Early Eukaryote Evolution Based on Mitochondrial Gene Order Breakpoints

David Sankoff * David Bryant * Mélanie Deneault * B. Franz Lang † Gertraud Burger †

Abstract

We present a general heuristic for the median problem for induced breakpoints on genomes with unequal gene content and incorporate this into a routine for estimating optimal gene orders for the ancestral genomes in a fixed phylogeny. The routine is applied to a phylogenetic study of an up-to-date set of completely sequenced protist mitochondrial genomes, confirming some of the recent sequence-based groupings which have been proposed and, conversely, confirming the usefulness of the breakpoint method as a phylogenetic tool even for small genomes.

1 Introduction.

The origin and early diversification of the eukaryotes is one of the fundamental problems of evolutionary theory. The widely accepted endosymbiotic origin of the mitochondrion and its consequent evolution, in key respects independent of the evolution of the nuclear genome, make it a natural focus of phylogenetic studies. Indeed phylogenies based on a number of mitochondrial genes have led to a far clearer understanding of the phylogeny of unicellular eukaryotes—the protists and the fungi—than was ever possible based on morphological classifications alone [3, 5, 6, 10, 11, 12, 13, 14, 23]. Nevertheless, this approach is limited by the relatively small number of genes present in all or most mitochondria, and the finite amount of phylogenetic information that can be extracted from the sequence comparison of any of these genes. For some time we have advocated the quantification of gene order changes within the mitochondrion as an independent measure of genomic divergence that can be used to supplement sequence comparison data [21].

Early work in the construction of phylogenies from gene order data used a distance based approach. In [18] the distance between two gene orders is estimated using a heuristic algorithm for minimizing a weighted measure of the number of reversals and transpositions of chromosome fragments, as well as the insertion and deletion of individual genes, necessary to transform one order into the other. An exact polynomial time algorithm for calculating the minimum number of reversals needed to transform one gene order into another was developed by Hannenhalli and Pevzner[8]. Distance matrix methods could then be used to reconstruct a phylogeny.

As attention focused on exact and efficient algorithms for rearrangement distances phylogenetic questions were somewhat neglected. This was partly due to the lack of realism in measuring genomic divergence in terms of reversals only, with all reversals of equal weight. It was also due to the inability of distance matrix methods to recover the nature of ancestral genomes, in contrast to methods such as parsimony or likelihood which avoid reducing the data to distances prior to phylogenetic reconstruction. Generalization of rearrangement distances to more than two genomes, necessary for non-distance approaches, has been shown to be NP-hard [4]. Moreover, even heuristic approaches to such a generalization work well only for very small problems (cf [7, 22]).

More recently, we have introduced the notion of breakpoint phylogeny [20]. For two genomes containing the same genes, breakpoints are simply pairs of consecutive genes g_1 and g_2 which occur in order g_1g_2 in one genome but not in the other. Adjacency is also considered disrupted if the two genes have different orientation to each other in the two genomes; e.g. g_2 succeeds g_1 but with opposite reading direction in one genome while they are adjacent on the same DNA strand in the other genome. The number of breakpoints correlates with the evolutionary divergence of the two genomes. In contrast to rearrangement distances, this is easily extended to three or more gene orders: the *median problem*, though it is computationally costly for large genomes [2, 16]. Solutions to the median problem can be combined and iterated to optimize the ancestral genomes in a given tree topology. For moderate numbers of genomes, all possible topologies can be evaluated to solve the phylogenetic problem. We have demonstrated the applicability of the method by showing the plausibility of breakpoint phylogenies constructed on the basis of the relatively small (37 genes) mitochondria of metazoans, from humans to nematodes [1].

Unfortunately the notion of breakpoint does not carry over in a straightforward way when the genomes being compared do not have the same set of genes. The shared genes in two genomes may be ordered in exactly the same way but because of intervening genes that belong to only one or the

*Centre de recherches mathématiques, Université de Montréal, CP 6128 succursale Centre-ville, Montréal, Québec H3C 3J7. E-mail: {sankoff,bryant,deneaulm}@crm.umontreal ca.

†Département de biochimie, Université de Montréal, CP 6128 succursale Centre-ville, Montréal, Québec H3C 3J7. E-mail: {langf,burgerg}@bch.umontreal.ca.

RECOMB 2000 Tokyo Japan USA

other, the number of breakpoints may be large. It is more appropriate in this context to consider *induced breakpoints*, the breakpoints remaining when the genes belonging to only one or the other genome are discarded. When comparing a set of three or more genomes which vary greatly in the number of genes they contain, it also becomes necessary to normalize the number of induced breakpoints between two genomes by the number of genes they share.

The main contributions offered in the present paper include a general heuristic for the median problem for (normalized) induced breakpoints on unequal genomes, the incorporation of this into a search for optimal phylogenies, and the application of this methodology to an up-to-date set of completely sequenced protist mitochondrial genomes, confirming some of the recent sequence-based groupings which have been proposed and, conversely, confirming the usefulness of the breakpoint method as a phylogenetic tool even for small genomes.

2 The evolution of the eukaryotes

Prior to plants, animals and fungi, a large number of mostly unicellular eukaryotes (protists) diverged from the common eukaryotic ancestor. Their classification has traditionally been difficult since they lack the differentiated tissues organized into organs that help categorize plants and animals. However, contrary to prokaryotes that virtually lack morphological character, protists can be classified based on ultrastructural features, such as features of the flagellar apparatus or the shape of mitochondrial cristae. Although not without exception, animals, fungi, plants, green and red algae all manifest flattened cristae, while *Euglena*, the trypanosomatids (like*Leishmania*) and heteroloboseans possess discoidal cristae. Still another large grouping, including the ciliates (such as *Parmecium*), the slime molds and the stramenopiles, have tubular cristae [6].

Among the organisms characterized by flattened cristae, sequence analysis of several mitochondrial genes has indicated a common ancestry for animals and fungi [14], a close relationship between red and green algae [3], and the origin of the land plants within the latter. Several of the subgroups within the discoidal cristae grouping can also be linked through gene sequence comparison. The same can be said within the tubular cristae group, particularly those within the stramenopiles, where links are evident between the chrysophytes, the synurophytes, the oomycetes and the bicosoecids. Within each of these large groupings, however, the earliest relationships remain unclear.

3 The data

GOBASE (http://megasun.bch.umontreal.ca/gobase/ gobase.html) is a relational organelle genome database, which integrates sequence data, information on evolution, taxonomy, biochemistry, RNA secondary structure, physical maps and more [9].

The major body of data contained in GOBASE consists of mitochondrial sequence data drawn from the Entrez database system and taxonomy data extracted from the NCBI Taxon database. These data are obtained on a regular basis by a custom-made tool, POP2, which reads the Entrez data in ASN.1 format, extracts the data relevant to the molecular features defined in GOBASE, and stores this information in GOBASE tables.

The production of gene orders from sequence data cannot be totally automated. Genes may overlap, may be fragmented and scattered across the genome (especially the rRNA genes), may be unrecognized or unannotated in the Entrez file, or annotated in an idiosyncratic way. As a first step, maps are produced from an entry, and a gene order with signed genes is derived from that. Maps are posted on the GOBASE website. Many of the mitochondrial sequences are produced in the laboratories affiliated with the Organelle Genome Megasequencing Program and the Fungal Mitochondrial Genome Project (e.g. [3, 13, 15, 14, 23]), and prepublication maps of these are also posted.

For the purposes of the analysis in the present paper, duplicate genes were excluded from some of the genomes because of the inability of our method to handle these duplicates. Most of these were tRNA genes. As far as possible, we tried to identify homologous sets of mitochondrial tRNA genes across as many protists as possible, taking into account the corresponding amino acid, the anticodon, the translation table appropriate to the organism and, in the few cases where it was possible, positional correspondences in closely related genomes. In the remaining instances where the duplicates remained indistinguishable, we deleted both from the gene order.

It will be seen in Section 4 that this introduces little bias into the comparison, though the loss of data does decrease the precision of the estimates. In other cases, where entire fragments of the genome were duplicated, we simply deleted the fragment which seemed the secondary one, based on comparisons with closely related genomes or by conforming to the strandedness of the majority of the genome.

For some genes in some genomes, part of the gene is located in one position and the rest elsewhere, in such a way that other genes intervene between two fragments. We retained only the position of the longer fragment. Where there was fragmentation of a gene into several pieces, the genome was excluded from the analysis.

Finally, we excluded all ORFs (hypothetical proteins) from the data unless there was good evidence that the same ORF appeared in two or more genomes.

Two other criteria served to exclude other genomes from the analysis; too few genes, such as in the trypanosomatids where, moreover, no tRNAs appear in the gene order, and an *a posteriori* filter where a genome turned out to bear no more than random resemblance to any of the other genomes in the data set.

The data we analyzed are summarized in Table 1.

4 Extending the notion of breakpoint to unequal genomes

For two genomes whose gene sets are not identical we introduce the notion of *induced breakpoints* (cf [19]) to capture the notion that the *shared* genes have parallel orders (or not). 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. This quantity is quite different from the original concept of breakpoint distance. 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. But if, as in Figure 1, there is a breakpoint between a_i and a_j in the reduced genome, where $j \neq i+1$, i.e. where $a_{i+1}, \cdots, a_{j-1}$ have been deleted in producing the reduced genome, then no ordinary breakpoint exists in the full genome, since a_i is not adjacent to a_j there. The measurement of induced breakpoints is thus a more subtle way of capturing the degree of parallelism of two gene orders than ordinary breakpoints.

Organism (Accession number)	Classification		Genes	tRNAs
TUBULAR CRISTAE				
1. *Acanthamoeba castellanii* (U12386)	lobose amoeba		56	16
2. *Chrysodidymus synuroideus*	stramenopile	(synurophyte)	53	19
3. *Ochromonas danica*	stramenopile	(chrysophyte)	57	22
4. *Phytophthora infestans*	stramenopile	(oomycete)	60	23
5. *Cafeteria roenbergensis*	stramenopile	(bicosoecid)	54	22
6. *Paramecium aurelia* (X15917)	alveolate	(ciliate)	39	3
7. *Dictyostelium discoideum* (D16466)	slime mold		48	17
8. *Jakoba libera*	jakobid		88	24
* *Plasmodium falciparum* (M76611)	alveolate	(apicomplexan)		
* *Plasmodium yoelii* (M29000)	alveolate	(apicomplexan)		
9. *Reclinomonas americana* (AF007261)	jakobid		97	26
10. *Tetrahymena pyriformis* (AF160864)	alveolate	(ciliate)	43	7
† *Theileria parva* (Z23263)	alveolate	(apicomplexan)		
11. *Thraustochytrium aureum*	stramenopile	(labyrinthulid)	53	19
DISCOIDAL CRISTAE				
‡ *Malawimonas jakobiformis*	malawimonad		68	25
12. *Naegleria gruberi*	heterolobosean		61	17
† *Leishmania tarentolae* (M10126)	trypanosomatid			
† *Trypanosoma brucei*	trypanosomatid			
FLATTENED CRISTAE				
13. *Marchantia polymorpha* (M68929)	land plant		69	24
§ *Arabidopsis thaliana* (Y08502)	land plant			
14. *Pedinomonas minor* (AF116775)	green alga		21	8
15. *Nephroselmis olivacea*	green alga		65	26
16. *Prototheca wickerhamii* (U02970)	green alga		63	26
* *Chlamydomonas eugametos* (AF008237)	green alga			
* *Chlamydomonas reinhardtii* (U03843)	green alga			
* *Chlorogonium elongatum* (Y13644)	green alga			
17. *Chondrus crispus* (Z47547)	red alga		50	23
18. *Cyanidioschyzon merolae* (D89861)	red alga		59	22
19. *Porphyra purpurea* (AF114794)	red alga		55	24
20. *Rhodomonas salina*	cryptophyte		67	27
‡ *Monosiga brevicollis*	choanoflagellate		50	22
‡ fungal,animal	fungal,animal			

Table 1: Sequenced mitochondrial genomes. Data from gene maps in GOBASE [9]. Gene numbers affected by the exclusion of some duplicate genes (see text). Organisms numbered 1-20 used in analysis. Other organisms excluded for the following reasons: * fragmented rRNA genes, † too few genes, ‡ no gene order resemblances with (other) protist mitochondrial genomes, § trans-spliced genes

Breakpoint measures and particularly induced breakpoint measures are robust against missing data, such as genes absent in some organisms or excluded for the methodological reasons invoked in Section 3. If a gene figures in an adjacency crucial for phylogenetic grouping, it is unlikely to have been one of two duplicate genes, since in most cases (all cases in our data) only one tRNA gene in a duplicate pair shares adjacencies with related genomes, and this was a criterion for deleting the other. Other duplicated genes excluded were part of longer duplicated fragments, and relevant adjacencies are conserved in the undeleted fragment. In any case, deleting genes which are not shared with neighboring genomes in a phylogeny has no effect on induced breakpoint measures, by definition.

For the purposes of phylogeny, the numbers of induced breakpoints can be misleading when the genomes vary significantly in gene content. Large genomes, even if they are fairly closely related, will tend to be placed far from each other in the tree. The savings from this stem from the deletion of non-shared genes from genomes in calculating the number of induced breakpoints. If small and large genomes are intermingled in the tree, many more deletions will be required when evaluating the branches of the tree than when the large genomes are close to each other. Fewer genes mean fewer total adjacencies and few potential breakpoints.

To eliminate this artifact, we normalize the number of induced breakpoints by the total number of genes shared by the two genomes.

5 A practical heuristic for the breakpoint median problem

Given three genomes A, B, C and a set $\mathcal{G}$ of genes (e.g. $\mathcal{G}$ = the union of the genes in A, B and C, or $\mathcal{G}$ = the intersection of the genes in A, B and C) we wish to find a genome S containing the genes in $\mathcal{G}$ that minimizes the sum of the breakpoints between A and S, B and S and C and S. Such a genome is called a **median genome** for A, B, C.

In the case where A, B, C and S have identical composition, there is a simple reduction from the breakpoint median problem to the traveling salesman problem (TSP) [19]. Let $d(X, Y)$ denote the breakpoint distance between X and Y, and let the weight $w(x, y)$ of an unordered pair of genes x, y

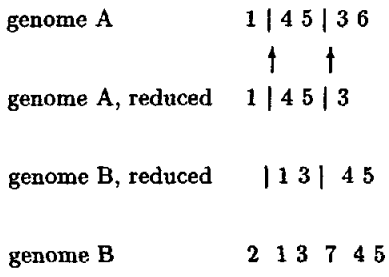

Figure 1: Induced breakpoints for (circular) genomes with different gene contents. Position of induced breakpoints (vertical strokes) found in reduced genomes with identical gene sets. This determines breakpoints between 1 and 4 and between 5 and 3 in genome A. There are no breakpoints in B.

be defined by

$$w(x,y) = |\{X \in \{A,B,C\} : \{x,y\} \text{ adjacent in } X\}|.$$

Setting

$$\Psi(S) = d(A,S) + d(B,S) + d(C,S)$$

and determining S that minimizes $\Psi(S)$ is equivalent to determining a tour of minimum length with respect to distance matrix δ with $\delta_{x,y} = 3 - w(x,y)$.

At first view one might ask the utility of the result. The TSP is an NP-hard problem. Indeed the median breakpoint problem is as well [2, 16]. NP-hardness, however, does not mean that a problem cannot be efficiently solved, it merely calls for a change in approach. A case in point is the TSP. While it is NP-hard, there exists an impressive selection of heuristics, lower bounds, branch and cut methods and iterative improvement techniques that can be used to solve the TSP for even moderately large problem instances (see [17] for a comprehensive survey of TSP methods). By the above reduction, the same techniques can be used on the breakpoint median problem, provided that the input genomes have the *same* gene set.

The reduction to TSP breaks down when we consider the *induced* breakpoint median problem for genomes with unequal gene sets. The equivalent TSP problem would be one with multiple distance matrices defined on overlapping sets of cities—the length of a tour being the summed length of the respective induced tours. Many of the standard lower bounds, heuristic and exact methods for TSP cannot be translated to this general framework (e.g. spanning tree and matching based methods, Lin-Kernighan local search, Held-Karp lower bound). Many of those that can be translated perform badly when compared to the alternatives (e.g. tour amalgamation methods or divide and conquer algorithms). While we were able to construct a linear programming formulation of the induced breakpoint problem, the large number of variables required prevented the practical application of branch and cut style methods.

Surprisingly, one of the simplest heuristics, and not one that is particularly effective in the equal gene sets context, greatly outperformed all other alternatives. In addition, the superiority of this heuristic increases as the input genomes contain increasingly different sets of genes. We studied the efficiency of the method with respect a number of lower bounds. Our experiments involved simulated data as well as some of the data in Table 1. In all cases studied we were able to find a near-optimal solution within a relatively small number of iterations.

We outline this heuristic in Section 5.1 and discuss an $O(n^2)$ time implmentation in Section 5.2 (where n is the number of genes).

5.1 A surprisingly effective heuristic for gene insertion

Let A, B, and C be signed genomes and put $\mathcal{G}$ = the union of the gene sets of A, B and C. The heuristic begins be arbitrarily restricting the problem to a small set of genes then progressively inserting new genes.

furthest insertion
  Randomly choose $g \in \mathcal{G}$.
  $med \leftarrow \langle g \rangle$
  $X \leftarrow \{g\}$
  while $X \neq \mathcal{G}$ **do**
   for each $x \in \mathcal{G} - X$ **do**
    Insert x into med so as to minimize $\Psi(med)$.
    Let $m[x]$ be the minimum value of $\Psi(med)$.
    Remove x from med.
   end for
   Choose x^* that maximizes $m[x^*]$ and then insert x^* into med so as to minimize $\Psi(med)$.
   Add x^* to X.
  end while
  return med
end.

Note that the heuristic can be applied when Ψ is the sum of either normalised, or un-normalised, breakpoint distances. It can be seen from the algorithm that the furthest insertions heuristic is an exact analog of the furthest insertion heuristic for the traveling salesman problem [17]. There is a subtle difference that makes the breakpoint median heuristic perform well whereas the original TSP heuristic performs quite badly. At each iteration we solve an *induced* subproblem, with induced genomes $A|_X$, $B|_X$ and $C|_X$. The induced subproblem captures far more important structural information of the whole problem than the simple restriction of a distance matrix to a subset of cities.

The furthest insertion heuristic performs better than the heuristic that randomly chooses which gene to insert next, but takes longer to execute. The random insertion heuristic in turn performs better than the greedy method that chooses the gene that gives the least increase. The difference between the three heuristics becomes marked when the number of genes common to all genomes decreases.

Examining the increase in Ψ at each iteration provides the explanation: the greedy heuristic first places genes that appear in only a few genomes. It then proceeds to slot in the genes belonging to the majority of the genomes. The value of Ψ increases slowly at first, then accelerates rapidly as the algorithm tries to fit new genes into a badly chosen subgenome. In contrast the furthest insertion heuristic places the genes belonging to the majority of the genomes first. It thus obtains a better subgenome to insert the remaining genes into. The heuristic with random insertion occupies the middle ground between the two other methods.

All three heuristics have some degree of randomness in the way they break ties. Improved genomes can be obtained by repeating the heuristics many times. The furthest insertion heuristic almost always outperformed the alternatives, finding genomes on the first repetition that are markedly better than the genomes found by the other heuristics even after a large number of repetitions.

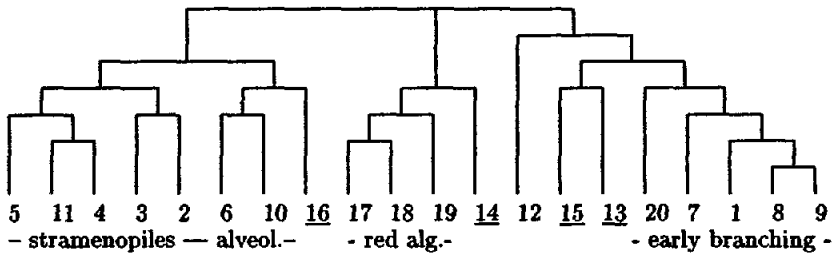

Figure 2: Distance-matrix analysis of protist evolution. Number keys in Table 1. Lengths of branches not to scale. Root is near *Reclinomonas* (9). Stramenopiles and alveolates cluster together. Jakobids and other early branching protists group together. Green algae and plants (underlined) scattered throughout phylogeny.

5.2 An $O(n^2)$ time algorithm for the furthest insertion heuristic

The furthest insertion heuristic as described runs in polynomial time but, if implemented directly, will have a running time of $O(n^4)$. This running time can be improved to $O(n^2)$. First we note that the optimal location to insert a gene x into the partially completed median will always be before or after a gene that is a successor or predecessor of x in one of the input genomes. This reduces the number of possible insertion points from $O(n)$ to $O(1)$. Second, we can maintain the successor tables (and therefore predecessor tables) of the input genomes and median genome so that:

- For each set X of already processed genes and each x not in X we can answer successor queries of the input genomes induced by $X \cup \{x\}$ in constant time.
- Successor queries for the partially completed median genome with gene set restricted to $\mathcal{G}(A)$, $\mathcal{G}(B)$, or $\mathcal{G}(C)$ can be answered in constant time.
- Each time we insert a new gene x^* into the set of processed genes the data structures can be updated in $O(n)$ time.

In this way, each main iteration of the furthest insertion algorithm takes $O(n)$ time giving $O(n^2)$ time for the entire algorithm.

6 Induced breakpoint phylogeny

The heuristic for the median problem developed in Section 5 is at the heart of our approach to phylogeny. Suppose we have N input genomes $G^1, \cdots, G^N$. For an unrooted binary branching tree T with leaves $\{V_1, \cdots, V_N\}$, each one associated with one input genome, and interior nodes $\{V_{N+1}, \cdots, V_{2N-2}\}$, the task is to find the genomes $G^{N+1}, \cdots, G^{2N-2}$ associated with the internal (ancestral) nodes, such that

$$\sum_{\text{all branches } V_i V_j \text{ of } T} \text{breakpoint measure between } G^i \text{ and } G^j$$

is minimized. Each of the internal nodes defines a median problem with its three adjacent nodes. Starting with an initial assignment of genomes to all the internal nodes (to be specified in Section 7.1), recalculation of the internal nodes is carried out one by one, each time using the most recently calculated versions of the neighboring internal nodes as input to the median problem. Iteration continues until no improvement can be made at any node. The sum of the branch lengths produced by this "steinerization" of the tree is then an evaluation of how well it accounts for the data.

For small N we can carry out this procedure for all possible trees (e.g. [1]), but for larger studies, such as the present analysis of $N = 20$ genomes, heuristics and local optimization must be invoked (see Section 7)

7 Experimental results: estimating the phylogeny of early eukaryotes

7.1 Distance matrix methods

Applying our measure to all pairs of genomes in Table 1 for which we could establish a usable gene order (indicated by an entry under "Genes"), resulted in a matrix which we used for the initialization of our phylogeny analysis. Two genomes, *Malawimonas* and *Monosiga*, manifested uniformly high values indicating random genome order with respect to all the other genomes, and were thus dropped from the analysis. The matrix remaining was submitted to two distance matrix analyses, neighbor-joining and the Fitch-Margoliash procedure, which both produced the tree in Figure 2.

These initial results indicate that the mitochondrial gene orders, as compared by our normalized breakpoints measure, contain a clear phylogenetic signal. The red algae form a monophyletic group; so do the stramenopiles. The large jakobid mitochondrial genomes, thought to most closely represent the ancestral form [13], group with other early-branching lineages. In addition, the ciliates group with the stramenopiles, a configuration which is sometimes seen in phylogenies constructed with single gene sequences. Only the plants and green algae, which, according to a great diversity of scientific evidence, should also form a monophyletic group, do not seem to have conserved sufficient commonality in their mitochondrial gene orders for them to be grouped together. This is, however, consistent with the rapid evolution of these orders known to occur among other green algae, such as those listed in Table 1.

More detailed phylogenetic techniques, to be discussed in the following sections, do not do any better in reconstructing the plant-green algae group. Indeed, the noise caused by the inclusion of the green algal genomes has a distorting effect on other parts of the tree, particularly the ciliate branching and possibly the "repulsion" of *Marchantia* from the red algal group. (It is thought that the plant-green algae group shares some common ancestry with the red algae.) For further investigation of protist phylogeny, then, we reduced our data set through the elimination of the *Prototheca* gene order, which seems highly derived, as well as that of *Pedinomonas*, which has a very small number of genes.

The distance matrix methods applied to the 18 remaining

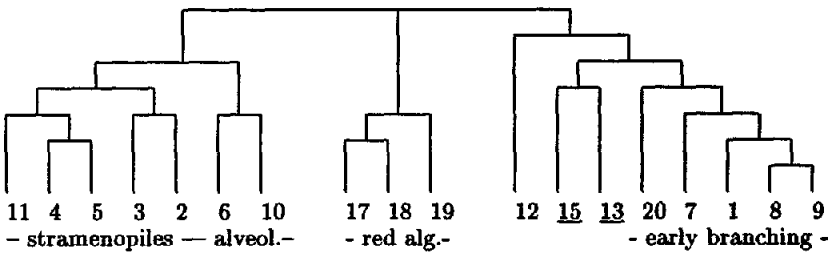

Figure 3: Distance tree without *Prototheca* and *Pedinomonas*. Hypothesis H_1.

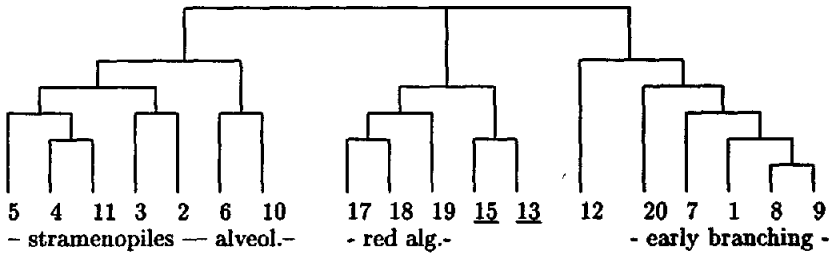

Figure 4: Hypothesis H_2 grouping plants and green algae with red algae

genomes produced the phylogeny in Figure 3. This is almost identical to what is obtained from the tree in Figure 2 by simply deleting the *Prototheca* and *Pedinomonas* branches. In the ensuing sections, we will refer to the evolutionary history implied by this tree as Hypothesis H_1.

Note that the what remains of the plant-green algae group still does not group with the red algae. Thus for the more detailed analyses below, we postulate another hypothesis, H_2, as represented in Figure 4.

Finally, we construct a more speculative hypothesis, H_3, which would place the cryptophyte *Rhodomonas salina* on the lineage leading to the red and green algae on plants, based on the possibility that flattened cristae may be monophyletic (Figure 5). This also has the effect that *Naegleria gruberi*, the sole representative of the organisms with discoidal cristae, now branches earlier, more in line with the ancient divergence thought to have occurred with this group.

7.2 Practical tree length evaluation

Despite the time complexity gains described in Section 5.2, computation time remains a major obstacle to extensive phylogenetic analysis. While the furthest insertion heuristic gives considerably better solutions to the breakpoint median problem and to the more general breakpoint phylogeny problem, the running time taken makes extensive tree searches impossible. For example, in carrying out a Nearest Neighbor Interchange (NNI) search for a local optimum starting with the phylogeny in Figure 2 (to be discussed below), we found that one pass through a single tree with 20 leaves with one iteration of random insertion at each node took approximately 10 seconds of CPU time on a Sparc Ultra 5. The same calculation with the furthest insertion heuristic takes roughly 7 minutes. Thus multiple long searches, requiring hundreds of tree evaluations, are not currently feasible using the latter.

In future it will be necessary to either improve the complexity of the furthest insertion heuristic, or modify the heuristic, as a time complexity of much more than $O(n)$ time appears to be impractical.

In the meantime, our approach to this problem is to calculate a rough upper bound on the tree, working with the assumption that the same rough method will preserve the relative order of tree lengths. Running the algorithm for a few passes over a tree, with a limited number of iterations at each node gives a rough upper bound of the total edge length - but we might hope that it over-estimate to the same degree for each tree. Furthermore, replacing the furthest insertion algorithm with one that chooses the insertion order randomly (and can thus be implemented in $O(n)$ expected time) speeds up the evaluation by orders of magnitude, again with uncertain effect on relative scores. This rough search, using a quickly evaluated upper bound as a criteria, is used to identify the possible neighborhoods of optimal trees. Optimal trees within these neighborhoods can be selected using a more time consuming, but more reliable, estimation procedure.

7.3 Evaluating the three hypotheses

To illustrate the possibilities and risks of these methods, we detail the results when they are applied using the hypotheses H_1, H_2 and H_3 in Figures 3, 4 and 5 as initial phylogenies in a local optimization. In Table 2, the total tree length of each hypothesis is presented after each of ten passes of the algorithm over the entire tree (the "outer" algorithm). In each pass the induced breakpoint median algorithm is iterated up to five times at each node. It can be seen that the total length criterion continues to improve for each iteration of the median algorithm and with each additional pass of the outer algorithm. Dramatically, a single application of the furthest insertion heuristic suffices to improve on the random insertion algorithm iterated five times, once there have been several passes of the outer algorithm.

Implementing our search for local optima using NNI, we limited ourselves to three passes of the outer algorithm and three iterations of the random-insertion median algorithm at each node. It can be seen that once the apparent local optima are evaluated more accurately, using the furthest-insertion method, for two of the hypotheses, H_1 and H_2, they turned out not to be optimal after all since their total lengths are greater than their respective initial trees.

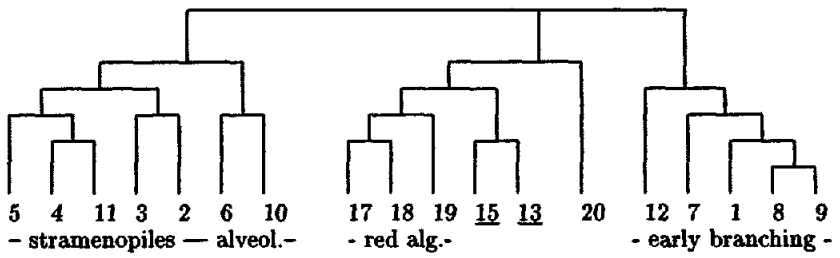

Figure 5: Hypothesis H_3 grouping all organisms with flattened cristae

In the case of H_3, four NNI's produced a tree which was clearly better than that of the initial one, and marginally the best tree of all we examined. The main differences between this tree and H_3 are the breakup of the stramenopile-alveolate cluster into three separate but successive branches, the loss of monophyly of the plant-green algae group into two successive branches from the lineage leading to the red algae, though the "flattened" group including *Rhodomonas* remains monophyletic.

Returning to the 20-species analysis, NNI produced a phylogeny clearly more economical than that in Figure 2. However, the artifactual effects of including *Prototheca* and *Pedinomonas* are worsened, so that, for example, *Prototheca* now groups with the alveolates as a subgroup *within* the stramenopiles. Thus we continue to disregard the results of the full data set.

8 Conclusions

This work highlights the potential of the induced breakpoint method for genomes with differing gene sets. The fact that we were able to obtain phylogenies as clear as many that are derived from sequence comparison attests to the amount of phylogenetic information which resides in gene order.

Until we can speed up the furthest-insertion algorithm to incorporate it into NNI, the results of our provisional phylogenetic analysis may be summarized as follows:

- The stramenopiles cluster together, usually but not always monophyletically, and there is a tendency for the ciliates to be a sister group.
- The jakobids, and other early branching (as previously revealed by sequence analysis) protists group together.
- The red algae group together.
- A tendency towards grouping the plants and green algae, and these with the other flattened cristae mitochondrial genomes, can only be detected by discarding the most highly diverged green algae from the analysis.
- The phylogenetic relationships at the earliest level – among the stramenopiles, alveolates and other tubular cristae mitochondrial genomes, and among the flattened, discoidal and tubular groups, remain uncertain, awaiting further mitochondrial sequences which fit the criteria for inclusion in our analyses.

There are a number of directions for further work. Iteration of the median problem can be replaced by a more efficient process of gene insertion at all the internal genomes simultaneously.

Though it is a useful exercise to consider gene order only, a more accurate approach might take into account both gene order and gene complement in a single measure.

Finally, this work underscores the interest inherent in the evolution of the mitochondrial genomes, especially at the earliest times.

Acknowledgments

Research supported by grants to the authors from the Natural Sciences and Engineering Research Council (NSERC) and the Medical Research Council of Canada. Thanks to C.J. O'Kelly for helpful comments. DS and BFL are Fellows, GB an Associate, and DB a postdoctoral fellow in the Evolutionary Biology Program of the Canadian Institute for Advanced Research. MD holds an NSERC summer studentship.

References

[1] Blanchette, M., Kunisawa, T. and Sankoff, D. 1999. Gene order breakpoint evidence in animal mitochondrial phylogeny. *Journal of Molecular Evolution* 49, 193-203.

[2] Bryant, D., Deneault, M. and Sankoff, D. 1999. The breakpoint median problem. Centre de recherches mathématiques, Université de Montréal. ms.

[3] Burger, G., Saint-Louis, D., Gray, M.W. and Lang, B.F. 1999. Complete sequence of the mitochondrial DNA of the red alga *Porphyra purpurea*: cyanobacterial introns, and shared ancestry of red and green algae. *Plant Cell* 11, 1675-1694.

[4] Caprara, A. 1999. Formulations and hardness of multiple sorting by reversals. In: Istrail, S., Pevzner, P.A. and Waterman, M. (eds) *Proceedings of the 3rd Annual International Conference on Computational Molecular Biology (RECOMB 99).* ACM, New York, pp 84-93.

[5] Gray, M.W., Burger, G. and Lang, B.F. 1999. Mitochondrial evolution. *Science* 283, 1476-81.

[6] Gray, M.W., Lang, B.F., Cedergren, R.J., Golding, B., Lemieux, C., Sankoff, D., Turmel, M., Brossard, N., Delage, E., Littlejohn, T.G., Plante, I., Rioux, P., Saint-Louis, D., Zhu, Y. and Burger, G. 1998. Genome structure and gene content in protist mitochondrial DNAs. *Nucleic Acids Research* 26, 865-878.

[7] Hannenhalli, S., Chappey, C., Koonin, E.V. and Pevzner, P.A. 1995. Genome sequence comparison and scenarios for gene rearrangements: a test case. *Genomics* 30, 299-311.

Iteration per node → ↓ per tree		Random order					Furthest	
		1	2	3	4	5	1 original	1 after NNI
1	H_1	20.05	17.03	17.77	16.57	16.47	18.52	18.96
	H_2	17.12	17.87	18.13	17.37	17.68	16.68	19.13
	H_3	18.29	18.30	16.91	16.64	16.38	17.80	15.58
2	H_1	15.74	12.25	12.08	11.89	11.89	14.30	13.84
	H_2	13.90	13.41	13.16	13.36	13.65	14.18	14.29
	H_3	14.98	13.91	13.06	11.78	11.78	15.15	11.99
3	H_1	13.49	11.27	11.31	10.90	10.95	12.20	11.94
	H_2	12.56	11.93	11.13	11.27	11.71	11.86	12.54
	H_3	12.34	11.71	11.28	10.95	10.83	12.33	10.23
4	H_1	11.63	11.00	10.93	10.47	10.73	10.71	10.65
	H_2	11.42	11.29	10.80	11.04	10.89	10.90	10.79
	H_3	11.69	11.32	10.88	10.87	10.70	10.99	9.92
5	H_1	11.33	10.98	10.64	10.25	10.57	10.23	10.24
	H_2	11.32	11.08	10.29	10.91	10.51	9.98	10.51
	H_3	11.33	11.15	10.84	10.76	10.64	10.70	9.89
6	H_1	10.88	10.80	10.57	10.23	10.46	9.78	10.24
	H_2	11.18	10.88	10.28	10.88	10.28	9.80	10.34
	H_3	10.93	11.06	10.61	10.71	10.64	10.68	9.72
7	H_1	10.84	10.80	10.50	10.22	10.44	9.78	9.85
	H_2	11.12	10.73	10.20	10.86	10.25	9.58	10.09
	H_3	10.89	10.95	10.56	10.70	10.64	10.54	9.62
8	H_1	10.81	10.78	10.42	10.21	10.39	9.63	9.73
	H_2	11.09	10.57	10.17	10.70	10.23	9.49	10.08
	H_3	10.85	10.95	10.56	10.70	10.61	10.40	9.54
9	H_1	10.76	10.76	10.38	10.20	10.34	9.53	9.73
	H_2	11.04	10.57	10.13	10.57	10.21	9.49	9.98
	H_3	10.76	10.95	10.46	10.68	10.60	10.31	9.52
10	H_1	10.67	10.76	10.38	10.20	10.27	9.52	9.67
	H_2	11.00	10.34	10.12	10.54	10.21	9.48	9.85
	H_3	10.76	10.91	10.44	10.65	10.60	10.29	9.47

Table 2: The effect of search parameters on tree length calculation. The median iterate algorithm was applied to trees representing Hypotheses H_1, H_2 and H_3 using the random insertion heuristic and the furthest insertion heuristic. The rows indicate the number of passes of the outer algorithm through the tree, and the columns give the number of iterations of the median heuristic at each node.

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

[9] Korab-Laskowska, M., Rioux, P., Brossard, N., Littlejohn, T.G., Gray, M.W., Lang, B.F. and Burger, G. 1998. The Organelle Genome Database Project (GOBASE). *Nucleic Acids Research* 26, 139-146.

[10] Lang, B.F., O'Kelly, C.J. and Burger, G. 1998. Mitochondrial genomics in protists, an approach to probing eukaryotic evolution. *Protist* 149, 313-322.

[11] Lang, B.F., Seif, E., Gray, M.W., O'Kelly, C.J. and Burger, G. 1998. A comparative genomics approach to the evolution of eukaryotes and their mitochondria. *Journal of Eukaryote Microbiology* 46, 320-326.

[12] Lang, B.F., Gray, M.W. and Burger, G. 1999. Mitochondrial genome evolution and the origin of the eukaryotes. *Annual Review of Genetics* 33, 351-97.

[13] Lang, B.F., Burger, G., O'Kelly, C.J., Cedergren, R.J., Golding, B., Lemieux, C., Sankoff, D., Turmel, M. and Gray, M.W. 1997. An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. *Nature* 387, 493-497.

[14] Paquin, B., Laforest, M.-J., Forget, L., Roewer, I., Zhang, W., Longcore, J. and Lang, B.F. 1997. The Fungal Mitochondrial Genome Project: evolution of fungal mitochondrial genomes and their gene expression. *Current Genetics* 31, 380-395.

[15] Paquin, B. and Lang, B.F. 1996. The mitochondrial DNA of *Allomyces macrogynus*: the complete sequence from an ancestral fungus. *Journal of Molecular Biology* 255, 688-701.

[16] Pe'er, I. and Shamir, R. 1998. The median problems for breakpoints are NP-complete. Electronic Colloquium on Computational Complexity Technical Report 98-071, http://www.eccc.uni-trier.de/eccc

[17] Reinelt, G. 1991. *The traveling salesman - computational solutions for TSP applications.* Springer Verlag.

[18] Sankoff, D. 1992. Edit distance for genome comparison based on non-local operations. In: Apostolico, A., Crochemore, M., Galil, Z. and Manber,

U. (eds) *Combinatorial Pattern Matching. 3rd Annual Symposium. Lecture Notes in Computer Science* 644. Springer Verlag, New York, pp 121-135.

[19] Sankoff, D. and Blanchette, M. 1997. The median problem for breakpoints in comparative genomics. In: Jiang, T. and Lee, D.T. (eds) *Computing and Combinatorics, Proceeedings of COCOON '97. Lecture Notes in Computer Science* 1276. Springer Verlag, New York, pp 251-263.

[20] Sankoff, D., Blanchette, M. 1998. Multiple genome rearrangement and breakpoint phylogeny. *Journal of Computational Biology* 5, 555-570.

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

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

[23] Turmel, M., Lemieux, C., Burger, G., Lang, B.F., Otis, C., Plante, I. and Gray, M.W. 1999. The complete mitochondrial sequences of *Nephroselmis olivacea* and *Pedinomonas minor*: two radically different evolutionary patterns within green algae. *Plant Cell* 11, 1717–1729.

[24] Watterson,G.A., Ewens, W.J., Hall,T.E. and Morgan, A. 1992. The chromosome inversion problem. *Journal of Theoretical Biology* 99, 1-7.