Alignments of Mitochondrial Genome Arrangements: Applications to Metazoan Phylogeny

Guido Fritzsch [b,c], Martin Schlegel [c,b] and Peter F. Stadler [a,b]

[a] *Bioinformatics Group, Department of Computer Science, Leipzig,*
Kreuzstraße 7b, D-04103 Leipzig, Germany
studla@bioinf.uni-leipzig.de

[b] *Interdisciplinary Center for Bioinformatics, University of Leipzig,*
Kreuzstraße 7b, D-04103 Leipzig, Germany
fritzsch@izbi.uni-leipzig.de

[c] *Molecular Evolution and Animal Systematics, Department of Biology II,*
Universität Leipzig,
Talstrasse 33, D-04103 Leipzig, Germany
schlegel@rz.uni-leipzig.de

Abstract

Mitochondrial genomes provide a valuable dataset for phylogenetic studies in particular of metazoan phylogeny because of the extensive taxon sample that is available. Beyond the traditional sequence based analysis it is possible to extract phylogenetic information from the gene order. Here we present a novel approach utilizing these data based on cyclic list alignments of the gene orders. A progressive alignment approach is used to combined pairwise list alignments into a multiple alignment of gene orders. Parsimony methods are used to reconstruct phylogenetic trees, ancestral gene orders, and consensus patterns in a straightforward approach. We apply this method to study the phylogeny of protostomes based exclusively on mitochondrial genome arrangements.

Key words: mitochondrial genome, gene order, cyclic alignment, protostome phylogeny, ancentral states.

1 Introduction

One of the prime challenges in metazoan evolution is the reconstruction of the so called Cambrian explosion. The abundance in fossil record dramatically increases in the Cambrian some 530 million years ago. Arthropods, annelids, molluscs, brachiopods, echinoderms, and chordates, i.e. representatives of most of the recent animal phyla appeared (and a considerable number of extinct forms). To date it remains a problem to resolve the order of ramification within these extant major lineages and thus, to understand the evolution of the various metazoan body plans, developmental processes, genome and proteome organisation as well as the phylogenetic position of important model organisms, such as Caenorhabdites, Drosophila, and Danio rerio (Zebrafish). Although besides comparative morphology and anatomy molecular data are increasingly applied in reconstruction of phylogenetic relationships, a consistent view of Cambrian phylogeny did not yet emerge.

The overwhelming part of the literature in molecular phylogenetics is based upon the analysis of nucleic acid and/or amino acid sequences of individual genes or groups of genes. With the advent of completely sequenced genomes these approaches are complemented by genome-wide comparisons of gene-contents (Fitz-Gibbon and House, 1999; Snel et al., 1999), gene orders (Boore and Brown, 1998; Coenye and Vandamme, 2003), or composition measures (Qi et al., 2004).

Mitochondrial genomes have been a particularly fruitful data set for phylogenetic reconstructions based on gene order, since the seminal papers of Watterson et al. (1982) and Sankoff et al. (1992). For a review we refer to Boore and Brown (1998). The mitochondrial genome has advantages over other molecular phylogenetic markers, in particular for the reconstruction of deep metazoan phylogenies. The gene content is almost invariant in metazoan mitochondria so that a consistent data set encompassing hundreds of taxa from all major lineages has accumulated. Stable structural rearrangements are rare events due to the fact that functional genomes have to be maintained. Back-mutations that restore a genome rearrangement are unlikely ($p = 1/N^2$, where N is the number of mitochondrial genes), hence there is little danger of homoplastic events hiding the true phylogeny. Furthermore, it is extremely unlikely that two or more taxa will independently converge to the same gene order.

Two broad classes of approaches have been used extensively: *break point distances* (Sankoff and Blanchette, 1998; Blanchette et al., 1999; Sankoff and Blanchette, 1999), as implemented e.g. in `GRAPPA` (Moret et al., 2002) can be computed efficiently and are used as input to standard distance methods such as neighbor-joining (Saitou and Nei, 1987). A Bayesian method of this type is described by Larget and Simon (2002). The second approach models the rearrangements by means of edit operations such as reversals and transpositions (Hannenhalli et al., 1995; Hannenhalli and Pevzner, 1999; Bourque and Pevzner, 2002; Bergeron et al., 2002). This approach can be used to infer ancestral gene-orders (Bourque and Pevzner, 2002), while break-point phylogenies may be used to identify ancestral blocks of genes (Blanchette et al., 1999).

Almost all the work cited above represents the mitochondrial gene order as signed circular permutations. As a consequence, deletion and duplications, which are relatively frequent for mitochondrial tRNAs (see e.g. Higgs et al. (2003) and the references therein), are hard to deal with. Furthermore, it is not clear how the different mobilities of protein-coding genes and tRNAs can be taken into account. Another practical problem is the lack of an intuitive way of representing the relationships among a larger number of mitochondrial gene orders. The visualization problem becomes more pressing with the rapid increase in sequenced mitochondrial genomes, more than 500 at the time of writing; for a specialized database see Jameson et al. (2003).

In this contribution we explore a different approach. Our main goal here is not the immediate reconstruction of a large scale phylogeny but rather the identification of consensus patterns in the gene orders that are characteristic for (most members) of a large metazoan clade — and to identify clades in which rapid rearrangements have wiped out any trace of an ancestral arrangement. The idea is to represent the mitochondrial gene arrangements as cyclically ordered lists. Just as strings, lists can be aligned to highlight common sublists, which, as we will see, allows a straightforward representation of similarities. A list alignment algorithm simply treats each list-entry as a letter in the alphabet over which the alignment is performed. Thus alignments of lists can be computed with standard sequence alignment algorithms provided suitable cost functions for matches, mismatches, insertions, and deletions can be meaningfully defined. The case of linearly ordered lists was investigated in the context of the *footprint sorting problem* by Fried et al. (2004). The case of mitochondrial gene arrangements is complicated by two facts: (i) the genes (list entries) have an orientation depending on their directionality on the the mitochondrial genome, and (ii) the mitochondrial genome is circular, hence we need to adapt algorithms for the alignment of cyclic sequences rather than ordinary sequence alignment algorithms.

Alignments of mitochondrial gene orders can readily be used to reconstruct phylogenetic trees as well as ancestral gene orders using standard parsimony approaches. We apply the alignment-based approach to a set of more than 100 mitochondrial genomes covering all major protostome phyla. Mitochondrial gene orders identify e.g. arthropods, annelids and platyhelminthes as monophyletic groups. A larger dataset covering all major metazoan clades confirms the separation of protostomes and deuterostomes. These results demonstrate that mitochondrial gene orders contain phylogenetic information that is potentially suitable to resolve at least partially the Cambrian radiation of the major metazoan groups.

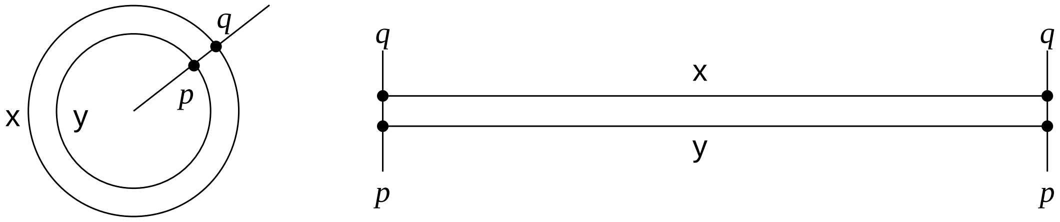

Fig. 1. An alignment of two cyclic strings that contains the (mis)match x_p, y_q is equivalent to a linear alignment of $\sigma^q(x)$ with $\sigma^p(y)$ with the constraint that x_p, y_q form a (mis)match. Note that — in constrast to the default of many alignment programs — we have to score "end-gaps" just as all other indels here.

2 Methods

In this section we will first introduce the cyclic sequence alignment problem and describe a polynomial solution for arbitrary cost functions. Then we discuss how the problem of directionality and the occurrance of duplicate mitochondrial genes is dealt with. In the final subsection we consider various possibilities of extracting consensus gene arrangments from cyclic alignments.

2.1 Cyclic Alignments

An alignment $\mathbb{A}$ of two strings x and y is a sequence of pairs of the form (x_i, y_j), $(x_i, -)$, and $(-, y_j)$ that preserves the order of sequence positions in both x and y. A pair (x_i, y_j) corresponds to a *substitution* of x_i by y_j, a pair $(x_i, -)$ represents the *deletion* of x_i, and $(-, y_j)$ is the *insertion* of y_j. A maximal subsequence consisting of deletions $(x_i, -), (x_{i+1}, -), \dots, (x_{i+q-1}, -)$ will be referred to as the deletion of the substring $x[i, i+q-1]$ of length q, and analogously for insertions. We consider here a cyclic variant $\tilde{\mathbb{A}}$ of the alignment in which we allow insertions and deletions of substrings to "wrap around" the ends of the alignments, so that e.g. $(x_1, -)$ and $(x_n, -)$ are part of the same deleted substring.

With each alignment we associate a cost function. We distinguish substitution costs $s(a, b)$ between two letters a and b and costs of insertions and deletions $g(\mathbf{a})$ for a substring $\mathbf{a}$. The cost function $g(\mathbf{a})$ is called the *gap cost function.* The total cost of an alignment is the sum of costs of the individual costs for each edit operation. We call the cost model *additive* if the gap cost functions are additive

$$g(\mathbf{a}) = \sum_{a_i \in \mathbf{a}} g(a_i) \tag{1}$$

Note that for additive gap costs the cost $f(\mathbb{A})$ of an alignment $\mathbb{A}$ and the cost $f(\tilde{\mathbb{A}})$ of its cyclic variant $\tilde{\mathbb{A}}$ is the same. Gap costs must be sub-additive, $g(\mathbf{a} \cup \mathbf{b}) \leq g(\mathbf{a}) + g(\mathbf{b})$. It follows that we have in general $f(\mathbb{A}) \geq f(\tilde{\mathbb{A}})$ since a "wrap-around" gap is cheaper than two separate end-gaps.

In the context of cyclic alignments one naturally considers the strings themselves as cyclic (Bunke and Bühler, 1993; Gregor and Thomason, 1993; Maes, 1990; Mollineda et al., 2002). Formally, cyclic strings are usually introduced as equivalence classes w.r.t. the cyclic shift operator σ that rotates a string by one position: $\sigma(x) = (x_2, \ldots, x_{n-1}, x_n, x_1)$. The cyclic string associated with an ordinary string x is thus the equivalence class $[x] = \{x, \sigma(x), \sigma^2(x), \ldots, \sigma^{n-1}(x)\}$. An alignment of two cyclic strings is simply a cyclic alignment of two representatives $\sigma^k(x)$ and $\sigma^l(y)$ of $[x]$ and $[y]$. Of course we are interested in those representatives that yield the optimal alignment, i.e., that minimize

$$f\left(\tilde{\mathbb{A}}([x],[y])\right) = \min_{k,l} f\left(\tilde{\mathbb{A}}(\sigma^k(x), \sigma^l(y))\right) \tag{2}$$

where $\tilde{\mathbb{A}}(p,q)$ denotes the cost-optimal cyclic alignment of the (non-cyclic) strings p and q. This problem can be solved in $\mathcal{O}(|x||y|\log(|x|+|y|))$ time and quadratic space in the case of additive cost functions (Maes, 1990; Gregor and Thomason, 1993), see also Landau et al. (1998). Unfortunately, this approach does not generalize to the problem at hand, which requires us to consider arbitrary cost functions.

A solution to the general problem can still be obtained with quadratic memory and in polynomial time. First we note that the optimal circular alignment of $[x]$ and $[y]$ is either the trival alignment with cost $g(x) + g(y)$ in which $[x]$ is deleted and $[y]$ is inserted (this is cheaper than deleting $[x]$ and inserting $[y]$ in multiple intervals because of the subadditivity of the gap cost function), or the optimal alignment contains at least one pair of match positions, say x_p and y_q. The cost $f(\tilde{\mathbb{A}}_{pq})$ of this alignment given by

$$f(\tilde{\mathbb{A}}_{pq}) = s(x_p, y_q) + f(\mathbb{A}(\sigma^p(x)[2..|x|], \sigma^q(y)[2..|y|]) \,. \tag{3}$$

To see this recall that the first position of $\sigma^p(x)$ is x_p so that we consider a non-cyclic alignment which in the very first position has the match (x_p, y_q) followed by the optimal alignment of the remainder of rotated string $\sigma^p(x)$ and $\sigma^q(y)$. Since the first position is a match, a possible end-gap cannot wrap around so that we have to consider an optimal non-cyclic alignment of the substrings $\sigma^p(x)[2..|x|]$ and $\sigma^q(y)[2..|y|]$, see Fig. 1. Each one of them can be computed in $\mathcal{O}(|x| \cdot |y| \cdot \max(|x|,|y|))$ operations with arbitrary cost functions, see e.g. Dewey (2001), or in $\mathcal{O}(|x| \cdot |y|)$ operations for affine gap cost functions (Gotoh, 1982). Thus we can compute the optimal cyclic alignment by computing $\tilde{\mathbb{A}}_{pq}$ for all pairs (p,q).

```
>NC_004419 Polyodon spathula
F 12S V 16S L2 ND1 I -Q M ND2 W -A -N -C -Y CO1 -S2 D CO2 K ATP8 ATP6 CO3 G ND3 R ND4L ND4 H S1 L1 ND5 -ND6 -E CYTB T -P
>NC_002639 Myxine glutinosa
F 12S V 16S L2 ND1 I -Q M ND2 W -A -N -C -Y CO1 -S2 D CO2 K ATP8 ATP6 CO3 G ND3 R ND4L ND4 H S1 L1 ND5 -ND6 -E CYTB T -P
>NC_001626 Petromyzon marinus
F 12S V 16S L2 ND1 I -Q M ND2 W -A -N -C -Y CO1 -S2 D CO2 K ATP8 ATP6 CO3 G ND3 R ND4L ND4 H S1 L1 ND5 -ND6 T -E CYTB -P
>NC_002177 Halocynthia roretzi
F G T ND6 L1 N G D CO3 ND4L C K 12S CO2 CYTB Y W I E ND2 H S1 R Q L2 ND5 M 16S ND1 ATP6 S2 CO1 ND3 A P ND4 V
```

Fig. 2. Example of input data. The abbreviations of mitochondrial gene names are listed in the supplemental material.

2.2 Encoding of Mitochondrial Genomes

Mitochondrial genomes clearly can be regarded as cyclically ordered lists of genes. In addition, however, the genes are oriented depending on whether they are located on the heavy or the light strand. This is taken into account in the framework of list alignments by considering the same gene in different orientations as different objects, Fig. 2.

Duplication and deletion of genes in mitochondrial genomes occured frequently during the evolution of the Metazoa. Recently, Paul Higgs and co-workers presented compelling evidence that duplication of a tRNA-Leu gene, followed by anticodon mutation, and subsequent deletion of tRNA-Leu genes has occurred at least five times during the evolution of the Metazoa (Higgs et al., 2003). Animal mitochondrial genomes, for example, usually have two transfer RNAs for both leucine and serine. While such duplicate genes are problematic in permutation-based approaches, they are naturally described in the list alignment model used here: There are simply two tRNA-Leu genes in the cyclic list. We simply use the same symbol, 'L', for both tRNA-Leu genes.

2.3 Scoring Model

Let us now turn to a more detailed description of the scoring functions that underlie the pairwise linear alignments. The (mis)match scores are trivial because it makes no sense to align non-homologous genes, i.e., non-identical list entries. We have simply $\sigma(x, y) = 0$ if $x = y$ and $\sigma(x, y) = \infty$ if $x \neq y$. Our knowledge about the mechanism of the genomic rearrangments must therefore be incorporated into the indel scores. Thus we have to compute only $\mathcal{O}(D \max(|X|, |Y|))$ pairwise linear alignments, where D is the maximum number of copies of a duplicated gene.

It is well known that tRNA genes are much more mobile than protein-coding mitochondrial genes (Boore, 1999). We therefore propose a scoring scheme that consists of two contributions for each inserted or deleted interval $\mathbf{a} = [a_i, a_{i+1}, \ldots, a_{j-1}, a_j]$ of the cyclic list. We define (1) an additive score contri-

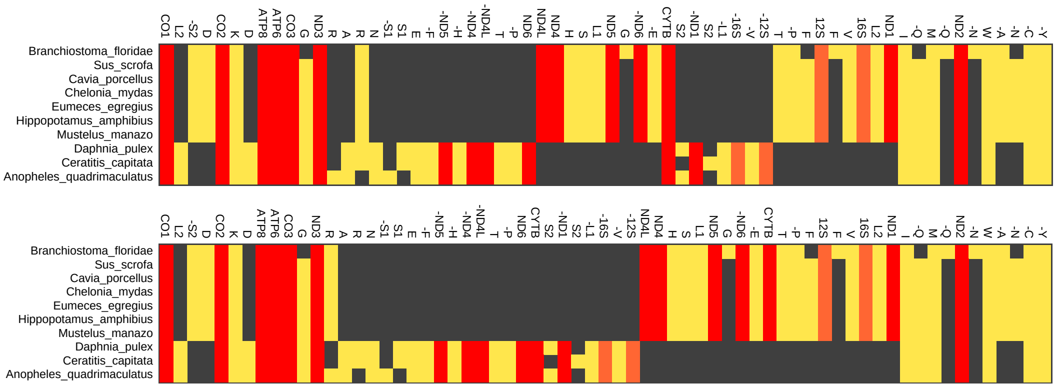

Fig. 3. Effect of scores: Top: additive model with $\delta(\mathsf{P}) = 20$, $\delta(\mathsf{r}) = 10$, $\delta(\mathsf{t}) = 3$ and $\eta(a) = 0$. Below: affine model with additive scores as above and $\eta(a_i) = 2\delta(a_i)$. As expected, the large one-time score, which essentially acts like a *gap-open penalty*, leads to alignments with a small number of large gaps.

bution $\delta(a_i)$ to which each deleted list entry a_i contributes independently and (2) a "one-time" contribution that allows us to distinguish between intervals that consist of tRNAs only and those that also contain proteins. This "one-time" score essentially plays the role of the gap-open penality in the usual models of sequence alignments, Fig. 3. We found it convenient to define the one-time score $\eta(a_i)$ for each list entry individually and to compute indel-score for the interval $\mathbf{a}$ as

$$g(\mathbf{a}) = \max_{i \leq k \leq j} \eta(a_k) + \sum_{k=i}^{j} \delta(a_k) \tag{4}$$

The default scores for mitochondrial genomes distiguish between three types of genes: proteins P, ribosomal RNA genes r, and tRNAs t. The downside of this scoring model is that we are forced to use a computationally expensive algorithm to compute the linear list alignments. Default values for this scoring model have been chosen so that a number of test datasets with well-established phylogenies were well reconstructed from the resulting alignments. In the following we use

$$\begin{array}{|c|ccc|} \hline & \mathrm{P} & \mathrm{r} & \mathrm{t} \\ \hline \delta & 4 & 3 & 3 \\ \eta & 6 & 4 & 4 \\ \hline \end{array} \tag{5}$$

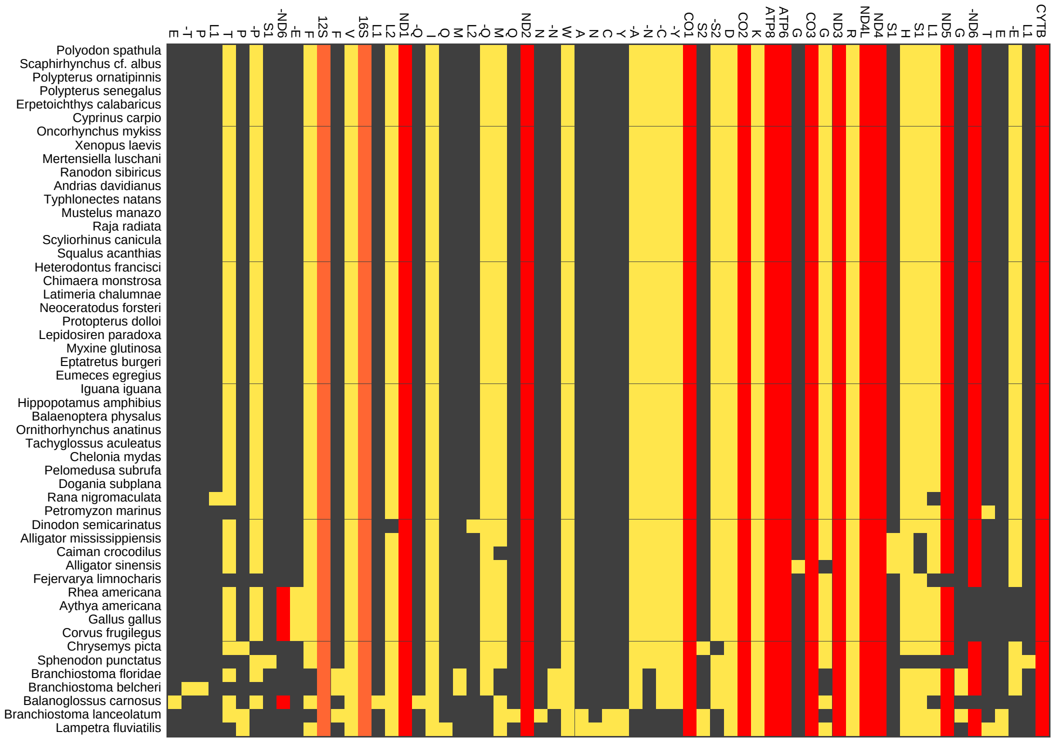

Fig. 4. Graphical display of an alignment of vertebrate mitocondrial genome arrangments. Proteins, rRNAs, and tRNAs are shown in red, orange, and yellow, resp. Most vertebrates share a common gene order, there are numerous small deviations that mostly involving transposed tRNA genes, however, in particular in the bird and reptile lineages, see e.g. Mindell et al. (1998) and Townsend and Larson (2002).

2.4 Multiple Cyclic Alignments

The pairwise cyclic alignment procedure outlined in the previous two subsection can be generalized to multiple sequences by means of the same *progressive alignment* approach that is used e.g. in `clustalw` (Thompson et al., 1994): We first compute all pairwise distances using the cyclic alignment procedure. From the resulting distance matrix we construct a guide tree, in our case using the WPGMA clustering method (Sokal and Michener, 1958). This guide tree is used to align profiles of aligned cyclic lists in the same way as individual lists. To this end, the scoring scheme is extended in the obvious way from individual lists to alignments: both the one-time score η and the additive score δ for a profile position is computed as the sum over all entries in each column of the two profiles (alignments) that are to be combined. Eq.(4) is then used to determine the indel/score for an interval.

2.5 Implementation

The algorithm described here is implemented in the program package `circal` which is written in `ANSI C`. The `circal` package is distributed under the GNU General Public License (GPL)[1]. The current implementation of `circal` produces a `nexus` format file of the alignment as well as a graphical overview in `PostScript` format, Fig. 4. Datasets of about 30 rather diverse gene arrangements can be computed within about a quarter of an hour on a PC (Linux operating system on a Dual-Pentium IV with two 2.4GHz CPUs and 1Gbyte RAM). The full protostome data set runs about 1 day on the same PC.

2.6 Tree Reconstruction

Phylogenetic trees can be inferred from the cyclic list alignments by any one of the usual approaches. One might simply use the multiple alignment to re-compute a pairwise distance matrix, possibly using a more sophisticated distance measure than those used for constructing the alignment. We do not pursue this approach here since the alignment contains much more information than just the mutual distances. Maximum likelihood methods are applicable in principle (Larget and Simon, 2002), albeit it seems non-trivial to derive a good rate model for mitochondrial genome arrangements. We therefore resort to maximum parsimony. Since each column of the alignment marks only the presence or absence of a gene in a particular alignment position, it is straightforward to apply standard programs such as `PAUP` (Swofford and Olsen, 1990) or `phylip` (Felsenstein, 1989) on the corresponding `0/1`-strings. Alternatively, the position of a gene in the list alignment can be interpreted as a distinct character state as described in subsection 2.8. This results in a string representation in which each mitochondrial gene corresponds to a single column. In practice we observe little difference between the two approaches. Bootstrap support values can of course be computed in the same way as for conventional sequence alignments.

2.7 Consensus Gene Arrangements

An apparent shortcoming of the list-alignment approach is that the same object (gene) may appear multiple times in the alignment, i.e., there are multiple columns of the final alignment that refer to the same protein or tRNA. We argue that this is actually an advantage. We can now identify a subgroup of

[1] Download from `http://www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/04-015/`.

```
> ancestral_nematode
+ND5 +E +S1 +ND2 +T +ND4 +CO1 +ND6 +R +Q +CYTB +L1 +CO3 +L2 +ND1 +I +G +CO2 +H +16S +ND3
```

Fig. 5. Reconstructed gene order of the ancestral nematode using 9 nematode genomes and 2 annelide genomes as outgroup. Only genes with unambiguously reconstructed positions are shown. The positions of the two proteins ND4L and ATP6, of the 12S RNA, and of most tRNAs (A, C, D, F, K, M, N, P, S2, V, W, and Y) remains ambiguous.

aligned genome arrangements (or use the whole alignment) to obtain a *consensus genome arrangement*. To this end, we simply compare all columns that refer to the same gene (in both orientations) and select the one with the most non-gap entries. In the case of duplicated tRNAs, say, we may take two most populated columns. The result is a "valid" mitochondrial genome arrangment that describes the consensus of the group in question.

By leaving out all columns that contain less than a minimum number of non-gap entries we can directly extract conserved parts of the gene order even if they do not correspond to conserved intervals.

2.8 Ancestral Genome Organization

It might be surprising at first glance that the multiple alignment can be used to reconstruct the ancestral genome organization since each gene k can appear in multiple columns of the list alignment. Suppose the number of these columns is m_k. Clearly, gene k is present in at most one of these columns in each taxon. A deletion of k is represented by the absence of k in all m_k columns belonging to gene k. Thus we can regard the position of k in the list alignment as a character with $m_k + 1$ possible states. Consequently, we can use standard parsimony approaches to obtain the ancestral state (position and orientation) of each gene, Fig. 5. On the other hand, if we allow duplicate genes, or, assuming a duplication-deletion mechanism for mitochonodrial genome rearrangements such as those proposed by Macey et al. (1998) and Lavrov et al. (2002), we may use the simple presence/absence patterns of genes in the list alignment itself to reconstruct ancenstral gene orders by means of maximum parsimony.

3 Metazoan Phylogeny from Mitochondrial Genomes

Analyses of mitochondrial genomes have significantly contributed to the reconstruction of metazoan deep phylogeny (Boore and Brown, 1998). For example, the phylogenetic position of Tentaculata (Lophophorata), either as pro-

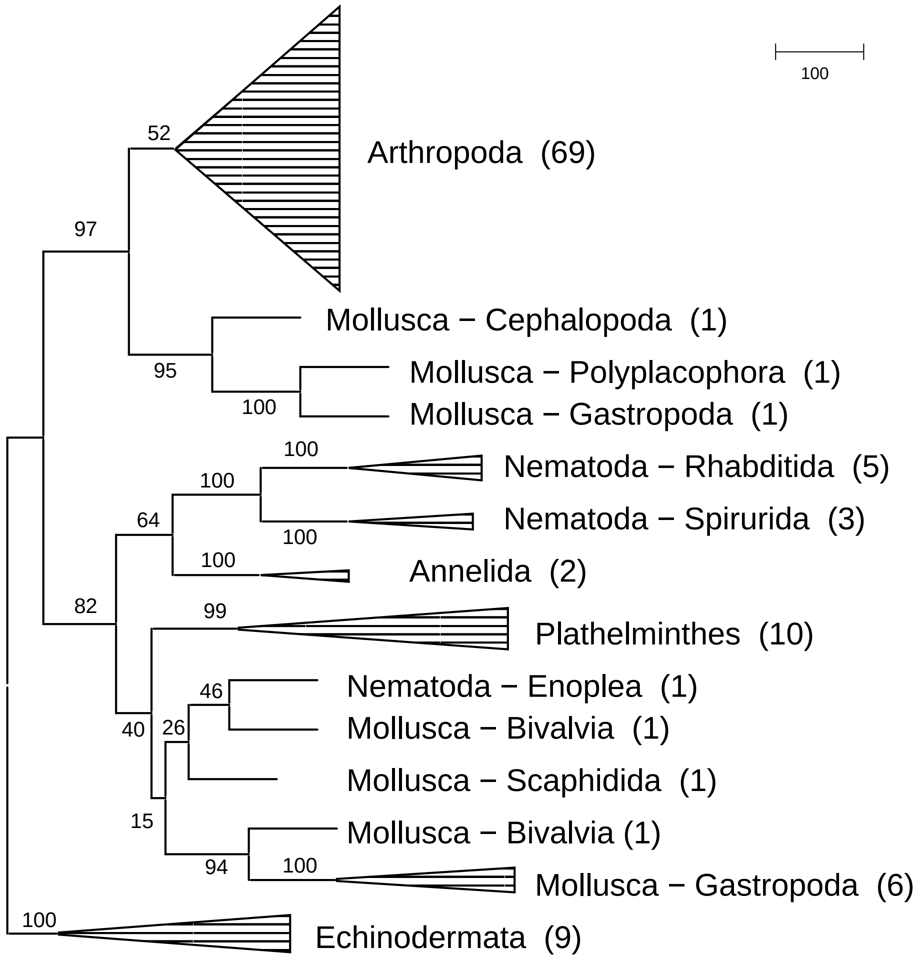

Fig. 6. Maximum parsimony tree of the mitochondrial gene order of the protostomian dataset. MP analysis was performed using PAUP with the heuristic search method (10 random stepwise additions and the TBR branch swapping, 100 bootstrap replicates). Triangles represent subtrees which are not shown in full resolution for clarity. Echinodermata are used as outgroup. Data and accession numbers are provided in the supplemental material.

tostomes, sister group deuterostomes, or even members of the deuterostomes was a matter of long and controversial debate. Mitochondrial genome analyses, both on gene order and nucleotide analyses of the brachiopod *Terebratulina retusa* convincingly support a protostome relationship with an affiliation to the spiral cleaving molluscs and annelids (Stechmann and Schlegel, 1999). However, some results on the analysis of mitochondrial genome comparisons challenge classical evidence, such as the monophyly of insects (Nardi et al., 2003).

We applied our novel method to a comprehensive dataset of metazoan mitochondrial genomes. Phylogenetic reconstructions reported here are performed by aligning the mitochondrial gene orders (proteins, rRNAs, and tRNAs) using `circal` and reconstructing maximum parsimony trees using `PAUP*`, version 4.0b4a (Swofford, 2002). Gaps were treated as missing data. Data were also bootstrapped using the MP method (100 replicates).

An analysis using 166 metazoan taxa showed a good separation between the deuterostomian and the protostomian lineages (data not shown). In Fig. 6 we show a phylogenetic reconstruction of 103 taxa from diverse protostome groups using 9 echinoderms as outgroup. Our approach supports the monophyly of arthropods, annelids, platyhelminthes and nematods, with the exception of a single taxon (Enoplea). The genome arrangements of molluscs are very variable (Hoffmann et al., 1992; Dreyer and Steiner, 2004). This a clearly a case where the ancestral gene order has been wiped out by rapid rearrangements. A multifurction within the arthropode clade (see online supplement) makes further systematic analyses impossible. Nevertheless some tendencies can be recognized: for example, we we find support for a clade of chelicerata (10 taxa) and myriapoda (4 taxa).

An ongoing debate in protostome phylogeny concerns the affiliation of the arthropods either to the annelids in the traditional articulates, or with the nematodes in the a clade Ecdysozoa (Adoutte et al., 2000). Our results do not support the latter hypothesis based on other molecular evidence, such as rRNA and *Hox* gene sequences.

We also analyzed 60 complete mitochondrial genomes of vertebrates, hemichordates, and cephalochordates. The vertebrate gene order underwent only very few rearrangments, see Fig. 4. Thus phylogenetic information cannot be recovered with the data set analyzed. It is noteworthy, however, that within the chordates, hemichordates share the only rearragement of a mitochondrial protein with the birds. Clearly, these are independent rearrangement events. This example suggests that changes in mitochondrial are not unbiased random events; multiple list alignments can be used to determine likelihood differences between rearrangements from sufficiently large datasets.

While testing our programs we observed that the resolution of the reconstructed trees and their agreement with well-established phylogenetic hypotheses improves with increasing taxon sampling. The reason is probably that a dense taxon coverage leads to smaller differences between the gene orders of adjacent taxa which improves quality of the multiple sequence alignment because the underlying pairwise alignments become less error prone. This contrasts the observation of Rosenberg and Kumar (2001) that increased taxon sampling does not lead to substantial improvements in sequence-based methods.

4 Discussion

We have described here a novel approach to reconstruct the evolution of mitochondrial gene orders based on multiple cyclic list alignments. The algorithm is fast as it can handle the complete available set of mitochondrial genomes on a single PC. The method produces plausible large scale phylogenies and allows to compute ancestral gene orders by means of maximum parsimony reconstruction. In addition, consensus gene orders for subgroupings can be reasily determined that are believed to be monophyletic based on external evidence. In contrast to most alternative approaches to the comparative analysis of gene orders the cyclic list alignment approach works well on datasets with different gene content. It could thus be readily applied to datasets that contain both metazoa and close unicellular relatives of animals such as choanoflagellate and ichthyosporean protists, whose mitochondria contain about twice as many proteins as those of higher metazoans (Burger et al., 2003). At present, the availability of such data is very limited, a situation that will change in the near future. Another natural field of applications are plastid genomes (Doyle et al., 1992; Odintsova and Yurina, 2003), for which a genome database has recently become available (Kurihara and Kunisawa, 2004).

The application of the list alignment approach to metazoan phylogeny demonstrates that the seemingly small set of only 37 mitochondrial genes can be used to recover e.g. the separation of protostomia and deuterostomia, and — within the protostome clade — to confirm arthropods, annelids, platyhelminthes and nematods (with a single exeception) as monophyletic groups.

The current implementation of a circular alignment method uses a straightforward generalization of the already expensive linear alignment algorithm for arbitray gap cost functions. More efficient algorithms can be devised and will be implemented to cope with the rapidly increasing amount of data. The model underlying the list-alignment used here does not distinguish between the heavy-strand and the light-strand of the mitochondrial genome. It is well known that nucleic acid substitution patterns of the two strands are not symmetric (Faith and Pollock, 2003). A modification of the alignment procedure that treats the two strands differently is conceivable; it is unclear at this point, however, whether the bias between heavy and light strand has a significant impact on the viable gene orders and hence affects phylogeny reconstructions.

Acknowledgements

Discussions with Dr. Detlef Bernhard and Dr. Axel Mosig are gratefully acknowledged. This work is supported by DFG Bioinformatics Initiative, BIZ-

6/1-2.

Supplemental Material

Supplemental material, including the C source code implementing the cyclic list alignment algorithm described here, is available at
`http://www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/04-015`.

References

Adoutte, A., N. Balavoine, G.and Lartillot, B. Lespinet, O.and Prud'homme, and R. de Rosa, 2000. The new animal phylogeny: reliability and implications. Proc. Natl. Acad. Sci. USA **97**:4453–4456.

Bergeron, A., S. Heber, and J. Stoye, 2002. Common intervals and sorting by reversals: A marriage of necessity. Bioinformatics **18**:S54–S63.

Blanchette, M., T. Kunisawa, and D. Sankoff, 1999. Gene order breakpoint evidence in animal mitochondrial phylogeny. J. Mol. Evol. **49**:193–203.

Boore, J. L., 1999. Animal mitochondrial genomes. Nucl. Acids Res. **27**:1767–1780.

Boore, J. L. and W. M. Brown, 1998. Big trees from little genomes: mitochondrial gene order as a phylogenetic tool. Curr. Opinion Gen. Devel. **8**:668–674.

Bourque, G. and P. A. Pevzner, 2002. Genome-scale evolution: Reconstructing gene orders in ancestral species. Genome Res. **12**:26–36.

Bunke, H. and U. Bühler, 1993. Applications of approximate string matching to 2D shape recognition. Patt. Recogn. **26**:1797–1812.

Burger, G., L. Forget, Y. Zhu, M. W. Gray, and F. B. Lang, 2003. Unique mitochondrial genome architecture in unicellular relatives of animals. Proc. Natl. Acad. Sci. USA **100**:892–897.

Coenye, T. and P. Vandamme, 2003. Extracting phylogenetic information from whole-genome sequencing projects: the lactic acid bacteria as a test case. Microbiology **149**:3507–3517.

Dewey, T. G., 2001. A sequence alignment algorithm with an arbitrary gap penalty function. J. Comp. Biol. **8**:177–190.

Doyle, J. J., J. I. Davis, R. J. Soreng, D. Garvin, and M. J. Anderson, 1992. Chloroplast DNA inversions and the origin of the grass family (poaceae). Proc. Natl. Acad. Sci. USA **89**:7722–7726.

Dreyer, H. and G. Steiner, 2004. The complete sequence and gene organization of the mitochondrial genome of the gadilid scaphopod *Siphonondentalium lobatum* (Mollusca). Mol. Phylog. Evol. **31**:605–617.

Faith, J. J. and D. D. Pollock, 2003. Likelihood analysis of asymmetrical mutation bias gradients in vertebrate mitochondrial genomes. Genetics **165**:735–745.

Felsenstein, J., 1989. Phylip – phylogeny inference package (version 3.2). Cladistics **5**:164–166.

Fitz-Gibbon, S. T. and C. H. House, 1999. Whole genome-based phylogenetic analysis of free-living microorganisms. Nucl. Acids Res. **27**:4218–4222.

Fried, C., W. Hordijk, S. J. Prohaska, C. R. Stadler, and P. F. Stadler, 2004. The footprint sorting problem. J. Chem. Inf. Comput. Sci. **44**:332–338.

Gotoh, O., 1982. An improved algorithm for matching biological sequences. J. Mol. Biol. **162**:705–708.

Gregor, J. and M. G. Thomason, 1993. Dynamic programming alignment of sequences representing cyclic patterns. IEEE Trans. Patt. Anal. Mach. Intell. **15**:129–135.

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

Hannenhalli, S. and P. A. Pevzner, 1999. Turning cabbage into turnip: Polynomial algorithm for sorting signed permutations by reversals. J. ACM **46**:1–27.

Higgs, P. G., D. Jameson, H. Jow, and M. Rattray, 2003. The evolution of tRNA-Leu genes in animal mitochondrial genomes. J. Mol. Evol. **57**:435–445.

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

Jameson, D., A. P. Gibson, C. Hudelot, and P. G. Higgs, 2003. OGRe: a relational database for comparative analysis of mitochondrial genomes. Nucl. Acids Res. **31**:202–206.

Kurihara, K. and T. Kunisawa, 2004. A gene order database of plastid genomes. Data Sci. J. **3**:60–79.

Landau, G. M., E. W. Myers, and J. P. Schmidt, 1998. Incremental string comparison. SIAM J. Comput. **27**:557–582.

Larget, B. and D. L. Simon, 2002. Bayesian phylogenetic inference from animal mitochondrial genome arrangements. J. Royal Statist. Soc. B **64**:681–693.

Lavrov, D. V., J. L. Boore, and W. M. Brown, 2002. Complete mtDNA sequences of two millipedes suggest a new model for mitochondrial gene rearrangements: Duplication and nonrandom loss. Mol. Biol. Evol. **19**:163–169.

Macey, J. R., J. A. Schulte II, A. Larson, and T. J. Papenfuss, 1998. Tandem duplication via light-strand synthesis may provide a precursor for mitochondrial genomic rearrangement. Mol. Biol. Evol. **15**:71–75.

Maes, M., 1990. On a cyclic string-to-string correction problem. Inform. Process. Lett. **35**:73–78.

Mindell, D. P., M. D. Sorenson, and D. E. Dimcheff, 1998. Multiple independent origins of mitochondrial gene order in birds. Proc. Natl. Acad. Sci.

USA **95**:10693–10697.

Mollineda, R. A., E. Vidal, and F. Casacuberta, 2002. Cyclic sequence alignments: approximate versus optimal techniques. Int. J. Pattern Rec. Artif. Intel. **16**:291–299.

Moret, B. M. E., J. Tang, L.-S. Wang, and T. Warnow, 2002. Steps toward accurate reconstructions of phylogenies from gene-order data. J. Comput. Syst. Sci. **65**:508–525.

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

Odintsova, M. S. and N. P. Yurina, 2003. Plastid genomes of higher plants and algae: Structure and functions. Mol. Biol. (Mosk.) **37**:649–662. Translated from *Molekulyarnaya Biologiya*, Vol. 37, No. 5, 2003, pp. 768-783.

Qi, J., B. Wang, and B.-l. Hao, 2004. Whole proteome prokaryote phylogeny without sequence alignment: a k-string composition approach. J. Mol. Evol. **58**:1–11.

Rosenberg, M. S. and S. Kumar, 2001. Incomplete taxon sampling is not a problem for phylogenetic inference. Proc. Natl. Acad. Sci. USA **11**:10751–10756.

Saitou, N. and M. Nei, 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol. Evol. **4**:406–425.

Sankoff, D. and M. Blanchette, 1998. Multiple genome rearrangement and breakpoint phylogeny. J. Comput. Biol. **5**:555–560.

Sankoff, D. and M. Blanchette, 1999. Phylogenetic invariants for genome rearrangements. J. Comput. Biol. **6**:431–435.

Sankoff, D., G. Leduc, N. Antoine, B. Paquin, B. F. Lang, and R. Cedergren, 1992. Gene order comparisons for phylogenetic inference: evolution of the mitochondrial genome. Proc. Natl. Acad. Sci. USA **89**:6575–6579.

Snel, B., P. Bork, and M. A. Huynen, 1999. Genome phylogeny based on gene content. Nature Genet. **21**:108–110.

Sokal, R. R. and C. D. Michener, 1958. A statistical method for evaluating systematic relationships. Univ. Kansas Sci. Bull. **38**:1409–1438.

Stechmann, A. and M. Schlegel, 1999. Analysis of the complete mitochondrial dna sequence of the brachiopod *Terebratulina retusa* places brachiopoda within the protostomes. Proc. R. Soc. Lond. B **266**:1–10.

Swofford, D. L., 2002. `PAUP*`: Phylogenetic Analysis Using Parsimony (*and Other Methods) Version 4.0b10. Sinauer Associates, Sunderland, MA. Handbook and Software.

Swofford, D. L. and G. J. Olsen, 1990. Phylogeny reconstruction. In D. M. Hillis and C. Moritz, eds., Molecular Systematics, pp. 411–501. Sinauer Associates, Sunderland MA.

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

Townsend, T. and A. Larson, 2002. Molecular phylogenetics and mitochondrial genomic evolution in the chamaeleonidae (reptilia, squamata). Mol. Phylog. Evol. **23**:22–36.

Watterson, G. A., W. J. Ewens, T. E. Hall, and A. Morgan, 1982. The chromosome inversion problem. J. Theor. Biol. **99**:1–7.