Accurate Annotation of Protein-Coding Genes in Mitochondrial Genomes

Marwa Al Arab[a,b,h], Christian Höner zu Siederdissen[a,b,c], Kifah Tout[h], Abdullah H. Sahyoun[a,b,h,i], Peter F. Stadler[a,b,c,d,e,f,g], Matthias Bernt[a,j,*]

[a] *Bioinformatics Group, Department of Computer Science University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany.*
[b] *Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany.*
[c] *Institute for Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria.*
[d] *Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany.*
[e] *Fraunhofer Institut für Zelltherapie und Immunologie, Perlickstraße 1, D-04103 Leipzig, Germany.*
[f] *Center for non-coding RNA in Technology and Health, University of Copenhagen, Grønnegårdsvej 3, DK-1870 Frederiksberg C, Denmark.*
[g] *Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501*
[h] *Doctoral School of Science and Technology, AZM Center for Biotechnology Research, Lebanese University, Tripoli, Lebanon*
[i] *TRON - Translational Oncology at the University Medical Center of the Johannes Gutenberg University Mainz gGmbH, Mainz, Germany.*
[j] *Parallel Computing and Complex Systems Group, Department of Computer Science, University of Leipzig, Augustusplatz 10 D-04103 Leipzig, Germany*

Abstract

Mitochondrial genome sequences are available in large number and new sequences become published nowadays with increasing pace. Fast, automatic, consistent, and high quality annotations are a prerequisite for downstream analyses. Therefore, we present an automated pipeline for fast *de-novo* annotation of mitochondrial protein-coding genes. The annotation is based on enhanced phylogeny-aware hidden Markov models (HMMs). The pipeline builds taxon-specific enhanced multiple sequence alignments (MSA) of already annotated sequences and corresponding HMMs using an approximation of the phylogeny.

[*] Corresponding author

Email addresses: marwa@bioinf.uni-leipzig.de (Marwa Al Arab), choener@bioinf.uni-leipzig.de (Christian Höner zu Siederdissen), ktout@ul.edu.lb (Kifah Tout), abdullah.sahyoun@tron-mainz.de (Abdullah H. Sahyoun), studla@bioinf.uni-leipzig.de (Peter F. Stadler), bernt@informatik.uni-leipzig.de (Matthias Bernt)

The MSAs are enhanced by fixing unannotated frameshifts, purging of wrong sequences, and removal of non-conserved columns from both ends. A comparison with reference annotations highlights the high quality of the results. The frameshift correction method predicts a large number of frameshifts, many of which are unknown. A detailed analysis of the frameshifts in *nad3* of the Archosauria-Testudines group has been conducted.

1. Introduction

The overwhelming majority of eukaryotes harbors mitochondria, organelles with their own genome. Mitochondrial DNA (mtDNA) is tiny compared to nuclear genomes (typically about 16.5 kb) and has a very limited gene content. Metazoan mitogenomes typically encode 13 protein-coding genes, 22 tRNAs, and 2 rRNAs. Complete mtDNA sequences are relatively easy and cost-effective to obtain even for “exotic” non-model organisms. They are frequently used for phylogenetic studies due to their peculiar evolutionary dynamics (Boore, 2006). Characteristic signals in mitochondrial protein-coding gene sequences have proved to be useful for resolving disputed phylogenetic relationships (Zardoya and Meyer, 1996; Bourlat et al., 2008), but the extraction of deep phylogenetic signals from protein-coding genes remains challenging (Bernt et al., 2013a).

The number of sequenced mitogenomes is increasing rapidly, creating the need for a fast, automatic, accurate, and reproducible annotation pipeline that requires little or no manual curation. A number of preliminary tools for annotating mitogenomes are available. DOGMA (Wyman et al., 2004) is a semi-automatic web tool for annotating mitogenomes and chloroplast genomes. Fully automated pipelines are implemented in MITOS for annotating metazoan mtDNA (Bernt et al., 2013c) and Mitoannotator for annotating fish mtDNA (Iwasaki et al., 2013).

All the tools mentioned above use BLAST for annotating protein-coding genes. The implemented simple strategies have several shortcomings. (i) In-correct annotations in reference databases such as RefSeq (Bernt et al., 2013c) require either additional manual curation to obtain an unbiased high quality set of trusted query sequences for the BLAST search or automatic methods to cope with misannotated or biased queries at the level of the search results. A manual curation step of the data base, as used e.g., in DOGMA and Mitoannotator, has the disadvantage that updates of the query database are labor intensive and therefore it is difficult to keep pace with the growth of the available data. So far all tools and databases that require any manual curation, such as MitoZoa (Lupi et al., 2010), are not updated anymore or were not sustainable. Therefore an automatic strategy is pursued in MITOS. Instead of curating the queries, the BLAST hits are aggregated and conflicts are resolved by what essentially amounts to majority voting. The results of the current version of MITOS occasionally need manual curation of the start and stop positions of protein-coding genes (Iwasaki et al., 2013; Cameron, 2014). (ii) Given statistical measures for the reliability of the annotations are either difficult to interpret (e.g., the aggregated e-values presented by MITOS) or do only represent the similarity with respect to a single sequence. (iii) Some very short and poorly conserved genes, in particular *atp8*, tend to be missed by BLAST-based approaches.

Furthermore, annotation methods need to consider the peculiar features of mitochondrial protein-coding genes that complicate the annotation of mitochondrial protein-coding genes: (i) The use of unusual genetic codes (Hyouta et al., 1987), which may even involve a re-interpretation of the universal stop codon UAG translated as Tyr, e.g., in the sponge *Clathrina clathrus* (Lavrov et al., 2013), for a detailed review see Bernt et al. (2013b); Wolstenholme (1992). (ii) The existence of overlap between genes (Wolstenholme, 1992). (iii) Ill-defined 3' ends with truncated stop codons which are sometimes not entirely encoded but completed by RNA polyadenylation (Attardi, 1996). For an overview about incomplete stop codons see Nagaike et al. (2005). (iv) The use of non-canonical start codons, e.g., in *cox1* of insect mitogenomes (e.g., UCG), see Stewart and Beckenbach (2009). (v) The presence of frameshifts in the open reading frame which is described in detail in the following.

The term *frameshift* (FS) designates a position-specific change of the frame in which the ribosome reads the mRNA sequence. Frameshifts are caused by specific sequence and/or RNA secondary structure elements that program the ribosome to shift the translation in the upstream direction (programmed −1 frameshift) or in the downstream direction (programmed +1 frameshift) (Farabaugh, 1996; Dinman, 2006). The presence of frameshifts complicates the computational analysis considerably since conceptual translations are effectively randomized downstream of the frameshift position.

Several frameshifts have been found in mtDNA of animals in different protein-coding genes. The majority of reported cases are a +1 frameshift at position 174 of the *nad3* gene which has been described first for *nad3* in ostrich (Harlid et al., 1997). A frameshift at the same position (*nad3*-174) has been reported for several turtles and birds (Mindell et al., 1998; Russell and Beckenbach, 2008; Parham et al., 2006). Whereas the African helmeted turtle (*Pelomedusa subrufa*) hosts a +1 frameshift at a different position *nad3*-135 and further ones in *nad4l* at positions 99 and 262 (Zardoya and Meyer, 1998). Pancake tortoise (*Malacochersus torneri*) has been found hosting an insertion in *nad4* (Parham et al., 2006). Further +1 frameshifts have been reported for other genes in species from various phyla, e.g., different sites in *cytb* of *Polyarchis* ants (Beckenbach et al., 2005) and oyster (Milbury and Gaffney, 2005), and *cox3* and *nad6* of glass sponge (Rosengarten et al., 2008). Recently Temperley et al. (2010) have shown that a −1 frameshift at the 3' end of *cox1* and *nad6* results in the use of the standard UAG stop codon.

In this study we present an improved automated method to annotate protein-coding genes in mtDNA using phylogeny-aware hidden Markov Models (HMMs). The HMMs are constructed from enhanced multiple sequence alignments (MSA) of available amino acid sequences. The enhancement is implemented by novel methods to remove sequences and rows that do not fit, e.g., due to annotation errors, and the correction of unannotated frameshifts. Furthermore we provide a comprehensive overview of the phylogenetic distribution of the widespread *nad3* frameshifts in *Archosauria*.

2. Materials and Methods

2.1. Overview of the Workflow

The starting point is a collection of known mitochondrial protein-coding genes and a phylogenetic tree. An initial MSA and corresponding HMM is constructed for each protein-coding gene and each clade of the given phylogeny (details in Section 2.3). The constructed models for the root of the phylogeny, i.e., Metazoa, are enhanced by (i) the correction of unannotated frameshifts and (ii) the removal of poorly aligned sequences and poorly conserved columns from both ends of the MSA (details in Section 2.4).

2.2. Data sets

The initial set of training data consists of the mitochondrial protein-coding genes in `RefSeq` (Pruitt et al., 2005), more precisely the annotated CDS features. The data of the 3842 complete metazoan mtDNA sequences that are contained in `RefSeq` release 63 serve as training data set, whereas the data of those 926 species that have been added in `RefSeq` release 69 afford a large collection of test data. For details on the taxonomic distribution of both training and test data check Supplement 1. The phylogeny-aware model building process described below requires a binary phylogenetic tree. We used the NCBI taxonomy database (Benson et al., 2008) (downloaded 06-Mar-2014) as starting point. Multifurcations were replaced as in (Sahyoun et al., 2015): The branching pattern is obtained from the neighbor joining tree computed from the alignment of the *nad5* gene sequences of one randomly selected representative of each child subtree of the multifurcation point. The *nad5* gene was chosen because it has shown good performance for phylogeny reconstruction (Havird and Santos, 2014).

2.3. Construction of Initial Models

The given binary phylogenetic tree T is used as a guide tree for the progressive construction of a multiple sequence alignment (MSA) for each node of

T. To this end we use `HMMER` version 3.1b2 (Eddy, 1998). For each leaf, i.e., for the species included in the initial set of sequences, a (trivial) HMM is constructed from the single input sequence using `hmmbuild`. For the progressive step consider an interior node i of T and its two children k and l for which the corresponding HMMs H_k and H_l and MSAs A_k and A_l are already available. We first compare the set of sequences S_l in the subtree below l with H_k and correspondingly S_k with H_l using `hmmsearch` and determine which of the two models scores better in these comparisons, i.e., which of the two models yields a better mean bit score. An MSA for the sequences in S_i, which contains the sequences of S_k and S_l, is constructed on the basis of the better model using `hmmalign`. This is implemented by using the better model to extend the alignment corresponding to the better model with the sequences of the other subtree. If, for instance, H_k scores better the sequences in S_l are added to A_k using H_k. Finally the model H_i is constructed from A_i with `hmmbuild`. Traversing T in a bottom-up fashion results in an MSA and a corresponding HMM for each node. In particular, we obtain the most general model at the root of T.

2.4. Model Enhancement

The construction of the root model (Metazoa) for each gene obtained from the previous steps was guided by phylogeny. However artifacts were included in the alignments while building the initial models. The main cause is that the input sequences contain wrong or misannotated sequences, usually sequences that were annotated too short or too long. Another frequent problem are unannotated frameshifts. To cope with these effects we modify the MSA in two ways: (i) identify and correct unannotated frameshifts, (ii) remove poorly conserved sequences and poorly conserved columns at both ends of the alignment.

2.4.1. Frameshifts Correction

In order to identify frameshifts we construct all possible frameshifted variants of each gene sequence. Each variant is generated as the conceptual translation of a nucleotide gene sequence from the training set where a nucleotide at posi-

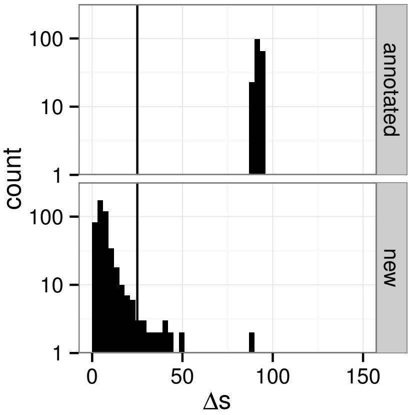

Figure 1: Distribution of bit score differences (Δs) of frameshifted and original sequence for annotated and newly detected putative frameshifts. The bold line at $\Delta s = 25$ marks the chosen threshold. Count is in log scale.

tion j is deleted. An iteration over all possible values of j yields all variants. Furthermore also all variants are generated where two consecutive nucleotides are removed. This makes it possible to identify shifts of two nucleotides and, indirectly, also missing nucleotides. Of these frameshift-translations we retain only those that do not include an early stop codon, i.e a stop codon before the end of the amino acid sequence. These are scanned with the query against the root HMM using `hmmsearch`. Since part of the frameshifted sequence is translated in a wrong reading frame, this part will not fit to the model and thus lead to a substantial decrease of the bit score. The Grubbs test (Grubbs, 1950) with $p \leq 0.01$ is employed to iteratively identify outliers from the distribution of the bit score differences of the original and variant sequences. These are plausible candidates for sequences frameshifts.

The outlier test predicted 325 putative frameshift candidates. Certainly, not all of these outliers are real frameshifts, but significant outliers are also possible for small bit score differences (Δs). Such instances are caused, for example, by sequences that are not homologous to the query at all. A comparison of the Δs values of previously annotated cases and the novel candidates, see Figure 1, suggested a threshold value of $\Delta s = 25$. At this level all previously annotated frameshifts are retained and the overwhelming majority of frameshift candidates with poor Δs are rejected.

2.4.2. Removal of Poorly Conserved Sequences and Columns

After the correction of the frameshifts, sequences that do not fit in the alignment and weakly conserved ends of the MSA are pruned. The row removal is done with the method implemented by `OD-Seq` (Jehl et al., 2015). `OD-Seq` uses a gap based distance counting the number of positions that have a gap in one sequence and not in the other. With this method rows are removed if they are outliers in the distribution of the mean gap distances with regard to the rest of the sequences. This test is implemented by a z-score threshold which was set to 3.5 standard deviations except for *nad5* and *cox3* genes where a value of 6 standard deviations was chosen because otherwise more than 1% of the sequences would have been removed. The majority of removed rows belongs to Mollusca, Arthropoda, Nematoda, and

Tunicates. The percentage of removed rows in each taxon, however, is gene specific. It never exceeds 0.6% of the sequences (Supplement 2). Weakly conserved ends of the alignment are removed by removing all columns with a low posterior probability, by default $\hat{p} \leq 0.4$, proceeding inwards from both ends of the alignment until $\hat{p} > 0.4$. This strategy removes up to 35% of the alignment. However, the average of gaps in removed columns is 92% . A final model is built from the cleaned alignment.

2.5. Annotation

The protein-coding gene annotations are obtained from the best hits of the search against the protein models. Therefore the annotation process starts with the translation of the complete unannotated mitogenomes in all six reading frames. Each of the six conceptual translations is then scanned against the final protein models with `hmmscan`. From the output the bit score and the coordinates are extracted for all hits with $E \leq 10^{-3}$. To accommodate known overlaps between mitochondrial genes (Wolstenholme, 1992), an *overlap* $< 20\%$ of the length of the shorter of two adjacent genes is tolerated in this step. For larger overlaps, the hit with larger e-value is discarded, in case of equality the hit with lowest bit score is discarded.

2.6. Benchmarking

In order to benchmark our annotation we compared the obtained predictions with the annotation available in `RefSeq` release 69. For each of our predictions the feature of the RefSeq annotation that overlaps most but by at least 10% is determined. A predicted gene is considered as equal if the overlapping pair is of the same gene, different if the pair consists of different genes, and over-predicted (OP) if no such overlap exists. Furthermore, annotated genes in RefSeq that are not included in any such pair are considered as under-predicted (UP).

To assess the quality of the generated alignments after the improvement steps, we compare the generated alignments separately with the alignments before improvement, the alignments obtained by `MAFFT` (Katoh and Standley, 2013), and `UPP` (Nguyen et al., 2015). As quality measures we employed the average percent pairwise alignments identity (APPI), the most unrelated pairwise identity (MUPI) as defined in the `Alistat` package (Eddy, 1998), and the sum of pairs score implemented (SP) in `MUSCLE` (Edgar, 2004). For a pairwise (sub)alignment of the MSA, denote by ℓ_1 and ℓ_2 the length of the two sequences and let q be the number of identities in the alignment. Then APPI is the average of the ratio, $q/min(\ell_1, \ell_2)$, MUPI is the minimum of q over the entire alignment, and SP is the sum of scores of all pairwise (sub)alignments in the MSA.

3. Results and Discussion

In order to evaluate the method, the obtained HMMs were used to annotate the mitochondrial protein-coding genes in the 926 species in `RefSeq` release 69 which were not included in the data set used to build the models (`RefSeq` release 63). On an an Intel® Core™2 Quad CPU Q9400 at 2.66GHz with 8 GB of memory the protein-coding genes can be annotated in less than 30 seconds in each of the 926 metazoan mitogenomes.

3.1. Annotation Quality

A comparison of our approach with `RefSeq` release 69 shows an agreement in 12013 out of 12132 (99%) cases. Of the remaining, 77 cases are over-predicted genes. As we will discuss below, at least 59 of them are true positives that are missing in the `RefSeq` annotations which leaves no more than 18 false positives. Additionally, there are 41 underpredicted genes, of which at least one is a true negative, 38 cases can be found by scanning the mitogenomes against more specific models (phylum, class, order, etc.). Finally, there is a single case in which our annotation differs from `RefSeq`. This case is an *atp8* in the strongylid worm *Ophagostomum columbianum* (`NC_023933`) that is predicted in the 5' region of a *cox1* in `RefSeq`. Despite a reasonable *e*-value of $6.5 * 10^{-5}$ this case is likely a false positive since an MSA with closely related species does not show conservation, see Supplement 3. Furthermore the known mitogenomes of Strongylida lack this gene. Although the enhanced models missed the annotation of 2 cases and over annotated 18 genes, the method corrected `RefSeq` in 60 cases.

Over-predicted genes. Among the 77 OP, 53 cases have an *e*-value $\leq 10^{-7}$. Among these hits 35 cases were certainly true predictions, since corresponding gene or `misc_feature` entries are present in the GenBank files but CDS features were missing whereby they are ignored by our GenBank parser. In 14 of these cases the annotations contained hints to pseudogenes. In the other 17 cases the region is annotated as non-coding, i.e., no gene was annotated, by `RefSeq`. Since the mitogenomes are compact and the alignment of these genes to the closely related species based to the general HMMs showed good quality, see Supplement 4, we are confident that these 17 cases are either pseudogenes or functional genes missed in the `RefSeq` annotation. The remaining high-scoring OP hit was the gene *nad4l* in the accession `NC_024927`. Here the product qualifier of `GenBank` entry incorrectly reads “NADH dehydrogenase subunit 2” instead of “NADH dehydrogenase subunit 4l". Among the remaining 24 cases with an e-value $> 10^{-7}$, six hits can be considered as true since they are either annotated by `RefSeq` as pseudogenes or they are annotated on the genome but no CDS is given.

Also note that, 29 of the OPs (all with e-values$\leq 8.210^{-6}$) are clustered in a few taxa: 14 cases in *Phasianus colchicus* (`NC_024152`) which has been removed in recent releases of RefSeq and 15 cases in three unpublished so called minichromosomes of *Liposcelis entomophila* (`NC_025503`, `NC_025504`, and `NC_025505`).

In summary only 18 predictions remain to be considered as OP. The 18 OP cases are shorter than the homologous query sequences and are either located in an unannotated part of the genome or inside other features (control regions or D-loop). Multiple sequence alignments with closely related species support that these predictions are false negatives. For details about OP by accession see Supplement 5.

Thus more than 75% of the OP cases are errors or misannotations in the reference which highlights the advantage of our method to overcome inconsistencies and errors in the reference database.

Under-predicted genes. The general models missed the prediction of 41 genes (UP genes). In one case, the *cox3* gene in *Loxioides bailleui* (`NC_025626`), an interval of only 18 nt is annotated in the reference; most likely this is a false positive in the reference. This accession has indeed been removed from the most recent version of `RefSeq`. The remaining 40 UP genes are distributed among the four short mitochondrial genes: *atp8* (16), *nad6* (12), *nad4l* (9), and *nad2* (3). For these cases we calculated the MSA of the UP amino acid sequences as given in the reference and closely related sequences with respect to the general model. The MSAs showed poor conservation in all cases (Supplement 6). Hence, the combination of small size and poor conservation seems to be the reason for missing these genes. We compared the length of the intersection of the UP genes with the consensus, to the average length of the intersection of the other sequences in the alignment with the consensus. In all 40 cases the

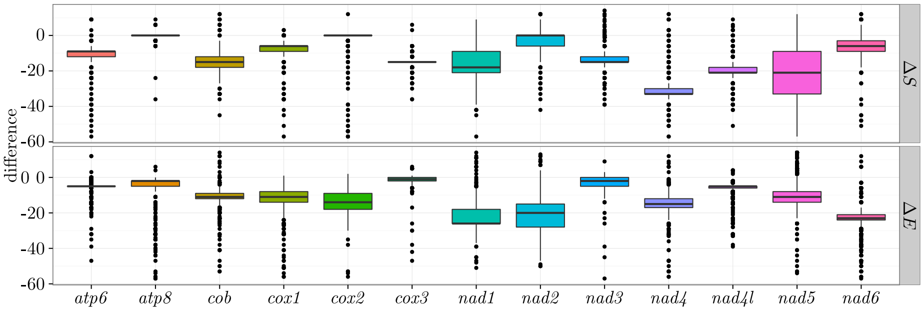

Figure 2: Differences of predicted start (ΔS) and stop (ΔE) positions with respect to `RefSeq` annotations. A positive difference corresponds to a prediction that is longer than the reference annotation.

`RefSeq` sequences have similar values as close relatives that are recognized by our approach, which indicates that they are false negatives of our approach. The majority of UP cases belong to fast evolving groups (19 Nematoda and 13 Chelicerata) although only one UP is only one in Tunicata and between zero and 2 in the remaining phyla (see Supplement 7). Likewise the UP cases are more prevalent in fast evolving genes i.e., *atp8, nad6* and *nad4l* (see Supplement 7). However, when scanning the UP genes against more specific models (closer to the leaves) almost all (38) of these cases are found. Hence choosing models of lower levels, i.e., phylum, class, or family, solves the fast evolving UP cases. Moreover other UP cases are caused by the overlap threshold mentioned in Section 2.5. The overlap problem can be solved by a column trimming strategy. Besides the considerable increase in run time from 30 sec with the root models to 743 sec with the family models. This can be solved by going to more specific models only in case of those species missing one of the mitochondrial genes. Note that, only a few more OP results from scanning more specific models (e.g., three when the family models are used) which can be eliminated by overlap with the non coding features of the mtDNA. For an overview on all the cases (equal, UP, OP, and different) check Supplement 8

3.2. Start and Stop Positions

To compare the two annotations we determined the differences in start and stop positions (ΔS and ΔE) for all genes that are present both in `RefSeq` release 69 and in our annotation. As expected, our annotations are systematically shorter due to the pruning of noisy columns from the ends of the alignments. However, the largest average ΔS and ΔE are 13.87 *nt* and 12.92 *nt* respectively, i.e., the difference is less than 5 *aa*. The distributions of ΔS and ΔE are shown in Figure 2. The differences depend systematically on the gene in question. However, the percentage of columns that is removed during the models enhancement phase is not consistent with ΔE and ΔS. For example, the start of *nad6* is trimmed more than the start (Supplement 9), yet ΔS is smaller than ΔE.

Additional automatic procedures for the precise detection of the start and stop positions would go beyond the scope of this paper since the handling of the multiple exceptions of the mitochondrial translation system, i.e., incomplete stop codons and non-canonical start codons, would be necessary. Nevertheless, even without such a method the start and stop positions are remarkably precise.

3.3. Alignment Quality

Let us first consider the effect of the improvement steps. We observe that the APPI increases in the overwhelming majority of genes between 1-4%, the MUPI gains are between 2-30%, and finally the SP score increases in all genes between 3-32 points, except for *cox1* where the SP decreases by 5. This indicates that the procedure is successful and in particular manages to identify and subsequently correct (frameshifts) or remove (wrong) divergent sequences.

Our enhanced alignments have either the same APPI as the alignments computed with `MAFFT`, or differ by 1% up or down. Again we gain with respect to measures that are sensitive to highly divergent sequences. The MUPI of `MAFFT` alignments is between 1-13% lower and the average SP is lower up to 77 (Table 1). Thus, in comparison with `MAFFT` the alignments of the majority of the genes have a higher quality with regard to the three measures.

Table 1: Alignment quality measures for the initial, enhanced, and MAFFT alignments. Shown are A: average percent pairwise alignments identity (APPI), M: most unrelated pairwise identity (MUPI), and S: sum of pairs (SP).

	Initial			Enhanced			MAFFT		
	A	M	S	A	M	S	A	M	S
atp6	44	1	223	48	7	236	47	4	211
atp8	27	0	51	29	0	54	30	0	55
cytb	65	0	658	66	23	665	66	10	656
cox1	76	0	974	80	30	969	79	17	953
cox2	61	16	343	62	18	351	62	15	342
cox3	68	2	467	71	18	499	70	7	462
nad1	59	15	430	59	19	431	60	16	416
nad2	38	1	349	40	4	355	40	5	315
nad3	52	0	151	52	12	157	52	3	145
nad4	49	0	525	48	11	533	48	10	494
nad4l	41	0	77	43	3	81	43	2	73
nad5	44	6	633	44	12	641	44	9	564
nad6	30	1	121	31	3	124	30	1	89

The comparison with UPP was possible for only one gene (*atp6*) since we were not able to install the software on our 32-bit machine. The measures show again an advantage of our alignment in the three measures. A gain in APPI of 1%, a raise of 2% in MUPI, and a raise of 7 in SP score can be observed.

3.4. Frameshifts

The frameshift detection method was applied on the data from RefSeq release 63. The method succeeded to detect all 200 frameshifts annotated in RefSeq[1]. In all these cases the location of the predicted frameshifts coincides

[1] Frameshift are indicated in RefSeq as "joined" parts of annotated CDS separated by 1 or 2 *nt*.

with the `RefSeq` annotation. With three exceptions (which were not included in the `RefSeq` set) and one difference, we recovered also all frameshifts that are mentioned in the literature (85). Annotated FS positions differ only slightly between `RefSeq` and our annotation: median 0.5, mean deviation 2.1 ± 7.29 nt, well within the range of ambiguities explained below.

The *nad3*-174 frameshift found in many but not all Archosauria-Testudines (Harlid et al., 1997; Russell and Beckenbach, 2008; Parham et al., 2006) is predicted at the position reported by Mindell et al. (1998). Other well-described examples include *nad3*-135 in *P. subrufa* (Zardoya and Meyer, 1998), *nad4* in *M. torneri* (Zardoya and Meyer, 1998), *cytb* in oyster (Milbury and Gaffney, 2005), *cox3* and *nad6* of glass sponge (Rosengarten et al., 2008). The polyarchis ants are not included in RefSeq. The double frameshift in *nad4l* of *P. subrufa* (Zardoya and Meyer, 1998) was not identified because our method only searches for cases with a single frameshift. We remark that one of two consecutive FSs in a single gene could still be reported by our method if no down-stream stop codon occurs. This is not the case in the *P. subrufa* example, however. Frameshifts that occur closer to the 3' end of the gene are harder to detect since the effect on the Δs becomes negligible. Therefore, it is not surprising that we missed the -1 frameshifts at the very end of human *cox1* and *nad6* predicted by Temperley et al. (2010).

Frameshifts in nad3. The *nad3* frameshifts events are restricted to the Archosauria-Testudines group (Figure 3). The frameshift at position 174 is widespread and is completely conserved in *Palaeognathae*. Whereas *nad3*-174 disappears completely in *Crocodilia*, *Passeriformes*, and *Pleurodira*. Some references consider *Squamata* as part of *Archosauria* (Tree of Life web project, 1996), however, it does not harbor the frameshifts. In the remaining taxa, the *nad3*-174 FS is conserved to a high degree, nevertheless it is absent in the minority of species of each group. Therefore as highlighted by Russell and Beckenbach (2008), this process of translational frameshift has been either (i) originated as a single event at *Archosauria-Testudines* and then multiple losses occurred at different sites in the tree, or (ii) the frameshifts arose independently. Note that for the cases where no *nad3*-174 frameshift was predicted also no candidate with $\Delta s < 25$ is found and also the alignment shown in Figure 3 supports the conclusion that the frameshift is absent in these sequences.

In Russell and Beckenbach (2008); Mindell et al. (1998) positions and different mechanisms were suggested for the frameshift. Both assume stalling of the ribosome while trnL is bound to `CUB` with `B` designating "not `A`", where the last base corresponds to position 172 (in the following all positions are reported w.r.t. our alignment) and the first position of the next 0-frame codon (AGN) corresponds to position 175. According to Russell and Beckenbach (2008) the mechanism for stalling is a 0-frame stop codon `AGN`, while according to Mindell et al. (1998) the cause is a stem-loop structure that starts at the `AGN` codon. Russell and Beckenbach (2008) suggested that the frameshift is caused by either "a re-pairing of the peptidyl site tRNA-Leu" (*Russel B*) with the codon 1nt downstream (i.e., frame shift at position 172) or the "occlusion of the first position of the amino-acyl site" (*Russel C*), which is the stop codon (i.e., a frameshift at position 175). The alternative explanation of Mindell et al. (1998) is that position 174 is ignored due to a +1 slippage or RNA editing. RNA editing as a possible cause was deemed unlikely by Mindell et al. (1998) and experimentally excluded by Russell and Beckenbach (2008).

The three possibilities affect only two codons (Figure 4). Due to the ambiguities of the genetic code, i.e., wobble pairing and – in this case – two codon boxes that are translated to Leu, the conceptual translations that correspond to these alternatives are equal and yield the same Δs value. Such ambiguities limit the precision with which the FS site can be located. Typically the ambiguous range covers only one or two codons. It is possible however, to construct even more ambiguous cases. In the theoretical worst case of a constant sequence such as `CCCC...` every position could be the FS site. It should be noted, however, that such ambiguities have no influence on our pipeline's ability to detect the frameshifted sequences, since it operates on the conceptually translated amino acid sequences – and these are by definition the same for all alternative FS

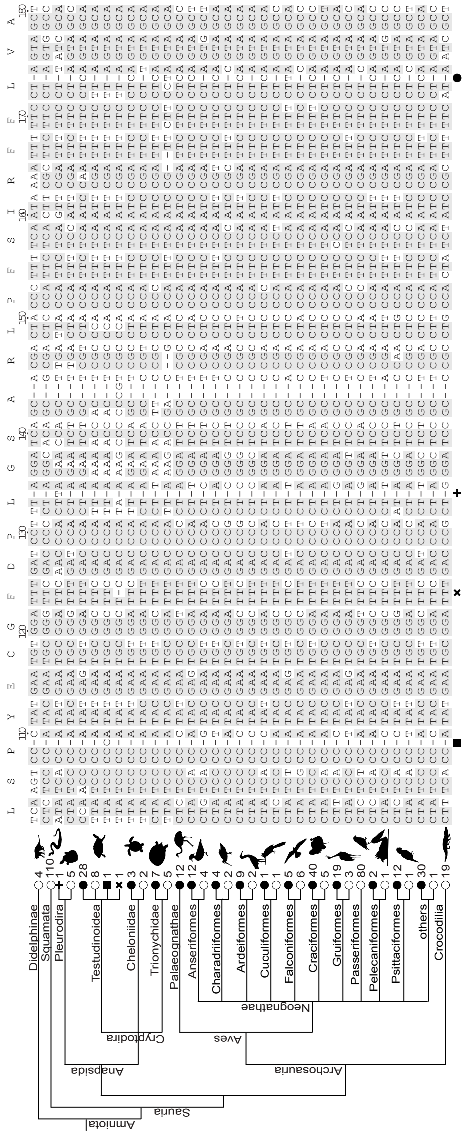

Figure 3: Distribution of the *nad3* frameshift in Archosauria and Anapsida on the NCBI taxonomy common tree (Benson et al., 2008). Symbols represent the absence (white circles) or presence of frameshifts (black circle: position 174, plus: position 134, square position 109, cross position 124). The number of species in each group is shown at the leaves. The MSA includes one random representative sequence of each group.

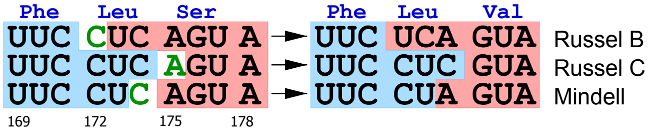

Figure 4: Alternative interpretations of the same frameshift event.

locations.

In order to find out which of the three cases is more likely a multiple sequence alignment of the Archosauria+Testudines has been created with ClustalX (Larkin et al., 2007). In contrast to the previous studies outgroup data is included (Squamata and Didelphinae as a more distant chordate group). The conservation pattern strongly favors the option suggested by Mindell et al. (1998). The sequence around the fame shift position is nearly perfectly conserved (consensus without position 174 is `TTC CTA GTA`). Also the frameshift position itself is well preserved being mostly `C` which constitute 86% of this column ignoring gaps. Moreover the `T` at position 173 is 100% conserved in all the species included in the alignment. Moreover at position 175 `A` is 91% conserved. A complete alignment can be found in the Supplement 10. Note that the position 174 is shifted to the position 181 due to other non-conserved insertions/deletions at earlier positions. Since the frameshift is in all groups more often present than absent an ancestral gain of the frameshift and infrequent loss seems to be a likely explanation.

Anapsida nad3 hosts, in addition, a non conserved frameshift mutation at three different positions 109, 124 and 134. The *nad3*-134 was annotated in `RefSeq` and described in Zardoya and Meyer (1998). The *nad3*-109 (occurs in *Cuora aurocapitata*) has been reported in (Bernt et al., 2013c). The translation in the open reading frame does not initiate an early stop codon, however the amino acid sequence translated in +1 frame is more similar to the consensus. The *nad3* gene in *Cuora aurocapitata* exhibit a deletion at position 124 and two extra nucleotides at positions 144-145.

To assess the frameshift detection method we compared our results for *nad3*

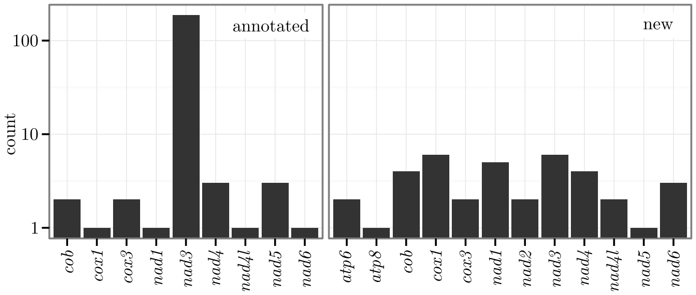

Figure 5: Number of detected frameshifts (log_{10} scale) for different genes shown separately for new (new) and already annotated frameshifts (annotated).

in the *Archosauria-Testudines* and *Anapsida* data sets to the results of `MACSE` on the same dataset (Ranwez et al., 2011). All the frameshifts detected by our method were reported also by `MACSE` (Supplement 11). However, for the FS at position 174, `MACSE` supports *Russel C* (Figure 4), i.e., an insertion at position 175. On the other hand two FS at the same positions were not detected with our method in the accessions `NC_017839` and `NC_022957` since these frameshifts would imply a stop codon before the end of the sequence. In addition, 68 frameshifts at the last position are predicted by `MACSE`. These are not detectable by our method because they do not result in a significant score difference Δs. The frameshifts are most likely spurious and are explained by incomplete stop codons (Attardi, 1996).

New frameshifts. The un-annotated frameshifts (new), are spread over all protein-coding genes except *cox2* (Figure 5). We found frameshifts in the genes: *atp8*, *atp6*, *cox1*, for which no frameshifts were previously annotated by RefSeq. These frameshifts are not phylogenetically conserved, see Supplement 12. Therefore we cannot determine whether real frameshifts have occurred in different sites in the tree, or the reading frame change is caused by sequencing or annotation errors.

To summarize, we found 98.5% of the annotated frameshifts at nearly the same positions. Furthermore we found 36 frameshifts which were not annotated in `RefSeq`. Those cases, if they are real frameshifts and not sequencing or annotation errors, shed light on the advantages of our method to automatically detect errors in the reference sequences.

4. Conclusions

In this paper an approach for the precise, automatic, and fast annotation of mitochondrial protein-coding genes has been presented. The implemented methods overcome known problems in the annotation of these genes, i.e., unannotated frameshifts, and misannotated genes. To this end we developed a fully automated pipeline for annotating mitochondrial protein-coding genes. The method creates taxon-specific hidden Markov models and their corresponding alignments from a set of annotated sequences and an approximation of the phylogeny. The pipeline incorporates several methods to improve the quality of the alignments and thereby also the quality of the model. That is, purging of sequences, removal of non conserved columns from both ends of the alignments, and correction of frameshifts. The method to detect frameshifts has been applied to all metazoan mitochondrial protein-coding genes which resulted in a large number of detected frameshifts, many of which have been unknown. A reanalysis of the frameshift in *nad3* of Archosauria-Testudines favors the position that was previously suggested by Mindell et al. (1998) instead of Russell and Beckenbach (2008).

The presented methods and the generated models are available from the authors upon request.

Future work is the precise annotation of the gene ends which is complicated by the peculiarities of the mitochondrial translations and integration with `MITOS`.

5. Acknowledgements

We thank Nam-Phuong Nguyen for kindly providing the UPP alignments for *atp6*. Marwa Al Arab was supported by Azm and Saade Association through

Lebanese University. This work was funded, in part, by the DFG project MI439/14-1.

References

Attardi, G., 1996. Mitochondrial Biogenesis and Genetics. Gulf Professional Publishing.

Beckenbach, A. T., Robson, S. K. A., Crozier, R. H., 2005. Single nucleotide +1 frameshifts in an apparently functional mitochondrial cytochrome b gene in ants of the genus Polyrhachis. Journal of Molecular Evolution 60 (2), 141–152.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Sayers, E. W., 2008. GenBank. Nucleic Acids Research 24 (37), D26–31.

Bernt, M., Bleidorn, C., Braband, A., Dambach, J., Donath, A., Fritzsch, G., Golombek, A., Hadrys, H., Jühling, F., Meusemann, K., Middendorf, M., Misof, B., Perseke, M., Podsiadlowski, L., von Reumont, B., Schierwater, B., Schlegel, M., Schrödl, M., Simon, S., Stadler, P. F., Stöger, I., Struck, T. H., 2013a. A comprehensive analysis of bilaterian mitochondrial genomes and phylogeny. Molecular Phylogenetics and Evolution 69 (2), 352–364.

Bernt, M., Braband, A., Schierwater, B., Stadler, P. F., 2013b. Genetic aspects of mitochondrial genome evolution. Molecular Phylogenetics and Evolution 69 (2), 328–338.

Bernt, M., Donath, A., Jühling, F., Externbrink, F., Florentz, C., Fritzsch, G., Pütz, J., Middendorf, M., Stadler, P. F., 2013c. MITOS: Improved *de novo* metazoan mitochondrial genome annotation. Molecular Phylogenetics and Evolution 69 (2), 313–319.

Boore, J. L., 2006. The complete sequence of the mitochondrial genome of *Nautilus macromphalus* (Mollusca: Cephalopoda). BMC Genomics 7, 182.

Bourlat, S. J., Nielsen, C., Economou, A. D., Telford, M. J., 2008. Testing the new animal phylogeny: a phylum level molecular analysis of the animal kingdom. Molecular Phylogenetics and Evolution 49 (1), 23–31.

Cameron, S. L., 2014. How to sequence and annotate insect mitochondrial genomes for systematic and comparative genomics research. Systematic Entomology 39 (3), 400–411.

Dinman, J. D., 2006. Programmed ribosomal frameshifting goes beyond viruses: Organisms from all three kingdoms use frameshifting to regulate gene expression, perhaps signaling a paradigm shift. Microbe 1 (11), 521–527.

Eddy, S. R., 1998. Profile hidden Markov models. Bioinformatics 14 (9), 755–763.

Edgar, R. C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32 (5), 1792–1797.

Farabaugh, P. J., 1996. Programmed translational frameshifting. Microbiological Reviews 30 (1), 507–528.

Grubbs, F. E., 1950. Sample criteria for testing outlying observations. The Annals of Mathematical Statistics 21 (1), 27–58.

Harlid, A., Janke, A., Arnason, U., 1997. The mtDNA sequence of the ostrich and the divergence between paleognathous and neognathous birds. Molecular Biology and Evolution 14 (7), 754–761.

Havird, J. C., Santos, S. R., 2014. Performance of single and concatenated sets of mitochondrial genes at inferring metazoan relationships relative to full mitogenome data. PLoS ONE 9 (1), e84080.

Hyouta, H., Haruhiko, M., Tatsushi, K., Takahisa, O., Izumi, K., Kin-ichiro, M., Kimitsuna, W., 1987. Unusual genetic codes and a novel gene structure for tRNA$^{Ser}_{AGY}$ in starfish mitochondrial DNA. Gene 56 (2-3), 219–230.

Iwasaki, W., Fukunaga, T., Isagozawa, R., Yamada, K., Maeda, Y., Satoh, T. P., Sado, T., Mabuchi, K., Takeshima, H., Miya, M., Nishida, M., 2013. MitoFish and MitoAnnotator: a mitochondrial genome database of fish with an accurate and automatic annotation pipeline. Molecular Biology and Evolution 30 (11), 2531–2540.

Jehl, P., Sievers, F., Higgins, D. G., 2015. OD-seq: outlier detection in multiple sequence alignments. BMC Bioinformatics 16, 269.

Katoh, K., Standley, D. M., 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Molecular Biology and Evolution 30 (7).

Larkin, M. A., Blackshields, G., Brown, N., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., et al., 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23 (21), 2947–2948.

Lavrov, D. V., Pett, W., Voigt, O., Wörheide, G., Forget, L., Lang, B. F., Kayal, E., 2013. Mitochondrial DNA of *Clathrina clathrus* (Calcarea, Calcinea): six linear chromosomes, fragmented rRNAs, tRNA editing, and a novel genetic code. Molecular Biology and Evolution 30 (4), 865–880.

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 (2), 192–199.

Milbury, C. A., Gaffney, P. M., 2005. Complete mitochondrial DNA sequence of the eastern oyster Crassostrea virginica. Marine Biotechnology 7 (6), 697–712.

Mindell, D. P., Sorenson, M. D., Dimcheff, D. E., 1998. An extra nucleotide is not translated in mitochondrial ND3 of some birds and turtles. Molecular Biology and Evolution 15 (11), 1568–1571.

Nagaike, T., Suzuki, T., Katoh, T., Ueda, T., 2005. Human mitochondrial mRNAs are stabilized with polyadenylation regulated by mitochondria-specific poly(A) polymerase and polynucleotide phosphorylase. The Journal of Biological Chemistry 280 (2), 19721–19727.

Nguyen, N. P., Mirarab, S., Kumar, K., Warnow, T., 2015. Ultra-large alignments using phylogeny-aware profiles. Genome Biology 16 (1), 124.

Parham, J. F., Feldman, C. R., Boore, J. L., 2006. The complete mitochondrial genome of the enigmatic bigheaded turtle (Platysternon): description of unusual genomic features and the reconciliation of phylogenetic hypotheses based on mitochondrial and nuclear DNA. BMC Evolutionary Biology 6, 11.

Pruitt, K. D., Tatusova, T., Maglott, D. R., 2005. NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Research 33 (suppl 1), D501–D504.

Ranwez, V., Harispe, S., Delsuc, F., Douzery, E. J. P., 2011. MACSE: Multiple Alignment of Coding SEquences Accounting for Frameshifts and Stop Codons. PLOS ONE 6 (9), e22594.

Rosengarten, R. D., Sperling, E. A., Moreno, M. A., Leys, S. P., Dellaporta, S. L., 2008. The mitochondrial genome of the hexactinellid sponge *Aphrocallistes vastus*: Evidence for programmed translational frameshifting. BMC Genomics 9, 33.

Russell, R. D., Beckenbach, A. T., 2008. Recoding of translation in turtle mitochondrial genomes: Programmed frameshift mutations and evidence of a modified genetic code. Journal of Molecular Evolution 67 (6), 682–695.

Sahyoun, A. H., Hölzer, M., Jühling, F., Höner zu Siederdissen, C., Al-Arab, M., Tout, K., Marz, M., Middendorf, M., Stadler, P. F., Bernt, M., 2015. Towards a comprehensive picture of alloacceptor tRNA remolding in metazoan mitochondrial genomes. Nucleic Acids Research 43 (16), 8044–8056.

Stewart, J. B., Beckenbach, A. T., 2009. Characterization of mature mitochondrial transcripts in drosophila, and the implications for the tRNA punctuation model in arthropods. Gene 445 (1), 49–57.

Temperley, R., Richter, R., Dennerlein, S., Lightowlers, R. N., Chrzanowska-Lightowlers, Z. M., 2010. Hungry codons promote frameshifting in human mitochondrial ribosomes. Science 327 (5963), 301.

Tree of Life web project, 1996. Squamata. Lizards and snakes. http://tolweb.orgSquamata/14933/1996.01.01inTheTreeofLifeWebProject,http://tolweb.org/, version 01 January 1996 (temporary).

Wolstenholme, D. R., 1992. Animal mitochondrial DNA: structure and evolution. Int. Rev. Cytol. 141, 173–216.

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

Zardoya, R., Meyer, A., 1996. Phylogenetic performance of mitochondrial protein-coding genes in resolving relationships among vertebrates. Molecular Biology and Evolution 13 (7), 933–942.

Zardoya, R., Meyer, A., 1998. Complete mitochondrial genome suggests diapsid affinities of turtles. Proceedings of the National Academy of Sciences of the United States of America 95 (24), 14226–14231.