Improved Automatic Annotation of Metazoan Mitochondrial Genomes

Alexander Donath[1], Fabian Externbrink[1], Frank Jühling[1], Matthias Bernt[2], Guido Fritzsch[1], Martin Middendorf[2], and Peter F. Stadler[1,3,4,5,6]

[1]Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University Leipzig, Germany
[2]Parallel Computing and Complex Systems Group, Faculty of Mathematics and Computer Science, University Leipzig, Germany
[3]Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany
[4]Fraunhofer Institut für Zell Therapie und Immunologie (IZI), Leipzig, Germany
[5]Department of Theoretical Chemistry, University of Vienna, Austria
[6]Santa Fe Institute, Santa Fe, New Mexico, USA

E-mail: {alex,fabian,frank,gfritzsch,studla}@bioinf.uni-leipzig.de {bernt,middendorf}@informatik.uni-leipzig.de
WorldWideWeb: http://www.bioinf.uni-leipzig.de http://pacosy.informatik.uni-leipzig.de

Current Situation

NCBI RefSeq is the most up-to-date source for mitochondrial genomes and their annotation. By July 13, 2009 the RefSeq contained 1,701 complete mitochondrial genome sequences of Metazoans. Large scale and automated analyses, e.g. phylogenetic and genome rearrangement analyses, require a high-quality annotation.

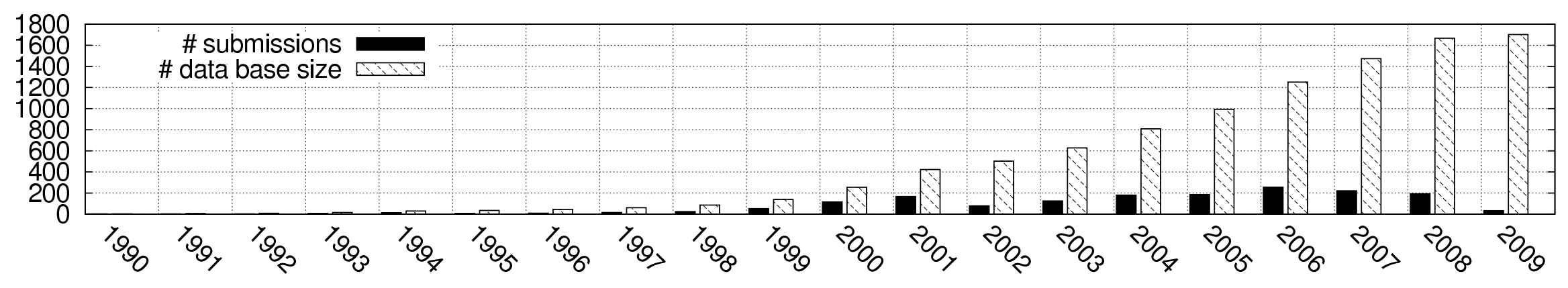


Fig. 1. Number of Metazoan mitochondrial genomes submitted and included in respectively the NCBI RefSeq database.

Though inconsistent methodology and manual annotation leads to a number of problems:

- Annotations on wrong strand
- Inconsistent gene descriptions
- Missing genes
- Genes are annotated too often
- Wrong start and end positions

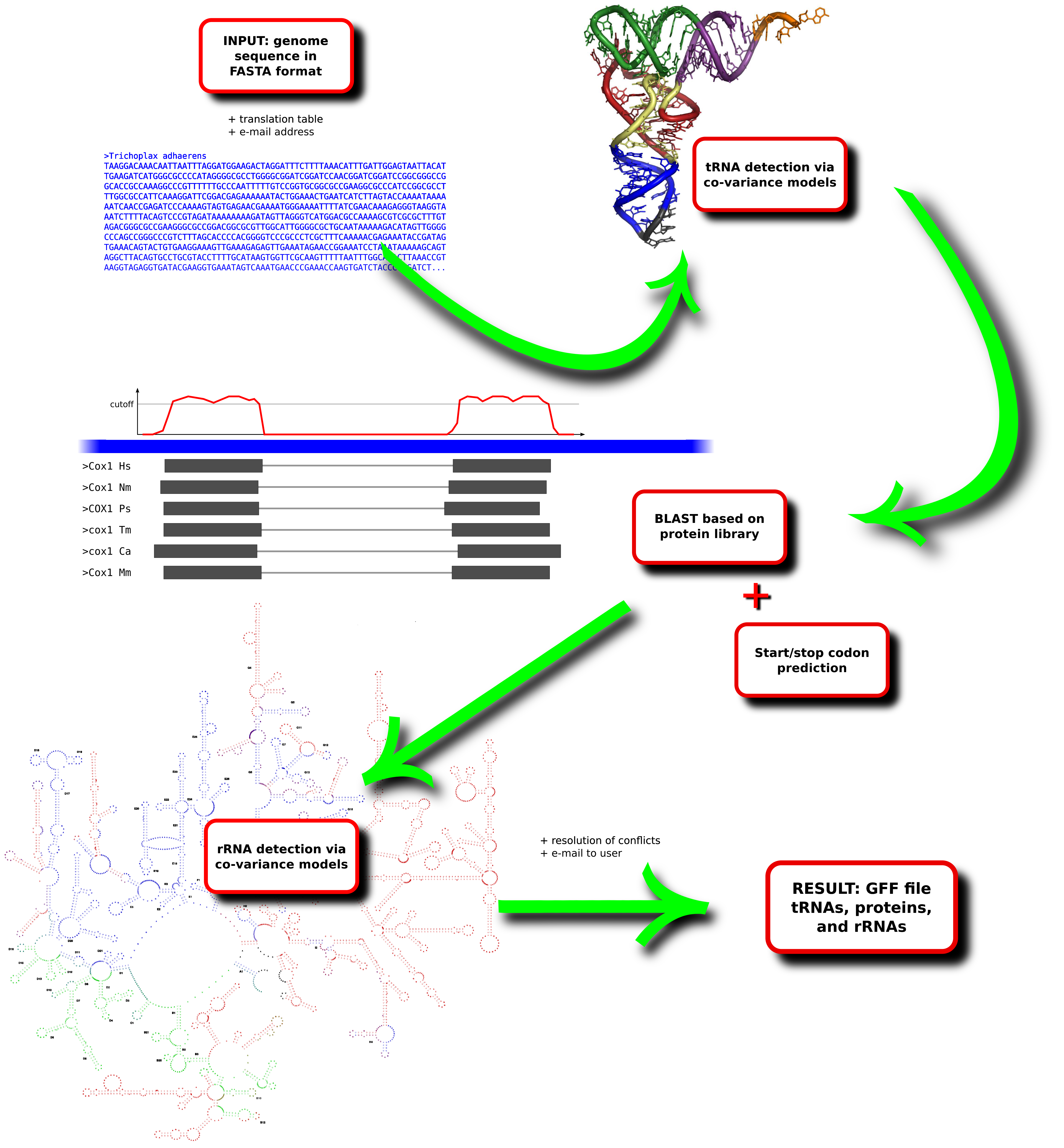
This highly influences the outcome of such analyses which rely on a correct feature annotation.

Approach

A complete automated re-annotation of the mitochondrial genomes is necessary that (a) reduces user interference to a minimum of input settings, (b) uses a standardized nomenclature, and (c) makes use of powerful, state-of the art methods.

Workflow:

1. Sequence upload in FASTA format; selection of translation table
2. tRNA detection via class-specific co-variance models (tRNAdb, INFERNAL)
3. Protein annotation (blastx) + Start and stop codon prediction
4. rRNA annotation via specific hand-curated co-variance models (European Ribosomal RNA database, INFERNAL)
5. Resolution of conflicts
6. E-mail to user with link to results


INPUT: genome sequence in FASTA format
+ translation table
+ e-mail address
tRNA detection via co-variance models
cutoff
>Cox1 Hs
>Cox1 Nm
>COX1 Ps
>cox1 Tm
>cox1 Ca
>Cox1 Mm
BLAST based on protein library
+
Start/stop codon prediction
rRNA detection via co-variance models
+ resolution of conflicts
+ e-mail to user
RESULT: GFF file tRNAs, proteins, and rRNAs

```
>Trichoplax adhaerens
TAAGGACAAACAATTAATTTAGGATGGAAGACTAGGATTTCTTTTAAACATTTGATTGGAGTAATTACAT
TGAAGATCATGGGCGCCCCATAGGGGCGCCTGGGGCGGATCGGATCCAACGGATCGGATCCGGCGGGCCG
GCACCGCCAAAGGCCCGTTTTTTGCCCAATTTTTGTCCGGTGCGGCGCCGAAGGCGCCCATCCGGCGCCT
TTGGCGCCATTCAAAGGATTCGGACGAGAAAAAATACTGGAAACTGAATCATCTTAGTACCAAAATAAAA
AATCAACCGAGATCCCAAAAGTAGTGAGAACGAAAATGGGAAAATTTTATCGAACAAAGAGGGTAAGGTA
AATCTTTTACAGTCCCGTAGATAAAAAAAAGATAGTTAGGGTCATGGACGCCAAAAGCGTCGCGCTTTGT
AGACGGGCGCCGAAGGGCGCCGGACGGCGCGTTGGCATTGGGGCGCTGCAATAAAAAGACATAGTTGGGG
CCCAGCCGGGCCCGTCTTTAGCACCCCACGGGGTCCCGCCCTCGCTTTCAAAAACGAGAAATACCGATAG
TGAAACAGTACTGTGAAGGAAAGTTGAAAGAGAGTTGAAATAGAACCGGAAATCCTA ATAAAAAGCAGT
AGGCTTACAGTGCCTGCGTACCTTTTGCATAAGTGGTTCGCAAGTTTTTAATTTGGC  CTTAAACCGT
AAGGTAGAGGTGATACGAAGGTGAAATAGTCAAATGAACCCGAAACCAAGTGATCTACC   ATCT...
```

Preliminary Results

tRNAs

~1000 tRNAs were wrongly annotated, missing, or annotated too often. More precisely, the results include:

- Correction of 203 tRNA families in 73 species (most error-prone: L1/L2 and S1/S2 tRNAs)
- Correction of multiplicity of 166 tRNAs in 107 species
- Identification of 108 missing tRNA genes
- Removal of up to 4 duplicated tRNAs in 58 cases
- Correction of the strand of 327 tRNAs in 195 species

rRNAs

We are able to identify split rRNA genes and unknown parts of rRNAs in a number of cases (confirmed by homology and EST data), such as 16S rRNA of *Trichoplax adhaerens*.

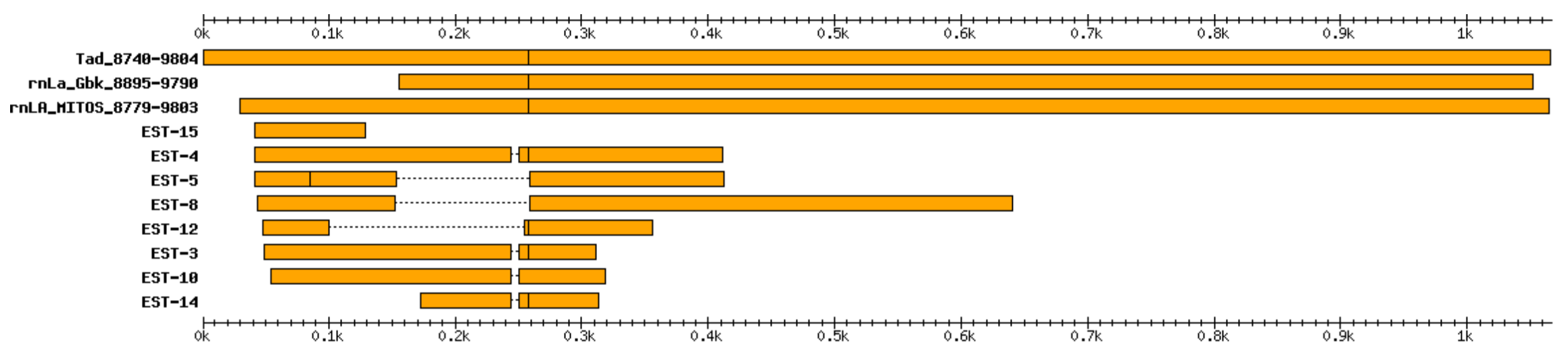


Fig. 2. Part A of the 16S rRNA of *Trichoplax adhaerens*. GenBank annotation vs. MITOS prediction vs. available EST data of *Trichoplax adhaerens* rnLA.

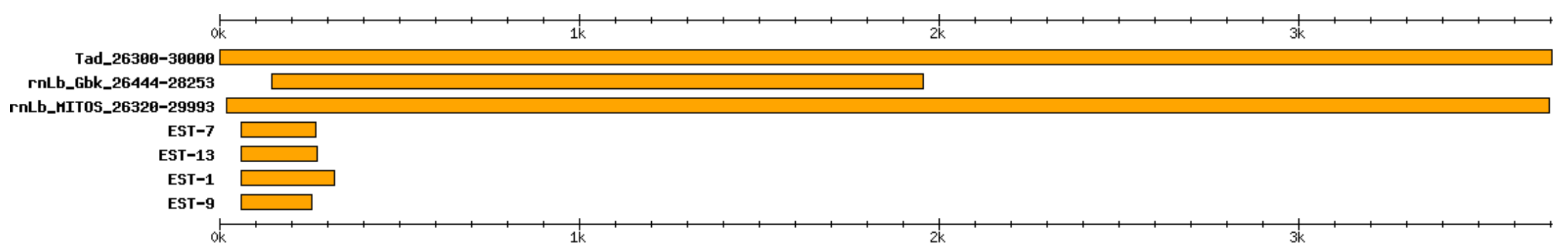


Fig. 3. Part B of the 16S rRNA of *Trichoplax adhaerens*. GenBank annotation vs. MITOS prediction vs. available EST data of *Trichoplax adhaerens* rnLB.

Proteins

We can confirm 21,981 of the 22,059 annotated proteins. Nevertheless, we can not validate 2,463 start and 8,118 stop codons. This could be caused by several factors, e.g. overlapping proteins, proteins followed by tRNAs etc..

Features in Final Version

- Graphical presentation of predictions
- Publication of completely re-annotated metazoan RefSeq mito-genomes
- Full GenBank and FASTA format output
- Group-specific rRNA models
- Automatic prediction of genomic code
- Protein annotation through HMMs and/or PSSMs
- Alternative tRNA prediction methods (tRNAscan-SE, ARWEN)
- Possibility to add novel tRNAs directly to tRNAdb

Availability:

a BETA version of the annotation server is publicly available at

http://mitos.bioinf.uni-leipzig.de/

Acknowledgements:

The authors want to thank Bernd Schierwater and Michael Eitel for *Trichoplax ad.* EST data. Financial support by the DFG SPP 1174 “Deep Metazoan Phylogeny” and SYNLET 043312.

UNIVERSITÄT LEIPZIG

Printed at the Universitätsrechenzentrum Leipzig.