CHLOROPLAST GENE ORDER AND THE DIVERGENCE OF PLANTS AND ALGAE, FROM THE NORMALIZED NUMBER OF INDUCED BREAKPOINTS

David Sankoff

Mélanie Deneault

David Bryant

Claude Lemieux

Monique Turmel

Normalizing the number of breakpoints between two genomes, previously reduced by deleting the genes specific to one or the other, gives a rapidly calculated index of gene order evolution which is not based on any assumptions about the relative importance of various possible rearrangement processes. This is used to compare chloroplast genomes with known gene orders, with a focus on recently sequenced members of the algal class Prasinophycae. Phylogenies based on pairwise comparisons highlight the paraphyly and polyphyly of this class, with *Mesostigma viride* branching near the time of divergence of the streptophytes and chlorophytes, possibly within the former, *Nephroselmis olivacea* branching near the base of the Chlorophyte lineage, and *Pedinomonas minor* diverging relatively recently from other Chlorophyte classes.

1 Introduction.

The origin and diversification of plants and algae, and their relationships with other chloroplast-containing organisms, are some of the fundamental problems of evolutionary theory. The widely accepted endosymbiotic origin of the chloroplast and its consequent evolution, in key respects independent of the evolution of the nuclear genome, make it a natural focus of phylogenetic studies, though

in a narrower range than the almost-ubiquitous eukaryote mitochondrion. Thus phylogenies based on the amino acid sequences of a number of proteins coded by organellar genes give a clearer understanding of the evolution of classes of green plants than was possible based on morphological classifications alone or on ribosomal RNA surveys [17, 16, 7]. In this note, we propose to study another type of chloroplast genome data, namely gene order, to see what this can contribute to the sequence-level analyses.

The key methodology used here is that of the normalized number of induced breakpoints, which allows the construction of phylogenies from gene order data using a distance-based approach. The idea of induced breakpoints was introduced by [13] and the normalized version was found to be most appropriate for phylogenetic purposes in a study of the early evolution of the protist mitochondrion [14].

2 The evolution of the green plants

Sequence comparisons suggest that all living green plants belong to one of two major phyla: Streptophyta (land plants and their closest green algal relatives, the charophytes); and Chlorophyta (the rest of green algae). Little is known about common ancestors of these two lines, except that among the "crown group" of eukaryotes they are most closely related to the red algae (phylum Rhodophyta), and more remotely to the glaucocystophytes, both of which diverged after the endosymbiotic event giving rise to the chloroplasts. They are possibly somewhat related to the cryptomonads (not on the basis of of the latter's plastids, which likely result from a secondary endosymbiosis, but on nuclear rRNA evidence [5]).

The earliest-branching chlorophyte class, the Prasinophycae, is paraphyletic, morphologically heterogeneous and consists of several orders stemming from a basal radiation. (Some of these, like the pedinophytes, are sometimes considered separate classes [9].) It is thus a likely locus for the search for ancestral forms. Organellar genomes from three prasinophytes *Mesostigma viride*, *Nephroselmis olivacea* and *Pedinomonas minor* have been sequenced in the laboratories of two authors (CL and MT), and the amino acid sequences of their genes have shown that their phylogenetic affiliations are very diverse, despite their erstwhile membership in the same class.

Thus on the basis of mitochondrial sequences, *Pedinomonas* groups with the derived chlorophyte class Chlorophyceae (containing, for example, *Chlamydomonas*), while *Nephroselmis* turned out to be the earliest branching chlorophyte known [16]. Analysis of chloroplast genes, summarized in Figure 1, confirms this early chlorophyte status of *Nephroselmis* [17] but places *Mesostigma viride* earlier than any known green plant [7], before the divergence of the streptophytes and the chlorophytes. This latter result is bolstered by independent evidence of *Mesostigma* chloroplast gene content and genome organization, which both clearly predate the streptophyte-chlorophyte split.

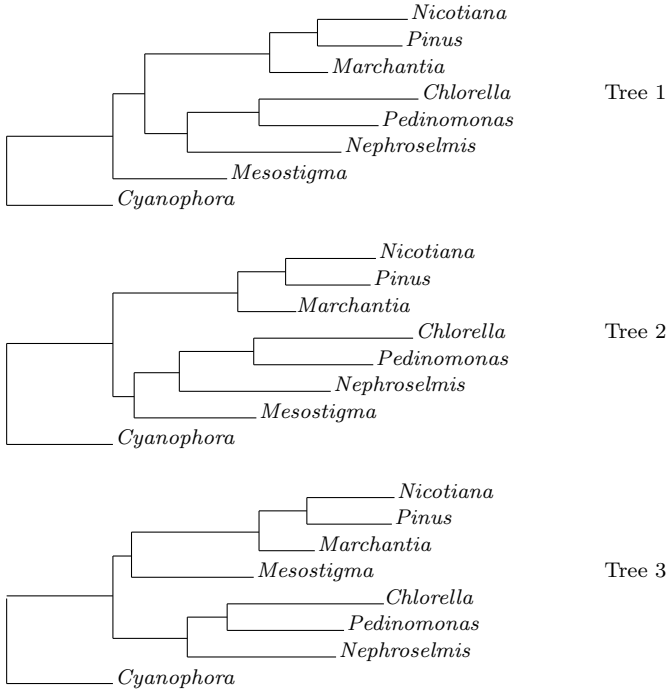

Figure 1: Trees representing three hypotheses for early branching of *Mesostigma*. These are the only trees with any support in a comprehensive study of protein sequences derived from 53 chloroplast genes [7]. Tree 1 (clearly favoured): Branching before streptophyte-chlorophyte split. Tree 2: Early branching within chlorophytes. Tree 3: Early branching within streptophytes

3 The data

For comparability, we analyzed the same chloroplast genomes as [7], summarized in Table 1. This includes the selection of the red alga *Cyanaphora paradoxa* as an outgroup. The gene orders of three of the chloroplasts (*Pedinomonas minor*, *Nephroselmis olivacea* and *Mesostigma viride*) are drawn from the primary data produced in the laboratories of these authors; gene orders for the other genomes were extracted from GenBank entries.

For the purposes of the analysis in the present paper, duplicate genes were excluded from the genomes. In the case of the inverted repeat seen in green plants, one of the two regions was deleted to avoid double weighting on the genes in our analysis. Other isolated duplicates (see Table 2) were deleted (both copies) because of the inability of our method to distinguish among paralogies. Methods for handling paralogies are available [12, 15], but cannot be applied in distance matrix methods in a consistent way.

Most of the deletions involved tRNA genes. As far as possible, we tried to identify homologous sets of chloroplast tRNA genes across as many of the genomes

Organism (Accession number)	Classification	Genes
Marchantia polymorpha (M68929)	land plant	86
Pinus thunbergii (D17510)	land plant	94
Nicotiana tabacum (Z00044)	land plant	107
Pedinomonas minor (AF116775)	green alga	106
Nephroselmis olivacea (AF137379)	green alga	130
Mesostigma viride (AF166114)	green alga	141
Chlorella vulgaris (AB001684)	green alga	109
Cyanophora paradoxa (U30821)	glaucocystophyte	175

Table 1: Chloroplast genomes studied. Gene numbers affected by the exclusion of some duplicate genes (see text and Table 2).

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 deletion introduces little bias into the comparison, though the loss of data does decrease the precision of the estimates.

Organism	Genes deleted
Chlorella	*trnG*(gcc), *trnV*(uac)
Cyanophora	*clpP**
Marchantia	*trnM*(cau)
Mesostigma	none
Nephroselmis	*trnL*(caa)
Nicotiana	none
Pedinomonas	*trnM*(cau)*
Pinus	*trnS*(gcu), *trnH*(gug), *trnT*(ggu), *trnI*(cau)

Table 2: Duplicate genes removed for the analysis. * indicates one of the duplicates was in the inverted repeat, so that all three copies were removed.

4 The notion of breakpoint and its extensions to unequal genomes

For two genomes X and Y containing the same genes, a breakpoint in X is simply a pair of genes g_1 and g_2 which are adjacent in reading order $g_1 g_2$ in X but are not adjacent, or not in this order, in Y. We use the minus sign $(-)$ to indicate a change in reading direction (i.e. change of DNA strand) consequent to the inversion of a chromosomal fragment. Thus the adjacent pair $-g_2 - g_1$ is considered to have the same order as $g_1 g_2$. We use $b_Y(X)$ to denote the number of breakpoints in X. Let $b(X, Y) = b_Y(X) = b_X(Y)$. The measure b tends to correlate with the evolutionary divergence of the two genomes.

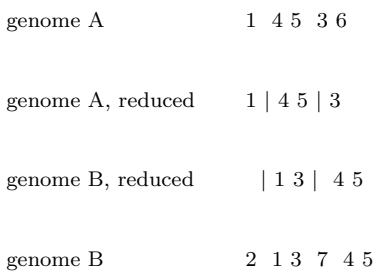

Figure 2: Induced breakpoints for (circular) genomes with different gene contents. Position of induced breakpoints (vertical strokes) found in reduced genomes with identical gene sets.

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 other, the number of breakpoints may be large. It is more appropriate in this context to consider *induced breakpoints*, (cf [13]), the breakpoints appearing when the genes belonging to only one or the other genome are discarded. As in Figure 2, 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.

Let X_Y be the genome X reduced by deleting all the genes not in Y. Then the number of induced breakpoints is $b_I(X, Y) = b(X_Y, Y_X)$. The measurement of induced breakpoints is a more subtle way of capturing the degree of parallelism of two gene orders than ordinary breakpoints.

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. First, deleting genes which are not shared with other genomes in a phylogeny has no effect on the induced breakpoint measure, by definition. A deleted duplicate gene is unlikely to figure in an adjacency crucial for phylogenetic grouping, since if it shares adjacencies with related genomes, it could be distinguished from its paralogs and would not be deleted. Other duplicated genes excluded were part of the inverted repeat, and relevant adjacencies are conserved in the undeleted copy.

For the purposes of phylogeny, the numbers of induced breakpoints can be misleading when the genomes vary significantly in gene content. Consider two small genomes having the same gene content and two much larger genomes having the same gene content, where the proportions of breakpoints are approximately the same within each pair. This is consistent with the hypothesis that the two pairs are equally divergent, but that by virtue of more potential sites the large genomes have had more opportunities to rearrange than the small genomes. The *absolute* number of breakpoints (or induced breakpoints) will then be much larger for the pair of large genomes, despite the equal divergence times for the two pairs.

To eliminate this artifact, we normalize the number of induced breakpoints by

the total number of genes shared by the two genomes. If $\mid X \mid$ is the number of genes in genome X, then the normalized number of induced breakpoints is

$$b_N(X,Y) = \frac{b_I(X,Y)}{\mid X_Y \mid}.$$

5 Experimental results: estimating the phylogeny of the green plants

Applying our measure $b_N(X,Y)$ to all pairs of genomes X and Y in Table 1, resulted in the matrix in Table 3, which we submitted to neighbor-joining and Fitch-Margoliash analyses. Both of these produced the tree in Figure 3, which is virtually identical to T_3 in Figure 1, except for the interchange of *Marchantia* and *Pinus*.

Chlorella	0	0.649	0.609	0.550	0.554	0.610	0.500	0.607
Cyanophora	0.649	0	0.618	0.598	0.653	0.640	0.625	0.607
Marchantia	0.609	0.618	0	0.461	0.597	0.092	0.554	0.176
Mesostigma	0.550	0.598	0.461	0	0.463	0.426	0.525	0.461
Nephroselmis	0.554	0.653	0.597	0.463	0	0.568	0.530	0.588
Nicotiana	0.610	0.640	0.092	0.426	0.568	0	0.543	0.169
Pedinomonas	0.500	0.625	0.554	0.525	0.530	0.543	0	0.519
Pinus	0.607	0.607	0.176	0.461	0.588	0.169	0.519	0

Table 3: Normalized numbers of induced breakpoints among chloroplast genomes.

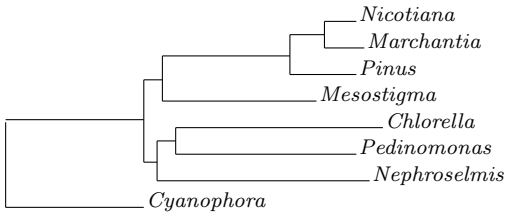

Figure 3: Result of applying neighbour-joining to the breakpoint measures in Table 3.

To confirm this result, we also applied the parsimony evaluation described by [14] to the three trees in Figure 1. This method estimates the minimum sum, over all branches in the tree, of the normalized number of induced breakpoints between the nodes at the two ends of the branch. This requires concurrently reconstructing hypothetical ancestral genomes at all interior nodes of the tree. Because current implementations of the method can sometimes produce local minima misleadingly larger than the true value, it is necessary to run the program many times with different starting configurations in order to have a measure of confidence in the

best result found. In twenty runs, we found the best value of T_3 to be 1.97, while the minimizing values for T_1 and T_2 were 2.25 and 2.26, respectively. (The worst result for each of the three trees was around 2.5.) That nine out of the 20 runs for T_3 scored less than 2.25 lends confidence to the conclusion that the latter is the most parsimonious tree. Although further runs might result in somewhat lower scores, it seems highly unlikely that either of the other two trees could score less than 1.97.

The position of *Mesostigma* within the streptophytes has some support from rRNA trees, from phylogenies based on actin genes, and from some structural similarities with the charophyte class of streptophytes, as cited in [7]. Nevertheless, the overwhelming weight of evidence presented in the latter paper, from over 50 chloroplast genes, place *Mesostigma* clearly before the streptophyte-chlorophyte split. And this is bolstered by many other features it shares with streptophytes, but which chlorophytes lack, and others it shares with chlorophytes only.

The appropriate conclusion to draw from Figure 3 pertains to the degree to which the phylogenetic signal is conserved in chloroplast gene orders. There is a local error in the streptophytes, due to the small number of rearrangements in land plants. The positions of *Mesostigma* and *Nephroselmis* as branching early and of *Pedinomonas* as branching late are clearly represented. And the general structure of the tree, including the position of the outgroup and the streptophyte-chlorophyte split, is correct.

6 Secondary endosymbionts

After comparing our gene-order results with the sequence-based phylogeny of [7], we can add the other chloroplast genomes currently available (Table 4) to our analysis.

Organism (Accession number)	Classification	Genes
Porphyra purpurea (U38804)	red alga	224
Cyanidium caldarium (AF022186)	red alga	225
Odontella sinensis (Z67753)	stramenopile	153
Epifagus virginiana (M81884)	land plant	41
Zea mays (X86563)	land plant	123
Oryza sativa (X15901)	land plant	92
Euglena gracilis (X70810)	euglenoid	86
Toxoplasma gondii (U87145)	alveolate (apicomplexan)	44

Table 4: Chloroplast genomes added to test set in Table 1.

As can be seen in Figure 4, the erroneous placement of *Marchantia* with the higher flowering plants persists in this analysis, but the other land plants are correctly configured. The red algae are shown to be monophyletic and to form a sister lineage to the green plants, so that the outgroup status of the glaucocystophyte genome is validated. The *Euglena* chloroplast is thought to result from a secondary endosymbiosis of a green alga by a primitive protist, and this is reflected in the

grouping of *Euglena* with the green plants. Conversely, the *Odontella* chloroplast is believed to reflect a secondary symbiosis of a red alga by a stramenopile as is consistent with its grouping with the red algae in Figure 4. The apicomplexan protist *Toxoplasma* plastid groups more loosely with the red algae and it has been argued that these parasites also arose through an endosymbiosis with a red alga [8], though other evidence points to green algal involvement [6, 11].

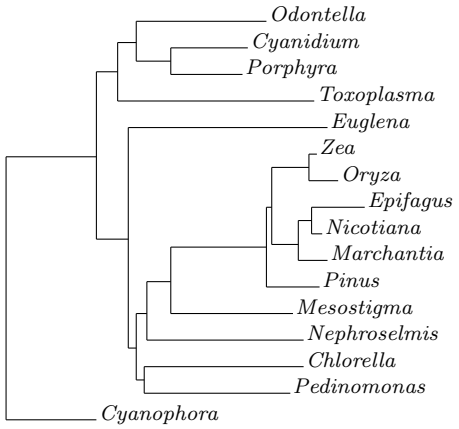

Figure 4: Tree showing origins of secondary endosymbiotic chloroplasts.

While all these results testify to the usefulness of our approach, we note that the pattern of similarities of the *Euglena* gene order with those of the green algae does introduce another error, disrupting the chlorophytes by rooting the green plant subtree within this group. This is the only inconsistency of this tree with that of Figure 3

7 Conclusions

Breakpoint methods have now been successfully applied to at least four data sets: animal mitochondia [1], protist mitochondria [14], the chloroplasts of the Campanulaceae [4] and the chloroplasts of green plants and protists (present paper).

The work reported here highlights the potential of the induced breakpoint method for genomes with differing gene sets. The fact that we are 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.

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 Gertraud Burger and Franz Lang for advice. DS is a Fellow, CL and

MT are Associates, and DB was a postdoctoral fellow (1999) in the Evolutionary Biology Program of the Canadian Institute for Advanced Research. MD held NSERC summer studentships (1999, 2000).

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] Boudreau, E. and Turmel, M. 1996. Extensive gene rearrangements in the chloroplast DNAs of *Chlamydomonas* species featuring multiple dispersed repeats. *Molecular Biology and Evolution* 13, 233-43.

[3] Buchheim, M.A., Lemieux, C., Otis, C., Gutell, R.R., Chapman, R.L. and Turmel, M. 1996. Phylogeny of the *Chlamydomonadales (Chlorophyceae)*: a comparison of ribosomal RNA gene sequences from the nucleus and the chloroplast. *Molecular Phylogenetics and Evolution* 5, 391-402.

[4] Cosner, M. E., Jansen, R. K., Moret, B. M. E., Raubeson, L. A., Wang, L.-S., Warnow, T. and Wyman, S. 2000. An empirical comparison of phylogenetic methods on chloroplast gene order data in Campanulaceae. This volume.

[5] Douglas, S.E., Murphy, C.A., Spencer, D.F. and Gray, M.W. 1991. Cryptomonad algae are evolutionary chimaeras of two phylogenetically distinct unicellular eukaryotes. *Nature* 350, 148-151.

[6] Köhler, S., Delwiche, C.F., Denny, P.W., Tilney, L.G., Webster, P., Wilson, R.J.M., Palmer, J.D., and Roos, D.S. 1997. A pastid of probable green algal origin in apicomplexan parasites. *Science* 275:1485–1488.

[7] Lemieux, C., Otis, C. and Turmel, M. 2000. Ancestral chloroplast genome in *Mesostigma viride* reveals an early branch of green plant evolution. *Nature* 403, 649-652.

[8] McFadden, G.I., Waller, R.F., Reith, M.E., and Lang-Unnasch, N. 1997. Plastids in apicomplexan parisites, pp. 261–287. *In* D. Bhattacharya (ed.), Origins of Algae and their Plastids. Springer-verlag, New York.

[9] NCBI Taxonomy browser.
http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/

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

[11] Roos, D.S., Crawford, M.J., Donald, R.G.K., Kissinger, J.C., Klimczak, L.J., and Striepen, B. 1999. Origin, targetting and function of apicomplexan pastid. *Current Opinion in Microbiology* 2:426–432.

[12] Sankoff, D. 1999. Genome rearrangements with gene families. *Bioinformatics* 15, 909-917.

[13] 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.

[14] Sankoff, D., Bryant, D., Deneault, M., Lang, B.F. and Burger, G. 2000. Early eukaryote evolution based on mitochondrial gene order breakpoints. In: Shamir, R., Miyano, S., Istrail, S., Pevzner, P. and Waterman, M. (eds) *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology (RECOMB 2000)* ACM Press, New York, pp 254-262.

[15] Sankoff, D. and El-Mabrouk, N. 2000. Duplication, rearrangement and reconciliation. This volume.

[16] Turmel, M., Lemieux, C., Burger, G, Lang, B.F., Otis, C., Plante, I. and Gray, M.W. 1999. The complete mitochondrial DNA sequences of *Nephroselmis olivacea* and *Pedinomonas minor*. Two radically different evolutionary patterns within green algae. *Plant Cell* 11, 1717-30.

[17] Turmel, M., Otis, C. and Lemieux, C. 1999. The complete chloroplast DNA sequence of the green alga *Nephroselmis olivacea*: insights into the architecture of ancestral chloroplast genomes. *Proceedings of the National Academy of Scences USA* 96, 10248-10253.

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

Département de biochimie et de microbiologie, Université Laval, Québec G1K 7P4
E-mail: {Monique.Turmel,Claude.Lemieux}@rsvs.ulaval.ca