Available online at www.sciencedirect.com
ScienceDirect
Molecular Phylogenetics and Evolution 44 (2007) 1130–1140
MOLECULAR PHYLOGENETICS AND EVOLUTION
www.elsevier.com/locate/ympev

Chromosomal character optimization

Ward C. Wheeler *

Division of Invertebrates, American Museum of Natural History, Central Park West at 79th Street, New York, NY 10024-5192, USA

Received 10 October 2006; revised 17 January 2007; accepted 20 January 2007
Available online 12 February 2007

Abstract

A method is presented to optimize chromosomal data on a cladogram. This procedure simultaneously considers variation at the nucleotide and locus levels including nucleotide substitution, insertion and deletion, locus insertions and deletion, and gene rearrangement. Locus labeling is not a requirement of the procedure and such annotation will result from the dynamic homology analysis of the chromosome data. An example of complete arthropod mtDNA sequences is presented.

Keywords: Optimization; Phylogeny; Chromosome; mtDNA; Systematics; Dynamic homology

1. Introduction

Comparative data sets are now available for complete chromosome sequences, but generally applicable tools for their systematic analysis are not. In the same way that sequences are arrays of nucleotides and undergo two basic types of transformation: substitution and insertion/deletion (indel), chromosomes are arrays of loci (=sequences) that undergo locus change and indel (the locus change being the sum of the substitutions and indels at the nucleotide level). In addition to these changes, the relative position of loci can vary resulting in rearrangement. In order to fully explain chromosomal variation, we need to accommodate all of these modes of transformation simultaneously (nucleotide substitution, nucleotide indel, locus indel, and locus rearrangement). To date these have not been simultaneously optimized in a character-based framework. The methods proposed here attempt to accomplish this.

2. The data

Complete chromosomal (whole genomes in some cases) data sets are available for a variety of taxa. Viral (e.g. Anderson et al., 2000) and bacterial genome sequences are available for hundreds and thousands of taxa (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi). There are close to 1000 complete mitochondrial (e.g. Nardi et al., 2003) as well as chloroplast sequences (http://www.ncbi.nlm.nih.gov/genomes/static/euk_o.html). These chromosomes vary from relatively short viral sequences (10 kb, 7 loci with little rearrangement for HIV; Anderson et al., 2000), to 16 kb for most animal mitochondrial DNA (with 35–40 loci), to low megabase sequences with hundreds of loci for some prokaryotic taxa (Fig. 1).

3. Existing approaches

The broadest array of data has been gathered from animal mtDNA and these exhibit both the strengths and weaknesses of current chromosomal analysis (Macey et al., 1997; Zardoya and Meyer, 1998; Inoue et al., 2001; Miya et al., 2001, 2003; Downton et al., 2003; Macey et al., 2004). Typically, two sorts of analysis are performed on these data: those that are based on the nucleotide sequences (e.g. Nardi et al., 2003), and those using gene order information (Boore et al., 1998). The major shortcoming of the nucleotide-based studies is their wholesale exclusion of data. Although these sequences are approximately 16 kb in length, frequently, many data go unused

* Fax: +1 212 769 5233.
E-mail address: wheeler@amnh.org.

1055-7903/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.ympev.2007.01.010

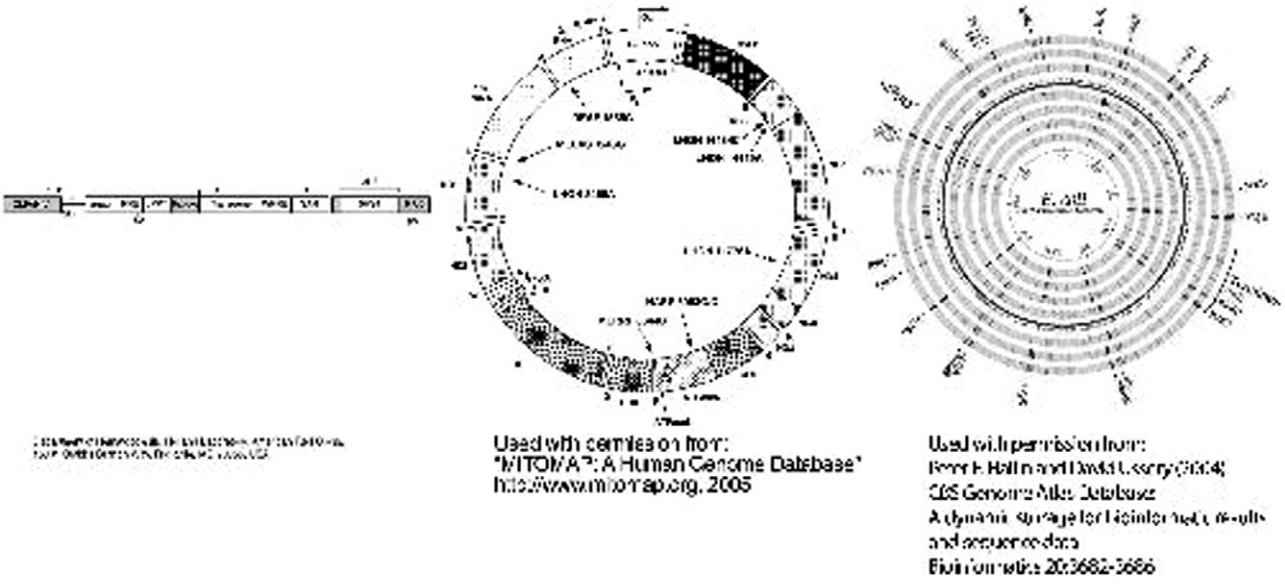

Fig. 1. Schematic representation of simple whole-genome chromosomes of HIV (left), human mtDNA (center), and *E. coli* (right).

in actual analysis (<40% by Miya et al., 2003; Ishiguro et al., 2003). The study of Nardi et al. is a case in point employing only a small percentage (4 of 13 gene regions) of the observations gathered in a concatenated multiple alignment. Not only is this extremely inefficient and wasteful of energy and resources, but the data exclusion process is prone to personal bias (Cameron et al., 2004).

Gene order information has been used by Boore et al. (1998) and others for both mitochondrial and chloroplast studies. Initially, this procedure involved hand-coding locus adjacencies as qualitative characters and analyzing them in standard fashion. The obvious difficulties in any hand analysis of such complex data, and the identification of rearrangement "hotspots," (Curole and Kocher, 1999; Mauro et al., 2006) have motivated less informal techniques. Recently, Sankoff et al. (1996), Moret et al. (2001), Bader et al. (2001), and Moret et al. (2002a,b) have applied algorithmic, optimality-based approaches to these data. These analyses are fine as far as they go, but they do not make use of the wealth of nucleotide variation (which might constitute orders of magnitude more information than gene order).

A generally useful method must synthesize these approaches and accommodate both these levels of variation.

4. The optimization approach

Methodologies of systematic analysis of molecular sequence data (optimality criterion aside) can be divided into two camps—those that are based on *a priori* homology statements and those that seek to identify homologies as transformations are optimized. Broadly, these can be referred to as alignment and optimization approaches, respectively (Wheeler, 2005). Alignment methods rely on some process to determine correspondences among nucleotides that are then further analyzed to create an optimal set of transformations on a cladogram or tree. Ideally, this homology process is explicit, optimality-based, and automatic such as with some multiple alignment programs (e.g. CLUSTAL Higgins and Sharp, 1988; Thompson et al., 1994; MALIGN Wheeler and Gladstein, 1994, 1991–1998). Optimality methods, on the other hand, create unique, topology-based, dynamic homology (Wheeler, 2001; Wheeler et al., 2005) schemes for each cladogram examined as they create their optimal transformation sets. In the context of single locus sequence analysis, both of these approaches can be viewed as heuristics to the general NP-Hard tree alignment optimization problem (Wang and Jiang, 1994).

Most analyses of chromosome data are alignment-type investigations (e.g. Nardi et al., 2003). Individual loci are homologized by some sort of annotation process (however well [Rawlings et al., 2003] defined) before sequence alignment and phylogenetic analysis. Then the nucleotides are concatenated into a larger alignment for subsequent treatment. The locus homology process might well involve rearranging chromosome sections such that appropriate segments are aligned in the concatenated structure. This rearrangement information is rarely used in tandem with nucleotide variation. Though not a requirement of this type of analysis, much hand-editing and modification of the data frequently occur at this step (data removal, by-eye alignment correction, and other manipulation; e.g. Nardi et al., 2003) which mar investigations and remove potentially informative variation (Harrison et al., 1985).

Objective analysis of locus order information is also something of an alignment technique in that locus homologies are specified before the optimization of rearrangements (Sankoff et al., 1996; Moret et al., 2001). These are viewed, however, as labels, and in that context the methods are optimization-based.

5. Criteria

The basic approach used here will be to create and examine alternate scenarios of chromosomal transformation on cladograms. Decisions will be made as to the relative quality of both the transformation scenarios and the cladograms upon which they rest. An objective criterion of quality is required to do this. The methods and discussions here use simplicity of explanation, i.e. parsimony, to decide between alternate schemata. This value is calculated as the weighted sum of all the transformation events required to explain the entirety of chromosomal variation on a cladogram (nucleotide substitutions and indels, locus indels, and rearrangements). The cladogram that minimizes this cost is the optimal solution for the problem.

Alternate optimality criteria exist to evaluate solution quality. Most prominent of these is the application of likelihood. In this context, explicit statistical models of nucleotide and chromosomal transformation would be required to evaluate the optimality of a cladogram-transformation scenario. Such models and methods do not currently exist, though a marrying of nucleotide (Thorne et al., 1991; Hein et al., 2003; Wheeler et al., 1996–2005) and likelihood-based rearrangement models (Larget et al., 2004) akin to what is accomplished here for parsimony, might lead to such a likelihood procedure.

The methods detailed here deal with the problem of determining the minimum cost scenario of chromosomal transformation for a single cladogram. In the context of phylogenetic search, many such cladogram optimizations would take place, calculating the same optimality value for each and choosing the best.

6. General approach

The basic approach to solving, in a heuristic sense, this complex problem is to build up a series of edit cost and median state (in essence ancestral hypothetical taxonomic unit—HTU) optimization heuristics from lower level problems. The three pieces of this construct are nucleotide optimization, locus indel determination, and locus rearrangement.

7. Sequence optimization

The first component of chromosomal optimization is the determination of the cost (and transformation events) required to transform one locus into another through their hypothetical ancestral locus (Fig. 2). This is the same tree-alignment problem first examined by Sankoff and Cedergren (1983) and the same fundamental methods can be employed. There are two general types of methods—estimation and search (Wheeler, 2005)—that can be used.

Estimation methods attempt to construct a minimum cost hypothetical ancestor from its two descendants (and perhaps its ancestor) using string-matching techniques. Direct optimization (DO—Wheeler, 1996) accomplishes

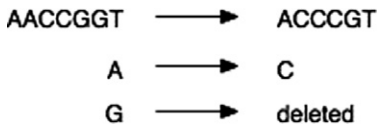

Fig. 2. Edit transformation from left sequence to right involving one substitution and one deletion.

this for two descendants and iterative-pass optimization (IP—Wheeler, 2003b) for the three including the ancestor of the two (Fig. 3a). Search procedures examine a pre-existing set of possible ancestral sequences and use dynamic programming to determine the optimal ancestral locus sequence. Examples of this include fixed-state optimization (FSO—Wheeler, 1999b) and the general search-based optimization (SBO—Wheeler, 2003c). Search methods can be rapid if the set of candidate ancestral sequences is small but the solutions suffer proportionately. The most restric-

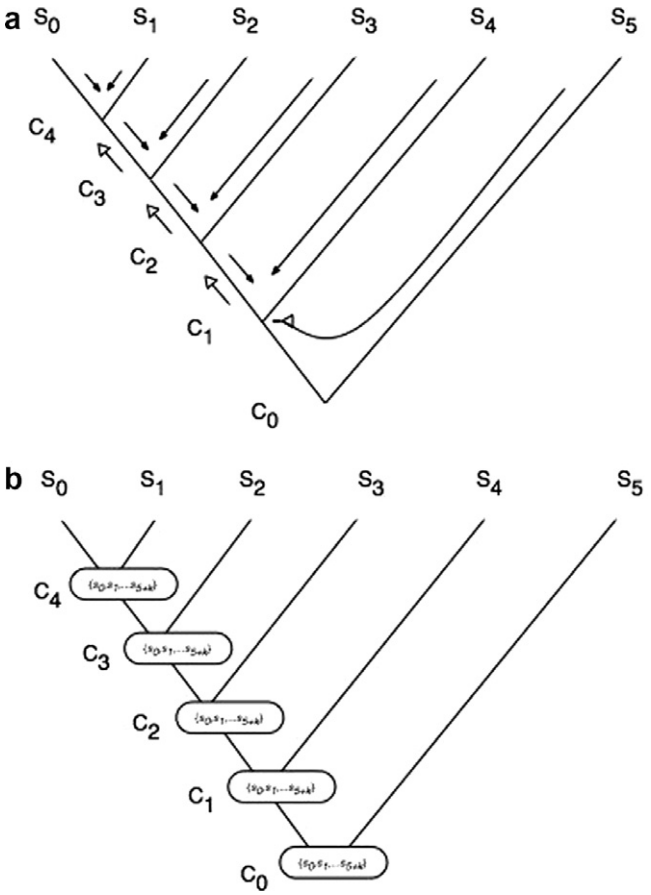

Fig. 3. Sequence median techniques to determine HTU states: (a) Direct (closed arrows) and iterative-pass optimization (closed and open arrows) sequence medians. The median sequences (C_i) are calculated based on either their two descendants, or their descendants and immediate ancestor. Multiple passes may be performed on the cladograms to update the vertex sequences and improve median quality. (b) Search-based optimization of a set of observed sequences (S_i) to determine vertex sequences (C_j) using a set of candidate sequences ($S_0, \ldots, S_{5+k}$). Fixed-states optimization would limit the vertex sequences to the observed S_i.

tive case of this is FSO where the sequence set is limited to those observed in terminal taxa (Fig. 3b).

8. Locus insertion–deletion

The second component of chromosomal optimization is the determination of the insertion–deletion pattern required to account for varying locus complement between two chromosomes which have the same relative positions of loci (no rearrangements).

This can be accomplished by the use of simple string matching modified as in DO, but for loci. In fact, the algorithm is identical except the locus case has a larger alphabet of objects to be matched (the loci in the two chromosomes to be compared versus A, C, G, T, and GAP). As with nucleotides, mismatch and indel costs are required. The mismatch cost of loci would be the DO transformation cost between loci (as described above). The cost of an indel could be specified as a simple single parameter for an indel of any locus size or could be more complex, allowing for differential costs of different size loci or contiguous multiple insertions.

9. Locus rearrangement

The third component of chromosomal optimization is the determination of the cost (and rearrangement events) required to convert one locus ordering into another (Figs. 4 and 5).

The median (HTU) problem (Fig. 6) for locus rearrangement is known to be NP-complete (Pe'er and Shamir, 1998). Whether one measures chromosomal rearrangement in terms of chromosomal breaks (changes in locus adjacency; Sankoff et al., 1996; Fig. 7a) or the more complex inversion distances (Bader et al., 2001; Fig. 7b), the basic problem is the determination of the minimum cost median arrangement among three chromosomes connected to a node.

Chromosomal breakpoints analysis was the first method proposed to discuss gene rearrangement in a phylogenetic context (Blanchette et al., 1997). Basically the locus adjacencies (irrespective of orientation) are listed as a series of pairs (Fig. 7a). Those pairs present in one chromosome, but not in another constitute the number of chromosomal breaks required to remove the original gene adjacencies.

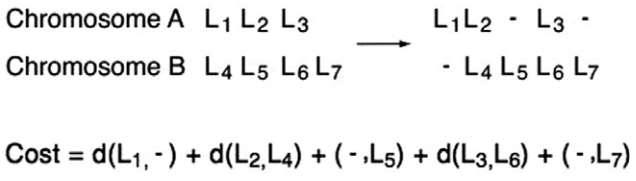

Fig. 4. Calculation of chromosome edit cost without rearrangement between chromosomes A and B based on the chromosomal "alignment" on the right. The total cost would be the sum of the three locus indels (L_1, L_5, and L_7) and the edit cost between loci L_2 and L_4, and L_3 and L_6.

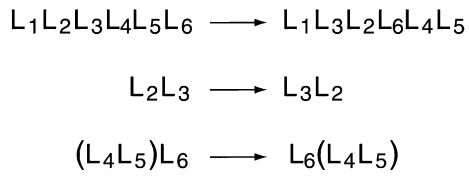

Fig. 5. Edits between two chromosomes (top). Rearrangement events involve the inversion of the relative positions of individual loci L_2 and L_3 (middle) and the two locus piece $L_4 + L_5$ (bottom).

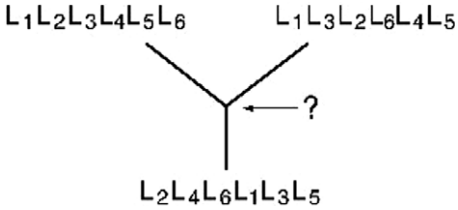

Fig. 6. Median HTU problem for three chromosomes.

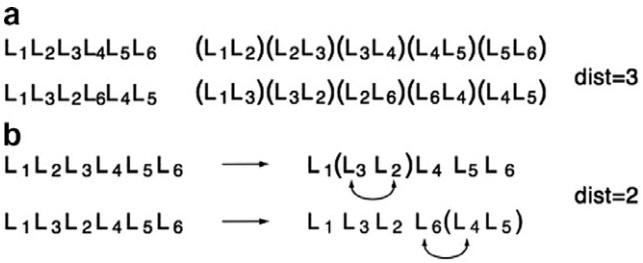

Fig. 7. Breakpoint distance (a) and inversion distance (b) between two chromosomes. (a) The adjacent loci in parentheses with three pairs found in the upper chromosome not found in the lower. (b) The two inversions required to edit one chromosome into the other.

The benefit of this metric comes in its ease of calculation, but it does not take into consideration the orientation of the loci. In order to use this information, the inversion distance was developed.

The inversion distance between two chromosomes is the minimum number of reversals required to transform one chromosome into another (Hanenhalli and Pevzner, 1995). The metric is more difficult to calculate, but allows orientation information (sign) to be used as well as relative location (Fig. 7b).

There are many heuristic approaches to solve these problems (Bader et al., 2001; Moret et al., 2001, 2002a,b) and many of these are implemented in GRAPPA (Bader et al., 2002). As mentioned above, these methods assume locus labels are known and that locus homology is known *a priori*.

10. Synthesis and methodology

The exact solution to this multi-level problem for just a single tree would involve the simultaneous solution of multiple NP-complete problems (tree alignment, inversion median) and is likely intractable for all but the smallest

of problems. The discussion here is limited to heuristic approaches which combine the three optimization components discussed above. There are two fundamental steps. The first is the determination of the edit cost (in terms of transformation events) between any two chromosomes, and second the median state optimization of three connected chromosomes to determine an unknown nodal chromosome. After a single node is determined, some sort of recursive revisiting of non-vertex (hypothetical taxonomic unit or HTU) nodes could be accomplished (Sankoff and Cedergren, 1983; Wheeler, 2003b) to optimize the entire cladogram, and from there to a complete cladogram search.

10.1. Edit distance

Given two chromosomes, what are the transformations and ensemble cost required to convert one chromosome into another? This conversion may involve nucleotide changes, nucleotide indels, locus indels, and locus rearrangements (as breakpoints or inversions). The edit cost distance between chromosomes A and B $[\Theta(A,B)]$ with M aligned loci, nucleotide substitution/indel edit costs σ, locus indel cost θ, and locus rearrangement cost ε:

$$\Theta(A,B) = \sum_{i=0}^{M} \left[\theta(A_i,B_i) + \left(\sum_{j=0}^{L_i} \sigma\left(A_i^j, B_i^j\right) \right) + \varepsilon(\{A_i, A_{i+1}\}, \{B_i.B_{i+1}\}) \right]$$

where

$$\theta(A_i,B_i) = \begin{cases} 0 & \text{if } A_i \neq \text{gap} \text{ and } B_i \neq \text{gap} \\ \text{locus\_gap\_cost} & \text{otherwise} \end{cases}$$

$$\varepsilon(\{A_i,A_{i+1}\},\{B_i.B_{i+1}\}) = \begin{cases} 0 & \text{if } \{A_i,A_{i+1}\} = \{B_i.B_{i+1}\} \\ \text{rearrange\_cost} & \text{otherwise} \end{cases}$$

The ancestral chromosome would be constructed from the locus indels and mismatches as in DO with ambiguities for those loci and indels that could not be uniquely determined.

This can be determined, in principle, by performing the chromosomal direct optimization for each of the $M!$ orderings of A on B, adding the rearrangement cost ε to the optimization cost (which included locus indels and change), and choosing the minimum value. This would guarantee the exact edit cost, but would be very time consuming for large numbers of loci. In practice, a heuristic subset of orderings is used.

Analogous to the Wagner (1961) procedure used to sequentially add taxa in building cladograms, chromosomes can be built by adding a locus at a time to each position to the nascent chromosome. For each placement, the edit cost between the complete (A) and incomplete (part of B) chromosomes is calculated, the best placement chosen, and the next locus added. This is done until all loci are added and the minimum edit cost determined (Fig. 8a). This could be refined by removing sets of loci and replacing them in all possible positions, and evaluating them in terms of improving the edit cost. Such a refinement step would be analogous to branch swapping in cladogram refinement (Fig. 8b).

Once the minimum edit cost is determined a candidate ancestor can be determined for the two chromosomes and used in a down-pass DO-type procedure to optimize the entire chromosome.

10.2. Median problem

The median problem can be broken into a series of edit distance calculations for a set of candidate HTU chromosomes (Fig. 9).

In essence, the HTU chromosome is compared to each of the three terminal chromosomes and the edit cost to each summed. This is repeated through an exhaustive or heuristic set of candidate ancestral chromosomes in order to find the HTU yielding the minimum cost.

In the same way that chromosomes could be "built" and refined for the edit cost calculation (above), a three-dimensional build can be defined where two of the three chromosomes are sequentially constructed with respect to the third. These builds could be sequential (yielding a $O(n) = 2n^2$ procedure) or simultaneous $[O(n) = n^4]$ for n loci. In each step of the build operation (and refinement if employed) a three-dimensional alignment of the chromo-

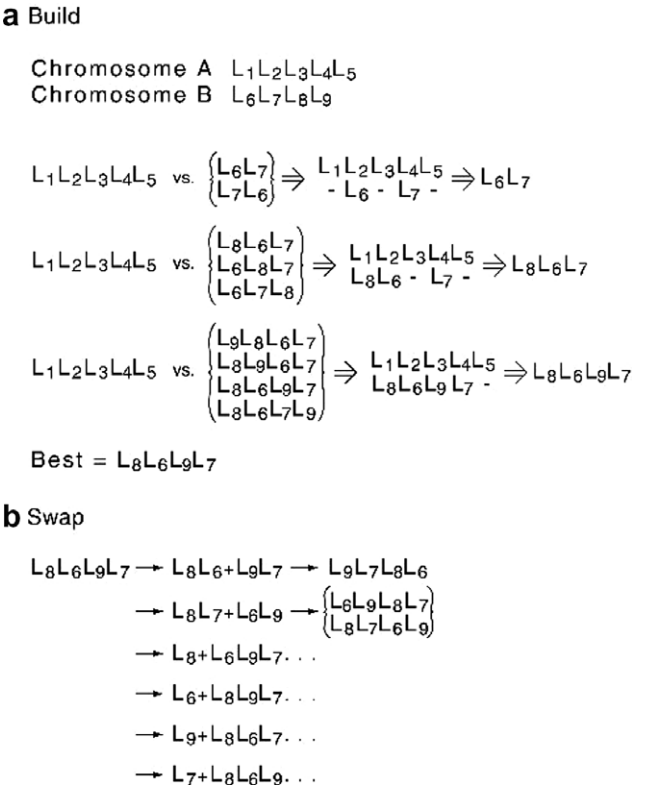

Fig. 8. Local search-type heuristics for chromosome edit cost modeled on tree search heuristics of initial build (a) and refinement (b).

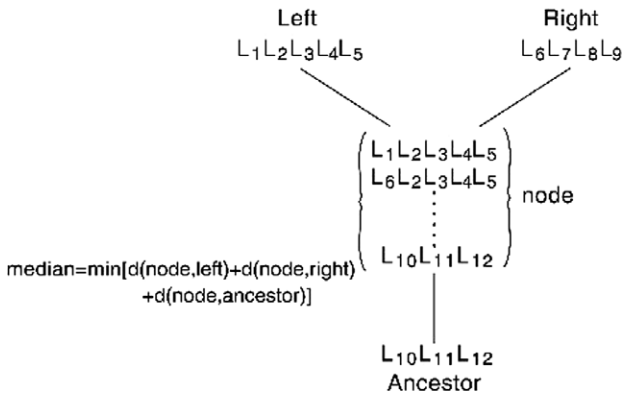

Fig. 9. Examination of candidate medians for three chromosomes.

somes would be performed and the joint indel-rearrangement scenario that minimized this cost retained.

As with the 2-D case above, after the best scenario for all three chromosomes is identified, a set of potential median chromosomes can be defined. The members of this set are evaluated by summing the pair-wise edit cost to each of the three adjacent chromosomes. The candidate median that minimizes this cost is retained.

11. Levels of heuristic

Several approaches to heuristic solutions are used to make the analysis of real data sets possible. The simplest of these is the FSO approach (Wheeler, 1999b). When chromosome ancestors are created in current implementations (POY—Wheeler et al., 1996–2005) via either 2-D or 3-D procedures (see above), the locus states are chosen from the set of loci in the input data. Under FSO novel sequence combinations are not calculated for locus ancestors. While the estimation of locus ancestor would likely improve the optimality values of cladograms, this would come at a premium in execution time.

A second example of this form of heurism would be to limit the possible set of ancestral chromosomes to those observed in the input (=leaf) data. This is truly a fixed-states approach to chromosomal optimization. This yields a relatively rapid heuristic cladogram cost, which could be very satisfactory for large data sets (known to be within factor of 2 of the minimum cost; since FSO is a superset of the "lifted" method, Gusfield, 1997). This method could also be used to generate an initial solution rapidly, which could then be progressively refined by more exhaustive procedures.

12. Simple example

Consider the example data of Fig. 10. There are four chromosomes, each with three or four loci that vary in nucleotide complement and number. Chromosomes"three" and "four" are rearranged in the second instance.

When the unrearranged data (Fig. 10) are analyzed using POY (Wheeler et al., 1996–2005) with base substitutions costing 1, indels 2, and locus indels 10, and locus breakpoint cost 1 using iterative pass optimization (these are somewhat arbitrary parameter values; Wheeler, 2003a), the resulting topology ((one (two (three four))); Fig. 11) had a cost of 20. This cost came from a single locus insertion (in chromosome three), along with eight nucleotide substitutions (two in loci 1–3) and one nucleotide indel in locus 3.

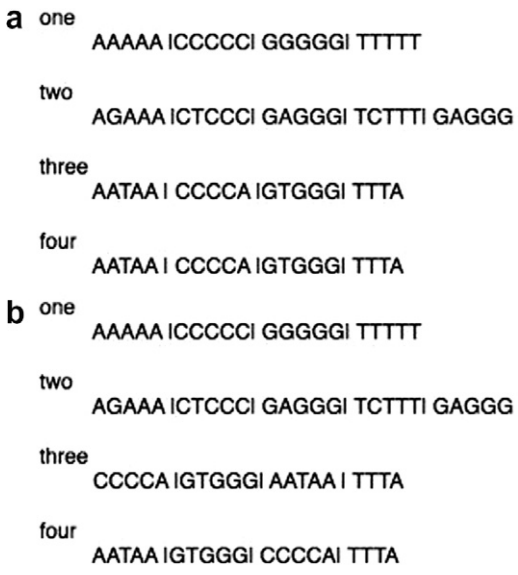

Fig. 10. Simple chromosomes (a) four chromosomes with 4 or 5 loci each, (b) data of (a) rearranged with locus 1 of taxon three moved to position 3 and locus 3 of taxon four in position 2. Pipes ("|") denote locus boundaries.

When chromosome three is reordered and no rearrangement is allowed, the cost increases to 34 from the extra nucleotide changes and indels at the locus level (Fig. 11b). When rearrangements are permitted, the cost reduces to 23—the original cost, with three extra events due to breakpoint cost of rearrangement along the lineage leading to chromosome three (Fig. 11c). When inversion costs are used (via the incorporation of GRAPPA; Bader et al., 2002), the cost is 26 since there are more inversions required than breakpoints in this example (this is partly due to patterns found in the short loci suggesting additional inversions; Fig. 11c).

13. Arthropod mtDNA

A test data set was created by downloading 108 arthropod complete mtDNA sequences (unique species) from GenBank (http://www.ncbi.nih.gov/). Based on annotation, there were 4587 "loci" (segments within and between annotated regions) defined, 4502 of which were unique and at least 20 nucleotides in length. These were analyzed using POY (Wheeler et al., 1996–2005, ported to a 64 bit system) with the *fixedstates* heuristic (Wheeler, 1999a) for rearrangements (only observed gene orders—the 108 input—considered for medians) under a variety of transformation

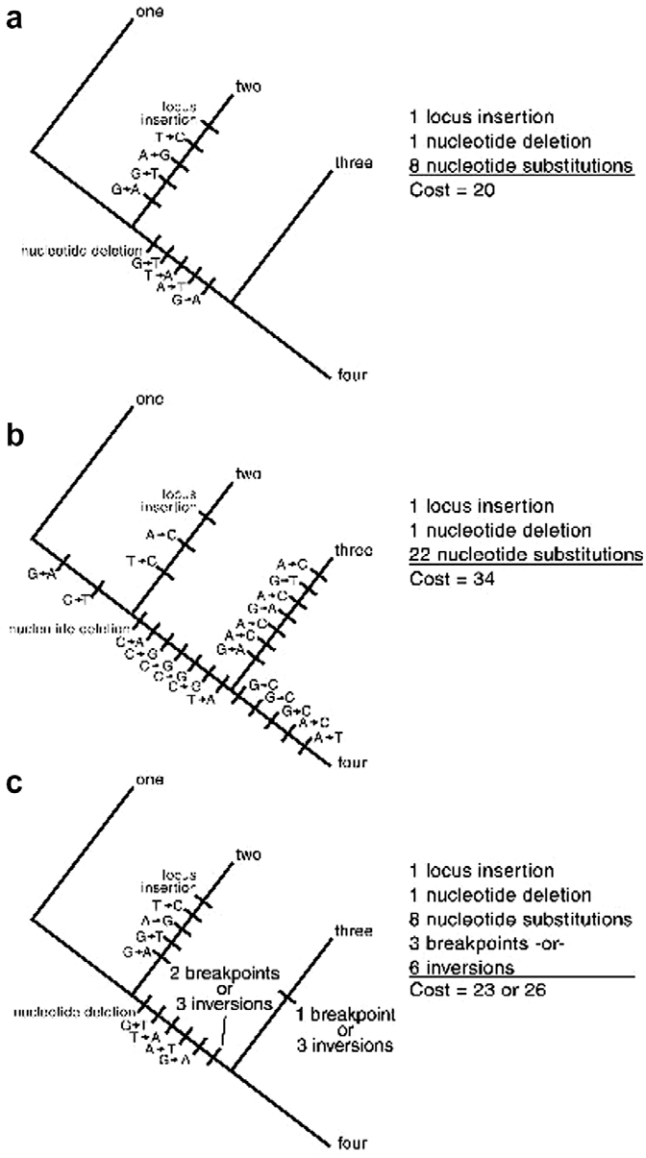

Fig. 11. Edits between four example chromosomes on three trees. (a) Sequence data of Fig. 10a. (b) Sequence data of Fig. 10b optimized without rearrangement—requiring many nucleotide changes; sequence data of Fig. 10b, allowing rearrangements measured by breakpoint and inversion requiring many fewer nucleotides changes at a cost of 3 or 6 rearrangements.

Table 1
Arthropod chromosome results

Rearrangement cost	Locus gap cost	Locus length cost	Locus "Swap"	Cladogram cost	Execution time
0	0	1	0	634,556	14,262
			100	630,258	58,327
		2	0	788,909	13,168
			100	698,967	64,717
	1000	0	0	721,373	26,793
			100	717,424	39,966
		1	0	752,902	12,898
			100	747,371	39,338
		2	0	1,078,173	13,882
			100	903,569	49,267
1000	0	1	0	731,112	13,999
			100	708,594	77,171
		2	0	851,607	13,214
			100	804,133	70,901
	1000	0	0	1,136,003	14,100
			100	1,083,837	52,297
		1	0	1,264,452	13,970
			100	1,192,635	47,081
		2	0	1,364,482	13,177
			100	1,254,597	89,607
100,000	0	1	0	731,723	14,010
			100	708,594	76,486
		2	0	851,607	13,077
			100	80,120	70,503
	1000	0	0	1,136,231	12,892
			100	1,087,239	51,889
		1	0	1,269,572	13,926
			100	1,202,557	45,436
		2	0	1,381,291	13,850
			100	1,250,420	44,437

Runs with locus gap and locus size gap both set to zero were not performed due to metricity constraints.

cost parameters. For each analysis, five replicates were performed each containing five Wagner build random replicates, TBR branch swapping, and tree fusing (Goloboff, 1999). Tree buffers were limited to 10 in all phases of analysis. All analyses treated the mtDNA chromosomes as circular with all sequence segments reversible and rearrangeable. Nucleotide substitutions were set to 1 and indels to 2 (linear cost) throughout. Individual runs varied in four parameters: locus build, whether chromosome medians were refined by "swapping" or not; cost of rearrangement events (0, 1000, 100,0000) measured by breakpoints (Blanchette et al., 1997); initial cost of a locus origin or loss (0, 1000); cost of locus origin or loss by length of segment (0, 1, or 2 length). Overall, 30 analyses were performed on an 8×2.2 GHz CPU AMD multiprocessor. The runs without locus swapping took from 12,898 to 14,262 s with one exception (26,793 s), those with swapping 39,338–77,171 s. Results are summarized in Table 1 and three sample cladograms are shown in Fig. 12.

As one might expect, more aggressive median refinement reduced overall cladogram cost. This improvement varied from less than 1% when rearrangements carried no cost—but still improved locus matching, to improvements of over 10% when rearrangements were costly (100,000). The taxonomic results of these analyses are not particularly satisfying, yet when combined with other data (e.g. morphological, nuclear) return results more in line with our notions of arthropod relationships (below).

14. Annotation

Annotation is the process of taking raw sequence data and identifying gene regions and other landmarks. Standard gene arrangement techniques presume loci are known

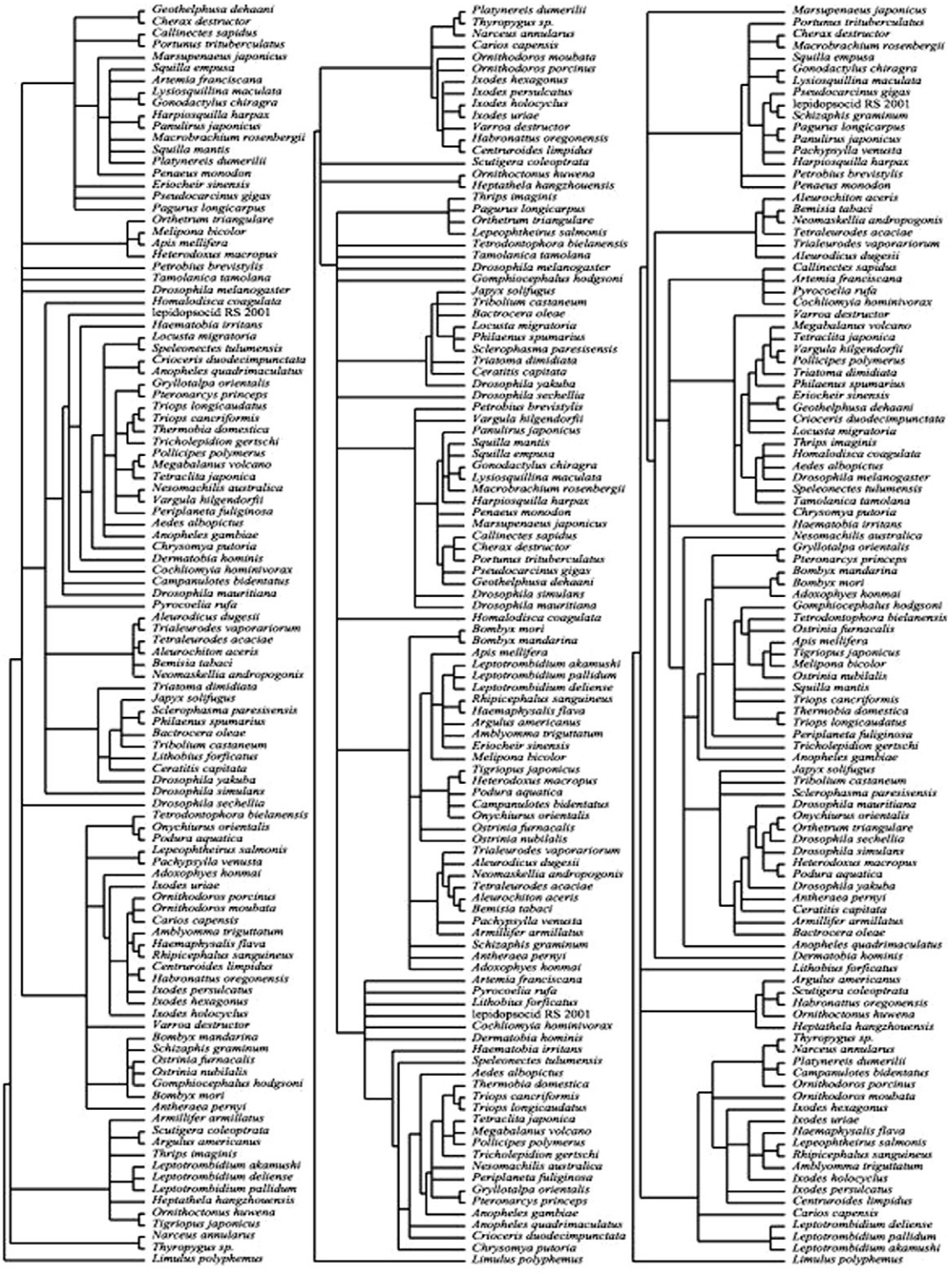

Fig. 12. Example arthropod complete mtDNA cladograms. The left cladogram resulted from analysis with rearrangement cost and initial locus indel cost set to zero, no locus refinement or swapping, and locus indel cost set equal to its length; the center cladogram was based on rearrangement cost and initial locus indel cost set to 1000, locus refinement of all locus segment sizes (<100), and locus indel cost set equal to its length; the right cladogram used a rearrangement cost of 100,000, initial locus indel cost of 1000, locus refinement of all locus segment sizes (<100), and locus indel cost set to twice its length.

and labeled. The method presented not only does not require *a priori* locus annotation, but produces one as an result (at least as far as gene labeling is concerned) in the proper context of historical homology. Since locus homology is dynamically determined, this aspect of annotation will be a topology specific, dynamic homology statement. A process akin to the implied alignment (Wheeler, 2003a) of nucleotides can be performed on the locus data yielding

locus homologies relative to the data at hand. If one or more of these chromosomes have been annotated previously, the absolute annotation can be extended to the other chromosomes (Fig. 13).

The most important aspect of annotation is not addressed with the optimization techniques described here and that is identifying the breaks between loci. These breaks not only delimit the gene regions, but also define the potential points of rearrangement. Ideally, such breaks would be able to occur at any position in the chromosomal sequence. Additional break locations can be added (at a cost of increasing the cost of rearrangement) to reduce some of the reliance on gene delimitation and its effects on phylogenetic analysis.

Annotation

one:	0|1|2|3|
two:	0|1|2|3|4|
three:	1|2|0|3|
four:	0|2|1|3|
HTU0:	0|1|2|3|
HTU1:	0|1|2|3|
HTU2:	0|2|1|3|

Presence-Absence

	0	1	2	3	4
one:	+	+	+	+	-
two:	+	+	+	+	+
three:	+	+	+	+	-
four:	+	+	+	+	-
HTU0:	+	+	+	+	-
HTU1:	+	+	+	+	-
HTU2:	+	+	+	+	-

Complete chromosome implied alignment

Locus 0

one	AAAAA
two	AGAAA
three	AATAA
four	AATAA

Locus 1

one	CCCCC
two	CTCCC
three	CCCCA
four	CCCCA

Locus 2

one	GGGGG
two	GAGGG
three	GTGGG
four	GTGGG

Locus 3

one	TTTTT
two	TCTTT
three	-TTTA
four	-TTTA

Locus 4

one	XXXXX
two	GAGGG
three	XXXXX
four	XXXXX

Fig. 13. Sequence data of Fig. 10b optimized with rearrangement showing annotation information. The ''Annotation'' portion shows locus homology to the first taxon, ''Presence–Absence'' the novel and deleted loci again with respect to the first taxon, and the ''Complete chromosome implied alignment'' the nucleotide homologies implied by the topology and locus-level variation.

15. Combined analysis

Ideally, data sets are analyzed in combination to bring the greatest amount of information to bear on a phylogenetic problem. Chromosomal characters, as treated here, can readily be combined with other information in a simultaneous analysis. As an example, the data of Giribet et al. (2005) (with partial mtDNA information deleted, but including seven nuclear loci and 352 morphological characters for 67 taxa) were augmented with 26 complete mtDNA sequences from GenBank (in some cases the data of closely related taxa were substituted when an exact match was unavailable). The analysis was performed using POY (Wheeler et al., 1996–2005) on a cluster of 20 pIV Xeon CPUs with 5 random replicates of 5 wagner tree builds per replicate, TBR branch swapping and tree fusing. Morphological changes and nucleotide indels were accorded a cost of 2, nucleotide substitutions 1, locus rearrangements (breakpoints) 100, locus indels cost 100 plus twice the locus length; locus refinement (swapping) was set to 100 (completion). The analysis examined 2,260,769 trees, taking 105,519 s yielding two cladograms at a weighted cost of 361,401 (Fig. 14). These results differ from those of Giribet et al. (2005) specifically by interdigitating several hexapod and non-hexapod taxa as is often seen in solely mitochondrial analyses.

16. Discussion

The strength of this approach to the phylogenetic analysis of chromosomal variation comes from its foundation in historical homology. This requires the simultaneous optimization of both sequence and locus level variation and in the definition of an explicit optimality criterion to choose among transformation scenarios and cladograms. Another benefit comes with the dynamic locus homology framework where loci are not required to be labeled (e.g. 18S rDNA) before analysis. The benefits are not without cost, however. The explicitness of the optimality criterion does require that additional analytical parameters be specified—locus indel and rearrangement costs—and the results will depend on the choice of these values (even though constrained by metricity). The annotation result of locus dynamic homology requires reference to a specific cladogram. As with nucleotide homology statements, these higher level correspondences are topology dependent.

A shortcoming of the method is its dependence on the a priori delimitation of the loci or at least the fragments which participate in rearrangement. To some extent this can be ameliorated by breaking up regions into smaller

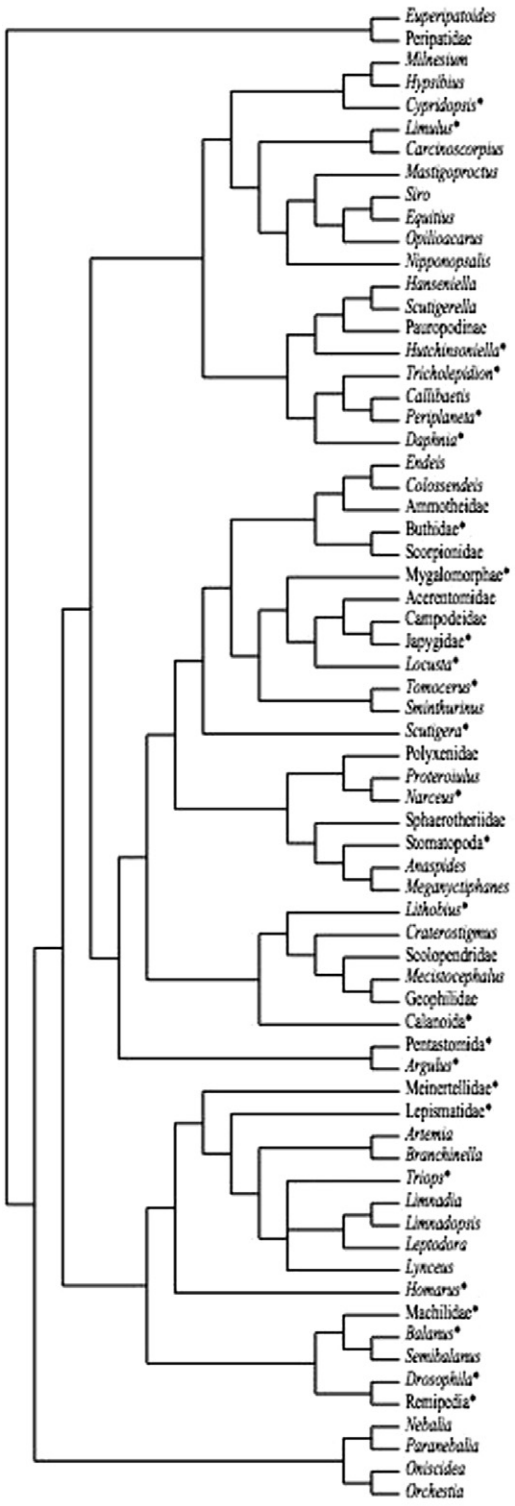

Fig. 14. Arthropod consensus cladogram based on combined analysis of morphology, nuclear, and mitochondrial data. The two constituent cladograms had a cost of 361,401. Taxa with "*" include complete mtDNA data.

fixed or randomly sized fragments, but such rearrangement analysis depends on the recognition of pieces that can be shuffled, and analytical times may increase significantly with the profusion of sequence fragments.

Acknowledgments

I thank Illya Bomash, Sidney Cameron, Louise Crowley, Julián Faivovich, Gonzalo Giribet, Pablo Goloboff, Taran Grant, Megan Harrison, Vinh Sy Le, Leo Smith, Andrés Varón, Michael Whiting, and an anonymous reviewer for discussion and manuscript commentary; Steven Thurston for the good artwork (as opposed to the rest); and the National Science Foundation and National Aeronautics and Space Agency for research support.

References

Anderson, J.P., Rodrigo, A.G., Learns, G.H., Madan, A., Delahunty, C., Coon, M., Girard, M., Osmanov, S., Hood, L., Mullin, J.I., 2000. Testing the hypothesis of a recombinant origin of human immunodeficiency virus type i subtype. J. Virol. 74, 10752–10765.

Bader, D.A., Moret, B.M.E., Yan, M., 2001. A linear-time algorithm for computing inversion distances between signed permutations with an experimental study. J. Comput. Biol. 8, 483–491.

Bader, D.A., Moret, B.M.E., Warnow, T., Wyman, S.K., Yan, M., Tang, J., Siepel, A.C., and Caprara, A., 2002. Grappa, version 2.0. http://www.cs.unm.edu/moret/grappa. Technical report, University of New Mexico.

Blanchette, M., Bourque, G., Sankoff, D., 1997. Genome Informatics, Chapter Breakpoint Phylogenies. In: Miyano, S., Takagi, T. (Eds.). Universal Academy Press, Tokyo.

Boore, J.L., Lavrov, D.V., Brown, W.M., 1998. Gene translocation links insects and crustaceans. Nature 392, 667–668.

Cameron, S.L., Miller, K.B., D'Haese, C.A., Whiting, M.F., Barker, S.C., 2004. Mitochondrial genome data alone is not enough to unambiguously resolve the relationships of Entognatha, Insecta and Crustacea sensu lato (Arthropoda). Cladistics 20, 435–557.

Curole, J.P., Kocher, T.D., 1999. Mitogenomics: digging deeper with complete mitochondrial genomes. Trends Ecol. Evol. 14, 394–398.

Downton, M., Castro, L.R., Campbell, S.L., Bargon, S.D., Austin, A.D., 2003. Frequent mitochondrial gene rearrangements at the hymenopteran nad3nad5 junction. J. Mol. Evol. 56, 517–526.

Giribet, G., Edgecombe, S.R.G.D., Wheeler, W.C., 2005. The position of crustaceans within the Arthopoda—evidence from nine molecular loci and morphology. In: Koenemann, S., Jenner, R.A. (Eds.), Crustacea and Arthropod Relationships. Taylor and Franscis, Boca Raton, pp. 307–352.

Goloboff, P., 1999. Analyzing large data sets in reasonable times: solutions for composite optima. Cladistics 15, 415–428.

Gusfield, D., 1997. Algorithms on Strings, Trees, and Sequences. Cambridge University Press, Cambridge, MA.

Hanenhalli, S., Pevzner, P.A., 1995. Transforming a cabbage into a 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.

Harrison, R.G., Rand, D.M., Wheeler, W.C., 1985. Mitochondrial DNA heteroplasmy within individual crickets. Science 228, 1446–1448.

Hein, J.C., Jensen, J.L., Pedersen, C.N.S., 2003. Recursions for statistical multiple alignment. PNAS 100, 14960–14965.

Higgins, D.G., Sharp, P.M., 1988. Clustal: a package for performing multiple sequence alignment on a microcomputer. Gene 73, 237–244.

Inoue, J.G., Miya, M., Tsukamoto, K., Nishida, M., 2001. A mitogenomic perspective on the basal teleostean phylogeny: resolving higher-level relationships with longer DNA sequences. Mol. Phyl. Evol. 20, 275–285.

Ishiguro, N.B., Miya, M., Nishida, M., 2003. Basal euteleostean relationships: a mitogenomic perspective on the phylogenetic reality of the protacanthopterygii. Mol. Phyl. Evol. 27, 476–488.

Larget, B., Simon, D.L., Kadane, J.B., Sweet, D., 2004. A bayesian analysis of metazoan mitochondrial genome arrangements. Mol. Biol. Evol. 22, 486–495.

Macey, J., Larson, A., Fang, N.A.Z., Papenfuss, T., 1997. Two novel gene orders and the role of light-strand replication in rearrangement of the vertebrate mitochondrial genome. Mol. Biol. Evol. 14, 91–104.

Macey, J.R., Papenfuss, T.J., Kuehl, J.V., Fourcade, H.M., Boore, J.L., 2004. Phylogenetic relationships among amphisbaenian reptiles based on complete mitochondrial genomic sequences. Mol. Phyl. Evol. 33, 21–31.

Mauro, D.S., Gower, D.J., Zardoya, R., Wilkinson, M., 2006. A hotspot of gene order rearrangement by tandem duplication and random loss in the vertebrate mitochondrial genome. Mol. Biol. Evol. 23, 227–234.

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

Miya, M., Takeshimab, H., Endoc, H., Ishigurob, N.B., Inoueb, J.G., Mukaib, T., Satohb, T.P., Yamaguchib, M., Kawaguchib, A., Mabuchib, K., Shiraid, S.M., Nishida, M., 2003. Major patterns of higher teleostean phylogenies: a new perspective based on 100 complete mitochondrial dna sequences. Mol. Phyl. Evol. 26, 121–138.

Moret, B.M.E., Wyman, S., Bader, D.A., Warnow, T., Yan, M., 2001. A new implementation and detailed study of breakpoint analysis. In: Proc. 6th Pacific Symp. On Biocomputing (PSB 2001), World Scientific Pub, Hawaii, pp. 583–594.

Moret, B.M.E., Siepel, A.C., Tang, J., Liu, T., 2002a. Inversion medians outperform breakpoint medians in phylogeny reconstruction from gene-order data. In: Lecture Notes in Computer Science 2452.

Moret, B.M.E., Tang, J., Wang, L.S., Warnow, T., 2002b. Steps toward accurate reconstruction of phylogenies from gene-order data. J. Comput. Syst. Sci., special issue on computational biology.

Nardi, F., Spinsanti, G., Boore, J.L., Carapelli, A., Dallai, R., Frati, F., 2003. Hexapod origins: monophyletic or paraphyletic? Science 299, 1887–1889.

Pe'er, I., Shamir, R., 1998. The median problems for breakpoints are np-complete. In: Elect. Colloq. on Comput. Complexity 71.

Rawlings, T.A., Collins, T.M., Bieler, R., 2003. Changing identities: tRNA duplication and remolding within animal mitochondrial genomes. PNAS 100, 15700–15705.

Sankoff, D.M., Cedergren, R.J., 1983. Simultaneous comparison of three or more sequences related by a tree. In: Sankoff, D.M., Kruskall, J.B. (Eds.), Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison. Addison Wesley, Reading, Massachesetts, pp. 253–263, chapter 9.

Sankoff, D., Sundaram, G., Kececioglu, J., 1996. Steiner points in the space of genome rearrangements. Int. J. Found. Comp. Sci. 7, 1–9.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673–4680.

Thorne, J.L., Kishino, H., Felsenstein, J., 1991. An evolutionary model for maximum likelihood alignment of dna sequences. J. Mol. Evol. 33, 114–124.

Wagner, W.H., 1961. Problems in the classification of ferns. In: Recent Advances in Botany. University of Toronto Press, Toronto, Ont., pp. 841–844.

Wang, L., Jiang, T., 1994. On the complexity of multiple sequence alignment. J. Comp. Biol. 1, 337–348.

Wheeler, W.C., 1996. Optimization alignment: the end of multiple sequence alignment in phylogenetics? Cladistics 12, 1–9.

Wheeler, W.C., 1999a. Fixed character states and the optimization of molecular sequence data. Cladistics 15, 379–385.

Wheeler, W.C., 1999b. Measuring topological congruence by extending character techniques. Cladistics 15, 131–135.

Wheeler, W.C., 2001. Homology and the optimization of DNA sequence data. Cladistics 17, S3–S11.

Wheeler, W.C., 2003a. Implied alignment. Cladistics 19, 261–268.

Wheeler, W.C., 2003b. Iterative pass optimization. Cladistics 19, 254–260.

Wheeler, W.C., 2003c. Search-based character optimization. Cladistics 19, 348–355.

Wheeler, W.C., 2005. Alignment, dynamic homology, and optimization. In: Albert, V. (Ed.), Parsimony, Phylogeny, and Genomics. Oxford University Press, Oxford, pp. 73–80.

Wheeler, W.C., Gladstein, D.S., 1991–1998. MALIGN: program and documentation available at http://research.amnh.org/scicomp/projects/malign.php. New York, NY. documentation by Daniel Janies and W.C. Wheeler.

Wheeler, W.C., Gladstein, D.S., 1994. Malign: a multiple sequence alignment program. J. Hered. 85, 417–418.

Wheeler, W.C., Gladstein, D.S., De Laet, J., 1996–2005. POY, 3.0.11 edition. ftp.amnh.org/pub/molecular/poy (current version 3.0.11). Documentation by D. Janies and W. Wheeler. Commandline documentation by J. De Laet and W.C. Wheeler.

Wheeler, W.C., Aagesen, L., Arango, C.P., Faivovich, J., Grant, T., D'Haese, C.A., Janies, D., Smith, W.L., Varon, A., Giribet, G., 2005. Dynamic Homology and Phylogenetic Systematics: a unified approach using POY. American Museum of Natural History.

Zardoya, R., Meyer, A., 1998. Complete mitochondrial genome suggests diapsid affinities of turtles. PNAS 95, 14226–14231.