Bioinformatics Methods for the Comparative Analysis of Metazoan Mitochondrial Genome Sequences

Matthias Bernt[a,*], Anke Braband[b], Martin Middendorf[a], Bernhard Misof[c], Omar Rota-Stabelli[d], Peter F. Stadler[e,f,g,h,i]

[a]Parallel Computing and Complex Systems Group, Department of Computer Science, University of Leipzig, Johannisgasse 26, D-04103 Leipzig, Germany
[b]LGC Genomics GmbH, Ostendstr. 25, 12459 Berlin
[c]Molekulare Biodiversitätsforschung, Zoologisches Forschungsmuseum Alexander Koenig, Adenauerallee 160, D-53113 Bonn, Germany
[d]Department of Sustainable Agro-Ecosystems and Bioresources, Istituto Agrario di San Michele all'Adige, Via E. Mach 1, I-38010 San Michele all'Adige (TN), Italy
[e]Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, Universität Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany
[f]Max-Planck-Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany
[g]Fraunhofer Institut für Zelltherapie und Immunologie Perlickstraße 1, D-04103 Leipzig, Germany
[h]Department of Theoretical Chemistry University of Vienna, Währingerstraße 17, A-1090 Wien, Austria
[i]Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA

Abstract

In this review we provide an overview of various bioinformatics methods and tools for the analysis of metazoan mitochondrial genomes. We compare available dedicated databases and present current tools for accurate genome annotation, identification of protein coding genes, and determination of tRNA and rRNA models.We also evaluate various tools and models for phylogenetic tree inference using gene order or sequence based data. As for gene order based methods, we compare rearrangement based and gene cluster based methods for gene order rearrangement analysis. As for sequence based methods, we give special emphasis to substitution models or data treatment that reduces certain systematic biases that are typical for metazoan mitogenomes such as within genome and/or among lineage compositional heterogeneity.

Keywords: Bioinformatics, Mitochondria, Genome Rearrangements, Substitution

[*]Corresponding author at: Parallel Computing and Complex Systems Group, Department of Computer Science, University of Leipzig, Johannisgasse 26, D-04103 Leipzig, Germany. Fax: +49 341 97 32252

Email addresses: bernt@informatik.uni-leipzig.de (Matthias Bernt), phylogenetics@arcor.de (Anke Braband), middendorf@informatik.uni-leipzig.de (Martin Middendorf), b.misof.zfmk@uni-bonn.de (Bernhard Misof), omar.rota@iasma.it (Omar Rota-Stabelli), studla@bioinf.uni-leipzig.de (Peter F. Stadler)

1. Introduction

In many aspects animal mitogenomes differ substantially from sources of other genomic data (Bernt et al., 2012a, in this special issue). With a size rarely exceeding 20kb and a nearly invariant gene inventory typically comprising 22 tRNAs, 2 rRNAs, and only 13 distinctive protein-coding genes they are amenable to computational techniques that are applicable to few other genetic systems. The small size of the genomes makes it possible to achieve nearly complete and quite accurate annotations in the absence of experimental transcript information. The high variability in particular of the tRNA sequences, on the other hand, requires specialised methods for tRNA gene finding. The constancy of the mitogenome gene content suggests to view mitochondrial gene orders as signed permutations. These signs indicate the reading direction, which is determined uniquely, since the two DNA strands of the circular mitogenome are usually distinguishable by their base composition.

In response to these atypical features, a large number of algorithms, pipelines, models, and software tools have been developed specifically for mitogenomic studies. Several databases are dedicated exclusively to make the results of such computational analyses available to the community. The comparison of mitochondrial gene orders benefits from their constant gene content which allows to apply the available methods developed for the permutation model that is of limited use for genetic systems with large differences in gene content, e.g., whole genome comparisons. Clearly, many computational tasks to analyse mitogenomes do not require specifically dedicated tools. For instance, generic software tools are applicable to assemble mitogenomic sequences (Boore et al., 2005), to construct sequence alignments, and to infer phylogenetic trees from sequence alignments, although for the latter specialised substitution models are required.

Here, we review only the bioinformatics infrastructure specialised to mitogenomics. We collect URLs and references for the databases, the software tools that are available for download, and the web services that are discussed throughout this contribu-

Table 1: Bioinformatics Resources discussed in the text.

Resource	URL	References
Databases		
GenBank	www.ncbi.nlm.nih.gov/genbank/	Benson et al. (2011)
GOBASE†	gobase.bcm.umontreal.ca/	O'Brien et al. (2009)
Mamit-tRNA	mamit-trna.u-strasbg.fr/	Pütz et al. (2007)
MamMiBase	www.mammibase.lncc.br/	de Vasconcelos et al. (2005)
MetAMIGA [a]	amiga.cbmeg.unicamp.br/	Feijão et al. (2006)
MitoZOA	mi.caspur.it/mitozoa/	Lupi et al. (2010)
NCBI O.R.	www.ncbi.nlm.nih.gov/genomes/ORGANELLES/organelles.html	Wolfsberg et al. (2001)
OGRe	drake.mcmaster.ca/ogre/	Jameson et al. (2003)
RefSeq	www.ncbi.nlm.nih.gov/RefSeq/	Pruitt et al. (2007)
tRNAdb	trnadb.bioinf.uni-leipzig.de/	Jühling et al. (2009)
Web Services		
ARWEN	www.acgt.se/online.html	Laslett and Canbäck (2008)
CREx	pacosy.informatik.uni-leipzig.de/crex/	Bernt et al. (2007)
DOGMA	dogma.ccbb.utexas.edu/	Wyman et al. (2004)
FACIL↓	www.cmbi.ru.nl/FACIL/	Dutilh et al. (2011)
GENESIS	www.uni-ulm.de/in/theo/research/genesis.html	Bader and Ohlebusch (2007)
GenDecoder	darwin.uvigo.es/software/gendecoder.html	Abascal et al. (2009)
MGR	nbcr.sdsc.edu/GRIMM/mgr.cgi	Bourque and Pevzner (2002)
MITOS	mitos.bioinf.uni-leipzig.de/	Bernt et al. (2012c) (this issue)
MOSAS	mosas.byu.edu/	Sheffield et al. (2010)
roci	bibiserv.techfak.uni-bielefeld.de/roci/	Stoye and Wittler (2009)
tRNAscan-SE ↓	lowelab.ucsc.edu/tRNAscan-SE/	Lowe and Eddy (1997)
UniMoG	bibiserv.techfak.uni-bielefeld.de/dcj/	Bergeron et al. (2006)
Software for Download		
ANGES	http://paleogenomics.irmacs.sfu.ca/ANGES/	Jones et al. (2012)
BADGER	www.badger.duq.edu/	Larget et al. (2005)
circal	www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/04-015/	Fritzsch et al. (2006)
EMRAE	www.gis.a-star.edu.sg/~bourque/download/EMRAE.zip	Zhao and Bourque (2009)
GRAPPA	www.cs.unm.edu/~moret/GRAPPA/	Moret et al. (2001)
inferCARs	www.bx.psu.edu/miller_lab/car/	Ma et al. (2006)
MiTFi	www.bioinf.uni-leipzig.de/Software/MiTFi/	Jühling et al. (2012)
PATHGROUPS	albuquerque.bioinformatics.uottawa.ca/lab/software.html	Zheng and Sankoff (2011)
PHYLO	http://www.uni-ulm.de/in/theo/m/alumni/bader.html	Bader et al. (2008)

[a] previously called AMIGA. † no longer updated. ↓ also available for download and local use.

tion (Tab. 1).

2. Databases

GenBank and RefSeq are the most commonly used repositories of sequence data. Whereas GenBank provides access to original sequence data and associated annotations including taxonomic and bibliographic information, RefSeq aims to provide expert-curated and largely non-redundant information based on original GenBank entries. The NCBI Organelle Resources database provides easy access to the mitochondrial genome data of RefSeq, to information such as the arrangement of the mitochondrial proteins, and allows specialised BLAST searches. Access to all three resources is provided through NCBI's gateway, hence we will, for brevity, refer to them collectively as NCBI.

The current GenBank (release 184) contains 13,916 complete metazoan mitogenomes of which 2,355 are included in RefSeq (release 48). Since NCBI provides data that were collected by different researchers aiming for different research questions and using different sets of tools, a certain amount of inconsistency in the available annotations is unavoidable despite all curatorial efforts (Boore, 2006a). GenBank and RefSeq display very uneven taxonomic coverage. RefSeq contains mitogenomes of 1,627 Chordata, 376 Arthropoda, and 104 Mollusca. The policy of RefSeq to provide "only one example of each natural biological molecule for major organisms" (Mizrachi, 2007) is a potential problem for taxa with doubly uniparental inheritance, e.g., many Mollusca or cryptic species (Lupi et al., 2010). Several databases have been built upon the GenBank or RefSeq with the aim of improving annotation and data quality. In many cases they also provide access to additional tools for data analysis. Important "derived" databases are OGRe, MetAMIGA, and MitoZOA (Tab. 2).

An important property of a database is the consistency of the data presented to the user. Among the six databases, only MetAMIGA uses a consistent naming convention for the genes for all species. MitoZOA internally uses a list of synonymous gene names for retrieving data, but presents search results using the original, inconsistent annotation. Although OGRe originally used a consistent naming scheme, inconsistencies were introduced with a recent update. The three NCBI databases and MitoZOA provide a

Table 2: Comparison of the databases; Meaning of the symbols: ● present, ○ absent, ⊛ partially available; Upper section: usability of data. Criteria are the consistency of gene names, the availability of (easily) parsable output, consistent designations of annotation items, and coherent definitions of gene boundaries. Middle section: improvements over primary data. Error screening (consistent, documented, traceable error screening, claimed and obvious improvements), anticodon annotation (available for all tRNAs), additional information (e.g., from literature); Lower section: versatility of search end data export. Options for searching data and downloading them in common formats: species set selection per search or from a list of species, sequences of features (as FASTA), gene orders (available for the selected species); complete database available for download; [†]: only HTML table, [*]: warns for potential problems, [§]: only list of unique gene orders

		NCBI	OGRe	MetAMIGA	MitoZOA
consistent	names	○	⊛	●	⊛
	format/annotation	●/○	○/●[†]	○/●[†]	●/○
	ORF/bounds	○/○	○/○	○/○	●/⊛[*]
	error screening	○	⊛	⊛	●
	tRNA anticodons	⊛	⊛	○	⊛
	additional information	○	○	○	●
	species selection (search/list)	●/●	●/●	●/●	●/○
download	feature seq (nt/aa)	●/●	●/●	●/●	●/○
	gene orders (w/wo tRNA)	○/●	●/●[§]	○/○	●/○
	complete database	●	○	○	●

parsable format for downloading the genome annotations (GenBank and EMBL format respectively), while OGRe and MetAMIGA display search results only as HTML-tables for viewing in a web browser. Only MitoZOA provides a consistent method for defining ORFs and implements rudimentary methods for determining gene boundaries. For instance, it warns of potential errors and non-canonical start codons. All three derived databases improve to some extent the annotations derived from NCBI, but only MitoZOA provides a well defined, semi-automatic procedure for constructing the enhanced database entries. Whereas in MitoZOA the anticodons of tRNAs are annotated for duplicated tRNAs, in OGRe this information is stored for all Leucine and Serine tRNAs, and MetAMIGA does not contain this information. Graphical representations of the mitogenome annotation are available from all databases except MitoZOA.

The databases do not only differ in the level of improvements and the amount of extensions of the annotation, but also in the versatility of their user interfaces. Beyond the ubiquitous search by species name, alphabetical or taxonomic browsing (OGRe, MetAMIGA), taxonomic search (MetAMIGA, MitoZOA) may be available. MitoZOA provides an advanced search function, allowing queries for genetic code, base composition, or genomic features such as a specific gene order. In all databases it is possible to export sequence data of single or multiple features for a previously selected set of species. A download of the complete database is only possible for GenBank, RefSeq, and MitoZOA. With the exception of MetAMIGA all databases can output gene order data. In the case of NCBI, this is restricted to protein coding genes.

Some of the databases offer additional useful functions. MetAMIGA, for instance, can display information related to codon- and nucleotide usage for the selected species. A similar feature is available in OGRe for selected species. NCBI in particular offers various sequence search tools. Pairwise comparisons of gene orders in terms of breakpoints can be performed in OGRe's interface.

In addition to these general-purpose databases, there are two curated specialised collections of mitochondrial tRNAs: Mamit-tRNA and tRNAdb provide high quality structural annotations of tRNAs for 150 Mammalia and 152 Metazoa, respectively. These databases provide secondary structure information, including consensus structures and information on deviations from typical folds. Another specialised database, MamMiBase, does not provide annotations or sequence data, but allows to compute alignments and phylogenetic trees of stored protein and nucleotide sequences from mammalian mitochondria.

Several databases described in the literature are no longer updated and maintained or have gone offline. Among the first group is GOBASE, compiling diverse data related to mitochondria and chloroplasts. Not available any more are MitBASE (Attimonelli et al., 1999), AMmtDB (Lanave et al., 1999), MitoNUC (Attimonelli et al., 2002), or

Mitome (Lee et al., 2008).

3. Annotation and Re-Annotation

3.1. tRNA Detection

Mitochondrial tRNAs frequently have aberrant sequence and structure features. In contrast to nuclear tRNAs, they are therefore often hard to find even in the small mitogenomes. Three specialised tools are available for tRNA annotation. tRNAscan-SE, the most commonly used software for this purpose, features a special mode for organelle genomes in which a special covariance model is employed by using the approach of Eddy and Durbin (1994) and the filtering of pseudogenes is disabled.

ARWEN is a heuristic approach that starts with a search for three adjacent hairpin loops surrounded by additional potential base pairs that meet constraints on stem and loop lengths typical for tRNA clover leaf structures. The compliance of the candidate with the pattern is translated into a score value. A major advantage of ARWEN is its speed.

While tRNAscan-SE and ARWEN use general models to recognise all tRNAs, MiTFi employs Infernal (Nawrocki et al., 2009) to search the mitogenome using a different covariance model for each of the tRNA families (a single model is employed for the two tRNA-Leu genes). One advantage of MiTFi is that Infernal computes reliable E-values. MiTFi makes it easy to detect duplicate tRNA genes since alternative non-conflicting predictions are also reported. It appears that MiTFi is more accurate than the other two tools (Jühling et al., 2012). Its false negative rate is much lower than that of tRNAscan-SE while it detects much fewer false positives compared to ARWEN.

3.2. Genome Annotation

Three web services offer full-fledged annotation pipelines for mitogenomes: DOGMA, MOSAS, and MITOS. All three use BLASTX searches against internal databases to identify protein coding genes, DOGMA and MOSAS also employ BLAST to detect rRNA genes. MITOS uses Infernal and covariance models of mitochondrial rRNAs to identify them in a sequence. Database sequences from a wide variety of Metazoa are used for the search in DOGMA and MITOS; MOSAS is currently restricted to insects. Both DOGMA and MOSAS use tRNAscan-SE for the identification of tRNA genes, while MITOS uses MiTFi for tRNA annotation. The MITOS pipeline attempts to improve the prediction of gene boundaries automatically.

All three tools provide graphical and tabular output and they let export Sequin formatted annotation files to facilitate the submission of new mitogenomes to GenBank. MOSAS and DOGMA assist the user in selecting from alternative annotations and fine tuning the annotation. The MOSAS system offers users to manage their own database of up to 2,000 annotated sequences, to perform sequence searches against the NCBI databases, and to produce and inspect multiple alignments. MITOS provides also direct access to gene order files.

3.3. Determination of the Genetic Code

Mitochondrial genetic codes can differ in the assignment of several codons from the standard genetic code, reviewed by (Knight et al., 2001). Currently available methods to estimate the genetic code use a comparative approach. GenDecoder uses a majority vote on the amino acids aligned with codon triplets. Highly variable alignment columns can be excluded. It requires an annotation of protein coding sequences for the input mitogenome. FACIL does not need any annotation. It uses a HMMER search of provisional translations against Pfam-fs protein domain models (Finn et al., 2010) for the mitochondrial proteins. FACIL also computes confidence scores for its codon assignments.

4. Gene order

The analysis of gene orders is a promising source of phylogenetic information (e.g., Boore et al., 1995; Sankoff et al., 1992). For mitogenomes, in particular, rearrangements are often considered to be infrequent and to affect mostly tRNA genes. Cases of reversion or convergent evolution are thought to be rare (Boore, 1999). Therefore they are particularly interesting for problems of deep phylogeny (Boore, 2006b). With increasing taxon sampling, however, more and more exceptions have been reported, such as the variable gene order of the Ascidians (Gissi et al., 2010) contrasting the well preserved deuterostome gene order or cases of convergent evolution in Hymenoptera (Dowton et al., 2009). Furthermore, due to the small size, the nearly conserved gene inventory, and the wide-spread availability gene order analysis is of particular interest for animal mitogenomes. Because the available tools have been considered as insufficient until recently (Grande et al., 2008) most of the published work on gene order comparisons is still carried out manually, a practice that is likely to change in the near future. Several methods automatising the comparative analysis have been developed in recent years and a variety of software tools implementing these approaches are becoming publicly available. In addition, manual analyses become impractical with the fast growing number of available mitochondrial genome (Boore, 1999). For a detailed presentation of the formal study of gene order evolution problems we refer to Fertin et al. (2009); we focus here on the available tools.

One may distinguish two main ways of treating gene order data. The historically older approach assumes that certain well-defined types of elementary "operations" are responsible for evolutionary changes in mitogenomic gene orders. These types of operations are *inversion* (e.g., Smith et al., 1990), *transposition* (e.g., Boore et al., 1995), *inverse transposition* (e.g., Boore et al., 1998), and *tandem duplication random loss* (TDRL) (e.g., Boore, 2000). Empirical evidence and mechanistic considerations are briefly reviewed in (Bernt et al., 2012a, in this special issue). Alternatively differences of gene orders are measured without recourse to specific operations, but instead properties of the gene orders itself, i.e., gene clusters, are used. In the following we will call the former *rearrangement-based* and the latter *gene-cluster-based* approach. We emphasise that this classification pertains to different ways of looking at gene order data. Both types can be analysed within the standard frameworks for phylogeny reconstructions, i.e., by means of distance methods, maximum parsimony, and maximum likelihood, as well as with different aims, e.g., reconstruction of phylogeny or ancestral states. It remains unclear, at present, which approach is preferable. Even though the rearrangement-based strategy seems more appealing from a biological point of view because of the underlying mechanistic model, it remains unknown how variable rates of rearrangements and prevalence of particular operations are over longer evolutionary time-scales. The advantage of the gene-cluster-based approach is that it is rearrangement-model-free, i.e., it does not make explicit assumptions on the type and frequency of rearrangement operations. But instead the phylogenetic significance of the considered gene clusters is assumed.

4.1. Rearrangement-Based Approaches

The eventual goal of the rearrangement-based approach is to reconstruct the actual sequence of rearrangement events that have led to the contemporary gene orders. Usually, one considers a maximum parsimony problem for a set of allowed rearrangement operations and associated costs. The simplest case is to minimise the *distance*, i.e., the weighted number of rearrangements, that transform a given gene order into another one.

Major progress has been made in developing algorithms for computing rearrangement distances for the case that only one particular type of operations occur. The best-studied cases are inversions (e.g., Hannenhalli and Pevzner, 1995) and TDRLs (Chaudhuri et al., 2006; Bernt et al., 2011). Nevertheless, there are still major unsolved problems that are relevant to the study of mitochondrial gene order evolution, including the lack of efficient exact algorithms for rearrangement problems incorporating transpositions, where only approximation algorithms are known, implemented for instance in `GENESIS`. Recently, the so called double cut and join (DCJ) rearrangement operation (Yancopoulos et al., 2005; Bergeron et al., 2006), i.e., the cut of a gene order at two places with a subsequent re-joining in a different order, was proposed to solve the problem more efficiently. This operation incorporates inversions, circularisation and linearisation, and block interchanges (a generalisation of transpositions). Thereby it circumvents the algorithmic complications of transpositions. A recent implementation is `UniMoG`. `CREx` is a heuristic tailored for studying rearrangement scenarios for pairs of mitochondrial gene orders (Bernt et al., 2007). It takes inversions, transpositions, inverse transpositions, and TDRLs into account. Also the other approaches discussed so far are based on pairwise comparisons and produce rearrangement distances that can be further analysed by distance matrix methods (e.g., Wang et al., 2006).

Alternatively, rearrangement parsimony problems have been considered. These approaches attempt to reconstruct ancestral gene orders in a given phylogenetic tree so that the sum of distances along the edges is minimised. In addition, the phylogenetic tree itself can be reconstructed by standard techniques enumerating phylogenies (Felsenstein, 2004). In contrast to the corresponding problem for sequences, already the problem of reconstructing ancestral gene orders for a given phylogeny is in most cases a computationally hard problem (e.g., Caprara, 2003; Tannier et al., 2009), but a few tractable cases have been reported recently (Feijao and Meidanis, 2011; Bernt et al., 2012b). The problem is studied most intensively for the case that only inversions are allowed: with GRAPPA, MGR, and amGRP three toolkits are available. Recent algorithmic advances include the fast heuristic approach PHYLO that is able to consider inversions and transpositions in a weighted fashion (Bader et al., 2008) and the heuristic PATHGROUPS that efficiently solves the problem to reconstruct ancestral states when DCJ operations are considered (Zheng and Sankoff, 2011). TreeREx (Bernt et al., 2008) and the method described recently in Bernt and Middendorf (2011) are based on CREx and allow to analyse genome rearrangements with or without a given phylogenetic tree. A Bayesian approach limited to inversions is available in the BADGER software. Recent progress on rearrangement-based probabilistic methods is reported in Miklós and Tannier (2012).

4.2. Gene-Cluster-Based Approaches

The gene-cluster-based approaches implicitly analyse the co-occurrence of consecutive groups of genes. Such *gene clusters* might also have to satisfy additional restrictions, e.g., a limited size. From a mathematical point of view, different definitions of consecutiveness have been investigated (Heber and Stoye, 2001; Luc et al., 2003). Empirically one observes that preservation of gene clusters can be phylogenetically informative. The number of gene clusters not shared by two gene orders thus provides a natural measure of dissimilarity for gene orders that is not based on any particular set of rearrangement operations.

Gene clusters of size two are most widely used and best understood. The *breakpoint distance* (Blanchette et al., 1997) is simply the number of adjacencies not shared between two gene orders. In general, the number of common gene clusters is a measure for the similarity of two gene orders. For gene orders with equal gene content this can be transformed into a distance measure by normalising (subtraction/division) it with the maximum possible number of common gene clusters, i.e., the number of gene clusters a gene order has in common with itself. Extra care has to be taken if the gene orders have different gene content. Bergeron and Stoye (2006) proposed to use the number of common gene clusters of size larger than two can be used to define a distance measure for gene orders. Jahn and Stoye (2009) tested different gene cluster definitions and weighting schemes for the generation of evolutionary distances being used in distance based tree reconstruction methods.

Instead of computing distances, binary characters can be derived from the absence/presence of gene clusters in gene orders. The `MPBE` method (Cosner et al., 2000) uses adjacencies, i.e., gene clusters of size two, for maximum parsimony analyses. The common theme of `inferCARs`, `roci`, as well as the approaches of Adam et al. (2007) and Chauve and Tannier (2008) is to derive binary character data from given gene orders, reconstruct ancestral binary states in a given phylogenetic tree, and to reconstruct ancestral gene orders from the reconstructed binary states. The latter presents the main algorithmic challenge in the gene-cluster-based approaches, because not all combinations of binary character states correspond to valid gene orders. Ma et al. (2006) used adjacencies by handling the predecessors and successors separately; Stoye and Wittler (2009), Adam et al. (2007) used consecutive gene sets without size restrictions; and Chauve and Tannier (2008) allows for gene sets such that the contained genes might be separated by a maximum number of other genes. Adam et al. (2007) reconstructs the ancestral binary states with the Fitch Algorithm and then do a greedy reconstruction of the gene order. Stoye and Wittler (2009) give an optimal algorithm that minimises the number of state changes. The tool `inferCARs` implements a heuristic approach to find a maximum weight subset of the ancestral gene clusters that can be realised by a valid gene order. This problem is solved optimally by the approach described in Chauve and Tannier (2008). `ANGES` implements a variety of Gene-Cluster-Based approaches that are suitable for analysing mitogenome gene orders.

Recently also gene-cluster-based maximum likelihood approaches have been introduced (Ma, 2010; Hu et al., 2011). A different approach for deriving phylogenetic characters from gene orders is implemented in `circal`, which computes circular alignments of gene orders. The approach allows to apply different costs for tRNAs, rRNAs, and protein coding genes.

4.3. Mixed Approaches

As discussed above both ways of looking at gene order data have advantages as well as drawbacks. The first available attempts to mix the approaches might be a solution to this dilemma. The foundation of these approaches is the observation that rearrangements correspond to certain patterns in the gene clusters. `EMRAE`, for instance, uses the correspondence between the margins of an inversion or transposition and the two, respectively three, breakpoints that are found in the gene orders separated by such an inversion or transposition. Given a phylogenetic tree, `EMRAE` determines pairs and triples of adjacencies corresponding to inversions and transpositions, respectively, that are present in the gene orders on one side of a phylogenetic split and absent on the other side. Thereby it reconstructs the rearrangements on the edges of the given phylogeny. Also `CREx` might be interpreted as mixed approach because it derives rearrangements from patterns of the relative order of common gene clusters.

Clearly, the dichotomy used above is fuzzy and has limitations. For instance, Feijao and Meidanis (2011) analysed the formal properties of a rearrangement model where the creation and destruction of a single adjacency are the only allowed kinds of rearrangement operations. Thereby it acts on the same objects as `inferCARs`, i.e., adjacencies, but instead of coding them as binary characters they are subject to the rearrangement operations.

5. Modelling MtDNA sequence evolution

The peculiarities of mitochondrial genome replication lead to biases in sequence composition (Bernt et al., 2012a, in this special issue). Together with the limited amount of sequence information, these systematic biases may introduce artefacts that impair the inference of a phylogeny. Systematic errors arise in particular from model misspecification, i.e., from discrepancies between the true substitutions patterns in mitogenomic sequence evolution and the model of substitutions employed by the phylogenetic inference software (Whelan and Goldman, 2001; Baurain et al., 2007).

Various methods and tools exist to select the most adequate model and promote more reliable mito-phylogenies. ProTtest (Abascal et al., 2005), jMODELTEST (Posada, 2008), and ModelGenerator (Keane et al., 2006) are popular programs that automatically select the most fitting substitution models for amino acids, nucleotides, or both. In general, the fit of a model to the data can be estimated in a Maximum Likelihood framework (Likelihood Ratio Test, formally applicable only for nested models) or by means of marginal likelihoods in a Bayesian framework (Bayes factor). Likelihoods are conveniently adjusted using simple statistics such as AIC and BIC to penalise models with many free parameters (Posada and Buckley, 2004; Abascal et al., 2005; Rota-Stabelli et al., 2009). Cross validation (Stone, 1974) is a more robust albeit computationally expensive way of comparing model fit to the data set. A dedicated pipeline is conveniently implemented in Phylobayes (Lartillot et al., 2009).

For mitochondrial coding genes, the best fitting model at the nucleotide level is likely to be nnGTR (General Time Reversible), which assumes different rates for the six possible substitutions and reversibility (Lanave et al., 1984). The model is “mechanistic”, i.e., all parameters are inferred from the data set. Its six parameters can easily be estimated by directed sampling even from data sets that are small both in terms of sequence lengths and number of taxa. Given the within codon-heterogeneity of the evolutionary process in mitochondrial sequences it may be advantageous to partition the data set by codon position and apply a different nnGTR model to each subset (Rota-Stabelli et al., 2010). Furthermore, it is common practice to exclude highly saturated third codon positions from the data set or alternatively, employ a recoding strategy, which can be further expanded to synonymous sites at first and second codon positions (Masta et al., 2009).

Although computationally demanding, amino acid substitutions can also be described by a mechanistic aaGTR model. The estimation of the 190 entries of the aaGTR matrix from the relatively short mitochondrial data sets may, however, introduce stochastic errors. It is advisable, therefore, to use a GTR model pre-estimated from large well curated data sets, the so-called empirical GTR models (Adachi and

Hasegawa, 1996). Various empirical models have been expressively designed for mitogenomic studies: MtREV (Adachi and Hasegawa, 1996), Mtmamm (Yang et al., 1998), MtArt (Abascal et al., 2007), MtPan (Carapelli et al., 2007), and MtZOA (Rota-Stabelli et al., 2009) are based on the analysis of mitochondrial proteins sets from vertebrates, mammals, arthropods, pancrustaceans, and all metazoans, respectively. Model fit to the data set will easily provide the most adequate empirical model for any given data set.

These empirical models of protein evolution assume that substitution rates are homogeneous among sites, hence treating all positions of the alignment equally. This may promote systematic artefacts, because models assume homogeneity where it does not exists. The heterogeneous CAT model (Lartillot and Philippe, 2004) and its empirical adaptation (Le et al., 2008) relax the homogeneity assumption and have been shown to yield more reliable phylogenies using mitochondrial genes (Rota-Stabelli et al., 2010). The CAT model can be effectively coupled with a GTR model (CAT-GTR) leading to significant improvements over either model (Philippe et al., 2011; Blanquart and Gascuel, 2011). Temporal heterogeneity of the substitution process (heterotachy) can be partially accounted for by the covarion approach as implemented in both MrBayes or PhyloBayes (Zhou et al., 2007).

A key problem in mitogenomics is the lineage-specific compositional heterogeneity, which expresses itself at the nucleotide level both as variation in G+C content and as imbalance of nucleotides between the two DNA strands (Saccone et al., 1999). These compositional biases can be so extreme that they also affect the amino acid content of the encoded proteins (Foster et al., 1997; Rota-Stabelli et al., 2010). It is responsible for serious systematic artefacts throughout all metazoans (Gibson et al., 2005; Hassanin, 2006; Jones et al., 2007). Heterogeneous models of evolution such as for example CAT-BP and the vector model implemented in P4 (Blanquart and Lartillot, 2008; Foster, 2004) allow the stationary frequencies to vary in different parts of the tree. This is computationally demanding but can indeed overcome systematic artefacts in mitogenomic studies (Bourlat et al., 2009; Blanquart and Gascuel, 2011).

6. Concluding Remarks

The currently available and still fast growing collection of mitochondrial sequence data contains highly valuable phylogenetic information. Adequate tools that a geared to the specifics of mitogenomic sequences, however, are required to unlock this treasure trove.

Comparative analyses of mitogenomes and phylogeny reconstruction requires the availability of reliable and consistent annotations for large sets of species. A problem that can be solved with one of the approaches described in the first part of the review. Still, there are open problems, e.g., precise structural annotation of rRNA genes or features of the mitochondrial control region. An issue of particular importance is knowledge on the precise properties of the evolutionary processes. For sequences this is treated by the methods to derive models for sequence evolution treated in the last section. For gene orders the gene-cluster-based approaches circumvents the problem of the unknown mode of rearrangement evolution. But the reconstruction of the rearrangements still calls for the rearrangement-based approach.

The present review of the bioinformatics methods for the analysis of mitochondrial genomes not only may serve as a comprehensive guide for future analyses but also a guide towards the pressing methodological issues that have remained unsolved so far.

7. Acknowledgments

We would like to thank Carmela Gissi and Paul Higgs for the helpful discussions on the data bases.

This work was supported by the German Research Foundation (DFG) through the project “Deep Metazoan Phylogeny” within SPP 1174 (projects STA 850/3-2, STA 850/2) and by a Marie-Curie/Province of Trento (COFOUND) fellowship to ORS.

References

Abascal, F., Posada, D., Zardoya, R., 2005. ProtTest: Selection of best-fit models of protein evolution. Bioinformatics 21, 2104–2105.

Abascal, F., Posada, D., Zardoya, R., 2007. MtArt: a new model of amino acid replacement for arthropoda. Mol. Biol. Evol. 24, 1–5.

Abascal, F., Zardoya, R., Posada, D., 2009. Genetic code prediction for metazoan mitochondria with GenDecoder. Methods Mol. Biol. 537, 233–242.

Adachi, J., Hasegawa, M., 1996. Model of amino acid substitution in proteins encoded by mitochondrial DNA. J. Mol. Evol. 42, 459–468.

Adam, Z., Turmel, M., Lemieux, C., Sankoff, D., 2007. Common intervals and symmetric difference in a model-free phylogenomics, with an application to streptophyte evolution. J. Comp. Biol. 14, 436–445.

Attimonelli, M., Altamura, N., Benne, R., Boyen, C., Brennicke, A., Carone, A., Cooper, J. M., D'Elia, D., de Montalvo, A., de Pinto, B., De Robertis, M., Golik, P., Grienenberger, J. M., Knoop, V., Lanave, C., Lazowska, J., Lemagnen, A., Malladi, B. S., Memeo, F., Monnerot, M., Pilbout, S., Schapira, A. H. V., Sloof, P., Slonimski, P., Stevens, K., Saccone, C., 1999. MitBASE: a comprehensive and integrated mitochondrial DNA database. Nucleic Acids Res. 27, 128–133.

Attimonelli, M., Catalano, D., Gissi, C., Grillo, G., Licciulli, F., Liuni, S., Santamaria, M., Pesole, G., Saccone, C., 2002. MitoNuc: a database of nuclear genes coding for mitochondrial proteins. Update 2002. Nucleic Acids Res. 30, 172–173.

Bader, M., Abouelhoda, M., Ohlebusch, E., 2008. A fast algorithm for the multiple genome rearrangement problem with weighted reversals and transpositions. BMC Bioinformatics 9, 516.

Bader, M., Ohlebusch, E., 2007. Sorting by weighted reversals, transpositions, and inverted transpositions. J. Comp. Biol. 14, 615–636.

Baurain, D., Brinkmann, H., Philippe, H., 2007. Lack of resolution in the animal phylogeny: closely spaced cladogeneses or undetected systematic errors? Mol. Biol. Evol. 24, 6–9.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Sayers, E. W., 2011. GenBank. Nucleic Acids Res. 39 (suppl 1), D32–D37.

Bergeron, A., Mixtacki, J., Stoye, J., 2006. A unifying view of genome rearrangements. In: Algorithms in Bioinformatics, 6th International Workshop, WABI 2006, Proceedings. Vol. 4175 of Lecture Notes in Bioinformatics. Springer, pp. 163–173.

Bergeron, A., Stoye, J., 2006. On the similarity of sets of permutations and its applications to genome comparison. J. Comp. Biol. 13 (7), 1340–1354.

Bernt, M., Braband, A., Schierwater, B., Stadler, P. F., 2012a. Genetic aspects of mitochondrial genome evolution. Mol. Phyl. Evol. under revision.

Bernt, M., Chao, K.-M., Kao, J.-W., Middendorf, M., Tannier, E., 2012b. Preserving inversion phylogeny reconstruction. In: WABI. LNBI. Accepted.

Bernt, M., Chen, K.-Y., Chen, M.-C., Chu, A.-C., Merkle, D., Wang, H.-L., Chao, K.-M., Middendorf, M., 2011. Finding all sorting tandem duplication random loss operations. J. Discr. Algorithms 9, 32–48.

Bernt, M., Donath, A., Jühling, F., Externbrink, F., Florentz, C., Fritzsch, G., Pütz, J., Middendorf, M., Stadler, P. F., 2012c. MITOS: Improved *de novo* metazoan mitochondrial genome annotation. Mol. Phyl. Evol. accepted subject to minor revision.

Bernt, M., Merkle, D., Middendorf, M., 2008. An algorithm for inferring mitogenome rearrangements in a phylogenetic tree. In: Comparative Genomics, International Workshop, RECOMB-CG 2008, Proceedings. Vol. 5267 of Lecture Notes in Bioinformatics. Springer, pp. 143–157.

Bernt, M., Merkle, D., Ramsch, K., Fritzsch, G., Perseke, M., Bernhard, D., Schlegel, M., Stadler, P. F., Middendorf, M., 2007. CREx: inferring genomic rearrangements based on common intervals. Bioinformatics 23 (21), 2957–2958.

Bernt, M., Middendorf, M., 2011. A method for computing an inventory of metazoan mitochondrial gene order rearrangements. BMC Bioinformatics 12 (Suppl 9), S6.

Blanchette, M., Bourque, G., Sankoff, D., 1997. Breakpoint phylogenies. In: Genome Informatics. Universal Academy Press, pp. 25–34.

Blanquart, S., Gascuel, N., 2011. Mitochndrial genes support a common origin of rodent malaria parasites and *Plasmodium falciparum* relatives infecting great apes. BMC Evol. Biol. 11, 70.

Blanquart, S., Lartillot, N., 2008. A site- and time-heterogeneous model of amino acid replacement. Mol. Biol. Evol. 25, 842–858.

Boore, J. L., 1999. Animal mitochondrial genomes. Nucleic Acids Res. 27 (8), 1767–1780.

Boore, J. L., 2000. The duplication/random loss model for gene rearrangement exemplified by mitochondrial genomes of deuterostome animals. In: Sankoff, D., Nadeau, J. H. (Eds.), Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and the Evolution of Gene Families. Vol. 1 of Computational Biology Series. Kluwer Academic Publishers, pp. 133–147.

Boore, J. L., 2006a. Requirements and standards for organelle genome databases. OMICS 10, 119–126.

Boore, J. L., 2006b. The use of genome-level characters for phylogenetic reconstruction. Trends Ecol. Evol. 21, 439–446.

Boore, J. L., Collins, T. M., Stanton, D., Daehler, L. L., Brown, W. M., 1995. Deducing the pattern of arthropod phylogeny from mitochondrial DNA rearrangements. Nature 376 (6536), 163–165.

Boore, J. L., Lavrov, D. V., Brown, W. M., 1998. Gene translocation links insects and crustaceans. Nature 392 (6677), 667–668.

Boore, J. L., R., M. J., Medina, M., 2005. Sequencing and comparing whole mitochondrial genomes of animals. Methods Enzymol. 395, 311–348.

Bourlat, S., Rota Stabelli, O., Lanfear, R., Telford, M. J., 2009. The mitochondrial genome of *Xenoturbella* is ancestral within the deuterostome. BMC Evol. Biol. 9, 107.

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

Caprara, A., 2003. The reversal median problem. INFORMS J. Computing 15 (1), 93–113.

Carapelli, A., Liò, P., Nardi, F., Van der Wath, E., Frati, F., 2007. Phylogenetic analysis of mitochondrial protein coding genes confirms the reciprocal paraphyly of Hexapoda and Crustacea. BMC Evol. Biol. 16, Suppl 2:S8.

Chaudhuri, K., Chen, K., Mihaescu, R., Rao, S., 2006. On the tandem duplication-random loss model of genome rearrangement. In: Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2006. ACM, pp. 564–570.

Chauve, C., Tannier, E., 2008. A methodological framework for the reconstruction of contiguous regions of ancestral genomes and its application to mammalian genomes. PLoS Comp. Biol. 4 (11), e1000234.

Cosner, M., Jansen, R. K., Moret, B. M. E., Raubeson, L. A., Wang, L.-S., Warnow, T., Wyman, S., 2000. A new fast heuristic for computing the breakpoint phylogeny and experimental phylogenetic analyses of real and synthetic data. In: Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology. AAAI, pp. 104–115.

de Vasconcelos, A. T. R., Guimarães, A. C. R., Castelletti, C. H. M., Caruso, C. S., Ribeiro, C., Yokaichiya, F., Armôa, G. R. G., Pereira, G. d. S. P., da Silva, I. T., Schrago, C. G., Fernandes, A. L. V., da Silveira, A. R., Carneiro, A. G., Carvalho, B. M., Viana, C. J. M., Gramkow, D., Lima, F. J., Corrêa, L. G. G., Mudado, M. d. A., Nehab-Hess, P., de Souza, R., Corrêa, R. L., Russo, C. A. M., 2005. Mam-

MiBase: a mitochondrial genome database for mammalian phylogenetic studies. Bioinformatics 21, 2566–2567.

Dowton, M., Cameron, S. L., Dowavic, J. I., Austin, A. D., Whiting, M., 2009. Characterization of 67 mitochondrial tRNA gene rearrangements in the hymenoptera suggests that mitochondrial tRNA gene position is selectively neutral. Mol. Biol. Evol. 26, 1607–1617.

Dutilh, B. E., Jurgelenaite, R., Szklarczyk, R., van Hijum, S. A., Harhangi, H. R., Schmid, M., de Wild, B., Françoijs, K., Stunnenberg, H. G., Strous, M., Jetten, M. S., Op den Camp, H. J., Huynen, M. A., 2011. FACIL: Fast and accurate genetic code inference and logo. Bioinformatics 27, 1929–1933.

Eddy, S. R., Durbin, R., 1994. RNA sequence analysis using covariance models. Nucleic Acids Res. 22 (11), 2079–2088.

Feijão, P. C., Neiva, L. S., Lima de Azeredo-Espin, A. M., Lessinger, A. C., 2006. AMiGA: the arthropodan mitochondrial genomes accessible database. Bioinformatics 22, 902–903.

Feijao, P., Meidanis, J., 2011. Scj: A breakpoint-like distance that simplifies several rearrangement problems. IEEE/ACM Trans. Comp. Biol. Bioinf. 8, 1318 –1329.

Felsenstein, J., 2004. Inferring Phylogenies. Sinauer Associates, Sunderland, Massachusetts.

Fertin, G., Labarre, A., Rusu, I., Tannier, E., Vialette, S., 2009. Combinatorics of Genome Rearrangements. MIT Press.

Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunesekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E. L., Eddy, S. R., Bateman, A., 2010. The Pfam protein families database. Nucleic Acids Res. 38, D211–D222.

Foster, P. G., 2004. Modeling compositional heterogeneity. Syst. Biol. 53, 485–495.

Foster, P. G., Jermiin, L. S., Hickey, D. A., 1997. Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. J. Mol. Evol. 44, 282–288.

Fritzsch, G., Schlegel, M., Stadler, P. F., 2006. Alignments of mitochondrial genome arrangements: Applications to metazoan phylogeny. J. Theor. Biol. 240, 511–520.

Gibson, A., Gowri-Shankar, V., Higgs, P. G., Rattray, M., 2005. A comprehensive analysis of mammalian mitochondrial genome base composition and improved phylogenetic methods. Mol. Biol. Evol. 22, 251–264.

Gissi, C., Pesole, G., Mastrototaro, F., Iannelli, F., Guida, V., Griggio, F., 2010. Hypervariability of ascidian mitochondrial gene order: Exposing the myth of deuterostome organelle genome stability. Mol. Biol. Evol. 27, 211–215.

Grande, C., Templado, J., Zardoya, R., 2008. Evolution of gastropod mitochondrial genome arrangements. BMC Evol. Biol. 8, 61.

Hannenhalli, S., Pevzner, P. A., 1995. Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals. In: Proceedings of the Twenty-Seventh Annual ACM Symposium on Theory of Computing. ACM, pp. 178–189.

Hassanin, A., 2006. Phylogeny of arthropoda inferred from mitochondrial sequences: strategies for limiting the misleading effects of multiple changes in pattern and rates of substitution. Mol. Phylogenet. Evol. 38, 100–116.

Heber, S., Stoye, J., 2001. Algorithms for finding gene clusters. In: Algorithms in Bioinformatics, First International Workshop, WABI 2001, Proceedings. Vol. 2149 of Lecture Notes in Computer Science. Springer, pp. 252–263.

Hu, F., Gao, N., Zhang, M., Tang, J., 2011. Maximum likelihood phylogenetic reconstruction using gene order encodings. In: Computational Intelligence in Bioinformatics and Computational Biology (CIBCB). pp. 1–6.

Jahn, K., Stoye, J., 2009. Approximative Gencluster und ihre Anwendung in der komparativen Genomik. Informatik-Spektrum 32 (4), 288–300.

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

Jones, B. R., Rajaraman, A., Tannier, E., Chauve, C., 2012. ANGES: Reconstructing ANcestral GEnomeS maps. Bioinformatics to appear.

Jones, M., Gantenbein, B., Fet, V., Blaxter, M., 2007. The effect of model choice on phylogenetic inference using mitochondrial sequence data: Lessons from the scorpions. Mol. Phylogenet. Evol. 3, 583–595.

Jühling, F., Mörl, M., Hartmann, R. K., Sprinzl, M., Stadler, P. F., Pütz, J., 2009. tRNAdb 2009: compilation of tRNA sequences and tRNA genes. Nucleic Acids Res. 37 (Database issue), D159–D162.

Jühling, F., Pütz, J., Bernt, M., Donath, A., Middendorf, M., Florentz, C., Stadler, P. F., 2012. Improved systematic tRNA gene annotation allows new insights into the evolution of mitochondrial tRNA structures and into the mechanisms of mitochondrial genome rearrangements. Nucleic Acids Res. 40 (7), 2833–2845.

Keane, T. M., Creevey, C. J., Pentony, M. M., Naughton, T. J., McInerney, J. O., 2006. Assessment of methods for amino acid matrix selection and their use on empirical data shows that *ad hoc* assumptions for choice of matrix are not justified. BMC Evol. Biol. 6, 29.

Knight, R. D., Freeland, S. J., Landweber, L. F., 2001. Rewiring the keyboard: evolvability of the genetic code. Nat. Rev. Genet. 2, 49–58.

Lanave, C., Attimonelli, M., De Robertis, M., Licciulli, F., Liuni, S., Sbis, E., Saccone, C., 1999. Update of AMmtDB: a database of multi-aligned metazoa mitochondrial DNA sequences. Nucleic Acids Res. 27, 134–137.

Lanave, C., Preparata, G., Saccone, C., Serio, G., 1984. A new method for calculating evolutionary substitution rates. J. Mol. Evol. 20, 86–93.

Larget, B., Kadane, J. B., Simon, D. L., 2005. A Bayesian approach to the estimation of ancestral genome arrangements. Mol. Phylogenet. Evol. 36, 214–223.

Lartillot, N., Lepage, T., Blanquart, S., 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25, 2286–2288.

Lartillot, N., Philippe, H., 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21, 1095–1109.

Laslett, D., Canbäck, B., 2008. ARWEN: a program to detect tRNA genes in metazoan mitochondrial nucleotide sequences. Bioinformatics 24, 172–175.

Le, S. Q., Gascuel, O., Lartillot, N., 2008. Empirical profile mixture models for phylogenetic reconstruction. Bioinformatics 24, 2317–2323.

Lee, Y. S., Oh, J., Kim, Y. U., Kim, N., Yang, S., Hwang, U. W., 2008. Mitome: dynamic and interactive database for comparative mitochondrial genomics in metazoan animals. Nucleic Acids Res. 36 (suppl 1), D938–D942.

Lowe, T. M., Eddy, S. R., 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25, 955–964.

Luc, N., Risler, J.-L., Bergeron, A., M., R., 2003. Gene teams: A new formalization of gene clusters for comparative genomics. Comp. Biol. Chem. 27, 59–67.

Lupi, R., de Meo, P. D., Picardi, E., D'Antonio, M., Paoletti, D., Castrignanó, T., Pesole, G., Gissi, C., 2010. MitoZoa: A curated mitochondrial genome database of metazoans for comparative genomics studies. Mitochondrion 10, 192–199.

Ma, J., 2010. A probabilistic framework for inferring ancestral genomic orders. In: Bioinformatics and Biomedicine (BIBM). pp. 179 –184.

Ma, J., Zhang, L., Bernard, B., Raney, B., Burhans, R., Kent, W., Blanchette, M., Haussler, D., Miller, W., 2006. Reconstructing contiguous regions of an ancestral genome. Genome Res. 16, 1557–1565.

Masta, S. E., Longhorn, S. J., Boore, J. L., 2009. Arachnid relationships based on mitochondrial genomes: asymmetric nucleotide and amino acid bias affects phylogenetic analyses. Mol. Phyl. Evol. 50, 117–128.

Miklós, I., Tannier, E., 2012. Approximating the number of double cut-and-join scenarios. Theor. Comput. Sci. 439, 30 – 40.

Mizrachi, I., 2007. GenBank: The nucleotide sequence database. In: McEntyre, J., Ostell, J. (Eds.), The NCBI Handbook. NCBI, Bethesda, MD, p. chap.1.

Moret, B. M. E., Wang, L.-S., Warnow, T., Wyman, S. K., 2001. New approaches for reconstructing phylogenies from gene order data. Bioinformatics 17, 165–173.

Nawrocki, E. P., Kolbe, D. L., Eddy, S. R., 2009. Infernal 1.0: inference of RNA alignments. Bioinformatics 25 (10), 1335–1337.

O'Brien, E. A., Zhang, Y., Wang, E., Marie, V., Badejoko, W., Lang, B. F., Burger, G., 2009. GOBASE: an organelle genome database. Nucleic Acids Res. 37, D946–950.

Philippe, H., Brinkmann, H., Copley, R. R., Moroz, L. L., Nakano, H., Poustka, A. J., Wallberg, A., Peterson, K. J., Telford, M. J., 2011. Acoelomorph flatworms are deuterostomes related to *Xenoturbella*. Nature 470, 255–258.

Posada, D., 2008. `jModelTest`: Phylogenetic model averaging. Mol. Biol. Evol. 25, 1253–1256.

Posada, D., Buckley, T. R., 2004. Model selection and model averaging in phylogenetics: advantages of Akaike Information Criterion and Bayesian approaches over likelihood ratio tests. Syst. Biol. 53, 793–808.

Pruitt, K. D., Tatusova, T., Maglott, D. R., 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 35, D61–D65.

Pütz, J., Dupuis, B., Sissler, M., Florentz, C., 2007. Mamit-tRNA, a database of mammalian mitochondrial tRNA primary and secondary structures. RNA 13, 1184–1190.

Rota-Stabelli, O., Kayal, E., Gleeson, D., Daub, J., Boore, J. L., Telford, M. J., Pisani, D., Blaxter, M., Lavrov, D. V., 2010. Ecdysozoan mitogenomics: evidence for a common origin of the legged invertebrates, the Panarthropoda. Genome Biol. Evol. 2, 425–440.

Rota-Stabelli, O., Yang, Z., Telford, M. J., 2009. MtZoa: a general mitochondrial amino acid substitutions model for animal evolutionary studies. Mol. Phylogenet. Evol. 52, 268–272.

Saccone, C., De Giorgi, C., Gissi, C., Pesole, G., Reyes, A., 1999. Evolutionary genomics in Metazoa: the mitochondrial DNA as a model system. Gene 238, 195–209.

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

Sheffield, N. C., Hiatt, K. D., Valentine, M. C., Song, H., Whiting, M. F., 2010. Mitochondrial genomics in orthoptera using MOSAS. Mitochondrial DNA 21, 87–104.

Smith, M. J., Banfield, D. K., Doteval, K., Gorski, S., Kowbel, D. J., 1990. Nucleotide sequence of nine protein-coding genes and 22 tRNAs in the mitochondrial DNA of the sea star pisaster ochraceus. J. Mol. Evol. 31, 195–204.

Stone, M., 1974. Cross-validatory choice and assessment of statistical predictions. J. R. Stat. Soc. Ser. B Stat. Methodol. 36, 111–147.

Stoye, J., Wittler, R., 2009. A unified approach for reconstructing ancient gene clusters. IEEE/ACM Trans. Comp. Biol. Bioinf. 6, 387–400.

Tannier, E., Zheng, C., Sankoff, D., 2009. Multichromosomal median and halving problems under different genomic distances. BMC Bioinformatics 10, 120.

Wang, L.-S., Warnow, T., Moret, B. M. E., Jansen, R. K., Raubeson, L. A., 2006. Distance-based genome rearrangement phylogeny. J. Mol. Evol. 63 (4), 473–483.

Whelan, S., Goldman, N., 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol. Biol. Evol. 18, 691–699.

Wolfsberg, T. G., Schafer, S., Tatusov, R. L., Tatusova, T. A., 2001. Organelle genome resources at NCBI. Trends Biochem. Sci. 26, 199–203.

Wyman, S. K., Jansen, R. K., Boore, J. L., 2004. Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20 (17), 3252–3255.

Yancopoulos, S., Attie, O., Friedberg, R., 2005. Efficient sorting of genomic permutations by translocation, inversion and block interchange. Bioinformatics 21, 3340–3346.

Yang, Z., Nielsen, R., Hasegawa, M., 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. Mol. Biol. Evol. 15, 1600–1611.

Zhao, H., Bourque, G., 2009. Recovering genome rearrangements in the mammalian phylogeny. Genome Res. 19, 934–942.

Zheng, C., Sankoff, D., 2011. On the PATHGROUPS approach to rapid small phylogeny. BMC Bioinformatics 12 (Suppl 1), S4.

Zhou, Y., Rodrigue, N., Lartillot, N., Philippe, H., 2007. Evaluation of the models handling heterotachy in phylogenetic inference. BMC Evol. Biol. 7, 206.