et la transposition sont des mutations intra-chromosomales (elles n'affectent qu'un seul chromosome), alors que la translocation est une mutation extra-chromosomale (elle affecte deux chromosomes). Par ailleurs, l'inversion renverse l'orientation des gènes contenus dans le segment inversé. L'orientation des gènes est habituellement connue, et elle constitue donc une information supplémentaire à l'ordre des gènes.

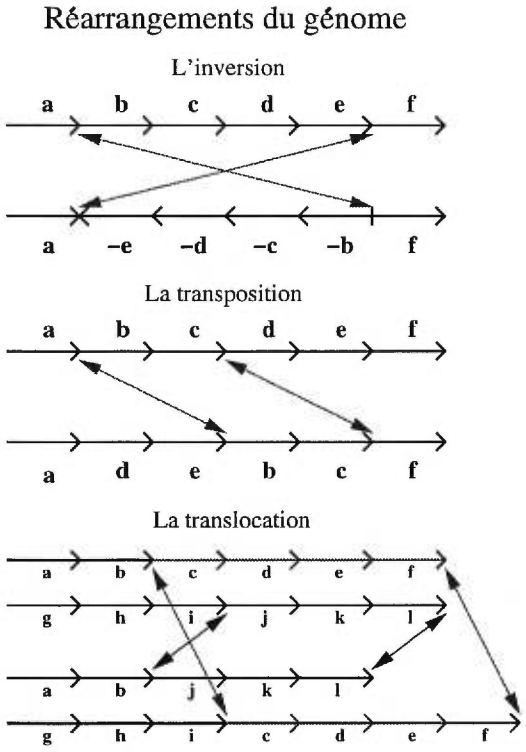

FIGURE 1. Les réarrangements du génome

1.4 Les mesures de distance

Toutes les méthodes de construction et d'évaluation d'arbres phylogénétiques sont basées sur la capacité d'évaluer la "distance évolutive" séparant deux espèces. Cette distance sera d'autant plus grande que les individus ont divergé depuis longtemps.

1.4.1 Les mesures de distance entre séquences

Dans le cas des séquences d'ADN, la distance sera évaluée après avoir **aligné** les deux séquences, en insérant dans les deux séquences des espaces vides (de façon à tenir compte des insertions et suppressions) dans chaque séquence, aux endroits

appropriés. Il existe un grand nombre d'algorithmes permettant d'aligner deux séquences de façon à optimiser certains critères ([70], chapitre 9), par exemple en tentant de maximiser le nombre de symboles bien alignés (voir, par exemple, la figure 7). Le score de l'alignement (le nombre d'insertions, suppressions ou substitutions associé à cet alignement) constituera alors une mesure de distance entre les deux séquences: deux séquences très éloignées auront un mauvais score, alors que deux séquences semblables s'aligneront bien.

1.4.2 Les mesures de distance entre génomes

Bien que la découverte des réarrangements remonte à 1936 ([66], [20]), l'étude plus systématique des réarrangements du génome ne débuta qu'en 1982, où l'idée d'utiliser le nombre minimal d'inversions entre deux ordres de gènes comme mesure de leur distance évolutive est proposée par Watterson et al. [71]. Suivant cette idée, ils suggèrent un premier algorithme d'approximation du nombre minimal d'inversions entre deux génomes circulaires, basé sur la notion de cassures du génome, présentée plus loin.

Par la suite, l'étude des réarrangements du génome se fait en deux directions parallèles. D'un côté, l'aspect biologique de ce type de mutations est étudié. En plus de très nombreuses recherches permettant le séquençage du génome de différentes espèces, les réarrangements sont étudiés d'un point de vue plus théorique, et cela met en évidence l'intérêt des réarrangements génomiques pour la construction d'arbres phylogénétiques. En particulier, les travaux de Palmer et al. ([48], [49], [46]) démontrent que les génomes des chloroplastes de différentes espèces de plantes sont très semblables quant au contenu des gènes (de 99 à 99.9% identiques), mais très différents quant à l'ordre dans lequel ces gènes apparaissent sur le chromosome. Cela démontre que l'analyse des réarrangements du génome peut fournir des renseignements que l'analyse de séquences ne peut révéler.

Dans d'autres cas, les réarrangements du génome sont beaucoup plus lents que les mutations au niveau de la séquence de nucléotides ([24]). Il pourrait donc arriver que l'analyse des réarrangements du génome puissent faciliter la construction d'arbres phylogénétiques où figurent des espèces très distantes au point de vue de la séquence d'ADN, mais encore quelque peu semblables au point de vue de l'ordre des gènes.

D'un autre côté, des informaticiens et des mathématiciens se sont intéressés au problème de trouver les suites minimales de réarrangements transformant un génome en un autre. Vu la complexité du problème à traiter, la plupart des travaux traitent séparément les inversions, transpositions et translocations.

Le problème consistant à trouver la plus courte suite d'inversions permettant de transformer un génome en un autre est d'abord traité par Watterson et al. ([71]). Par la suite, Kececioglu et Sankoff ([39], [40]) développent un ensemble de bornes inférieures et supérieures sur le nombre minimal d'inversions, qui permettent de solutionner exactement, à l'aide d'un algorithme de "Branch-and-Bound", des problèmes de taille intéressante (une trentaine de gènes). Le problème est aussi traité par Pevzner et al. ([4]), qui améliorent les bornes de Sankoff et al. Finalement Hannenhalli et al. ([30]) découvrent un algorithme polynomial permettant de trouver le nombre minimal d'inversions entre deux génomes, dans le cas où l'orientation des gènes est connue. L'algorithme sera par la suite amélioré par Berman et al. ([7]) pour le rendre quadratique. Cet algorithme est non seulement intéressant du point de vue de sa complexité, mais aussi assez efficace une fois implanté; il peut donc être appliqué à des problèmes réels. Une version simplifiée et plus efficace est donnée par Kaplan et al. ([37]). Il est intéressant de noter que le cas où l'orientation des gènes est inconnue a été récemment démontré NP-Complet par Caprara [15]. Notons par ailleurs que Hannenhalli et al. ont étendu leur algorithme pour inclure les translocations ([28], [30]).

Le cas des transpositions semble plus complexe, car sa complexité est

toujours inconnue. Des bornes et des algorithmes pour les transpositions sont dévellopées par Bafna et al. ([6]), mais un algorithme exact (ou, à défaut, une preuve de complexité) se fait toujours attendre.

La version la plus générale du problème, celle incluant tous les types de réarrangements (inversions, transpositions et translocations), ne semble pas encore avoir été traitée. Cependant [10] et [62] décrivent des méthodes heuristiques pour traiter le cas restreint aux réarrangements intra-chromosomaux (inversions et transpositions), où chaque type de réarrangements est pondéré par un coût représentant la fréquence relative de chacun.

D'un point de vue plus mathématique, [35] étudie le problème à l'aide de la théorie des groupes, avec l'étude des générateurs minimaux d'un groupe.

Plusieurs des algorithmes présentés jusqu'ici ont été appliqués à des cas réels, souvent avec des résultats intéressants. Dans [62], les auteurs appliquent leur programme DERANGE à l'inférence de la phylogénie du génome mitochondrial des animaux. Dans [5], les auteurs traitent le génome des plasmides de certaines espèces de plantes, et [29] étudient l'évolution du virus de l'herpès.

1.5 Les méthodes de construction d'arbres phylogénétiques

Dans le domaine de la phylogénétique basée sur les réarrangements du génome, tout comme dans celle basée sur les séquences, deux méthodes se distinguent pour l'inférence d'arbres phylogénétiques. La première est basée sur le calcul préalable de distances entre des **paires** de génomes, alors que la seconde évite cette simplification et la perte d'information qui s'en suit en utilisant l'ensemble des génomes disponibles.

1.5.1 Les méthodes basées sur les matrices de distances

Cette méthode, la plus simple des deux, se base sur la capacité d'évaluer une distance évolutive entre deux espèces. Toutes les mesures de distance présentées jusqu'à présent sont donc susceptibles d'être utilisées. Le processus se déroule alors en deux étapes. Tout d'abord, évaluer la distance entre chaque paire d'espèces qu'on désire inclure dans notre arbre. Ensuite, chercher l'arbre qui respecte le mieux possible les distances calculées.

Plus formellement, soit $F = (f_1, f_2, ..., f_n)$ un ensemble d'espèces et soit $d : F \times F \to \mathbb{R}$ une fonction de distance entre deux taxons. Soit $T = (F \cup A, E)$ un arbre dont les feuilles sont F et les noeuds internes sont A. Soit $c : E \to \mathbb{R}^+$ une fonction associant une longueur aux arètes de T. Pour $v_1, v_2 \in F \cup A$, on définit alors

$$d_T(v_1, v_2) = \sum_{e \in \text{chemin de } v_1 \text{ à } v_2} c(e)$$

la distance **sur l'arbre** entre v_1 et v_2.

On cherche alors l'arbre T et la fonction c telle que

$$\sum_{i,j=1...n} \Delta(d(f_i, f_j), d_T(f_i, f_j))$$

soit minimale. Ici, la fonction $\Delta : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ est le critère qu'on désire optimiser. Par exemple, on choisit souvent les fonctions comme la somme des carrés des différences. L'arbre trouvé est alors celui qui respectera le mieux les distances $d(f_i, f_j)$.

Bien que ce problème soit en général NP-complet, il existe des algorithmes heuristiques assez efficaces pour trouver cet arbre T et ces longueurs d'arcs c à partir de la matrice de distances D ([22]).

L'avantage de cette méthode est de séparer l'analyse en deux étapes: le calcul des distances dépend du type d'information utilisé, mais la seconde partie en est

indépendante. La principale difficulté réside donc dans la définition et l'évaluation d'une fonction de distance entre génomes qui soit appropriée biologiquement et calculable en un temps raisonnable.

Cette approche présente cependant quelques faiblesses. La principale vient du fait que l'arbre n'est pas construit directement à partir des données mais plutôt à partir de la matrice de distances calculée, une version "digérée" des données originales. La perte d'information qui en découle réduit d'autant l'intérêt de la méthode. Par ailleurs, le résultat obtenu est un arbre dont les longueurs de branches sont connues, mais qui ne contient aucune information quant aux génomes ancestraux, à part leurs distances hypothétiques à leurs voisins.

Cela porte à chercher des méthodes plus informatives à cet égard.

1.5.2 Les méthodes de reconstruction

Dans ce cas-ci, on tente non seulement de trouver la topologie de l'arbre phylogénétique, mais aussi d'associer des génomes à chaque noeud interne de l'arbre (les génomes ancestraux). Le problème est alors le suivant: Soit S un espace (celui des génomes), et soit $d : S \times S \rightarrow \mathbb{R}$, une métrique sur cet espace. Il s'agit donc d'un problème d'arbre de Steiner:

Donné : $F = \{f_1, f_2, ..., f_n\}, f_i \in S \forall i = 1...n$

Question : Trouver $A = \{a_1, a_2, ..., a_{n-2}\} \in S$ et un arbre $T = (F \cup A, E)$ avec F pour feuilles, tels que: $\sum_{(v_i,v_j)\in E} d(v_i, v_j)$ soit minimale.

Ce problème est bien connu dans une grande variété de domaines, des télécommunications à la bio-informatique. Dans la plupart des espaces métriques, il est NP-Complet ([69]).

Dans les cas qui nous intéressent, l'espace S sera celui des génomes

(séquences d'ADN, ordre des gènes, etc.) , et les feuilles F seront les espèces contemporaines connues. Voici maintenant quelques méthodes basées sur ce principe.

1.5.3 Les méthodes de parsimonie

Avec la méthode de parsimonie, on travaille dans l'espace des séquences Σ^* (souvent, $\Sigma = \{A, C, G, T\}$). On connaît, pour un certain nombre d'espèces, les séquences de certains gènes. La métrique utilisée est basée sur le score de l'alignement de deux séquences.

On cherche donc les séquences ancestrales qui minisent le nombre de mutations le long de l'arbre. Dans certains cas, par exemple celui où on ne considère que des substitutions (donc où l'alignement est évident), on peut facilement déterminer les génomes ancestraux associés aux noeuds internes de l'arbre à l'aide d'un algorithme de programmation dynamique, pour une topologie donnée. Par contre, l'optimisation sur la topologie est beaucoup plus complexe.

Par ailleurs, bien que ce modèle mathématique soit attrayant, son intérêt biologique est limité. En effet, tous les types de mutations ne sont pas équiprobables ([59]).

1.5.4 Les arbres de vraisemblance maximale

Cela nous amène à modifier les métriques définies sur l'espace des séquences de façon à ce qu'elles représentent plutôt la probabilité de passer d'une séquence à l'autre. Ceci consistera donc à trouver l'arbre et les génomes ancestraux de vraisemblance maximale, ce qui semble plus près du processus biologique sous-jacent. Évidemment, le processus d'optimisation s'en trouve passablement complexifié.

1.5.5 Les méthodes de parsimonie pour les réarrangements du génome

Le problème des arbres de Steiner dans l'espace des génomes, sous la métrique des réarrangements minimaux, est évidemment très complexe. Pour cette raison, il a été assez peu étudié. Seul le problème le plus simple, celui de déterminer **un** génome ancestral à trois autres génomes de façon à minimiser la distance vers ces trois génomes, dans le cas des inversions, a été étudié dans ([64]). Ceux-ci proposent une méthode heuristique pour ce problème, qui a été récemment démontré NP-complet dans ([15]).

Dans les chapitres qui suivent, une méthode de reconstruction se basant sur une nouvelle métrique, le nombre de "cassures" entre deux génomes, est proposée.

1.6 Présentation du mémoire

Ce mémoire est présenté sous la forme de 4 articles. L'ensemble de ceux-ci présente les divers aspects de la construction d'arbres phylogénétiques basés sur les cassures du génome.

Le Chapitre 2, présente l'article "The median problem for breakpoints in comparative genomics". Il introduit l'idée de construire des arbres phylogénétiques basés sur les cassures du génome, c'est-à-dire utilisant l'adjacence de paires de gènes dans le génome pour établir une métrique. Cette mesure de distance est beaucoup plus facile à calculer que les séquences minimales de réarrangements, et elle contient presque autant d'information. La notion de cassure du génome est donc présenté formellement, pour des génomes orientés et non-orientés. On propose, par ailleurs, la notion de cassure cachée, nécéssaire dans le cas où les génomes des différentes espèces ne sont pas formés du même ensemble de gènes.

On introduit ensuite le problème de la médiane, qui consiste à trouver l'ordre des gènes d'**un** ancêtre commun à trois espèces ou plus, de façon à ce que la somme des distances (toujours en terme de cassures) entre l'ancêtre et chaque espèce soit minimale. On montre alors comment réduire ce problème à un problème de commis-voyageur (PCV). Comme le PCV auquel on se réduit possède une structure bien particulière, on peut espérer une solution plus efficace que dans le cas d'un PCV général. C'est ce qu'on tente d'obtenir à l'aide d'un algorithme de type "Branch-and-bound" qui tente de tirer parti de cette structure. On présente donc la borne utilisée et l'algorithme qui en découlent, pour le cas de génomes non-orientés et constitués d'un même ensemble de gènes. On étend ensuite la réduction au cas des gènes orientés, pour lequel un algorithme de "Branch-and-Bound" semblable s'applique.

On s'attaque finalement au cas où les génomes ne sont pas tous formés du même ensemble de gènes. Ici, il ne semble pas y avoir de réduction simple vers le PCV. Cependant, une adaptation de l'algorithme de "Branch-and-Bound" permet la résolution du problème avec un ensemble de gènes quelconques.

Dans le troisième chapitre on présente l'article intitulé "Breakpoints phylogenies". On présente des méthodes heuristiques pour inférer l'ordre des gènes dans les génomes ancestraux associés à un arbre phylogénétique donné. À la différence du Chapitre 2, où on ne cherchait qu'**un** ancêtre commun à un ensemble de génomes, on cherche maintenant **un ensemble** d'ancêtres, chacun correspondant à un point de divergence d'un arbre donné. Chaque méthode est basée sur l'algorithme de la médiane de trois génomes, décrit au Chapitre 2. Les méthodes présentées se déroulent toutes en deux étapes: l'initialisation des génomes ancestraux, puis l'application itérative de l'algorithme de la médiane sur chaque génome ancestral, jusqu'à ce qu'on n'observe plus d'amélioration dans la taille totale de l'arbre.

Ce qui distingue les trois méthodes présentées est donc la méthode d'initia-

lisation utilisée. La première technique se contente d'initialiser chaque génome ancestral au génome connu le plus près, dans l'arbre donné. Les deux autres méthodes nécéssitent la résolution d'un PCV (qui ne comporte pas nécéssairement la structure du PCV associé à un problème de médiane) à chaque noeud interne de l'arbre. La seconde méthode exploite l'idée que, si chaque génome devait être au "centre" de ses trois voisins dans l'arbre, les graphes associés à chaque noeud interne devraient aussi être au "centre" (en terme de poids sur les arcs) des graphes associés à ses trois voisins.

Finalement, la dernière méthode proposée est basée sur l'idée que le coût d'inclure une certaine paire de gènes adjacents dans un génome ancestral donné devrait dépendre du nombre minimal d'évènements de création ou de destruction de cette paire nécéssaires pour satisfaire aux contraintes de présence ou d'absence dans les feuilles de l'arbre. Un algorithme de programmation dynamique permet de calculer ces coûts.

On étudie finalement, à l'aide de simulations, différents aspects de l'utilisation de ces trois méthodes, comme la relation entre la taille de l'arbre reconstruit et celle de l'arbre réel ayant mené aux données. On regarde aussi l'efficacité de chaque méthode d'initialisation, avant et après l'optimisation par itérations de médianes. Enfin, on tente d'évaluer l'optimalité de ces méthodes en observant le nombre de fois où les trois techniques s'accordent sur la taille de l'arbre optimal.

Le quatrième chapitre, "Multiple genome rearrangement" dresse un parallèle entre l'alignement de séquences multiples et la méthode basée sur les réarrangements du génome. On y étudie l'unicité du génome ancestral dans une étoile à n branches, puis, dans le cas d'un arbre binaire complet sans racine à 12 feuilles.

Dans le dernier chapitre, "Gene order breakpoint evidence in animal mitochondrial phylogeny", on fournit une justification et une application d'un point

de vue biologique, de la méthode utilisée. Dans un premier temps, on rappelle les méthodes classiques, basées sur des matrices de distances, pour inférer un arbre phylogénétique à partir des réarrangements du génome. Les avantages et inconvénients de ces dernières sont mis en évidence. Nous montrons alors comment les phylogénies basées sur les cassures du génome peuvent réduire (voire éliminer) certains de ces problèmes.

Pour démontrer l'utilité de notre méthode, on tente de reconstruire l'arbre phylogénétique des métazoaires (animaux), ce qui pose de graves problèmes aux méthodes classiques. Les résultats obtenus montrent que l'analyse des cassures du génome peut, au moins dans ce cas, donner des résultats beaucoup plus précis que les méthodes classiques. Un des aspects les plus intéressants de cette méthode, à savoir la capacité d'inférer une partie de l'ordre des gènes dans les génomes ancestraux est alors discutée et mise en pratique sur 11 espèces de métazoaires représentant différentes branches du monde animal.

CHAPITRE 2

The median problem for breakpoints in comparative genomics [57]

David Sankoff [1] Mathieu Blanchette[2]

Abstract

During evolution, chromosomal rearrangements, such as reciprocal translocation, transposition and inversion, disrupt gene content and gene order on chromosomes. We discuss algorithmic and statistical approaches to the analysis of comparative genomic data. In a phylogenetic context, a combined approach is suggested, leading to the *median problem for breakpoints*. We solve this problem first for the case where all genomes have the same gene content, and then for the general case.

[1]Centre de recherches mathématiques, Université de Montréal, CP 6128 Succursale Centre-Ville, Montréal, Québec H3C 3J7 sankoff@ere.umontreal.ca

[2]Laboratoire de biologie informatique et théorique, Département d'informatique et de recherche opérationnelle, Université de Montréal, CP 6128 Succursale Centre-Ville, Montréal, Québec H3C 3J7, blanchem@iro.umontreal.ca

2.1 Introduction

During biological evolution, inter- and intrachromosomal exchanges of chromosomal fragments disrupt the order of genes on a chromosome and, for multichromosomal genomes, the partition of genes among these chromosomes.

When comparing two evolutionarily diverging species, any (maximal) contiguous region of the genome in which gene content and order have been conserved in both species is called a *conserved segment.* Between any two adjacent conserved segments is a *breakpoint.* The number of conserved segments increases as they are disrupted by new events, so that they tend to become shorter over time. The number of chromosomal segments conserved during the divergence of two species, or equivalently, the number of breakpoints, can be used as a rough measure of their genomic distance.

Two approaches, the algorithmic and the statistical, have been taken to the reconstruction of genomic history based on the comparison of chromosomal gene content and order in two or more genomes. The first attempts to infer a most economical sequence of rearrangement events to account for the differences among the genomes, based only on the breakpoints, and neglects the contents of conserved segments. The second approach ignores the details of rearrangement history and assumes that a random model (the Nadeau-Taylor model) accounts for the differences in chromosomal gene content and order. In this paper, we discuss the strengths and weaknesses of the two approaches.

In the phylogenetic context, a compromise approach can be adopted, algorithmic, but not attempting to infer precise details of hypothesized evolutionary events. This leads to a new, tractable, problem, the *median problem for breakpoints.* We give a solution to the version of this problem where all genomes have the same gene content, and extend it to the case where the median and other genomes involved may have partially different gene sets.

2.2 The algorithmic approach

The algorithmic study of comparative genomics has focused on inferring the most economical explanation for observed differences in gene orders in two or more genomes in terms of a limited number of rearrangement processes. For single-chromosome genomes, this has been formulated as the problem of calculating an edit distance between two linear orders on the same set of objects, representing the ordering of homologous genes in two genomes. In the most realistic version of the problem, a sign (plus or minus) is associated with each object in the linear order, representing the direction of transcription, or strandedness, of the corresponding gene. The elementary edit operations may include one or more of:

1) inversion, or reversal, of any number of consecutive terms in the ordered set, which, in the case of signed orders, also reverses the polarity of each term within the scope of the inversion. Kececioglu and Sankoff [40] considered the problem of computing the minimum reversal distance between two given permutations in the unsigned case, including approximation algorithms and an exact algorithm feasible for moderately long permutations. Bafna and Pevzner [4] gave improved approximation algorithms for this problem. Recently, Caprara [15] showed this problem to be NP-complete. Kececioglu and Sankoff [39] also found tight lower and upper bounds for the signed case and implemented an exact algorithm which worked rapidly for long permutations. Indeed, Hannenhalli and Pevzner [30] showed in 1995 that the signed problem is only of polynomial complexity, and an improved polynomial algorithm was given by Kaplan, Shamir and Tarjan [37].

2) transposition of any number of consecutive terms from their position in the order to a new position between any other pair of consecutive terms. This may or may not also involve an inversion. Computation of the transposition distance between two permutations was considered by Bafna and Pevzner [5]. Sankoff *et al.* [10, 55, 62] implemented and applied heuristics to compute an edit distance

which is a weighted combination of inversions, transpositions and deletions.

In addition, for multi-chromosome genomes, a major role is played by:

3) reciprocal translocation. Kececioglu and Ravi [38] began the investigation of translocation distances, and Hannenhalli [28] has shown that a formulation is of polynomial complexity. A relaxed form of translocation distance was proposed by Ferretti *et al.* [23] and the complexity of its calculation was shown to be NP-complete by DasGupta *et al.* [18].

2.2.1 Shortcomings of the algorithmic approach.

It would seem to be an advantage of the algorithmic approach that it actually constructs an optimizing series of events that accounts for the rearrangement of one genome with respect to another. There are two problems with this, however. One is the non-uniqueness of the solution, especially when all rearrangement events are weighted equally - one event equals one unit of the objective function being minimized - or even if the weights are integral multiples of some common factor. Though this problem is reduced with suitable event weights, a serious measure of arbitrariness is thereby introduced. Some progress has recently been made in estimating appropriate weights empirically [10, 55].

The advantage of reconstructing a feasible history is thus diminished, since this history likely has no particular status with respect to many other equally parsimonious solutions. This problem is somewhat attenuated in the context of the median problem, to be discussed later. A more serious problem with reconstructed solutions is that when the number of steps approaches a certain proportion of the number of breakpoints, this number is almost certainly a serious underestimate [40]. Again, having a reconstructed history is a dubious advantage, since it inevitably contains some wrong steps and omits even more true events.

Finally, the algorithmic approach is very sensitive to errors and other small changes in the data. These are especially numerous when gene order has been determined by mapping techniques other than complete sequencing [61].

2.3 The statistical approach

Our formulation of the Nadeau-Taylor model of genomic divergence assumes that each reciprocal translocation breaks chromosomes at random points on two randomly chosen chromosomes. As a consequence when we compare two divergent genomes, the endpoints of the conserved segments making up each chromosome are uniformly and independently distributed along its length (spatial homogeneity of breakpoints). We also assume that which genes of a genome are discovered and mapped first does not depend on their position on the chromosome (spatial homogeneity of gene distribution), nor on their proximity to each other (independence of map positions).

In trying to count the number of conserved segments for the quantification of evolution, we must deal with underestimation due to conserved segments in which genes have not yet been identified in one or both species. This is particularly important if there are relatively few genes common to the data sets for a pair of species, so that many or most of the conserved segments are not represented in the comparison, and genomic distance may be severely underestimated. Nadeau and Taylor [46] in 1984 could only treat 13 segments out of the 100-200 now known to exist.

We model the genome as a single long unit broken at n random breakpoints into $n + 1$ segments, within each of which gene order has been conserved with reference to some other genome. (Little is lost in not distinguishing between breakpoints and concatenation boundaries separating two successive chromosomes [60].)

It is remarkable that to estimate n from m and the number of segments n_r observed to contain r genes, for $r = 1, 2 \ldots$, only the number of non-empty segments $a = \sum_{r>0} n_r$ is important [50]. The variable a is a sufficient statistic for the estimation of n.

2.3.1 Estimating n from a

To estimate n, we study $P(a, m, n)$, the probability of observing a non-empty segments if there are m genes and n breakpoints. Combinatorial arguments give

$$P(a, m, n) = \frac{\binom{m-1}{a-1}\binom{n+1}{a}}{\binom{n+m}{m}}$$

After observing m and a it is an easy matter to find the value of n which maximizes P, i.e. the maximum likelihood estimate.

2.3.2 Weaknesses of the statistical approach.

One weakness of the statistical approach is that it does not estimate a specific series of events, although we have discussed how this advantage of the algorithmic approach is dubious. The number of breakpoints (or conserved segments) cannot be deterministically converted into a number of events, since different types of rearrangement produce different numbers of breakpoints, and even a single type of event does not always produce the same number of breakpoints.

Perhaps the greatest potential weakness of this approach is that it depends on the applicability of a particular probabilistic model. This is a temporary problem, however, in that it gives rise to further research on better models [44].

2.4 The median problem.

For phylogenetic purposes, it is useful to solve the following sort of problem: Given a distance or dissimilarity d, three genomes A, B and C, and a set of genes Σ, we want to find a genome S containing all the genes in Σ such that

$$d(S, A) + d(S, B) + d(S, C)$$

is minimized. How a solution to this problem can be the key to solving phylogenetic problems involving many genomes is discussed in [23, 64].

The set Σ may be determined by the phylogenetic problem under study, or may be defined by the analyst. For example, if A, B and C all contain the same set of genes, then it is natural to use this set for Σ. Another example is drawn from organelle evolution, where genes tend to be lost from the genome and not re-inserted. Then if A is ancestral to both B and C, the set Σ will be the union of the set of genes in B and the set of genes in C. Still another possibility, where the direction of evolution is less clear, is to include in Σ just those genes that are in at least two of A, B and C.

2.4.1 Breakpoints

Consider two genomes $A = a_1 \dots a_n$ and $B = b_1 \dots b_n$ on the same set of genes $\{g_1, \dots, g_n\}$. We say a_i and a_{i+1} are adjacent in A (and a_n and a_1 are adjacent as well in circular genomes). If two genes g and h are adjacent in A but not in B, they determine a breakpoint in A. The number of breakpoints in A is clearly equal to the number of breakpoints in B.

For two genomes whose gene sets are not identical, to calculate the breakpoints, we first remove all genes that are present in only one of the genomes. We then find the breakpoints for the reduced genomes, now of identical composition. The positions of the breakpoints are well-defined in the reduced genomes. In the

full genomes, there is a breakpoint between a_i and a_{i+1} only if this is a breakpoint for the reduced genome. If, as in Figure 2, there is a breakpoint between a_i and a_j in the reduced genome, where $j \neq i+1$, then there is a corresponding breakpoint in the full genome, but its position is ambiguous. We call it a *hidden* breakpoint; it is somewhere between a_i and a_j, which are not adjacent.

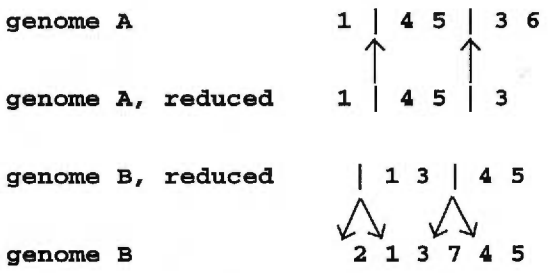

FIGURE 2. Defining breakpoints for (circular) genomes with different gene contents. Position of breakpoints (vertical strokes) found first in reduced genomes with identical gene sets. This unambiguously determines breakpoints between 1 and 4 and between 5 and 3 in genome A. Breakpoint between 5 and 1 in genome B is "hidden" by gene 2; that between 3 and 4 is hidden by gene 7.

2.4.2 The median problem for a fixed gene set Σ

Now define d to be the number of breakpoints derived from the comparison of two genomes containing the same genes Σ. The median problem becomes one of finding the genome S on Σ that determines the fewest total breakpoints between itself and the three genomes A, B and C. In contrast to other genomic distances, this problem seems relatively tractable, although its computational complexity remains to be determined. We proceed by reduction to the Traveling Salesman Problem (TSP).

It will be convenient to describe genomes in graph-theoretical terms. The genes will be represented by the vertices of the graph and adjacency of two genes will be indicated by the existence of a corresponding edge in the graph. Thus, only graphs consisting of a single complete cycle of the vertices represent (circular)

genomes.

We first define G to be the complete graph whose vertices are the elements of Σ. For each edge gh in $E(G)$, let $u(gh)$ be the number of times g and h are adjacent in the three genomes. Set $w(gh) = 3 - u(gh)$. Then the solution to TSP on (G, w) traces out an optimal genome S on Σ, since if g and h are adjacent in S, but not in A, for example, then they form a breakpoint in S.

2.4.3 A lower bound

To solve this restricted form of TSP, we resort to a branch-and-bound algorithm based on the following lower bound:

Let the *edge-pool* $P \subseteq E(G)$, be disjoint from the *fragment* $F \subseteq E(G,)$ and let $score = \sum_{gh \in F} w(gh)$. Define $a(g)$, the *availability* of $g \in V(G)$, to be 2, 1 or 0, depending on whether g is incident to zero, one, or more than one edge in F, respectively. Let $\mu(g)$ be the sum of the $a(g)$ smallest weights of edges in P incident to g. ($\mu(g)$ is undefined if there are fewer than $a(g)$ such edges.)

If there is a TSP solution cycle S of weight W_S which includes all the edges in the fragment F and some additional edges drawn from the edge-pool P, let $\nu(g)$ be the sum of the weights of the exactly $a(g)$ edges of S in P incident to g. (In this case $\mu(g)$ is always defined.) Clearly $\mu(g) \leq \nu(g)$.

Now,

$$W_S = score + \sum_{gh \in E(S) \cap P} w(gh) = score + \frac{1}{2} \sum_{g | gh \in E(S) \cap P} w(gh)$$

since each edge in $E(S) \cap P$ is counted twice in the sum. Thus

$$W_S = score + \frac{1}{2} \sum_{g | gh \in E(S) \cap P} \nu(g).$$

Defining

$$L(P) = \frac{1}{2} \sum_{g|gh \in E(S) \cap P} \mu(g),$$

$$score + L(P) \leq score + \frac{1}{2} \sum_{g|gh \in E(S) \cap P} \nu(g) = W_S.$$

We use $L(P)$ as a lower bound in the branch-and-bound algorithm in Section 2.4.4. When $P = E(G)$ and $F = \emptyset$ this as a well-known bound on TSP (see, e.g., pp. 272-273 in [45]). There are a number of other bounds which can be used for the TSP, but this one is of particular interest in that it can be modified for use in the median problem with more general genomes as discussed in Section 2.5.2.

2.4.4 Algorithm BBF

input: weighted complete graph (G, w)
output: solution S to the TSP on (G, w)

initialization
$V(S) \leftarrow V(G)$
$F \leftarrow \emptyset$
$P \leftarrow E(G)$
score$\leftarrow 0$
best$\leftarrow \infty$

procedure BBF(P, F, S, score,best)
if $|F| = |G|$ **and** score $<$ best **then**
store $S = F$ as current best solution
best$\leftarrow$score
if $|F| < |G|$ **then**
if $L(P)$ + score $<$ best **then**
choose $gh \in P$ to try to add to F so that
$a(g) > 0, a(h) > 0$ and $w(gh)$ is as small as possible,
and $F \cup \{gh\}$ is not a cycle on less than $|G|$ vertices.
BBF($P - \{gh\}, F \cup \{gh\}, S$,score$+ w(gh)$,best)
BBF($P - \{gh\}, F, S$,score,best)

The recursion functions as a "greedy" search until it first finds a cycle, which is necessarily an upper bound. If its cost $U = L(E(G))$, it is optimal.

2.4.5 Genomes with directionality

In the case of directed genomes, the notion of breakpoint must be modified to take into account the polarity of the two genes. If gh represents the order of two genes in one genome, then if another genome contains gh or $-h-g$ there is no breakpoint involved. However, between gh and hg there is a breakpoint, similarly between gh and $-g-h, g-h, -gh, h-g$ or $-hg$, and so sdjacency is no longer commutative. The reduction of the median problem to TSP must be somewhat different to take into account that the median genome contains g or $-g$ but not both. Let G be a complete graph with vertices $V = \{-g_n, \ldots, -g_1, g_1, \ldots, g_n\}$, where $\Sigma = \{g_1, \ldots, g_n\}$. For each edge gh of G, let $u(gh)$ be the number of times $-g$ and h are adjacent in the three genomes A, B and C, and $w(gh) = 3 - u(gh)$, if $g \neq -h$. If $g = -h$, we simply set $w(gh) = -M$, where M is large enough to assure that a minimum weight cycle must contain the edge $-gg$.

Proposition: If $s = s_1, -s_1, s_2, -s_2, \ldots, s_n, -s_n$ is a solution of the TSP on (G, w), then a median is given by $S = s_1 s_2 \ldots s_n$.

Proof: $d(S,A) + d(S,B) + d(S,C) = \sum_{gh \in S, g \neq -h} w(gh)$

$$= nM + \sum_{gh \in s} w(gh).$$

Thus S minimizes $d(S,A) + d(S,B) + d(S,C)$ iff s is of minimal weight. ∎

The same bound $L(G)$ may be constructed as before, though for directed genomes $\mu(g) = -M+$ smallest weight of any edge incident to g.

An implementation of the algorithm we have described finds the median of three directed genomes of size 50 in one minute, on average on an Origin 200 computer with a RISC 10000 processor. Random genomes are easily processed

since $L(G)$ tends to be a fairly tight bound. Three similar genomes are also rapidly treated since the first $|G|$ "greedy" recursive steps are likely to produce an optimal solution. It is between these extremes that longer execution times are encountered.

2.4.6 Larger stars

The median problem can also be defined for $k > 3$ genomes. When all these genomes have identical gene sets, the BBF procedure is directly applicable to finding the median, the only difference being in the calculation of the weights where $w(gh)$ becomes $k - u(gh)$.

2.5 The case of a more general median gene set

2.5.1 Extension of the previous method

If the differences among the sets of genes in A, B, C and Σ consist of very few genes, the bound and algorithm in Section 2.4.2 can be adapted to function relatively efficiently. We redefine $w(gh)$ =(number of genomes containing both g and h) $- u(gh)$. In the algorithm in Section 2.4.4, in the call

BBF$(P - \{gh\}, E(S) \cup \{gh\}$,score $+ w(gh)$,best),

"score $+ w(gh)$" must be replaced by "score $+ w(gh) + z(gh)$" where $z(gh)$ counts the "hidden" breakpoints (cf. Section 2.5.2) caused by the addition of gh to the solution.

2.5.2 A better bound

In this section, we develop a bound designed for the situation where the gene content of Σ can differ considerably from that of A, B and/or C.

We assume all genes in A, B or C are also in Σ, and each gene in Σ is in at least one of A, B or C, since only these can contribute to the weight of a cycle. There will, however, generally remain genes in Σ which are absent from some, but not all, of A, B and C, and as we shall see, this is the crux of the difficulty.

The bound in Section 2.4.3 was based on the fact that each vertex on a cycle is incident to two edges, and it was easy to bound the sum of their two weights. In the present context, when examining each vertex g on a cycle, we have to take into account that its incident edges may not be relevant to the breakpoint calculations with respect to one or more of the given genomes; we may have $gi \in S, ih \in S$, but i absent from A and g not adjacent to h in A. The breakpoint between g and h in S is hidden by gene i.

Suppose we wish to bound the contributions, to the cost of a cycle, of the edges"near" g in S, since the individual edges directly incident may not be relevant to all of A, B and C, as we have seen. We arbitrarily impose a directionality on S. If g is in genome X, let l_X and r_X be the closest vertices to the left and right of g in S that are also present in genome $X, X \in \{A, B, C\}$. If g is not in genome X, it cannot be involved in a breakpoint. The cost of the edges near g, summed over all $g \in \Sigma$, is then

$$W = \frac{1}{2} \sum_{X \in \{A,B,C\}} \sum_{g \in \Sigma \cap X} w(l_X g) + w(g r_X)$$

where $w(l_X g) = 0$ if l_X is adjacent to g in X, and $w(l_X g) = 1$ otherwise; similarly for $w(r g_X)$.

What is the configuration of the l_X and r_X around g in S? To the left of g

we may have $l_{Y(1)} \ldots l_{Y(2)} \ldots l_{Y(3)}$, where $(Y(1), Y(2), Y(3))$ is a permutation of (A, B, C), and "rightward exclusion" prevails: $l_{Y(1)}$ is in genome $Y(1)$ but not in genomes $Y(2)$ or $Y(3)$; $l_{Y(2)}$ is in genome $Y(2)$ but not in genome $Y(3)$; $l_{Y(3)}$ is in genome $Y(3)$. If g is absent from one or two of genomes A, B or C, then there will be at most two or one l terms, respectively.

Other possibilities are that we may have only $l_{Y(1)} \ldots l_{Y(2)}$ left of g, and one of these genes is in two of the genomes A, B and C (rightward exclusion still obtains), or that there is only one l gene common to the three genomes.

A similar accounting of the possibilities can be made for the r genes, involving the notion of "leftward exclusion".

Then a lower bound on the cost of S is found by choosing, for each $g \in \Sigma$,

- up to three (depending on how many of A, B, C contain g) genes l_A, l_B, l_C, not necessarily distinct, each l_X in genome X, and some permutation $(Y(1), Y(2), Y(3))$ of (A, B, C) such that $l_{Y(1)} l_{Y(2)} l_{Y(3)}$ (or $l_{Y(1)} l_{Y(2)}$ if there are only two distinct genes) is rightward exclusive, and

- up to three (depending on how many of A, B, C contain g) genes r_A, r_B, r_C, not necessarily distinct, each r_X in genome X, and some permutation $(Z(1), Z(2), Z(3))$ of (A, B, C) such that $r_{Z(1)} r_{Z(2)} r_{Z(3)}$ (or $r_{Z(1)} r_{Z(2)}$ if there are only two distinct genes) is leftward exclusive, such that

$$t(g) = \sum_{X \in \{A,B,C\}, g \in X} w(l_X g) + w(g r_X)$$

is minimized by the cycle fragment $l_{Y(1)} l_{Y(2)} l_{Y(3)} g r_{Z(1)} r_{Z(2)} r_{Z(3)}$.

Then a lower bound on W is given by $\Lambda = \frac{1}{2} \sum_{g \in \Sigma} \min t(g)$, since each breakpoint is counted at most twice in the sum.

2.5.3 Adapting the bound for the stepwise construction of a cycle.

Suppose, somewhat differently from Section 2.4.3, a candidate fragment of a path $F = s_1\, s_2\, \ldots s_j$ of a solution cycle has already been constructed, and a pool Q of vertices remain to be tested for possible addition to the cycle. We define the availability $\alpha_X(g)$ to be 0 if g is not in X, and otherwise to be 2, 1 or 0, depending on whether g is not in F, g is the leftmost or rightmost in F of $V(X) \cap V(F)$, or is some other gene in F, respectively.

Then $\Lambda(Q) = \frac{1}{2}\sum_{g \in Q} \min t(g)$, is a lower bound on the weight of the remainder of the cycle, where the search for the minimizing cycle fragment for each g is constrained to respect the order of genes already in F and to use both an l_X and an r_X only if $\alpha_X(g) = 2$. Only one of l_X or r_X can be used if if $\alpha_X(g) = 1$.

2.5.4 Algorithm BBG

input: genomes A, B, C, median gene set Σ
output: solution S to the median problem

initialization
  $F \leftarrow g$ (arbitrary choice)
  $Q \leftarrow \Sigma - g$
  score$\leftarrow 0$
  best$\leftarrow \infty$
  for $X = A, B, C$
    $last(X) = g$ if g is in X, otherwise $last(X) = \Phi$
    $\alpha_X(g) = 2$ if g is in X, otherwise $\alpha_X(g) = 0$

procedure BBG$(Q, F, S, \alpha, last$,score,best)
  if there are $|\Sigma|$ edges in F **and** score $<$ best **then**
    store $S = F$ as current best solution
    best$\leftarrow$score

  if there are less than $|\Sigma|$ edges in F **then**
    if $\Lambda(Q)$ + score $<$ best **then**
      choose $h \in Q$ to try to add to F, i.e.add edge s^*h, where $s^* = s_{|F|}$
      (except if there are $|\Sigma| - 1$ edges in F: here we add s^*g)

for $X = A, B, C$ if $h \in X$
  $\alpha_X(last(X)) = \alpha_X(last(X)) - 1$
  $last'(X) = last(X)$
  $last(X) = h$
BBG$(Q - \{h\}, F \cup s^*h, S, \alpha, last$,score$(F \cup s^*h)$,best)
for $X = A, B, C$ if $h \in X$
  $\alpha_X(last(X)) = \alpha_X(last(X)) + 1$
  $last(X) = last'(X)$
  remove s^*h from further consideration
BBG$(Q, F, S, \alpha, last$,score,best)

Note that in the first recursive call of **BBG**,

$$\text{score}(F \cup s^*h) = \text{score} + \sum_{X|h \in X} w(last(X)h)$$

unless F contains $|\Sigma| - 1$ edges, in which case

$$\text{score}(F \cup s^*h) = \text{score} + \sum_{X|h \in X} w(last(X)h) + \sum_{X|h \in X} w(first(X)h).$$

2.6 Discussion

The problems of non-uniqueness and underestimation inherent in parsimonious analyses of genomic distance (or sequence distance) are attenuated when more than two genomes (or sequences) are compared. The median problem is the archetype of this effect: triangulation increases accuracy. With other methods of genomic distance, however, the median problem turns out to be much more difficult than a pairwise comparison [64].

The number of breakpoints between two genomes is not only the most general measure of genomic distance, requiring no assumptions about the mechanisms of genomic evolution underlying the data, but it is also the easiest to calculate. We might then expect the median problem for breakpoints to be more tractable than for other measures, and the preliminary work reported here supports this hope.

The relatively easy extension from 3 to k genomes is also a positive indication of its feasibility for phylogenetics.

2.7 Acknowledgements

Research supported by grants to DS from the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Canadian Genome Analysis and Technology program, and a NSERC fellowship for graduate studies to MB. DS is a Fellow of the Canadian Institute for Advanced Research.

CHAPITRE 3

Breakpoint phylogenies [8]

Mathieu Blanchette[1] Guillaume Bourque [2] David Sankoff [3]

Abstract

We describe a number of heuristics for inferring the gene orders of the hypothetical ancestral genomes in a fixed phylogeny. The optimization criterion is the minimum number of breakpoints (pairs of genes adjacent in one genome but not the other) in the gene orders of two genomes connected by an edge of the tree, summed over all edges. The key to the method is an exact solution for trees with three leaves (the median problem) based on a reduction to the Travelling Salesman Problem.

[1]Laboratoire de biologie informatique et théorique, Département d'informatique et de recherche opérationnelle, Université de Montréal, CP 6128 Succursale Centre-Ville, Montréal, Québec H3C 3J7, blanchem@iro.umontreal.ca

[2]Centre de recherches mathématiques, Université de Montréal, CP 6128 Succursale Centre-Ville, Montréal, Québec H3C 3J7, bourque@crm.umontreal.ca

[3]Centre de recherches mathématiques, Université de Montréal, CP 6128 Succursale Centre-Ville, Montréal, Québec H3C 3J7, sankoff@ere.umontreal.ca

3.1 Introduction

There have been a number of investigations of phylogeny of $N > 2$ genomes based on the pairwise comparison of the gene orders of these genomes, followed by distance matrix methods (e.g. [62]). Treeing methods based on the direct comparison of all N gene orders, which infer gene order at ancestral nodes [31, 64], have been little used because of the difficulty in generalizing measures of genomic distance to more than two genomes – there are no algorithms available, aside from rough heuristics, for handling even three relatively short genomes. Besides this technical problem,there are conceptual problems inherent in the use of rearrangement-event types of edit-distance, or their $N-$genome generalizations, for the purposes of reconstructing evolutionary history.

This include unwarranted assumptions as to the relative importance (i.e. costs) of reversals, transpositions, translocations and other rearrangement events (cf. [10]) and the fallacy that calculation of an edit distance allows the recoverability of the "true" history of genomic divergence – in fact, there is a proliferation of of optimal edit paths (and severe underestimation of the total number of events generating the divergence, cf. [40]) for moderate or large gene-order distances.

These problems all militate in favour of extending gene-order comparisons to three or more genomes through a much simpler and model-free metric, namely the number of breakpoints.

Consider two genomes $A = a_1 \ldots a_n$ and $B = b_1 \ldots b_n$ on the same set of genes $\{g_1, \ldots, g_n\}$. We say a_i and a_{i+1} are adjacent in A (and a_n and a_1 are adjacent as well in circular genomes). If two genes g and h are adjacent in A but not in B, they determine a breakpoint in A. We define $\Phi(A, B)$ to be the number of breakpoints in A. This is clearly equal to the number of breakpoints in B.

The number of breakpoints between two genomes is not only the most general measure of genomic distance, requiring no assumptions about the mechanisms of genomic evolution (inversion versus transposition versus translocation) underlying the data, but it is also the easiest to calculate.

In this paper we offer a number of solutions to the problem of inferring ancestral gene order by minimizing the number of breakpoints associated with each edge of a given phylogenetic tree, summed over the entire tree. These involve the solution of the Travelling Salesman Problems (TSP) at each internal vertex of the tree, and an iterative approach to optimizing the entire tree. The approaches differ only in the initialization of the set of genomes associated to the internal vertices. Simulation experiments show that better initialization reduces the chances of converging to a non-global solution.

3.2 Steiner Points under the Breakpoints Metric.

The problem is formulated as follows: Let T=(V,E) be an unrooted binary tree with $N \geq 3$ leaves and $\Sigma = \{g_1, \cdots, g_n\}$ be a set of genes. Suppose $\{V_1, \cdots, V_N\} \subset V(T)$ are the leaves of the tree and $\{V_{N+1}, \cdots, V_{2N-2}\}$ are the internal vertices of the tree. The data consist, for each leaf $V_i, i = 1, \cdots, N$, of a circular permutation $G^i = g_1^i \cdots g_n^i$ of the genes in Σ, representing a contemporary genome. The task is to find the permutations $G^{N+1}, \cdots, G^{2N-2}$ associated with the internal (ancestral) vertices $V_{N+1}, \cdots, V_{2N-2}$, such that $\sum_{V_iV_j \in E(T)} \Phi(G^i, G^j)$ is minimized.

3.3 The Median and the Travelling Salesman Problem.

The smallest problem of this type is that of finding the median, when $N = 3$: Given three genomes A, B and C, containing the genes in Σ, we want to find **median**(A, B, C), a genome S containing the genes in Σ such that $\Phi(S, A) + \Phi(S, B) + \Phi(S, C)$ is minimized.

This can be reduced to the TSP as follows [57]. We define Γ to be the complete graph whose vertices are the elements of Σ. For each edge gh in $E(\Gamma)$, let $u(gh)$ be the number of times g and h are adjacent in the three genomes. Set $w(gh) = 3 - u(gh)$. Then the solution to TSP on (Γ, w) traces out an optimal genome S on Σ, since if g and h are adjacent in S, but not in A, for example, then they form a breakpoint in S.

3.3.1 Genomes with directionality

Our simulations will involve directed genomes; we assume we know the strandedness, or direction of transcription, of each gene in each genome in the data set. In this case, the notion of breakpoint must be modified to take into account the polarity of the two genes [57]. If gh represents the order of two genes in one genome, then if another genome contains gh or $-h - g$ there is no breakpoint involved. However, between gh and hg there is a breakpoint, similarly between gh and $-g - h, g - h, -gh, h - g$ or $-hg$, so adjacency is no longer commutative. The reduction of the median problem to TSP must be somewhat different to take into account that the median genome contains g or $-g$ but not both. Let Γ be a complete graph with vertices $V(\Gamma) = \{-g_n, \ldots, -g_1, g_1, \ldots, g_n\}$. For each edge gh in $E(\Gamma)$, let $u(gh)$ be the number of times $-g$ and h are adjacent in the three genomes A, B and C, and $w(gh) = 3 - u(gh)$, if $g \neq -h$. If $g = -h$, we simply set $w(gh) = -Z$, where Z is large enough to assure that a minimum weight cycle

must contain the edge $-gg$.

Proposition: If $s = s_1, -s_1, s_2, -s_2, \ldots, s_n, -s_n$ is the solution of the TSP on (Γ, w), then the median is given by $S = s_1 s_2 \ldots s_n$.

Proof: $\Phi(S,A) + \Phi(S,B) + \Phi(S,C) = \sum_{gh \in S, g \neq -h} w(gh)$

$$= nZ + \sum_{gh \in s} w(gh).$$

Thus S minimizes $\Phi(S,A) + \Phi(S,B) + \Phi(S,C)$ iff s is of minimal weight.

3.4 Median Algorithm Applied Iteratively to Phylogeny Decomposed into Overlapping Triples.

A general method for the inference of ancestral genomes on a binary tree with a fixed topology is the iterative improvement method of [59], as adapted for the genomics context in [23,64]. Each of the $N-2$ internal vertices, together with its three neighbors, defines a 3-star. The solution to the Steiner point problem will have a reconstructed genome associated with each such vertex, which must be a solution to the median problem determined by these neighbors.

Then the following algorithm, in which we leave unspecified how to set up the initial TSP for each genome to be reconstructed, converges to a (local) optimum:

algorithm optimize_tree
input $G^1, \cdots, G^N$
cost $\leftarrow \infty$
extremities $\leftarrow \{1, \cdots, N\}$
internal $\leftarrow \{N+1, \cdots, 2N-2\}$
do for $M = N+1, \cdots, 2N-2$,
  set_up_TSP for G^M
  solve TSP for G^M
  remove the neighbors of V_M preceding it in the vertex numbering from

extremities
transfer V_M from *internal* to *extremities*
enddo

routine iterate_median
output $G^{N+1}, \cdots, G^{2N-2}$

In each of Sections 3.4.2, 3.4.3 and 3.4.4 below, the **set_up_TSP** instruction will be replaced by a specific routine. The **iterate_median** routine is independent of the set-up strategy in the initialization; in fact all three approaches to be used are identical for 3-leaf trees (i.e. the median problem).

routine iterate_median
while $C = \sum_{V_iV_j \in E(T)} \Phi(G^i, G^j) < cost$,
$cost \leftarrow C$
do for $M = N+1, \cdots, 2N-2$,
$G^* \leftarrow$**median**(G^h, G^j, G^k), where V_h, V_j and V_k are the neighbors of V_M
if $\sum_{x \in \{h,j,k\}} \Phi(G^*, G^x) \leq \sum_{x \in \{h,j,k\}} \Phi(G^M, G^x)$
$G^M \leftarrow G^*$
endif
enddo
endwhile

3.4.1 Initialization strategies.

The output of this algorithm is not necessarily a global optimum. The main factor in directing convergence towards a global optimum, and the focus of this paper, is the how the initialization is carried out.

A promising initialization, which makes use of the most pertinent input data for each internal node, bases the initial TSP on the three nearest data genomes. In Section 3.4.2 we will use this latter idea as the basis of one of our heuristics, **three_nearest**. In addition, in Section 3.4.3, we define an initial TSP at each internal node, where the edge-weights are the average of the corresponding edge-

weights at the three neighbouring nodes of the tree under consideration, found by solving a system of linear equations. Finally, in Section 3.4.4, we introduce an initialization method which involves setting up and solving an initial TSP at each internal node, where the edge-weights are calculated by dynamic programming, minimizing the number of times a given adjacency has to be created or disrupted within the tree to be present or absent, respectively, at that node.

It can be seen in **optimize_tree** that rather than initializing all internal nodes at once, they are initialized more "cautiously", i.e. one at a time, starting with an internal node with two terminal node neighbours. Once it is initialized, it is treated as a terminal node as the initialization proceeds, and its two neighbours are disregarded.

Without loss of generality, we may assume that the internal vertices are numbered in such a way that of the three neighbors of each vertex, two either precede it in the list or are leaves. This assures that if genomes for the internal vertices are inferred one by one according to this numbering, the set of untreated vertices, as it shrinks, at all times forms a connected tree.

3.4.2 Triangulation.

We can replace the **set_up_TSP** instruction in **optimize_tree** by the following:

routine three_nearest
let V_h, V_j, V_k be the three vertices in *extremities* closest to V_M
on three disjoint paths leading from V_M
define TSP for G^M, based on V_h, V_j, V_k.

3.4.3 Trees of TSPs.

Instead of setting up the TSP at each internal vertex as a function of the three closest previously solved genomes, we can define a TSP on the basis of the three immediately neighboring TSPs. For each leaf V_M of the tree, we set

$$w_M(gh) = \begin{cases} 1 & \text{if } gh \text{ is not in } G^M \\ 0 & \text{if } gh \text{ is in } G^M \end{cases}$$

We then determine the weights for the internal vertices as follows:

$$w_M(gh) = \frac{1}{3}(w_h(gh) + w_j(gh) + w_k(gh)),$$

for each $gh \in \Gamma$, where V_h, V_j and V_k are the three neighbors of V_M. The weight system $\mathbf{w}$ can then all be easily found by solving the system of simultaneous equations derived from all the internal vertices of T.

We can replace the **set_up_TSP** instruction in **optimize_tree** by the following:

routine average_TSP
calculate **w** for the vertices in *internal* based on the vertices in *extremities*

3.4.4 Minimizing Adjacency Disruptions.

Our third heuristic focuses first on each pair of genes in Σ and tries to minimize the number of times this pair is inferred to have been directly affected by rearrangement of the genome. Dynamic programming is used to calculate the weights for the TSP.

For any internal vertex V_M, suppose we have already calculated a genome for vertices $V_{N+1}, \cdots, V_{M-1}$ and we wish to do so for V_M. We impose a direction on all edges of the tree, namely the direction leading to V_M. Then V_M has three

edges leading to it, all other internal vertices have two, and leaves have none. The dynamic programming routine included in the set-up routine below follows this direction towards V_M.

routine adjacency_parsimony
direct all edges in $E(T)$ towards M
do for $i \in$ *extremities* and all $gh \in \Gamma$
$w_i^+(gh) \leftarrow 0$ if $ij \in G^i$, $w_i^+(gh) = 1$ if $ij \notin G^i$.
$w_i^-(gh) \leftarrow 1$ if $ij \in G^i$, $w_i^-(gh) = 0$ if $ij \notin G^i$.
enddo
remain $\leftarrow$ *internal*
while *remain* $\neq \Phi$
find $i \geq M, i \in$ *remain*, such that for all vertices j leading to $i, j \notin$ *remain*
do for all $gh \in \Gamma$
$w_i^+(gh) \leftarrow \Sigma_{V_j \text{ leads to } V_i} \min(w_j^+(gh), 1 + w_j^-(gh))$
$w_i^-(gh) \leftarrow \Sigma_{V_j \text{ leads to } V_i} \min(w_j^-(gh), 1 + w_j^+(gh))$
enddo
remove i from *remaining*
endwhile
do for all $gh \in \Gamma$
$w_M(gh) \leftarrow w_M^+(gh) - w_M^-(gh)$
enddo

3.5 The Simulations

To assess and compare the three approaches to initializing the iteration of the median algorithm, a series of simulations were carried out. The parameters were N, the number of terminal vertices in the tree, n, the number of genes in each genomes, and r, the total number of breakpoints between all pairs of adjacent genomes in the tree. Here, we illustrate with the results for $N = 7$ and $n = 20$. The total number of rearrangements r was varied from 20 to 300 in steps of 10.

For each value of r, 10 trees were generated, each starting with genome $(12 \cdots n)$ at one vertex and generating neighboring vertices with the appropriate random number of rearrangements until all internal and terminal vertices were assigned a genome. Each rearrangment was randomly chosen to be a transposition

or an inversion (cf [10]), of random length.

Once all genomes were generated, the breakpoints on each edge were counted, and the tree was retained only if r was one of our target values for which we had not yet our quota of 10 examples. The genomes from the terminal vertices only then served as input for each of our three algorithms separately.

For solving our TSP problems we used C. Hurwitz' **tsp_solve** software on an Origin 200 computer with a RISC 10000 processor.

3.6 Results

It can be seen from Figure 3, that at when the average number of breakpoints per edge approaches $\frac{1}{2}n$, the algorithm tends to reconstruct evolutionary histories more parsimonious than those actually responsible for the data. After $\frac{2}{3}n$, the number of of reconstructed breakpoints actually levels off sharply.

The accuracy of our initializations can be assessed in Figure 4, which gives the improvement to the objective R obtained by the iteration step as a function of r for the three heuristics. This improvement is generally less than $\frac{1}{2}\%$, reaching more than 1% for the **average_TSP** initialization only for values of r where, as we shall see, this routine performs relatively poorly.

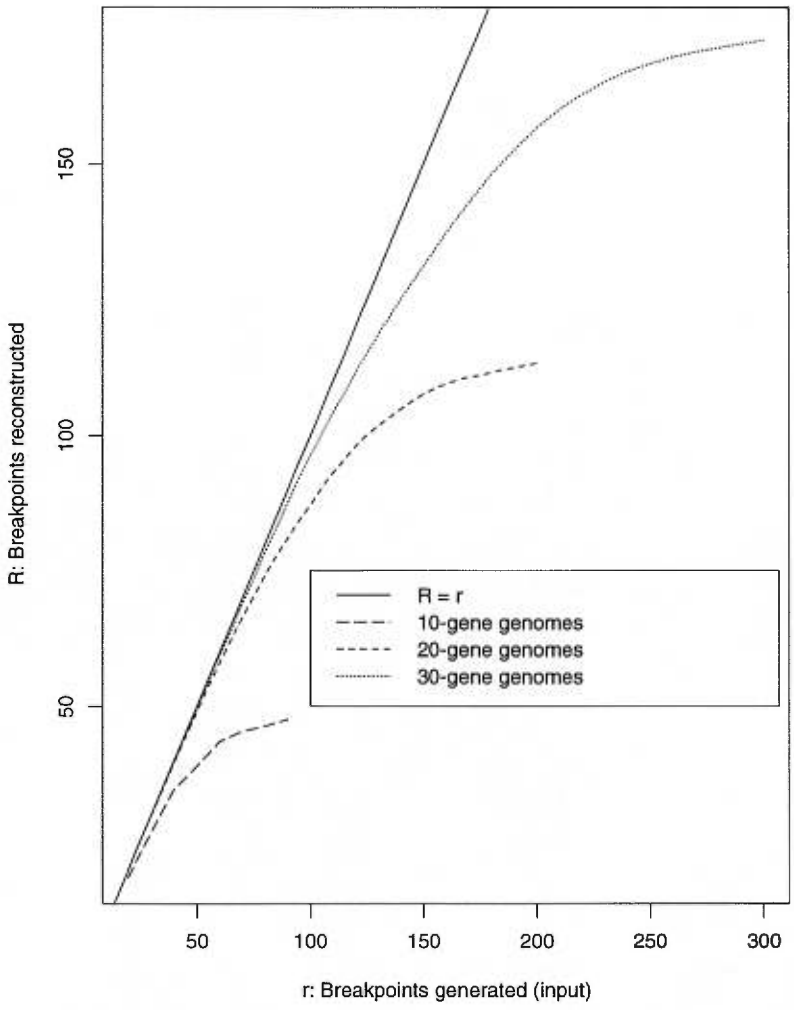

FIGURE 3. Number of reconstructed breakpoints R (best of three heuristics) as a function of number of breakpoints generated in the input data, for 10-gene, 20-gene and 30-gene genomes. Number of leaves $N = 7$, number of branches, $2N - 3 = 11$. The results of each n replaced by a smooth curve (generated using splines in the SPLUS package).

Figure 5 compares the performance of the two heuristics **average_TSP** and **adjacency_parsimony** (both outperform **three_nearest**) over a range of evolutionary divergences. It is striking that for small r, **adjacency_parsimony** performs distinctly better, even after both initializations benefit from the iterative

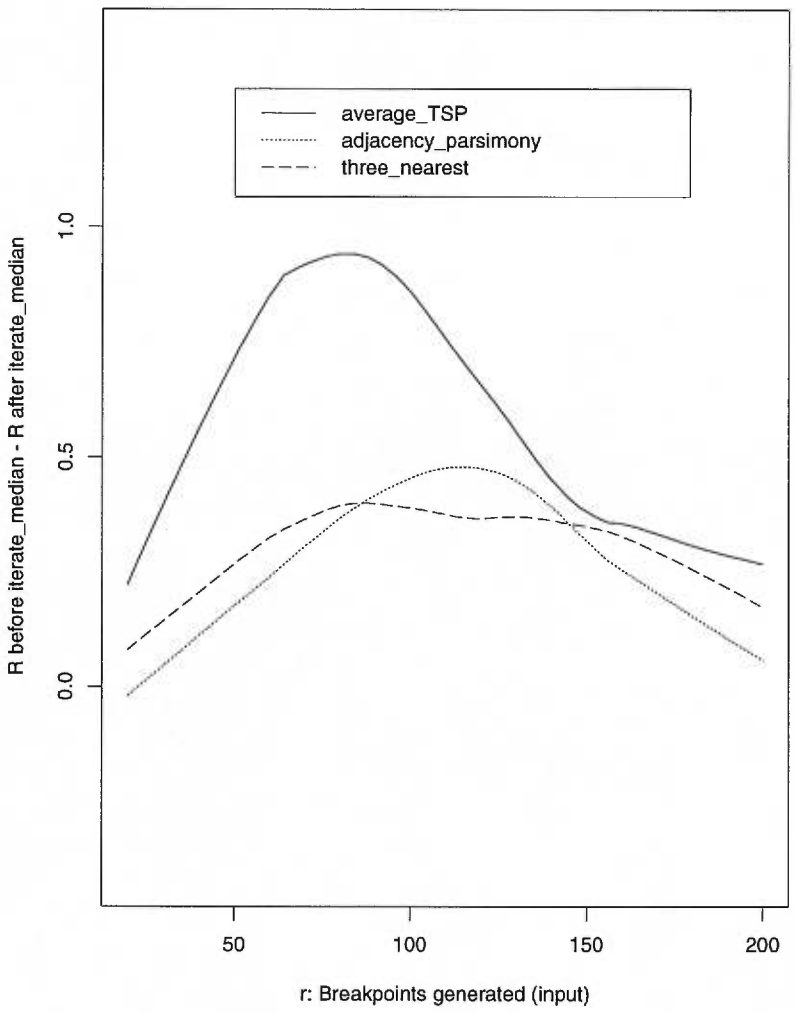

FIGURE 4. Decrease in number of reconstructed breakpoints R for each heuristic obtained through iteration step, as a function of number of breakpoints generated in the input data. $n = 20, N = 7$. Results for each heuristic replaced by a spline fit.

improvements, while for large r it is the **average_TSP** which is clearly superior.

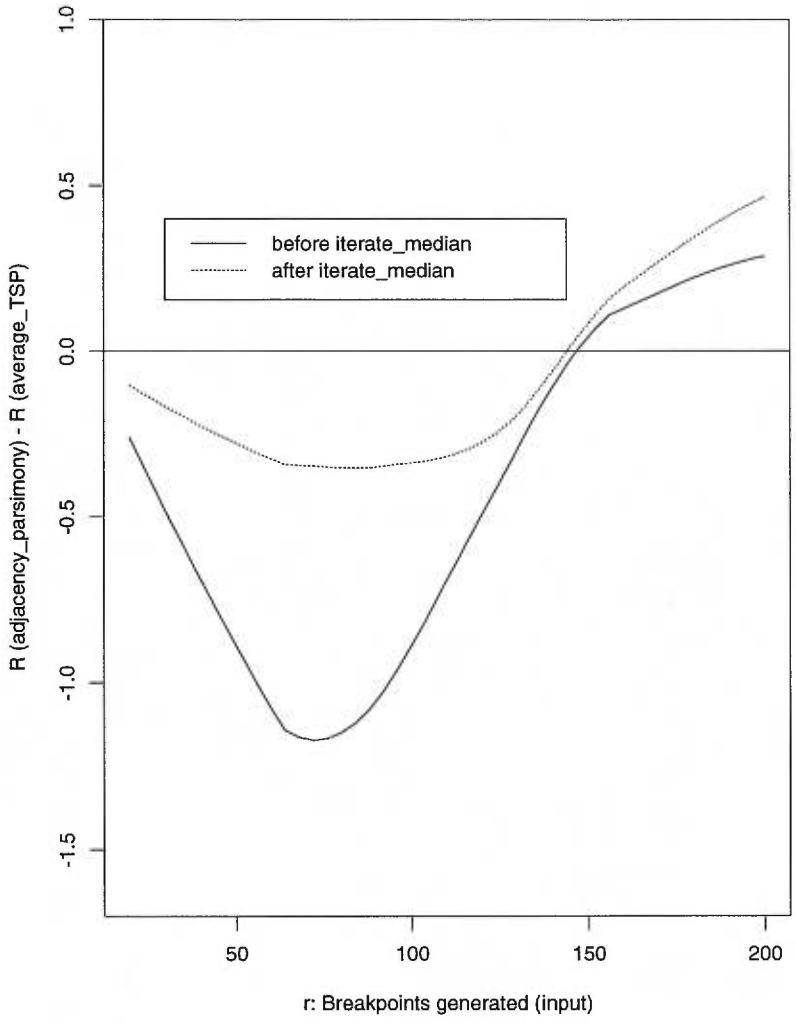

FIGURE 5. Difference between results of **adjacency_parsimony** and **average_TSP** as a function of r, before and after iterative improvements. $n = 20, N = 7$. Results for each set of differences replaced by a spline fit.

To address the question of global optimality, we count how many heuristics give the minimum solution for R. In Figure 6, we see that (except for genomes that have diverged very little) around 1.6 heuristics, on the average, seem to obtain the minimum. Assuming a doubly-attained minimum is a global solution (not always valid, of course), and since **adjacency_parsimony** and **average_TSP** are the

ones that tend to achieve the lowest values, we can estimate that individually they attain global optimality about half of the time.

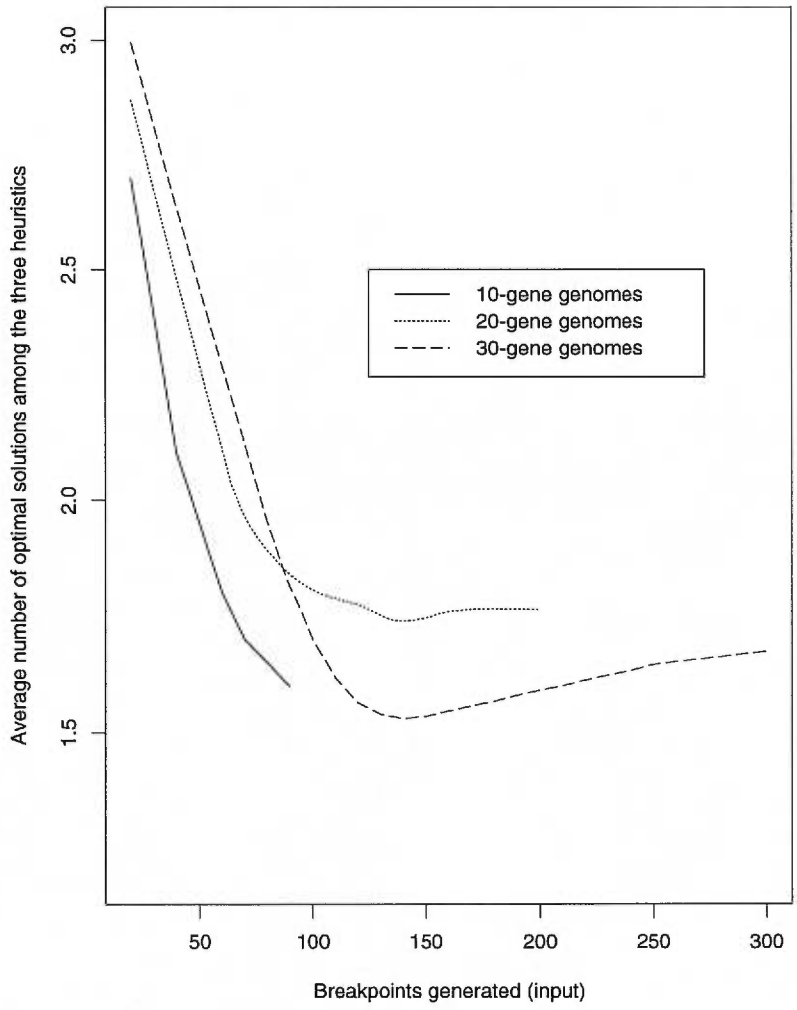

FIGURE 6. Number of heuristics (out of three) attaining optimal solution as a function of number of breakpoints generated in the input data, for 10-gene, 20-gene and 30-gene genomes. $N = 7$. The results for each n replaced by a spline fit.

3.7 Summary and Conclusions.

We have proposed and tested three initializations for solving the breakpoint phylogeny problem by iterative improvement. We showed that the initializations were very precise, within one percent or so of the best solution. The obverse of this is that the iterative step leads to a small, but non-negligible, improvement.

We were able to identify one initialization which worked better for low-divergence data and one which is superior for high-divergence data. Studying the rate of coincidental solutions among the three heuristics enabled us to assess how frequently the methods are likely to achieve global optima.

We have found at what point parsimony leads to underestimation of the number of events generating the data. In another paper [58], we analyze the multiplicity of equivalent local minima and the breakpoint distances amonst them, as an assessment of the reliability of reconstructed gene orders.

An important assumption in this work has been the fixed set of genes present in the data genomes. This is unrealistic in many contexts, but relaxing it makes the median problem, and hence, phylogenetic reconstruction, much more difficult [57]. Further work involves non-binary trees, as reported in [58].

Acknowledgements

Research supported by grants to DS from the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Canadian Genome Analysis and Technology program, and a NSERC fellowship for graduate studies to MB. DS is a Fellow of the Canadian Institute for Advanced Research.

CHAPITRE 4

Multiple genome rearrangement [58]

David Sankoff [1] Mathieu Blanchette [2]

[1]Centre de recherches mathématiques, Université de Montréal, CP 6128 Succursale Centre-ville, Montréal, Québec H3C 3J7. E-mail: sankoff@ere.umontreal.ca.

[2]Centre de recherches mathématiques, Université de Montréal, CP 6128 Succursale Centre-ville, Montréal, Québec H3C 3J7. E-mail: blanchem@iro.umontreal.ca.

4.1 Introduction

Multiple alignment of macromolecular sequences, an important topic of algorithmic research for at least 25 years [56, 63], generalizes the comparison of just two sequences which have diverged through the local processes of insertion, deletion and substitution. Recently there has been much interest in gene-order sequences which diverge through non-local genome rearrangement processes such as inversion (or reversal) and transposition (reviewed in [65] ch.7, [53], [27] ch. 19 and [10]). What would be the analog of multiple alignment under these models of divergence? In this introduction we first review some formulations of multiple alignment and show which have counterparts in multiple rearrangement. We then discuss the difficulties inherent in edit-distance formulations of multiple rearrangement, referring to relevant work, and argue for a potentially simpler approach based on "breakpoint analysis".

4.1.1 Multiple sequence alignment

The goal of multiple sequence alignment is to align the terms of N sequences into a number of relatively homogeneous columns through the judicious insertion of one or more null terms, or gaps, between consecutive terms in some or all of the sequences, as in Figure 7, so as to optimize an objective cost function.

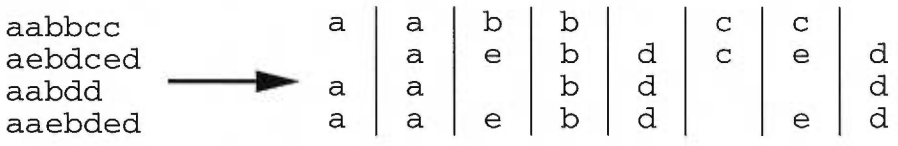

FIGURE 7.

In the simplest case, this objective is just the sum of column costs across all

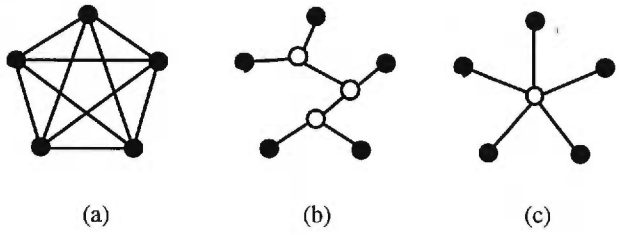

FIGURE 8.

columns of the alignment. Each column cost measures how different the terms in that column are among themselves.For example, in a "complete" comparison the column cost is the number of pairs of sequences which differ in that column, as represented in (a) in the figure 8,in which every vertex (sequence) is compared to every other.

Another definition of column cost depends on a given phylogenetic tree, as in (b) in the figure 8, in which the leaves are the data sequences and the internal nodes (open dots) are hypothetical ancestral sequences reconstructed by some method from the contemporary sequences. Both data and reconstructed sequences must be aligned and the column cost is just the number of tree branches where the sequences at the two ends have different elements in the column. A special case of the tree-based comparison is the "consensus" comparison, represented in (c), where there is just one reconstructed sequence.

There are many other formulations of the problem which we will not discuss, e.g. the column cost may be $N - M$, where M is the number of occurrences of the most frequent term in a column.

4.1.2 The analogy with genome rearrangement

A key difference between sequence comparison and gene-order comparison is that in the former, algorithms try to identify corresponding terms in the two sequences being compared and the number of divergence steps then falls out directly, whereas in the latter the correspondence (i.e. alignment) is given and it is the number of steps which must be calculated.

Thus version (a) of the multiple alignment problem has no analog in gene-order rearangement, since there is nothing to optimize once the pairwise distances are given. On the other hand, in versions (b) and (c) of the problem, there is something to optimize, namely the ancestral gene orders represented by the white dots in Figure 8. This is the focus of this article.

4.1.3 Difficulties and a solution

There have been a number of investigations of phylogeny based on the algorithmic comparison of gene order within a number of genomes, using *pairwise* comparisons followed by *distance matrix* methods (e.g. [62]). However, treeing methods which involve the optimal reconstruction of gene order at ancestral nodes [31, 64] have been little used because of the computational difficulty in generalizing measures of genomic distance to more than two genomes. Caprara has recently shown that the most promising case, reversal distance for only three signed permutations, is NP-hard [15]. In [57] and [8] we argued that

(i) this computational difficulty – there are no algorithms guaranteeing exact solutions for even three relatively short genomes – together with

(ii) unwarranted assumptions as to the relative importance of different rear-

rangement events implicit in genome distances such as minimum reversal distance, minimum transposition distance, minimum translocation distance and even distances combining these (cf. [10]), as well as

(iii) the fallacy that calculation of an edit distance allows the recoverability of the "true" history of genomic divergence – in fact, the severe non-uniqueness of the optimal edit path for moderate or large gene-order distances has much worse (i.e. non-local) consequences than with the classical multiple alignment problem, and

(iv) the bias in simulations, where calculation of genome distance severely underestimates the actual number of events generating moderate or large gene-order differences,

all militate in favour of extending gene-order comparisons to three or more genomes through a much simpler and model-free metric. In this paper we suggest the number of breakpoints as just such a metric. We show how breakpoint analysis addresses all four of these problems.

In the Section 5.3.2 we define this measure for unoriented and oriented genomes, and set up the analogy to multiple alignment modes (b) and (c) above. In Section 4.3 we review how (c), consensus-based multiple rearrangement, can be solved exactly through reduction to a version of the Travelling Salesman Problem, as proposed in [57]. In Section 4.4, we use this exact solution applied to simulated data to show how the problem of non-uniqueness is attenuated with increasing numbers of data genomes. In Section 4.5, we report how (b), tree-based multiple alignment, can be achieved to a great degree of accuracy by decomposing the tree into a number of overlapping 3-stars centred on the non-terminal nodes, and solving the consensus-based problem iteratively for these nodes until convergence. The accuracy depends on very careful initializations at the non-terminal nodes, as

assessed through simulation in [8]. In Section 4.6 we investigate non-uniqueness by means of simulation and accurate heuristics again, this time focusing on the effect of the position of the node in the tree in terms of path length to the terminal vertices.

4.2 Breakpoint analysis

Consider two genomes $A = a_1 \dots a_n$ and $B = b_1 \dots b_n$ on the same set of genes $\{g_1, \dots, g_n\}$. We say a_i and a_{i+1} are adjacent in A. We also consider that a_1 is adjacent to the genome "start" and a_n is adjacent to the "end". For circular genomes, it suffices to consider that a_n and a_1 are adjacent. If two genes g and h are adjacent in A but not in B, they determine a breakpoint in A. We define $\Phi(A, B)$ to be the number of breakpoints in A. This is clearly equal to the number of breakpoints in B.

The number of breakpoints between two genomes is not only the most general measure of genomic distance, requiring no assumptions about the mechanisms of genomic evolution (inversion versus transposition versus translocation) underlying the data, but it is also the easiest to calculate. In addition, it has proven relations to the edit distances; e.g. half the number of breakpoints is a lower bound on the reversal distance.

4.2.1 Oriented genomes

Our simulations will involve directed, or oriented, genomes; we assume we know the strandedness, or direction of transcription, of each gene in each genome in the data set. In this case, the notion of breakpoint must be modified to take into account the polarity of the two genes [57]. If gh represents the order of two

genes in one genome, then if another genome contains gh or $-h-g$ there is no breakpoint involved. However, between gh and hg there is a breakpoint, similarly between gh and $-g-h, g-h, -gh, h-g$ or $-hg$.

4.2.2 Tree-based multiple genome rearrangement

The problem is formulated as follows. Let T=(V,E) be an unrooted binary tree with $N \geq 3$ leaves and $\Sigma = \{g_1, \cdots, g_n\}$ be a set of genes. Suppose $\{V_1, \cdots, V_N\} \subset V(T)$ are the leaves of the tree and $\{V_{N+1}, \cdots, V_L\}$, where $N < L \leq 2N-2$, are the internal vertices of the tree. The data consist, for each leaf $V_i, i = 1, \cdots, N$, of a circular permutation $G^i = g_1^i \cdots g_n^i$ of the genes in Σ, representing the genome of a contemporary species. The task is to find the permutations $G^{N+1}, \cdots, G^L$ associated with the internal (ancestral) vertices $V_{N+1}, \cdots, V_L$, such that

$$\sum_{V_iV_j \in E(T)} \Phi(G^i, G^j)$$

is minimized.

4.2.3 Binary tree- versus consensus-based multiple genome rearrangement

We will concentrate on two extreme cases: that of completely resolved, or binary, trees, where $L = 2N-2$ and all non-terminal nodes are of degree 3; and that of completely unresolved trees, or "stars", where $L = N+1$ and the single non-terminal node has degree N.

4.3 Consensus-based rearrangement

Though all these breakpoint-based multiple rearrangement problems seem NP-hard, as reported for $N = 3$ by Pe'er and Shamir [51], for moderate n they are tractable, since they may be reduced to a number of interconnected instances of the Traveling Salesman Problem (TSP). In the case of consensus-based rearrangement, the solution, involving just one TSP, is globally optimal.

We define Γ to be the complete graph whose vertices are the elements of Σ. For each edge gh in $E(\Gamma)$, let $u(gh)$ be the number of times g and h are adjacent in the N data genomes. Set $w(gh) = N - u(gh)$. Then the solution to TSP on (Γ, w) traces out an optimal genome S on Σ, since if g and h are adjacent in S, but not in V_1, for example, then they form a breakpoint in S.

For oriented genomes, the reduction of the median problem to TSP must be somewhat different to take into account that the median genome contains g or $-g$ but not both. Let Γ be a complete graph with vertices $V(\Gamma) = \{-g_n, \ldots, -g_1, g_1, \ldots, g_n\}$. For each edge gh in $E(\Gamma)$, let $u(gh)$ be the number of times $-g$ and h are adjacent in the N data genomes and $w(gh) = N - u(gh)$, if $g \neq -h$. If $g = -h$, we simply set $w(gh) = -Z$, where Z is large enough to assure that a minimum weight cycle must contain the edge $-gg$.

Proposition: If $s = s_1, -s_1, s_2, -s_2, \ldots, s_n, -s_n$ is the solution of the TSP on (Γ, w), then the median is given by $S = s_1 s_2 \ldots s_n$.

Proof: $\sum_{i=1}^{N} \Phi(S, V_i) = \sum_{gh \in S, g \neq -h} w(gh)$

$$= nZ + \sum_{gh \in s} w(gh).$$

Thus S minimizes $\sum_{i=1}^{N} \Phi(S, V_i)$ iff s is of minimal weight.

4.4 The uniqueness of the consensus

The reduction to TSP allows us to obtain global solutions for moderate-sized problems in reasonable time – typically 5 seconds for reconstructing the consensus of three or more scrambled genomes with 20 genes, on an Origin 200 computer with a RISC 10000 processor. This enables us to undertake systematic simulation studies. To assess the uniqueness of the solutions to the problem as a function of the number of genomes simultaneously rearranged, we constructed N genomes, each by applying a number R of random reversals to a common ancestor $(1, 2, \cdots, 20)$, and then solved the consensus problem. (Each reversal reverses the order of a number of consecutive terms, and also changes the sign of each of these terms. It adds at most two new breakpoints. We used reversals not because we privilege them as a model of biological evolution but simply as a convenient way of scrambling permutations.)

Though the TSP software we used (C.Hurwitz' tsp-solve) only produces one result, we were able to search for other results by permuting the labels of the 20 genes. We repeated each example with 10 different gene labelling.We compared all 10 solutions obtained by averaging their pairwise distances (i.e. the number of breakpoints between the two solution genomes). For each value of N between 2 and 16, and each value of R between 1 and 12, we repeated the experiment for 10 different examples, and averaged their results, running the programme 100 times – 10 examples times 10 gene labellings per example.

The figure 9 above shows the results of these experiments. We note first that the curves for large R are systematically "worse" than those for low R; the more scrambled are the data genomes, the less likely they are to have a unique consensus.

More interesting perhaps, is the rapid, almost linear, decrease in non-unicity

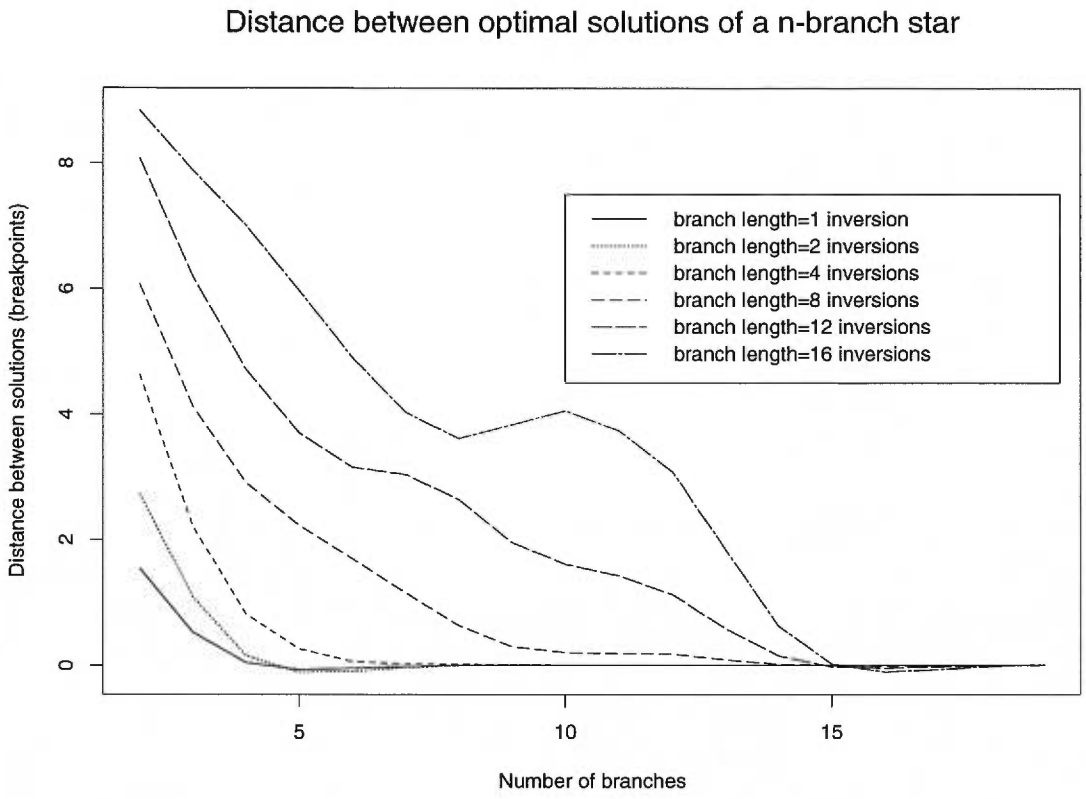

FIGURE 9.

for each R as the number of data genomes increases. For 15 branches or more, there seems to be a unique consensus, no matter how large R is. And in all cases examined, this consensus order was none other than $(1, 2, \cdots, 20)$. Of course, for larger genomes, we could expect a larger cut-off point.

4.5 Binary tree-based rearrangement

A general method for the inference of ancestral genomes on a fixed binary tree is the iterative improvement method of [59], as adapted for the genomics context in [23, 64]. Each of the $N - 2$ internal vertices, together with its three

neighbors, defines a 3-star. The solution to the tree-based multiple rearrangement problem will have a reconstructed genome associated with each such vertex, which must be a solution to the consensus-based problem determined by these neighbors.

The strategy is to start with an initial tree where some genome is assigned to each internal genome, then to improve one of these ancestral genomes at a time by solving the consensus problem for the 3-star consisting of its immediate neighbours (the "median problem"), iterating across the tree until convergence. Of course, there is no guarantee that convergence will occur at a global optimum. The following algorithm formalizes this approach.

algorithm optimizetree
input $G^1, \cdots, G^N$
initialize each of $G^{N+1}, \cdots, G^{2N-2}$ to some genome.
$cost \leftarrow \infty$
routine iteratemedian
output $G^{N+1}, \cdots, G^{2N-2}$

routine iteratemedian
while $C = \sum_{V_iV_j \in E(T)} \Phi(G^i, G^j) < cost$,
    $cost \leftarrow C$
    do for $i = N+1, \cdots, 2N-2$,
        $G^* \leftarrow$**median**(G^h, G^j, G^k), where V_h, V_j, V_k
                are the three neighbors of V_i
        if $\sum_{\{h,j,k\}} \Phi(G^*, G^x) < \sum_{\{h,j,k\}} \Phi(G^i, G^x)$
            $G^i \leftarrow G^*$
        endif
    enddo
endwhile

The **median** routine is just the solution of the consensus-based rearrangement problem for 3 genomes, based on the reduction to the TSP in Section 4.3.

The main factor in directing convergence towards a global optimum is the how the initialization is carried out. We can identify at least six distinct approaches to initialization, which can be grouped into three levels of increasing

likelihood that they fall into the domain of attraction of a global optimum. Thus each internal genome can be assigned:

"arbitrarily",

(1) a fixed, arbitrary permutation, e.g.$(1, 2, \cdots, n)$, or

(2) a different random permutation

"reasonably",

(3) the permutation representing a nearest data genome, or

(4) the consensus of three nearest data genomes

"with much effort",

(5) by setting up and solving an initial TSP at each internal node, where the edge-weights are calculated by dynamic programming, minimizing the number of times a given adjacency has to be created or disrupted within the tree to be present or absent, respectively, at that node [8], or

(6) by setting up and solving an initial TSP at each internal node, where the edge-weights are the average of the corresponding edge-weights at the three neighbouring nodes, found by solving a system of linear equations [8].

In addition, for the "reasonable" and "much effort" modes, all internal nodes can be initialized "recklessly", i.e. at once, or they can be initialized "cautiously", i.e. one at a time, starting with any internal node with two terminal node neighbours. Once it is initialized, it is treated as a terminal node as the initialization proceeds, and its two neighbours are disregarded.

In [8], we investigated methods (4),(5) and (6) above, the latter two incorporating the "cautious" approach. Simulations on trees generating seven to fifteen 20- and 30-gene data genomes showed that "much effort" paid off with one to two percent better results (i.e. fewer breakpoints over the entire tree) for moderate to highly divergent data. The dynamic programmming approach (5) led to better results than method (6) for moderately divergent data while the latter was superior for highly divergent data, approaching randomness, in each case by about one half of one percent.

In addition, we found that once the number of breakpoints generated per tree branch reached about half the number of genes, underestimation began to be manifested, rapidly worsening so that when the number of breakpoints per branch reached two thirds the number of genes, this number was underestimated by about 30%.

4.6 Uniqueness in tree-based rearrangement

In our investigation of multiple optima in reconstructed genomes, we simulated genomic rearrangement in the tree of figure 10. We again used genomes of size 20 and generated trees containing a total of B breakpoints over all their edges. We applied methods (5), (6) and (7) of Section 4.5 to calculate the multiple rearrangement for each tree. When at least two heuristics found the same total cost (which happened at least 70% of the time, even for highly divergent data), we calculated the distances between the genomes reconstructed by each method at each internal node. This experiment was repeated 10 times (and the results averaged) for each value of B between 40 and 230. The results appear in the figure 11.

The results of these simulations indicate that multiple solutions for peripheral

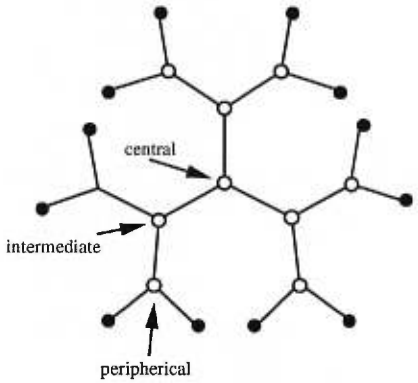

FIGURE 10.

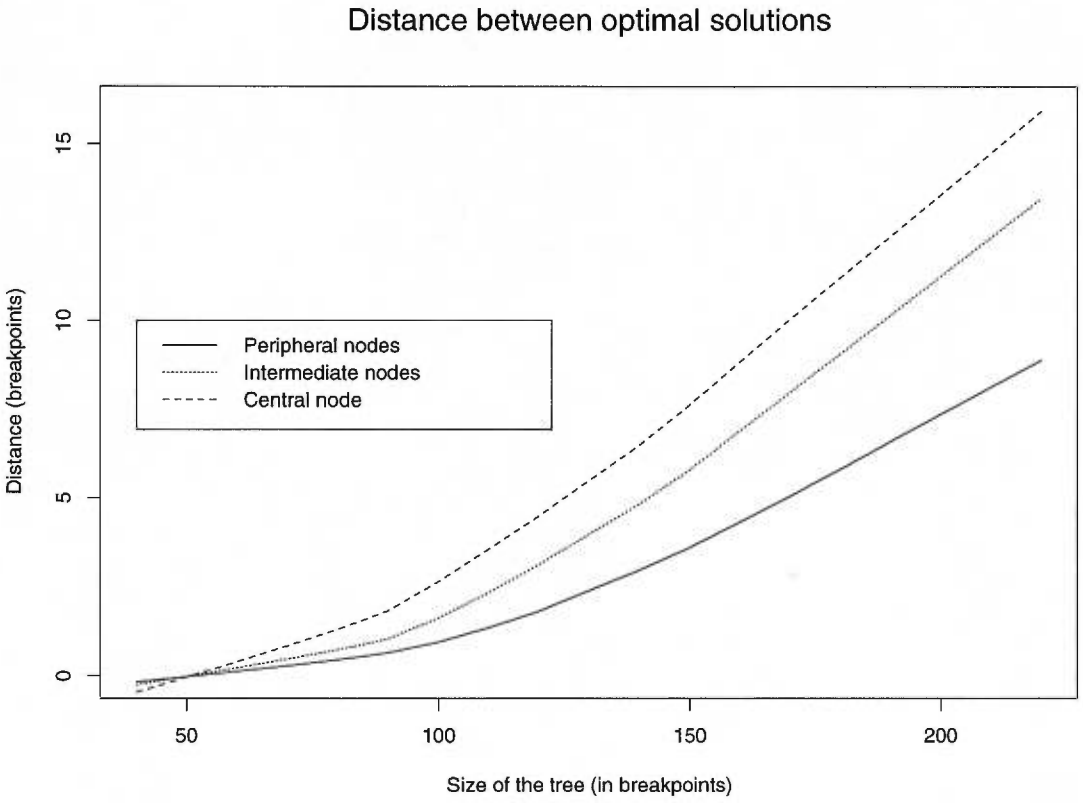

FIGURE 11.

nodes are relatively close to each other for low and moderate divergence. But with 10 breakpoints per branch, on an average, somewhat more than 200 breakpoints in all, multiple solutions could be non-negligibly far from each other – about 7 breakpoints between them on the average. The situation is progressively worse as we get deeper into the tree, so that there are considerably more breakpoints (around 15) between two solutions for the central genome than between two neighbouring nodes on the tree.

4.7 Conclusions.

Our previous work has established the feasibility of breakpoint analysis as a method of multiple genome rearrangement [57] compared to the difficulties in edit distance-based approaches, and assessed the relative reliability of various heuristics in achieving a global optimum [8]. To feasibility and reliablity, we add here the study of accuracy of genomic reconstruction, in terms of an analysis if the multiplicity of equivalent local minima and the breakpoint distances amongst them. Non-uniqueness remains a major consideration in genomic reconstruction, but we would conjecture that this is less of a problem in breakpoint analysis than with other approaches.

An important assumption in this work has been the fixed set of genes present in the data genomes. This is unrealistic in many contexts, but relaxing it makes multiple rearrangement and genomic reconstruction, much more difficult [57].

Acknowledgements

Research supported by grants to DS from the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Canadian Genome Analysis

and Technology program, and a NSERC fellowship for graduate studies to MB. DS is a Fellow of the Canadian Institute for Advanced Research.

CHAPITRE 5

Gene order breakpoint evidence in animal mitochondrial phylogeny [9]

Mathieu Blanchette [1] Takashi Kunisawa [2] David Sankoff [3]

Abstract

Multiple genome arrangement methodology, based on minimization of total breakpoints, is used in an investigation of animal phylogeny through the gene order on mitochondrial genomes. Breakpoint distance is situated in the context of genomic distances in general and its significance is evaluated in relationship to other measures when applied to the data. We compare a number of theories of metazoan evolution to phylogenies reconstructed from the breakpoint distance matrix or from ancestral genome optimization. We also discuss the effects of small gene (i.e. tRNA) mobility and the effect of "long branches". We introduce the notion of "unambiguously reconstructed segments" as a way of extracting the invariant aspects of multiple solutions for a given ancestral genome, and use this to characterize the evolution of non-tRNA mitochondrial gene order.

[1]Laboratoire de biologie informatique et théorique, Université de Montréal, CP 6128 Succursale Centre-Ville, Montréal, Québec H3C 3J7. E-mail: blanchem@iro.umontreal.ca.

[2]Department of Applied Biological Sciences, Science University of Tokyo, Noda 278,Japan. E-mail: kunisawa@jipdalph.rb.noda.sut.ac.jp.

[3]Centre de recherches mathématiques, Université de Montréal, CP 6128 Succursale Centre-ville, Montréal, Québec H3C 3J7. E-mail: sankoff@ere.umontreal.ca. Research Supported by grants from the Natural Sciences and Engineering Research Council of Canada and the Canadian Genome Analysis and Technology program. DS is a fellow of the Canadian Institute for Advanced Research.

5.1 Introduction.

In comparative genomics, the quantitative comparison of gene order differences can be used for phylogenetic inference about a set of organisms. As in quantitative phylogenetics more generally, there is a fundamental distinction between methods that start with a matrix of pairwise distances (or differences, or similarities) among the organisms, and then infer the phylogeny through some type of cluster analysis (e.g. [55]), either stepwise or globally computed, and methods which reconstruct some aspect of the genome at each of its ancestral nodes as an essential part of the optimization of a global objective function on the space of trees (e.g. [64]). The first approach is easier, especially when dealing with gene orders, since only pairwise distances need to be calculated and the resulting matrix input into any one of a number of efficient tree construction algorithms. The second approach has the advantage that it does not reduce the multidimensional comparison between two genomes into a single number before building the tree, retaining information potentially very pertinent to the tree structure being inferred. In addition, this latter approach locates the ancestral nodes in the same space as the data nodes; these locations are optimized simultaneously with the tree topology.

Most of the distances of comparative genomics (e.g. the minimum edit distances calculated by Hannenhalli and Pevzner, [30], [31]) are more conducive to the distance-matrix approach to phylogenetic inference, simply because the generalization to the comparison of more than two genomes has not proved computationally feasible for even moderately sized genomes ([15]). An exception to this is the breakpoint distance ([71]). Though an NP-hard problem ([51]), generalization of breakpoint distance to the simultaneous comparison of three or more genomes – multiple genome rearrangement – can be reduced to an instance of the Traveling Salesman Problem (TSP) which is quite tractable for moderate

size genomes ([57]).

We have previously shown how to incorporate multiple genome arrangement into an iterative heuristic for the optimization of ancestral genome reconstruction on a fixed topology phylogenetic tree, and demonstrated through simulation the precision that can be achieved through careful initialization, as well the relative proximity of the different optimal solutions ([8], [58]). The present paper represents the first application of this methodology to real data, an investigation of animal phylogeny through the gene order on mitochondrial genomes.

We discuss metazoan phylogeny, including ongoing debates about the relationships among the major metazoan branches, in Section 2, as well as the available genomic data. We then situate breakpoint distance in the context of genomic distances in general (Section 3) and evaluate its significance in relationship to other measures when applied to the data. In Section 4 we compare a number of theories of metazoan evolution to phylogenies reconstructed from the breakpoint distance matrix or from ancestral genome optimization. Here we also discuss the effects of small gene (i.e. tRNA) mobility and the effect of "long branches", i.e. highly divergent genomes. In Section 5, we introduce the notion of "unambiguously reconstructed segments" as a way of extracting the invariant aspects of multiple solutions for a given ancestral genome, and apply this to characterize rather closely the evolution of non-tRNA mitochondrial gene order.

5.2 The mitochondrial genome and problems in animal phylogeny.

Our goal here is to investigate gene order evidence pertinent to the phylogenetic relationships among the following major metazoan groupings: chordates (abbreviation: CHO), echinoderms (ECH), arthropods (ART), molluscs (MOL), annelids (ANN), and nematodes (NEM). Aspects of metazoan phylogeny are

controversial; among the groupings analyzed here, only the link of echinoderms and chordates seems undisputed. Many scholars would group annelids and molluscs as sister taxa, with arthropods related to these at a deeper level. Nematodes would represent the earliest branch on the metazoan phylogeny. But Rouse and Fauchald ([54]) have proposed reviving a traditional grouping ("Articulata") of annelids and arthropods as sister taxa which had been discredited by Eernisse ([21] and others. Lake has recently advocated a radical change linking arthropods and nematodes ([1]). Some of the published phylogenies are schematized in Figure 12. Given all the types of data and argument that have been marshalled in favour of various opposing positions, it is hardly our ambition to definitively resolve metazoan phylogeny on the basis of the order of some thirty-odd genes in the mitochondrial genome. It is rather to investigate what aspects of gene order evolution are consistent with gene-level and other phylogenetic evidence, and to what extent early genomes can be reconstructed during phylogenetic inference. Not that gene order is not valid phylogenetic evidence – we will indicate wherever pertinent how it discriminates among various theories of metazoan evolution.

As of April 1998, mitochondrial gene order is known for 56 metazoan species, including 36 chordates, seven arthropods, one annelid, five echinoderms, three molluscs and three nematodes. Thirty-seven genes are present in all species, except for ATPase 8, absent from the nematode genomes, and one of the two tRNA-Ser and one the two tRNA-Leu genes, absent from the snail Cepaea nemoralis.

To simplify the analysis and presentation, while retaining as much phylogenetic information as possible, we included in our analyses exemplars of the most diverse members of each group, excluding closely related species with identical or nearly identical mitochondrial gene orders. For example, the various chordate mitochondrial gene orders differ from the human order by one or two inversions

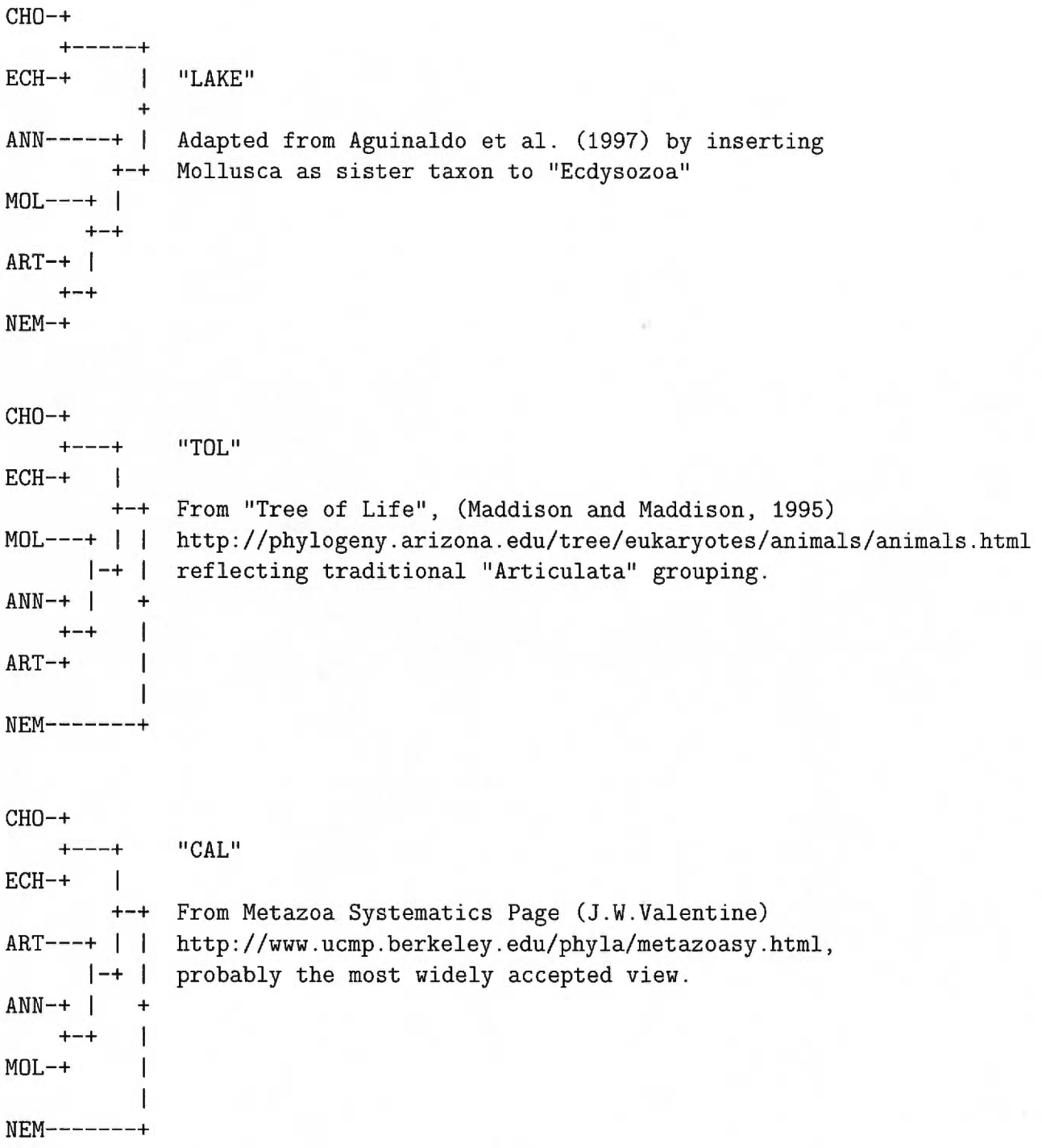

FIGURE 12. Three alternative views of metazoan evolution.

ORGANISM	GROUP
HU Human	CHO chordate
SS Asterina pectinifera (sea star)	ECH echinoderms
SU Strongylocentrotus purpuratus (sea urchin)	
DR Drosophila yakuba (insect)	ART arthropods
AF Artemia fransiscana (crustacean)	
AC Albinaria coerulea (snail)	
CN Cepaea nemoralis (snail)	MOL molluscs
KT Katharina tunicata (chiton)	
LU Lumbricus terrestris (earthworm)	ANN annelid
AS Ascaris suum	NEM nematodes
OV Onchocerca volvulus	

TABLE I. Mitochondrial genomes compared in this investigation, with assumed monophyletic groupings. Citations: [1] [2], [3], [12], [13], [11], [16], [32], [34], [41], [47], [52] and [67].

or transpositions, so we retained only the human one for our analysis. Similarly we selected only one insect, two echinoderms, and two nematodes. The species studied are listed in Table I. The table also presents the major groups for which we assume monophyly in some of our analyses to reduce the size of phylogenetic computations.

5.3 Genome rearrangement distances.

5.3.1 Edit distances

The algorithmic study of comparative genomics has focused on inferring the most economical explanation for observed differences in gene orders in two genomes in terms of a limited number of rearrangement processes. For single-chromosome genomes such as in the mitochondrion, this has been formulated as the problem of calculating an edit distance between two circular permutations of the same set of genes. For these purposes, degradation of homology at the sequence level of individual genes is not pertinent; once homology is established, the two genes are considered to be identical. A sign (plus or minus) is associated with each gene in a genome, representing the direction, or orientation, of its transcription. The elementary edit operations may include:

inversion, or reversal, of any number of consecutive terms, which also reverses the polarity of each term within the scope of the inversion. Increasingly efficient exact algorithms for this problem have been given by Kececioglu and Sankoff ([39]), Hannenhalli and Pevzner ([30]), Berman and Hannenhalli ([7]), and Kaplan *et al.* ([37]), whose version runs in quadratic time.

transposition of any number of consecutive terms from their position in the order to a new position between any other pair of consecutive terms. This may or may not also involve an inversion. No efficient exact algorithm is available for this problem. Sankoff *et al.* ([62]), Sankoff ([55]) and Blanchette *et al.* ([10]), Gu *et al.* ([26])) implemented and applied heuristics to compute edit distances which combine inversions, transpositions and deletions, in some cases allowing differential weighting of these operations.

There are a number of problems associated with the use of these distances. The first is that there is no *a priori* reason for using one of them, say transposition distance, versus another, perhaps inversion distance. Even for combined distances, the appropriate weights for the different operations may differ from context to context. Second, the reconstruction of evolutionary history implicit in calculating the distance is biased towards too few events, and is highly non-unique. Third, no exact algorithm is known for extending the distances to three or more genomes.

5.3.2 Breakpoint analysis.

Consider two genomes $A = a_1 \ldots a_n$ and $B = b_1 \ldots b_n$ on the same set of genes $\{g_1, \ldots, g_n\}$, where each gene is signed (+ or −). We say a_i precedes a_{i+1} in A, and a_n precedes a_1. If gene g precedes h in A and neither g precedes h nor $-h$ precedes $-g$ in B, they determine a breakpoint in A. We define $\Phi(A, B)$ to be the number of breakpoints in A. This is clearly equal to the number of breakpoints in B.

For two genomes whose gene sets are not identical, to calculate the breakpoints, we first remove all genes that are present in only one of the genomes. We then find the breakpoints for the reduced genomes, now of identical composition. The positions of the breakpoints are well-defined in the reduced genomes. In the full genomes, there is a breakpoint between a_i and a_{i+1} only if this is a breakpoint for the reduced genome.

The number of breakpoints between two genomes is not only the most general measure of genomic distance, requiring no assumptions about the mechanisms of genomic evolution (inversion versus transposition) underlying the data, but it is also the easiest to calculate (linear in n).

	AS-35.1										
AS		OV-35.3									
OV	25		HU-30.0								
HU	36	36		SS-33.4							
SS	36	36	27		SU-33.3						
SU	36	36	26	6		DR-30.9					
DR	36	36	21	33	33		AF-31.3				
AF	36	36	23	33	33	6		AC-35.1			
AC	35	36	36	35	35	35	35		CN-33.0		
CN	33	34	34	33	33	33	33	7		KT-29.1	
KT	34	33	29	34	34	22	23	34	32		LU-31.7
LU	34	35	32	34	34	29	30	34	31	24	

AS		21	35	33	36	36	35	35	32	33	33
OV	21.0		35	33	35	33	34	35	33	32	33
HU	34.3	35.3		28	24	20	23	35	33	28	31
SS	32.5	31.6	26.1		5	33	34	35	32	33	34
SU	34.3	34.2	24.2	3.0		32	33	34	31	32	32
DR	34.2	33.4	19.4	31.2	32.2		6	35	32	21	28
AF	35.2	33.3	21.2	32.0	31.3	4.0		34	31	21	28
AC	34.1	34.1	34.2	33.3	32.2	33.4	32.5		8	34	34
CN	31.3	31.3	33.2	30.5	29.5	30.6	31.2	5.1		31	29
KT	32.2	30.3	28.2	32.3	32.4	19.3	21.2	32.3	31.2		22
LU	32.5	32.4	30.4	32.1	32.2	26.3	26.5	33.2	29.4	22.2	

TABLE II. Distance matrices for all genes. Top to bottom, breakpoints, minimal inversion, minimal i + 2.1 x t. Average breakpoint distance to non-group members (e.g. excluding MOL co-members CN and KT for AC, excluding ECH co-member SU for SS, no exclusions for HU) on upper diagonal. Note that AS and OV each have only 36 genes and CN has 35, while the other genomes have 37.

5.3.3 Empirical comparison of the distances.

There are 55 comparisons among the genomes in the data. Table II contains the results of calculating the number of breakpoints, the minimal inversions distance, and minimal inversions or transpositions distance, the latter with a relative cost of 2.1 imposed on transpositions (see Blanchette *at al.* ([10]) for a justification of this parameter value). It can be seen in the number of breakpoints that many of the gene orders seem to be random or near random permutations of each

	AS-13.3										
AS		OV-13.4									
OV	6		HU-9.1								
HU	13	14		SS-12.0							
SS	14	14	6		SU-11.8						
SU	13	13	6	2		DR-10.1					
DR	13	13	5	10	10		AF-10.1				
AF	13	13	5	10	10	0		AC-14.3			
AC	14	14	14	15	15	14	14		CN-13.6		
CN	14	14	13	14	14	13	13	3		KT-8.9	
KT	13	13	6	11	11	5	5	14	13		LU-11.4
LU	13	13	9	14	14	8	8	14	14	7	

AS		5	11	12	11	11	11	12	13	12	11
OV	4.2		13	12	11	11	11	12	12	12	11
HU	11.4	13.5		5	6	3	3	12	12	5	7
SS	12.3	12.2	4.1		1	7	7	14	12	9	11
SU	11.1	11.5	4.2	1.0		7	7	14	13	9	11
DR	11.3	11.3	3.0	7.1	7.1		0	12	12	4	7
AF	11.3	11.3	3.0	7.1	7.1	0.0		12	12	4	7
AC	12.4	12.1	12.2	14.2	14.2	12.2	12.2		3	11	12
CN	13.3	12.1	12.1	12.3	13.2	12.3	12.3	2.1		11	10
KT	12.2	12.2	5.1	9.1	9.2	3.1	3.1	11.2	11.2		5
LU	11.3	10.5	7.2	11.2	11.3	7.2	7.2	12.2	10.2	5.1	

TABLE III. Distance matrices for all genes except tRNAs. Top to bottom, breakpoints, minimal inversion, minimal i + 2.1 x t. Average breakpoint distance to non-group members (e.g. excluding MOL co-members CN and KT for AC, excluding ECH co-member SU for SS, no exclusions for HU) on upper diagonal. Note that AS and OV each have only 14 genes, while the other genomes have 15.

other. (Random genomes with n genes would have $n - \frac{1}{2}$ breakpoints with each other, on the average.) Some of this is simply due to the high rates of rearrangement in some groups – nematodes, echinoderms, snails – as roughly indicated by their average breakpoint distance to non-group members. Another contribution to randomness is the relatively great mobility of tRNA genes within the genome. When the calculations are repeated after deleting the tRNAs, the results are in Table III. Figure 13 is a scattergram of the data in both Tables II and III of the

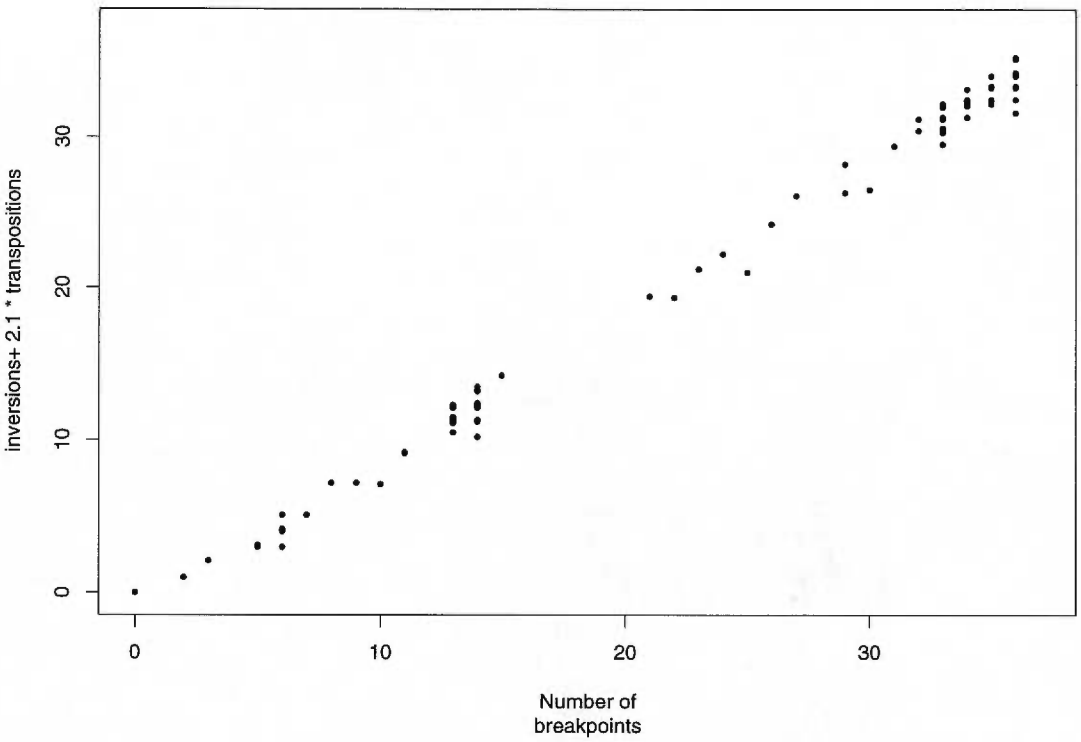

FIGURE 13. Relationship of breakpoint distance to minimal inversions + 2.1 x transpositions

breakpoint and l =minimum inversion + 2.1 transposition distance. An almost identical pattern is revealed with simple inversion distance versus breakpoints. Clearly the number of breakpoints b is highly predictive of l, though theoretically all that can be said is that $\frac{b}{2} \leq l$, ([71]).

5.4 Tree inference.

In this section we compare, in the light of theories of metazoan evolution, three criteria for optimum tree topology: neighbour joining, Fitch-Margoliash nor-

malized sum of squared errors, and minimum breakpoint. The first two operate on the genome data as reduced to the breakpoint distance matrix in Table II, somewhat modified, the third is based on the gene orders themselves. The first two produce ancestral nodes characterized only by their linear distance from neighbouring (colinear) nodes, the third actually reconstructs hypothetical ancestral genomes which are jointly optimal with the tree topology.

A modified data set is used for these analyses so that each genome contains the same number of genes. This means adding two tRNA genes to the CN genome – it is relatively clear where these should go to minimize changes in the pattern of distances, through analogy with the related AC genome – and adding an atp8 gene to the nematode genomes or deleting this gene from all the other genomes. Since different choices of insertion location for atp8 in the nematodes has different consequences for the entire distance matrix, we repeated our calculations, first with atp8 deleted from all genomes, and then with atp8 inserted in the nematode genomes directly adjacent to the atp6 gene, where it is most often located in the other genomes, including the conservative Katharina tunicata genome. There are two reasons for these modifications: first, it removes biases in all the analyses due to unequal numbers of genes – genomes with fewer genes would otherwise have systematically lower breakpoint distances and could thus tend to gravitate to a more central part of the tree than where they should be located. The second reason is purely technical and will be discussed in Section 5.4.2 below.

All the methods produce unrooted trees, though for interpretability we present them graphically as if they were rooted.

5.4.1 Neighbour-joining and Fitch-Margoliash

Neighbour-joining analysis, either with or without the atp8 data, produces the tree in Figure 14.

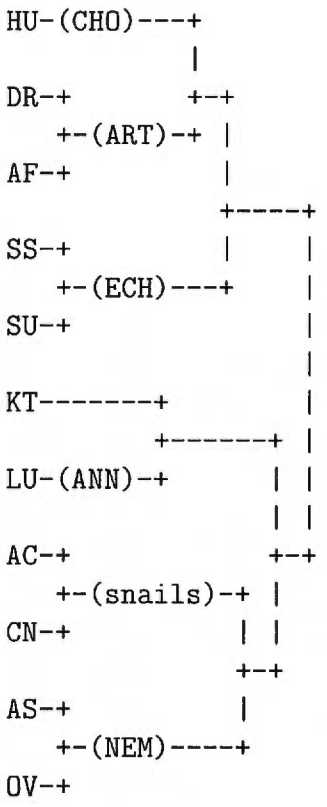

FIGURE 14. Neighbour-joining tree.

This tree disrupts the deuterostomes by grouping the arthopods with the human genome and, less problematic, disrupts the molluscs by grouping Katharina tunicata with the annelid Lumbricus terrestris.

The Fitch-Margoliash routine, which minimizes the sum of squared differences between distance matrix entries and total path length on the tree between two species, divided by the square of the matrix entry, produces the tree in Figure 15. The same tree is produced whether or not the atp8 data are included.

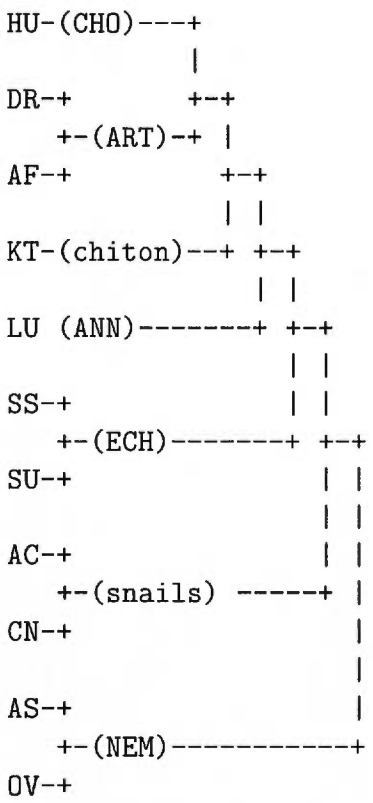

FIGURE 15. Best tree according to Fitch-Margoliash method.

This tree also disrupts the deuterostomes by grouping the arthopods with the human genome and disrupts the molluscs by branching the echinoderms between the snails and the annelid. Indeed, the Fitch-Margoliash tree must be considered considerably worse than the neighbour-joining one; its branching order reflects little more than the overall rate of evolution of the lineages as measured by the average breakpoint distance in Table II. The rapidly evolving lineages, nematodes, snails and echinoderms, are grouped together, and the more conservative lineages are grouped together, thus completely disrupting both the deuterostome (CHO+ECH) grouping and the MOL grouping.

Without atp8 data, the Fitch Margoliash normalized sum of squared differences is 0.401 for this tree, while for the trees in Figure 12, it is 0.759 for CAL,

0.764 for TOL and 0.765 for LAKE. With the atp8 data included, it is 0.359 for the tree in Figure 17, 0.694 for CAL, 0.702 for TOL and 0.708 for LAKE.

5.4.2 Minimal Breakpoint phylogeny

A minimum breakpoint tree is one in which a genome is reconstructed for each ancestral node, the number of breakpoints is calculated for each pair of nodes, ancestral or given, directly connected by a branch of the tree, and the sum is taken over all branches, where this sum is minimal over all possible trees. This problem may be decomposed into an inner and outer component.

The inner problem starts with a given topology, or branching structure, for the tree, and optimizes the ancestral genomes. In our method, the key to this solution is a technique for multiple genome rearrangement consensus: given three or more known genomes, find the median – the genome such that the sum of the number of breakpoints between it and each of the given genomes is minimal. Our solution to this ([57]) is based on a reduction to the Traveling Salesman Problem. The given phylogeny is then decomposed into a set of overlapping multiple genome rearrangement consensus problems, each one defined by one of the ancestral nodes as the median, to be found, with all the colinear nodes, ancestral or given, as the "known" genomes. With suitable initialization of the ancestral genomes ([8]), successive solution of all the overlapping consensus problems leads, after very few iterations, to convergence. In this study, we calculate the ancestral genomes on all trees using three different initializations and five passes of the successive optimization procedure. In case the three results were not identical, we retain the best one.

To solve the outer problem, we simply evaluate every possible tree on the set of given data genomes. Since we are not questioning on the basis of our data

the major groupings in Table I, we discard all trees that disrupt them, leaving a total of 105 unrooted binary branching trees. Auxiliary tests indicate that all the assumed groupings are robust in any case, with the exception of the conservative Katharina tunicata in MOL, which does not necessarily group with the highly diverged snails.

As with the neighbour-joining and Fitch-Margoliash methods, we first carried out our analysis on the genome data without the atp8 gene, and then repeated it with this gene appropriately inserted in the nematode genomes and restored to its original position in the other genomes.

5.4.3 Minimal breakpoint phylogeny for Metazoans

The scores – minimal number of breakpoints – for the 105 trees evaluated are distributed from 199 to 218 as in Figure 16 (data without atp8).

The two trees in Figure 17 are clearly optimal (both with and without the atp8 data), neither biologically plausible, because they give the impression either that the deuterostomes (CHO+ECH) are a late-branching sister taxon of the arthropods or, rooted differently, that nematodes constitute a late-branching sister taxon to the annelids (or of ANN+MOL). Note, however, that in both of these trees, the deuterostome taxon is correctly found, while it did not emerge from either neighbour-joining or Fitch-Margoliash.

What of the trees in Figure 12? All are clearly suboptimal; without atp8, there are no trees with scores between 199 and 202, one tree at 202, and score(CAL)=203, score(TOL)=204 and score(LAKE)=206. Including the atp8 data, there are no trees with scores between 201 and 205, but score(CAL)=205, score(TOL)=206 and score(LAKE)=209. Is clear that these data favour CAL, and TOL somewhat less, but the LAKE tree does not account for the configura-

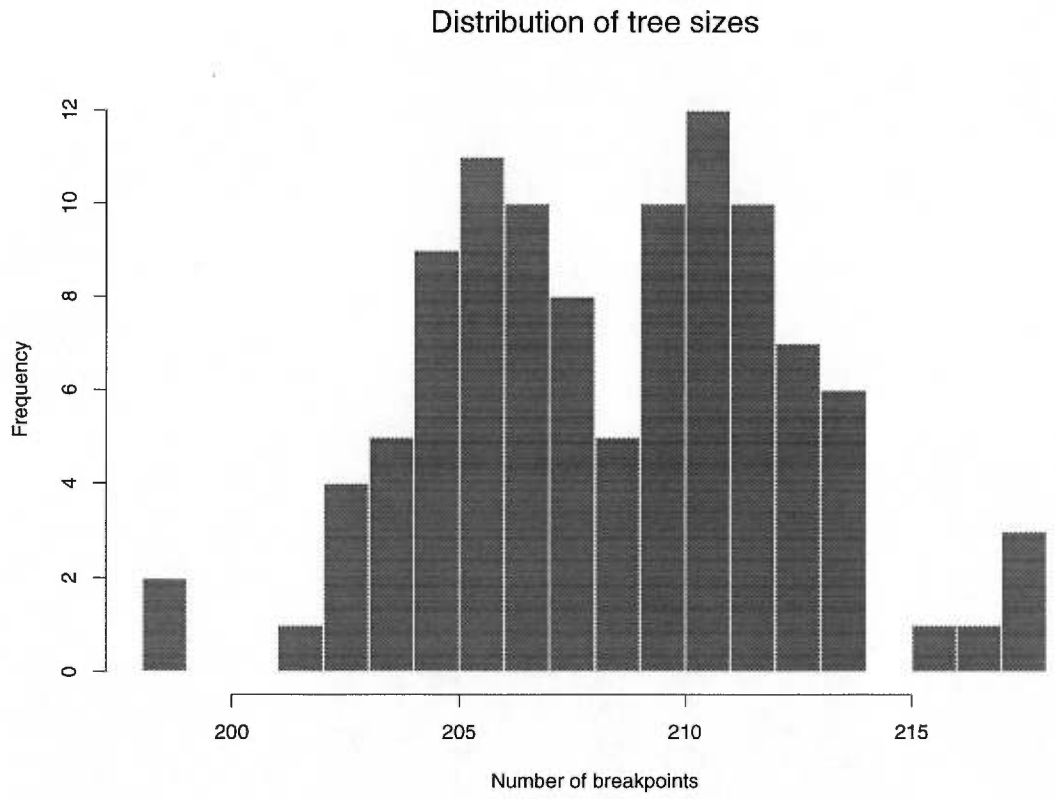

FIGURE 16. Distribution of number of breakpoints in the 105 possible trees.

tion of gene orders any better than a random tree.

To verify to what extent these results may be the result of tRNA gene mobility blurring out conserved order of the protein coding and rRNA genes, the analysis was repeated using only these latter 14 genes. Here there was little to distinguish all the trees we have discussed, with the two in Figure 17 and CAL scoring an optimal 59, while TOL and LAKE score 60. Including the atp8 data or not had no effect.

Returning to the full set of genes, what are we to make of the two trees in Figure 17? And what significance can a few points difference in the total number

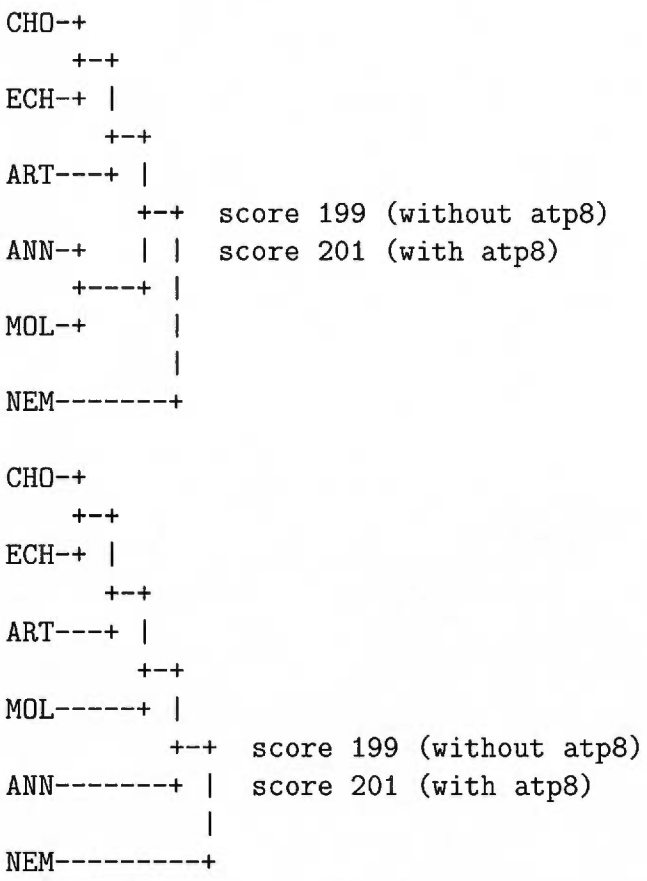

FIGURE 17. Minimal breakpoint trees. See Figure 18 for more details.

of breakpoints have? To answer the first question, the fact that not both of deuterostomes and nematodes can be deep-branching, no matter how the trees are rooted, seems to reflect the conservative versus rapidly-evolving distinction among the groups, as with the neighbour-joining and especially the Fitch-Margoliash criteria. But surprisingly, since minimal breakpoints is a parsimony criterion, this method seems less affected than the other two by the "long branches attract" artifact, since it successfully identifies the CHO+ECH grouping. The echinoderm line has diverged almost to randomness, but the link with the human gene order is too strong for them to be detached. This result must be seen to be a strong point of the breakpoint phylogeny method.

Our imposition of monophyly on the molluscs forces Katharina tunicata to group with the snails. However, when this constraint is relaxed, the trees in Figure 17 remain optimal. Five other trees also score 199 when Katharina tunicata is unconstrained, but four of these involve only local restructuring of the tree configuration Lumbricus terrestris-Katharina tunicata. Just one of the seven optimal trees shows further susceptibility to the "long branches attract" artifact, as the snails break away from the MOL grouping and become attached to the echinoderms.

The second question can be answered through reference to the breakpoint distances in Table II. All that distinguishes the comparison between Ascaris suum on one hand and Katharina tunicata and Lumbricus terrestris on the other from the complete randomness (36 breakpoints) of most other nematode comparisons are two instances of adjacent genes which occur in these three genomes. Boore *et al.*, ([13]) find this fact alone is important evidence of arthropod monophyly. Note that the minimal breakpoint trees in Figure 17 are both consistent with this fact, while neither the neighbour-joining nor the Fitch-Margoliash trees are.

5.4.4 Non-uniqueness

The study of genomic rearrangement inevitably encounters the problem of non-uniqueness. There are often many distinct solutions, all optimal, and many different ways of arriving at these results.

We illustrate with two aspects of non-uniqueness, that of the branch lengths of reconstructed trees and that of the reconstructed genomes.

Consider the first tree in Figure 17 with score 199. It is reproduced in Figure 18 with the branch lengths to scale, as consistent with one particular set of choices of optimally reconstructed ancestor genomes. On each branch we have

superimposed the range of lengths we obtained over ten different sets of choices of optimal genomes.

The greatest variability occurs with branches leading to terminal taxa – for example, if there are two sister terminal taxa, a number of different optimal genomes can be assigned to their immediate common ancestor, as one branch contracts and the other lengthens.

There are other sources of variability; the three succesive short but variable branches in the lineage of Katharina tunicata are consistent with the fact that most of the variation among optimal trees involves different positioning of this species with respect to Lumbricus terrestris. Both types of local transformation (branch length, neighbouring node interchange) in this region of the tree are compatible with optimality.

A consequence of this variability means that any single representation of the tree which attempts to portray branch lengths may be very misleading – of two branches having the same length, one may be constrained to always have this length and another may vary in different solutions betwen zero and twice the length, so any comparison of the two is arbitrary.

In the next section, we investigate the other aspect of non-uniqueness, the reconstructed genomes, and propose a solution to the representation of these genomes which escapes the problem of arbitrariness.

5.5 Reconstructing ancestral genomes

The genomes reconstructed at the ancestral nodes of a phylogeny are not generally unique; unless the genomes associated with the three adjacent nodes are all very similar, there will generally be many different optimal solutions for their

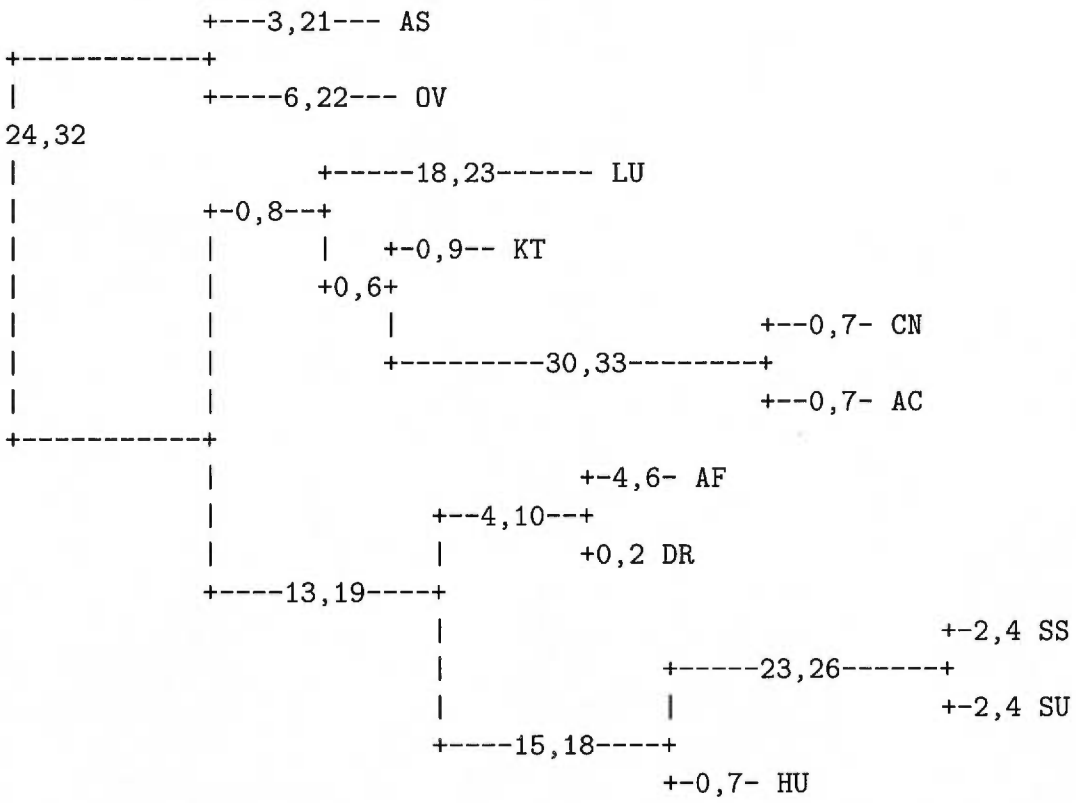

FIGURE 18. Tree with variable branch lengths.

median. And if there are several connected ancestral nodes, the set of optimal genomes for each such node will differ depending on the specific optima chosen for the others.

It will generally be the case, however, that some aspects of the optimal genome for a node are invariant over all solutions, independent even of the particular solutions chosen for phylogenetically adjacent nodes. We can be quite certain that these aspects are correctly reconstructed.

For a given tree, we can construct a large a number of optimal solutions for each node by running the algorithm on the same data but with different numbering of the genes, and then extract any strings of contiguous genes which reoccur in all these solutions. Since it represents the biologically plausible tree

most consistent with gene order data, we will illustrate with the CAL model.

When the ordering of all mitochondrial genes are used as input, the number of invariant segments obtained for each node is given in Figure 19. It is clear that for several of the ancestral nodes, there is a great deal of uncertainty about the gene order, represented by a large number of predominantly short invariant segments.

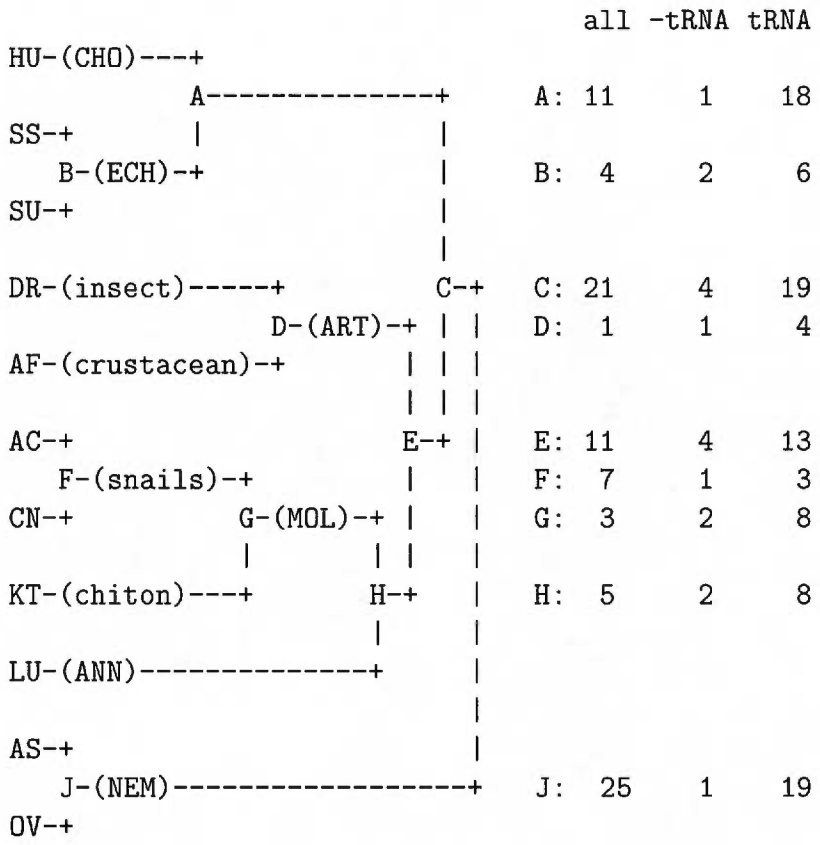

FIGURE 19. Number of unambiguously reconstructed segments, using all genes, without tRNA genes, and using tRNA genes only.

When the 22 tRNA genes are excluded from the same analysis, however, a different picture emerges, as indicated in the second column Figure 19. There are few, much longer segments, so that the ancestral genomes can be reconstructed

up to just an alternative orderings of these segments. The segments are shown in Figure 20. The small numbers of unambiguously reconstructed segments are

```
Ancestral nematode: unique reconstruction
nad6 cob cox3 nad4L rns nad1 atp8 atp6 nad2 nad4 cox1 cox2 rnl nad3 nad5

Protostome/deuterostome ancestor: 4 segments
(nad2 cox1 cox2 atp8 atp6 cox3 nad3)(nad4L nad4 nad5)(nad6 cob)(rns rnl nad1)

Ancestral protostome: 4 segments
(nad2 cox1 cox2 atp8 atp6 cox3 nad3)(nad4L nad4 nad5)(nad6 cob)(rns rnl nad1)

Ancestral deuterostome: unique reconstruction
cox1 cox2 atp8 atp6 cox3 nad3 nad4L nad4 nad5 -nad6 cob rns rnl nad1 nad2

Ancestral echinoderm: 2 segments
(cox1 nad4L cox2 atp8 atp6 cox3 nad3 nad4 nad5 -nad6 cob rns)(nad1 nad2 rnl)

Ancestral arthropod: unique reconstruction
nad6 cob -nad1 -rnl -rns nad2 cox1 cox2 atp8 atp6 cox3 nad3 -nad5 -nad4 -nad4L

Annelid/mollusc ancestor: 2 segments
(nad6 con)(-nad1 -rnl -rns cox3 nad3 nad2 cox1 cox2 atp8 atp6 -nad5 -nad4 -nad4L)

Ancestral mollusc: 2 segments
(nad6 cob)(-nad1 -rnl -rns cox3 nad3 nad2 cox1 cox2 atp8 atp6 -nad5 -nad4 -nad4L)

Ancestral snail: unique reconstruction
nad6 nad5 nad1 nad4L cob cox2 -atp8 -atp6 -rns -nad3 -cox3 nad4 nad2 cox1 rnl
```

FIGURE 20. Unambiguously reconstructed segments

somewhat surprising, but it is confirmed by inspection of Figure 20, which reveals patterns quite contrary to the impressions left by Tables II and III and first column in Figure 19. The non-tRNA mitochondrial gene order is seen to be relatively conservative.

Indeed, although it is perhaps not worth presenting the details for this limited data set, the reconstructed gene orders are relatively robust against small changes in the phylogeny, except of course in regions of the tree which are reconfigured and there is no one-to-one correspondence between the ancestors in the

original and modified trees.

This result, i.e. a reduced number of segments, is not just due to a reduced number of genes. When only the 22 tRNA genes are included, as in the third column of Figure 8, the number of segments is proportionately still quite elevated.

It is the increased mobility of the tRNA genes which is responsible for proliferation of alternative gene orders in the reconstructions, and the consequent decrease in the length of invariant segments and increase in their number.

5.6 Conclusion

5.6.1 Breakpoint distance

The relatively easy computability of multiple breakpoint distance makes it possible for the first time to do a systematic parsimony-type of phylogenetic study based on genome rearrangements. The very high correlation between this distance and the edit distances hitherto used to compare genomes, further validates the present approach.

5.6.2 Interpretation of phylogenetic results.

How much phylogenetic history is contained in metazoan mitochondrial gene orders? What methods are most apt to correctly infer this history? Finally, what theory of metazoan evolution best accounts for the genomic data?

It is clear from the matrices in Table II and III, that a large proportion of the pairwise genomic comparisons reveal no trace of common ancestry, as far as gene order is concerned. Only the more conservative genomes retain deep phylogenetic parallels. Nevertheless, through these latter genomes and through the connections

to one or other of them of the more rapidly evolving lines, all of the phylogenetic history can be inferred, with the exception of the earliest branching nematode lineage. More genomic data from early branching metazoan lineages, perhaps platyhelminthes, sponges and other nematodes and more divergent deuterostomes, would resolve this difficulty.

Aside from the high rate of evolution in comparison to the metazoan time scale, the other major impediment to phylogenetic reconstruction is the difference in this rate from lineage to lineage, together with the lack of " tree additivity" in measures like breakpoint distance which attain a maximal upper value after a certain amount of evolution. Superficial linearizations of these measures do not compensate for the complete loss of resolution when divergences reach the level of random genomes. All the phylogenetic reconstructions we tested were susceptible to the "long branches attract" artifact resulting from these problems, especially the Fitch-Margoliash technique. The minimal breakpoint phylogeny was most resistant to this effect, and neighbour-joining was intermediate. Only the minimal breakpoint reconstruction produced results at all compatible with any major theory of metazoan evolution.

Among these theories, the currently most acceptable view represented by the CAL tree is the one most supported by the genomic data. Not only does it have near optimal scores in terms of minimal breakpoints, with or without the atp8 data, with or without the tRNA genes, it also scores better than TOL, and especially LAKE, when tested with the Fitch-Margoliash criterion, and among the three theories, it most resembles the neighbour-joining tree. While the LAKE hypothesis clearly has no support whatsoever from the gene order data, the TOL tree is only slightly disfavoured by comparison with CAL. We emphasize again that the present study focuses entirely on mitochondrial gene orders; we do not suggest that these are superior to gene sequence evidence or other kinds of evidence; in any case a number of additional genomes from diverse lineages are

needed before confidence can be placed in this type of inference.

5.6.3 Unambiguously reconstructed segments

The algorithmic reconstruction of ancestral genomes has been plagued by the problem of non-uniqueness. It is very difficult to assimilate, display and interpret all the many different solutions which may be simultaneously optimal. The technique we introduce here, focusing on segments which appear in all these solutions, and not trying to account for the sometimes combinatorially prohibitive number of ways they may be assembled, essentially solves this problem, at least for the kind of data studied here.

In addition, this approach provides an unexpectedly dramatic characterization of the differential evolutionary mobility of tRNA genes versus rRNA and protein-coding genes. Whereas the 60 % of genes which code for tRNA account for about 70 % of the evolutionary divergence, as measured by including and then excluding them from the calculation of breakpoint distance, the decrease in the number of unambiguusly reconstructed segments when they are excluded is well over 80 %. As a result, the relatively conservative pattern of gene order in metazoan mitochondrial genomes is highlighted.

CHAPITRE 6

Conclusion

Les recherches présentées dans ce mémoire proposent une méthode entièrement nouvelle pour inférer des arbres phylogénétiques à partir des réarrangements du génome. Les méthodes utilisées jusqu'alors réduisaient l'information contenue dans chaque génome à une matrice donnant une distance "évolutive" entre chaque paire de génomes, puis utilisaient ces distances pour construire un arbre phylogénétique. Les arbres construits ainsi présentent plusieurs problèmes. En particulier, ils dépendent beaucoup de la métrique utilisée et ne donnent aucune information quant aux génomes ancestraux, à part leur distance hypothétique à leurs voisins.

La méthode proposée ici résoud une partie de ces problèmes, en considérant le problème de phylogénétique comme un problème d'arbres de Steiner dans l'espace des génomes. Les génomes ancestraux associés aux noeuds internes d'un arbre sont donc situés dans le même espace que les génomes connus. Les arbres ainsi obtenus s'en trouvent consolidés, sans compter qu'on dispose alors d'information sur les génomes ancestraux eux-mêmes.

Les problèmes d'arbres de Steiner sont habituellement NP-complets, et celui proposé ici ne fait pas exception. Cependant, la métrique proposée, le nombre de cassures entre deux génomes, permet des solutions exactes ou approximatives en un temps intéressant. En particulier, le cas où l'arbre de Steiner ne contient qu'un seul noeud interne (appelé problème de la médiane) peut être réduit à un PCV, dont la structure particulière rend possible une solution assez efficace. La

réduction au PCV, dans le cas de génomes orientés et non-orientés, est présentée au chapitre 2. On présente des bornes inférieures permettant d'appliquer un algorithme de "Branch-and-Bound". On étend les notions de cassures du génome aux cas où certains gènes sont absents d'un des génomes, et on présente l'extension des algorithmes permettant de traiter ce cas.

Le problème de la médiane, permet de développer des méthodes heuristiques permettant une approximation d'arbres de Steiner plus imposants. Appliquée de façon répétitive aux différents noeuds internes de l'arbre, la méthode de la médiane converge rapidement vers un optimum local. La qualité de cet optimum local dépendra des génomes auxquels sont initialisés les différents ancêtres. Trois méthodes d'initialisation sont été proposées au chapitre 3, dont deux, "AdjacencyParsimony" et "AverageTSP", donnent des résultats intéressants. Des simulations faites avec ces différentes techniques d'initialisation permettent d'évaluer la qualité de chacune, et d'évaluer aussi la qualité globale de la solution, avant et après le processus de médianes itératives.

Le Chapitre 4 dresse un parallèle entre l'alignement multiple de séquences d'ADN et le problème de réarrangements multiples du génome. On s'y intéresse aussi à la multiplicté des génomes ancestraux optimaux et à leur distance entre eux, en fonction du nombre de branches dans un arbre en "étoile", et en fonction de la position dans l'arbre, dans le cas d'arbres complets avec 12 espèces.

Le Chapitre 5 illustre finalement une application des algorithmes développés précédemment à un cas réel, celui du génome mitochondrial des métazoaires (animaux). On étudie dans un premier temps les arbres obtenus de manière classique, c'est-à-dire basés sur des matrices de distances, en termes de réarrangements, entre chaque paire d'espèces. Les arbres produits par ces techniques sont clairement en conflit avec les connaissances biologiques dans ce domaine, en particulier en ce qui concerne la position des vertébrés par rapport aux échinodermes. En comparaison, les arbres à cassures minimales obtenus par la méthode proposée

dans ce mémoire sont, sans être parfaits, beaucoup plus plausibles biologiquement. De plus, la méthode des arbres à cassures minimales permet de départager quelque peu les différentes hypothèses émises sur la phylogénie des animaux. En particulier, l'arbre phylogénétique le plus généralement admis se situe parmi les meilleurs, alors que d'autres arbres plus révolutionnaires se classent beaucoup moins bien.

Un des points les plus intéressants de la méthode d'analyse des cassures du génome est la capacité d'extraire de l'information sur les génomes ancestraux. La notion de "segments invariants" ("unambiguously reconstructed") est alors présentée pour traiter le problème des solutions multiples. Cette méthode permet d'inférer l'adjacence de certaines paires de gènes qui ont de très bonnes chances d'avoir été adjacentes chez un ancêtre donné. La technique est donc appliquée aux génomes mitochondriaux pour faire des hypothèses sur les ancêtres des différents groupes de métazoaires. Cette technique est aussi utilisée pour confirmer l'hypothèse selon laquelle certains gènes de petite taille (les ARN de transfert) auraient plus tendance à se déplacer que les gènes codant des protéines, de taille plus imposante.

6.1 Développements futurs

L'étude des arbres phylogénétiques basée sur les cassures du génome demeure embryonnaire. Il reste énormément de travail à faire, tant du côté algorithme que de celui des applications. D'un point de vue algorithmique, le fait que Pe'er et al. ([51]) aient démontré la NP-Complétude du cas le plus simple des arbres phylogénétiques, celui de la médiane de trois génomes, réduit presque à néant les espoirs de trouver un algorithme efficace (polynomial) pour ce cas et les cas plus complexes. Cependant, il y a NP-Complet et NP-Complet! Dans le cas de la médiane, on dispose d'une réduction linéaire vers un PCV, lequel est suffisament étudié pour disposer de solutions relativement efficaces. L'algo-

rithme de "Branch-and-Bound" présenté au chapitre 2 pourrait probablement être améliroré en travaillant sur les bornes utilisées. Par ailleurs, il est très probable que des méthodes tout-à-fait différentes, c'est-à-dire sans passer par un PCV, puissent donner des résultats intéressants. De plus, le problème de la médiane de génomes ne possédant pas le même ensemble de gènes n'est toujours pas résolu de façon satisfaisante. Une meilleure solution à ce problème serait intéressante d'un point de vue pratique, car, dans bien des cas (génomes chloroplasmiques, génomes mitochondriaux des protistes), les génomes étudiés ont des gènes assez variés. A l'heure actuelle, l'algorithme du chapitre 2 ne permet pas de s'attaquer à ces problèmes. En ce qui concerne des arbres ayant plusieurs génomes ancestraux à déterminer, la porte est aussi ouverte. D'abord, il n'existe toujours pas d'algorithmes exacts meilleurs que la simple énumération des solutions. L'intérêt pratique de cet algorithme serait probablement limité (à cause de sa complexité), mais il permettrait d'évaluer des méthodes heuristiques plus rapides. Deux approches semblent possibles: améliorer les méthodes d'initialisation présentées au chapitre 3, ou encore, utiliser une méthode complètement différente. L'intérêt de ces heuristiques est évident dans la pratique, puisqu'ils pourraient permettre de construire des arbres plus gros, plus rapidement et surtout avec de meilleurs garanties d'optimalité.

D'un autre côté, on ne s'est intéressé dans ce mémoire qu'aux réarrangements intra-chromosomaux (inversions et transpositions). Cela limite donc les applications aux génomes ne possédant qu'un seul chromosome, comme la plupart des génomes mitochondriaux et chloroplasmiques. Cependant, il ne devrait pas être trop difficile d'étendre les algorithmes développés ici aux cas multi-chromosomaux.

D'un point de vue appliqué, les défis ne manquent pas non plus. D'abord, les génomes mitochondriaux ont été séquencés non seulement pour les métazoaires utilisés au chapitre 5, mais aussi pour les champignons, les levures, les plantes et

les protistes, et on connaît aussi les génomes chloroplasmiques d'une quinzaine d'organismes. Chacun des cas pourrait être traité avec la méthode décrite dans ce mémoire. Des expériences préliminaires ont par contre démontré la difficulté de telles analyses dans le cas des protistes et celui des génomes chloroplasmiques, à cause du haut niveau de divergence à l'intérieur de ces groupes.

La capacité de traiter les génomes multi-chromosomaux ouvrirait aussi la porte à de nombreuses applications, et apporterait probablement une information plus complète que celle venant seulement d'un seul chromosome. Bien entendu, l'échelle du problème s'en trouverait aussi décuplée, puisque, par exemple, le génome humain contiendrait de 50 000 à 100 000 gènes, alors que ceux traités jusqu'à présent n'en contiennent que 200 au plus. Par ailleurs, à cause de leur taille beaucoup plus grande, les génomes nucléiques sont beaucoup plus longs et difficiles à séquencer, ce qui fait que, actuellement, les données permettant l'analyse des cassures de ces génomes sont très peu nombreuses.

Par conséquent, le travail stimulant ne manque pas pour qui le désire, et les résultats, tant théoriques qu'appliqués, qui devraient en découler seront sans doute intéressants pour les biologistes, les mathématiciens et les informaticiens.

RÉFÉRENCES

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

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

[3] S. Asakawa, H. Himeno, K. Miura et K. Watanabe. Nucleotide sequence and gene organization of the starfish *Asterina pectinifera* mitochondrial genome. inédit, 1993.

[4] V. Bafna et P.A. Pevzner. Genome rearrangements and sorting by reversals. *SIAM Journal of Computing*, 25: 272-289, 1996.

[5] V. Bafna et P.A. Pevzner. Sorting by reversals: Genome rearrangements in plant organelles et evolutionary history of X chromosome. *Molecular Biology and Evolution*, 12: 239-246, 1995.

[6] V. Bafna et P.A. Pevzner. Sorting by transpositions. *Proceedings of the 6th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 95),* 614-623, 1995.

[7] P. Berman et S. Hannenhalli. Fast sorting by reversal. *Proceedings of the seventh annual symposium on Combinatorial Pattern Matching (CPM)*, 1996.

[8] M. Blanchette, G. Bourque et D. Sankoff. Breakpoint phylogenies. *Genome Informatics 1997,* Tokyo: Universal Academy Press, 25-34, 1997.

[9] M. Blanchette, T. Kunisawa et D. Sankoff. Gene order breakpoint evidence in animal mitochondrial phylogeny. Soumis à *Journal of Molecular Evolution*, 1998.

[10] M. Blanchette, T. Kunisawa et D. Sankoff. Parametric genome rearrangement. *Gene,* 172, GC:11-17, 1996.

[11] J.L. Boore, T.M. Collins, D. Stanton, L.L. Daehler et W.M. Brown. Deducing the pattern of arthropod phylogeny from mitochondrial DNA rearrangements. *Nature*, 376: 163-165, 1995.

[12] J.L. Boore et W.M. Brown. Complete DNA sequence of the mitochondrial genome of the black chiton, *Katharina tunicata. Genetics*, 138: 423-443, 1994.

[13] J.L. Boore et W.M. Brown. Complete sequence of the mitochondrial DNA of the annelid worm *Lumbricus terrestris. Genetics*, 141: 305-319, 1995.

[14] E. Boudreau, C. Otis et M. Turmel. Conserved gene clusters in the highly rearranged chloroplast genomes of *chlamydomonas moewusii* and *chlamydomonas reinhardtii. Molecular Biology*, 24: 585-602, 1994.

[15] A. Caprara. Sorting by Reversals is Difficult. *Proceedings of the First Annual International Conference on Computational Molecular Biology (RECOMB 97),* 75-83, 1997.

[16] D.O. Clary et D.R. Wolstenholme. The mitochondrial DNA molecular of *Drosophila yakuba*: nucleotide sequence, gene organization, and genetic code. *Journal of Molecular Evolution* 22: 252-271, 1985.

[17] C.H. Darwin. *On the origin of species by means of natural selection in the struggle for life.* Londres: Murray, 1859.

[18] B. DasGupta, T. Jiang, S. Kannan, M. Li et Z. Sweedyk. On the complexity and approximation of synthenic distance. *Proceedings of the First Annual*

International Conference on Computational Molecular Biology (RECOMB 97), New-York: ACM-Press 99-108, 1997.

[19] T. Dobzhansky. *Genetics of the Evolutionary Process.* Columbia University Press, New York, 1970.

[20] T. Dobzhansky et A. H. Sturtevant. Inversions in the chromosomes of *drosophila pseudoobscura. Genetics*, 23: 28-64, 1938.

[21] D.J. Eernisse, J.S. Albert et F.E. Anderson. Annelida and Arthropoda are not sister taxa. A phylogenetic analysis of spiralian metazoan morphology. *Systematic Biology*, 41: 305-330, 1992.

[22] J. Felsenstein, *PHYLIP version 3.3*, University of Washington, 1990.

[23] V. Ferretti, J.H. Nadeau et D. Sankoff. Original synteny. *Proceedings of the Seventh Annual Symposium on Combinatorial Pattern Matching (CPM)*, Springer Verlag Lecture Notes in Computer Science, 1075: 159-167, 1996.

[24] N. Franklin. Conservation of genome form but not sequence in the transcription antitermination determinants of bacteriophages λ, $\phi 21$ et $P22$. *Journal of Molecular Evolution*, 181: 75-84, 1985.

[25] W. H. Gates et C. H. Papadimitriou. Bounds for sorting by prefix reversal. *Discrete Mathematics*, 27: 47-57, 1979.

[26] Q.-P. Gu, K. Iwata, S. Peng et Q.-M. Chen. A heuristic algorithm for genome rearrangements. *Genome Informatics 1997,* Tokyo: Universal Academy Press: 268-269, 1997.

[27] D. Gusfield. *Algorithms on strings, trees, and sequences: computer science and computational biology.* Cambridge University Press, 1997.

[28] S. Hannenhalli. Polynomial algorithm for computing translocation distance between genomes. *Proceedings of the 6th Symposium on Combinatorial*

Pattern Matching (CPM 95), Springer Verlag Lecture Notes in Computer Science: 162-176, 1995.

[29] S. Hannenhalli, C. Chappey, E. Koonin et P.A. Pevzner. Genome sequence comparison and scenarios for gene rearrangements : A test case. *Genomics*, 30(2): 199-211, 1995.

[30] S. Hannenhalli et P.A. Pevzner. Transforming cabbage into turnip. (polynomial algorithm for sorting signed permutations by reversals). *Proceedings of the 27th Annual ACM-SIAM Symposium on the Theory of Computing,* 178-189, 1995.

[31] S. Hannenhalli et P.A. Pevzner. Transforming men into mice (polynomial algorithm for genomic distance problem). *Proceedings of the IEEE 36th Annual Symposium on Foundations of Computer Science,* 581-592, 1995.

[32] E. Hatzoglou, G.C. Rodakis et R. Lecanidou. Complete sequence and gene organization of the mitochondrial genome of the land snail *Albinaria coerulea. Genetics,* 140: 1353-1366, 1995.

[33] J. Holland. *Adaptation in Natural and Artificial Systems.* University of Michigan Press, 1975.

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

[35] M. Jerrum. The complexity of finding minimum-length generator sequences. *Theoretical Computer Science*, 36: 265-289, 1985.

[36] C.J. Jolly et R. White. *Physical Anthropology and Archaeology.* McGraw-Hill, 1995.

[37] H. Kaplan, R. Shamir et R.E. Tarjan. Faster and Simpler Algorithm for Sorting Signed Permutations by Reversals. *Proceedings of the Eighth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 97),* 1997.

[38] J. Kececioglu et R. Ravi. Of mice and men. Evolutionary distances between genomes under translocation. *Proceedings of the 6th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 94),* 604-613, 1995.

[39] J. Kececioglu et D. Sankoff. Efficient bounds for oriented chromosome inversion distance. *Proceedings of the 5th Symposium on Combinatorial Pattern Matching (CPM)*, Springer Verlag Lecture Notes in Computer Science, 807: 307-325, 1994.

[40] J. Kececioglu et D. Sankoff. Exact and approximation algorithms for sorting by reversals, with application to genome rearrangement. *Algorithmica,* 13: 180-210, 1995.

[41] E.M. Keddie et T.R. Unnasch. Complete sequence of mitochondrial genome of *Onchocerca volvulus.* Inédit.

[42] K. von Linné. *Systema naturae*, Lipsiae : Impensis G. Kiesewetteri, 1748.

[43] D. Maddsion et W. Maddison. Tree of Life metazoa page, http://phylogeny.arizona.edu/tree/eukaryotes/animals/animals.html, 1995.

[44] I. Marchand. *Généralisations du modèle de Nadeau et Taylor sur les segments chromosomiques conservés.* Mémoire de maîtrise, Département de mathématiques et de statistiques, Université de Montréal, 1997.

[45] E. Minieka, *Optimization Algorithms for Networks and Graphs*, Industrial Engineering vol.1, New York: Marcel Dekker, 272-273, 1978.

[46] J.H. Nadeau et B.A. Taylor. Lengths of chromosomal segments conserved since divergence of man and mouse. *Proceedings of the National Academy of Sciences USA*, 81: 814-818, 1984.

[47] R. Okimoto, J.L. Macfarlane, D.O. Clary et D.R. Wolstenholme. The mitochondrial genomes of two nematodes, *Caenorhabditis elegans* and *Ascaris suum*. *Genetics*, 130: 471-498, 1992.

[48] J. D. Palmer et L. A. Hebron. Plant mitochondrial DNA evolves rapidly in structure, but slowly in sequence. *Journal of Molecular Evolution*, 27: 87-97, 1988.

[49] J. D. Palmer, B. Osorio et W. F. Thompson. Evolutionary significance of inversions in legume chloroplasts DNAs. *Current Genetics*, 14: 65-74, 1988.

[50] M.-N. Parent. *Estimation du nombre de segments vides dans le modèle de Nadeau et Taylor sur les segments chromosomiques conservés.* Mémoire de maîtrise, Département de mathématiques et de statistiques, Université de Montréal, 1997.

[51] I. Pe'er et R. Shamir (communication personnelle, janvier 1998).

[52] M.L. Perez, J.R. Valverde, B. Batuecas, F. Amat, R. Marco et R. Garesse. Speciation in the *Artemia* genus: mitochondrial DNA analysis of bisexual and parthenogenetic brine shrimps. *Journal of Molecular Evolution* 38: 156-168, 1994.

[53] P. Pevzner et M.S. Waterman. Open combinatorial problems in computational molecular biology. In *Proceedings of the Third Israel Symposium on the Theory of Computing and Systems*: 158-173, 1995.

[54] G.W. Rouse et K. Fauchald. The Articulation of Annelids. *Zoologica Scripta* 24: 269-301, 1995.

[55] D. Sankoff. Edit distance for genome comparison based on non-local operations. *Proceedings of the 3rd Symposium on Combinatorial Pattern Matching (CPM 92)*, Springer Verlag Lecture Notes in Computer Science, 644: 121-135, 1992.

[56] D. Sankoff. The early introduction of dynamic programming into computational biology. *Advances in the Mathematical Sciences - CRM's 25 years*, CRM Proceedings and Lecture Notes, vol. 11: 403-413, 1997.

[57] D. Sankoff et M. Blanchette. The median problem for breakpoints in comparative genomics. *Proceedings of Computing and Combinatorics (COCOON '97)*, Lecture Notes in Computer Science, 1276: 251-263, 1997.

[58] D. Sankoff et M. Blanchette. Multiple genome rearrangement. *Proceedings of the First Annual International Conference on Computational Molecular Biology (RECOMB 98)*, New York: ACM Press, 243-247, 1998.

[59] D. Sankoff, R.J. Cedergren et G. Lapalme. Frequency of insertion-deletion, transversion, and transition in the evolution of 5S ribosomal RNA. *Journal of Molecular Evolution*, 7: 133-149, 1976.

[60] D. Sankoff et V. Ferretti. Karotype distributions in a stochastic model of reciprocal translocation. *Genome Research*, 6: 1-9, 1996.

[61] D. Sankoff, V. Ferretti et J.H. Nadeau. Conserved segment identification. *Proceedings of the First Annual International Conference on Computational Molecular Biology (RECOMB 97)*. New York: ACM Press, 252-256, 1997.

[62] D. Sankoff, G. Leduc, N. Antoine, B. Paquin, B.F. Lang et R. Cedergren. Gene order comparisons for phylogenetic inference: Evolution of the mitochondrial genome. *Proceedings of the National Academy of Sciences USA*, 89: 6575-6579, 1992.

[63] D. Sankoff, C. Morel et R.J. Cedergren. Evolution of 5S RNA and the non-randomness of base replacement. *Nature New Biology,* 245: 232-234, 1973.

[64] D. Sankoff, G. Sundaram et J. Kececioglu. Steiner points in the space of genome rearrangements. *International Journal of the Foundations of Computer Science*, 7: 1-9, 1996.

[65] J. Setubal et J. Meidanis. *Introduction to Computational Molecular Biology.* PWS Publishing, 1997.

[66] A. H. Sturtevant et T. Dobzhansky. Inversions in the third chromosome of wild races of *drosophila pseudoobscura*, and their use in the study of the history of the species. *Proceedings of the National Academy of Sciences USA*, 22: 448-450, 1936.

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

[68] J.W. Valentine, Metazoa Systematics Page, http://www.ucmp.berkeley.edu/phyla/metazoasy.html, University of California Museum of Paleontology.

[69] P. Winter. Steiner problem in networks: A survey. *Networks*, 17: 129-167, 1987.

[70] M.S. Waterman, *Introduction to computational biology*, Chapman & Hall, 1995.

[71] G. A. Watterson, W. J. Ewens, T. E. Hall et A. Morgan. The chromosome inversion problem. *Journal of Theoretical Biology*, 99: 1-7, 1982.