Molecular phylogenetic inferences on the position of the Mollusca within the Lophotrochozoa

Dissertation
Zur Erlangung des Grades
Doktor der Naturwissenschaften

Am Fachbereich Biologie
der Johannes Gutenberg-Universität Mainz

Achim Meyer
geb. am 22.10.1965 in Essen

Mainz, 2009

Dekan: [redacted]
1. Berichterstatter: [redacted]
2. Berichterstatter: [redacted]

Tag der mündlichen Prüfung: 16.12.2009

Ludwig H. Plate (1901) S. 528: *Allgemeine Reflexionen über die Evolution der Chitonen*:

„Da man zur Zeit nicht selten der Ansicht begegnet, das Aufstellen von Stammbäumen sei eine billige und überflüssige Spielerei, so sei hier betont, dass ganz im Gegenteil solche Constructionen die logische Consequenz der Abstammungslehre sind und mit dieser stehen und fallen. Es kann sich im speziellen Falle immer nur um die Frage handeln ob schon ein genügendes Beobachtungsmaterial zu derartigen theoretischen Schlüssen vorliegt.“

This dissertation is based on the following manuscripts:

Chapter 2: Hausdorf, B., Helmkampf, M., Meyer, A., Witek, A., Herlyn, H., Bruchhaus, I., Hankeln, T., Struck, T. H., Lieb, B. (2007). *Spiralian Phylogenomics Supports the Resurrection of Bryozoa Comprising Ectoprocta and Entoprocta*. Mol Biol Evol 24, 2723-2729.

Chapter 3: Mwinyi, A., Meyer, A., Bleidorn, C., Lieb, B., Bartolomaeus, T., Podsiadlowski, L. (2009). *Mitochondrial genome sequence and gene order of* Sipunculus nudus *give additional support for an inclusion of Sipuncula into Annelida*. BMC Genomics 10, 27.

Chapter 4: Meyer, A. and Lieb, B. (submitted). *Respiratory proteins in* Sipunculus nudus *Linnaeus 1766 – implications for phylogeny and evolution of the hemerythrin family.* Comp Biochem Physiol.

Chapter 5: Meyer, A., Todt, C., Mikkelsen, N. T., Lieb, B. (submitted). *Fast evolving 18S rRNA sequences from Solenogastres (Mollusca) resist standard PCR amplification and give new insights into mollusk substitution rate heterogeneity.* BMC Evol Biol.

Chapter 6: Meyer, A., Todt, C., Lachnit, H., Witek, A, Lieb, B. (submitted). *Selecting ribosomal protein genes for phylogenetic invertebrate inferences - How many genes to resolve the Mollusca?* Mol Biol Evol.

Chapter 7: Meyer, A. (in preparation). *Realizing broad taxon sampling with large datasets: Probe detection of ribosomal proteins from cDNA libraries.*

The contributions of the different authors were as follows:

Chapter 2: Manuscript writing was done by [redacted]. I did the laboratory work on *Barentsia elongata* and *Sipunculus nudus.* I accomplished data analyses together with [redacted]. [redacted] and [redacted]. Discussion of the results was done by all authors.

Chapter 3: I generated and analysed the EST data. [redacted] did the remaining lab work. [redacted] and [redacted] analysed the PCR data and wrote the main part of the manuscript. Interpretation and discussion of the results was done by all authors.

Chapter 4: I wrote the manuscript and performed all laboratory work. I discussed the results with [redacted].

Chapter 5: I wrote the main part of the manuscript with contributions from [redacted] and [redacted]. I analysed the data and performed all laboratory work. I discussed the results together with [redacted].

Chapter 6: I wrote the manuscript, analysed the data and performed all laboratory work, but the construction of the *Lepidochitona cinerea* cDNA library was done by [redacted] and most of the PCR experiments (sampling of 12 genes) were done by [redacted]. I discussed the results together with [redacted] and [redacted].

Contributions by further persons are mentioned in the acknowledgement section of each chapter.

Eidesstattliche Erklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen direkt oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet.

Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.

Mainz, im September 2009

Achim Meyer

Contents

1. General Introduction

The term Mollusca was initially introduced by Jonston (1657), when he collated a group consisting of cephalopods and barnacles. During following times nearly every soft bodied invertebrate group had been temporarily included into the Mollusca (reviewed in: Brusca and Brusca, 2003). The last important rearrangements leading to the current genealogical understanding were the exclusion of Brachiopoda (Caldwell, 1882) and the recognition of Solenogastres as mollusks (e.g.: Thiele, 1902). Today's dispute about the inclusion of certain taxa within the Mollusca is restricted to incomplete preserved fossils for example *Kimberella* or *Wiwaxia* (Caron et al., 2007). A few autapomorphies define the Mollusca explicitly, such as the mantle with mantle grove (Ax, 2000), but most of them are not present throughout all taxa, for example the radula and the ventral differentiation into head and foot are lost in bivalves. All molluscan species exhibit the pericardium, which performs ultrafiltration and release germ cells in some taxa (Brusca and Brusca, 2003). This small coelomic cavity is a complex structure that accounts for additional autapomorphic characters such as renopericardial ducts (Bartolomaeus, 1997).

The recovery of the Mollusca is a remarkable achievement of zoologists keeping in mind the enormous variation in morphology and life strategies. To give just a random flashlight on their diversity: The body size vary between up to 20m in giant squids to less than 1mm in some interstitial slugs (Lindberg et al., 2004); there are filter feeding bivalves reaching nearly 400 year life span (Schöne et al., 2005), chitons with unique light sense organs (Eernisse, 2007) and vermiform deep sea Solenogastres with carnivorous lifestyle (Salvini-Plawen, 2008). Many mollusks are economically important as food, cultural objects, hosts for human parasites, or pests.

Eight separate extant lineages find general agreement among the scientific community. The current literature consensus tree is depicted in figure 1.1 (e.g.: Hyman, 1967; Salvini-Plawen, 1980; Lauterbach, 1983; Beesley et al., 1998; Ax, 2000; Haszprunar, 2000; Ponder and Lindberg, 2008). The two vermiform taxa Solenogastres (= Neomeniomorpha) and Caudofoveata (= Chaetodermomorpha) are generally stated as most basal groups (but see Waller, 1998) and were sometimes united as Aplacophora (Scheltema, 1993; Ivanov, 1996), whereas others reject this grouping as paraphyletic (Haszprunar, 2000; Salvini-Plawen, 2003; Lieb and Todt, 2008). The robust determination of the systematic position of Solenogastres is hindered by their mixture of plesiomorphic (e.g.: ventral fold with locomotory cilia and vermiform body) and derived features (e.g. carnivorous lifestyle) but is indispensable to understand

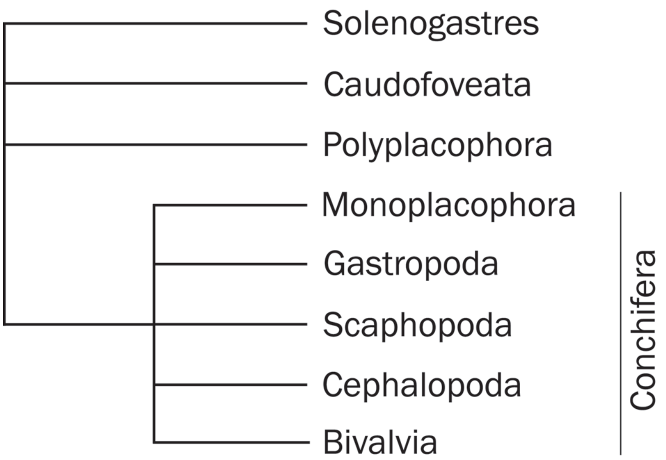


Figure 1.1: Consensus tree of molluscs inferred from the cited literature. The taxa shown here comprise all extant lineages and are considered monophyletic, but see Ponder and Lindberg (1997) and Giribet et al. (2006) for a different view regarding the position of Monoplacophora. The root is generally assumed near the vermiform molluscs Solenogastres, Caudofoveata or their last common ancestor (e.g.: Nielsen, 2001).

molluscan evolution. Agreement can be found for the monophyletic shell bearing Conchifera separated from mollusks with spicules ('Aculifera'), the latter comprising the vermiform grade or clade Aplacophora and the Polyplacophora (Salvini-Plawen, 1990; Haszprunar, 2000). Aculifera have been stated as paraphyletic assemblage of basal molluscan lineages without systematic justification (Salvini-Plawen and Steiner, 1996; Haszprunar, 2000), but see Sigwart and Sutton (2007) and Ivanov (1996) for a different view. Extant chitons (Polyplacophora) consist of two recognized lineages, where the Lepidopleuridae are considered to be basal to the Chitonida the latter covering the majority of app. 900 extant species (Okusu et al., 2003). The earliest fossilized molluscan radulae are supposed to originate from chitons (Butterfield, 2008).

Conchifera comprise monoplacophora (Tryblidia), limpets, snails and sea slugs (Gastropoda), clams and mussels (Bivalvia), tusk shells (Scaphopoda) as well as octopuses and squids (Cephalopoda). The evolutionary history within the Conchifera is unsettled. Monoplacophora are stated as either basal to the gastropod mollusks (Nielsen, 2001; Brusca and Brusca, 2003) or as basal branch to all remaining Conchifera (Scheltema, 1993; Haszprunar, 2000). Extant Cephalopoda belong either to the Coleoidea or Nautiloidea. Whereas Nautiloidea comprise only five extant species, Coleoida underwent a radiation to more than 800 exclusively marine species living today (Lindgren et al., 2004). Coleoidea can be divided in two subgroups: Decapodiformes (squids and cuttlefish) and Octopodiformes; a taxon comprising among others the well known octopuses (e.g.: Strugnell et al., 2006). Gastropods and bivalves underwent an enormous radiation leading to 62,000 described living species (Lindberg et al., 2004). Their ingroup relationships are not all-encompassing resolved but some apparently monophyletic clades in both taxa are commonly accepted. Bivalvia comprise Protobranchia, Palaeoheterodonta, Pteriomorpha, Heterodonta and Anomalodesmata (Giribet and Wheeler, 2002); and the most important subtaxa of Gastropoda are: Patellogastropoda, Vetigastropoda, Neritimorpha, Caenogastropoda and Heterobranchia (Aktipis et al., 2008). This thesis has the focus on the basal branching pattern of Mollusca and their possible sistergroup, thus a detailed introduction into the comprehensive bivalve and gastropod ingroup relationships is set aside.

Whereas the monophyly of Mollusca found general agreement since more than 100 years, the identification of their sistertaxon has a controversial debate (e.g.: Brusca and Brusca, 2003). Molecular data robustly support a clade named Lophotrochozoa (introduced by Halanych et al., 1995), a term that is not without criticism, because of the lophophorate taxa Brachiopoda, Phoronida and also Bryozoa sharing some deuterostome characters (e.g.: Schmidt-Rhaesa, 2007). Despite robust molecular support for the clade Lophotrochozoa, the phylogenetic analyses neither achieve convincing support for a distinct lophotrochozoan internal branching pattern, nor recovered the immediate sister taxon of mollusks. The problems are assumed to result from an ancient rapid speciation in the Cambrian (Adoutte et al., 2000; Passamaneck et al., 2004). In particular three trochozoan taxa are currently discussed as sistergroup of mollusks, largely based on morphological data: (i) Annelida, (ii) Sipunculida and (iii) Kamptozoa.

(i) Annelids share several characters with mollusks, but their ingroup relationships are not convincingly resolved and hamper the identification of ancestral character states (Bleidorn, 2007). Annelids possess a planktotrophic trochophoran larvae with a downstream collecting system (Nielsen, 2001), anteriorly positioned ferrous oxide structures as teeth and jaws and a cross configuration of micromeres during early development (Lindberg et al., 2004). The latter character in turn, combined with similarities of larval locomotary and feeding structures caused Scheltema (1993; 1996) to postulate a sistergroup relationship between sipunculids and mollusks.

(ii) The unsegmented peanut worms (Sipunculida) comprise about 150 species of marine worms (Cutler, 1994). They are mostly deposit feeders but some species developed a tentacular crown for filtration (*Themiste*). The body is divided in trunk and introvert and possess a nuchal organ and a ventral nerve cord. Sipunculans have a variety of epidermal structures such as papillae, hooked chaetae and shields. The intestine is organized in a coiled loop ending up dorsally on the anterior end of the trunk. Most of the diagnostic characters are nearly unmodified since Cambrian times (Huang et al., 2004). Morphological and molecular data show congruent results regarding their monophyletic status and their internal phylogeny (Maxmen et al., 2003; Schulze et al., 2007). Both studies suggested *Sipunculus nudus* as sistertaxon to all remaining sipunculids.

(iii) Kamptozoa are sessile filter feeders with a cup shaped calyx on a stalk surrounded with a whorl of ciliated tentacles. 150 species inhabit mostly marine habitats, but the species *Urnatella gracilis* (Barentsiidae) is found in freshwater (Emschermann, 1995). They are subdivided in Solitaria with the subtaxon Loxosomatidae containing app. 100 species, and the Coloniales, comprising Barentsiidae, Loxokalypodidae and Pedicellinidae (Emschermann, 1972). A number of morphological character robustly support a sistergroup relationship of Kamptozoa (=Entoprocta) and Mollusca: A chitinous body wall cuticula that is not moulted, blood sinuses, the larval ciliary gliding sole with pedal gland and a tetraneural nerveous system (Bartolomaeus, 1993; Haszprunar, 1996; Ax, 2000; Wanninger et al., 2007). In contrast Emschermann (1982) suggested a neotenic origin from an annelid larvae, thus altering the idea of a modified trochophoran larvae becoming fixed and sexually mature (Balfour, 1880; Jägersten, 1964). The latter author suggested molluscan affinities but Emschermann (1982) stated an annelid kamptozoan sistergroup relationship due to similar photoreceptors of polychaete and loxosomatid larvae. In contrast Nielsen (1971) proposed to revive the traditional subsuming of Kamptozoa and Bryozoa (introduced by Ehrenberg, 1834; Clark, 1921) based on developmental similarities.

From the outlined phylogenetic frame the position of mollusks and sipunculids is of particular importance. Associated with the reconstruction of ancestral character states is the possible segmentation of the molluscan ancestor (Jacobs et al., 2000; Nielsen, 2001; Friedrich et al., 2002; Giribet et al., 2006). Segmentation means the serial arrangement of several (!) organ systems in parallel along the anterior-posterior body axis. Annelids and arthropods are textbook examples of segmented organisms, and also chordates are segmented. Molluscs are not segmented. Though in some taxa repetitive patterns are observed and controversially discussed: Monoplacophora exhibit serial branchiae, eight retractor muscles and up to six pairs of nephridia (Schaefer and Haszprunar, 1996). Caudofoveats and chitons have eight rows of scales or plates and cephalopods have two pairs of branchiae (Nautiloidea) (Ponder and Lindberg, 2008). Are these and other structures remains of segmentation of a segmented ancestor? Hence the robust systematic position of unsegmented sipunculids is of significant importance. Morphological characters are limited in sipunculids and thus molecular data are a valuable source for phylogenetic inferences. Unfortunately single or few gene analyses failed to resolve the position of Sipuncula convincingly (e.g.: Bleidorn et al., 2006). Additionally not a single molecular study shed light on the branching pattern of the molluscan classes and even one of the best sampled metazoan genes, the SSU 18S rRNA, lacks adequate data for two molluscan class level taxa: Monoplacophora and Solenogastres. The Monoplacophora (= Tryblidia) are rare deep sea inhabitants, but recently an accessible population has been discovered and hopefully will lead to the publication of molecular data soon (Wilson et al., 2009). Difficulties had also been described to amplify the 18S from Solenogastres (Okusu and Giribet, 2003).

Due to constantly decreasing sequencing costs and technical enhancements large molecular datasets are currently assembled. These ‘phylogenomic’ analyses base predominantly on Expressed Sequence Tags (EST) and benefit from more genes and a continuously growing taxon sampling. Not just data mining but also the analytical tools used in multi gene inferences are constantly enhanced using better models (e.g. CAT model: Lartillot and Philippe, 2004) and efficient computation of thorough search strategies (e.g. Raxml: Stamatakis and Ott, 2008). Whereas the majority of published trees are based on concatenated alignments, a second possible route is to resume inferences of single gene trees using Bayesian models (Edwards et al., 2007) or parsimony (Wehe et al., 2008). These attempts are perhaps enforced by the observation that single gene phylogenies can present unresolved and conflicting branching patterns, and a closer look at the underlying signal is considered as helpful. Concatenated alignments and supertree methods usually aim to use as much data as possible (Rokas et al., 2003; Bourlat et al., 2006), but certainly all studies additionally use gene selection criteria aiming to discard paralogous genes, or at least conducting BLAST searches (Altschul et al., 1997) with a defined threshold (E-value) to identify homologues genes.

The resolution of phylogenomic analyses is much “deeper” back in time than for example 18S inferences due to a better signal to noise ratio and unambiguous alignment of amino acid residues (Philippe and Telford, 2006). The drawbacks are similar to small scale analyses mostly described as long branch effects (Felsenstein, 1978) and systematic errors (Philippe et al., 2005). Interestingly Jeffroy et al. (2006) underlined the benefit of selecting only data (=genes) that contain minimal non-phylogenetic signal takes full advantage of phylogenomics and markedly reduces incongruence.

The economic and ecological importance of bivalves and gastropods led to several published EST datasets available in the trace archives (Genbank). This database offers the opportunity to extract ribosomal protein sequence data which are well suited for phylogenetic analyses, but such data for basal mollusks and taxa possibly representing their sistergroup are lacking (GenBank 2006). To trace the evolution of the Mollusca using phylogenomic data three EST projects were generated within this study: *Sipunculus nudus* (Sipuncula), *Barentsia elongata* (Kamptozoa) and *Lepidochitona cinerea* (Polyplacophora, Mollusca) to be combined with data from public databases.

The chapters two to four within my thesis investigate the phylogenetic position of *Sipunculus nudus* (Sipuncula) using different and independent molecular markers. Chapter two: *Spiralian phylogenomics supports the resurrection of Bryozoa comprising Ectoprocta and Entoprocta* focus on the relationship of kamptozoans (*Barentsia elongata*) and bryozoans (*Flustra foliacea*). Additionally ribosomal protein genes from the generated *Sipunculus nudus* ESTs are included. To test the results of the sipunculan affinities inferred from nuclear genes sequence data, the independently inherited mitochondrial genome of *S. nudus* is examined in chapter three: *Mitochondrial genome sequence and gene order of Sipunculus nudus give additional support for an inclusion of Sipuncula into Annelida*. Supplementary to the mitochondrial genome and the ribosomal proteins one housekeeping gene family was analysed. The rise of atmospheric oxygen is closely tied to biological evolution. Since Cambrian times all mass extinction events and following times of rapid evolution are closely linked to varying O_2 regimes (Berner et al., 2007). Chapter four takes a look at expressed respiratory hemerythrin genes*: The monomeric hemerythrin protein family from Sipuncula has annelid affinity and hints to cryptic species in the cosmopolitan Sipunculus nudus, Linnaeus 1766.*

Mollusca are the second largest metazoan phylum beyond arthropods with an approximated species number of 200,000 (Heywood, 1995). This diversity demand a broad taxon sampling which is currently only feasible using single gene analyses. Three newly generated 18S sequences from the phylogenetic crucial taxon Solenogastres are analyzed together with more than 800 molluscan species from Genbank and the results are presented in chapter five: *Fast evolving 18S rRNA sequences from Solenogastres (Mollusca) resist standard PCR amplification and give new insights into mollusk substitution rate heterogeneity*. Here sequences for solitary Kamptozoa and Cycliophora are included. Cycliophora has been described with rank of a phylum comprising two commensalic species inhabiting mouth parts of crustaceans, and assumed to be related to Bryozoa or Kamptozoa (Funch and Kristensen, 1995; Obst et al., 2006).

Finally the two following chapters aim to optimize the prospective data sampling strategy adapted to molecular inferences on the evolution of the major molluscan lineages. Chapter six: *Selecting ribosomal protein genes for phylogenetic invertebrate inferences - How many genes to resolve the Mollusca?* evaluates the number and assortment of genes needed for reliable class level reconstruction of the Mollusca. Here genes of a single well defined macromolecular structure are used exclusively: The ribosome. The taxon sampling presented in chapter six is currently the most comprehensive for Mollusca, respectively Trochozoa (Annelida + Sipuncula + Echiura + Mollusca + Kamptozoa). Here the phylogenetic position of Kamptozoa is again analyzed and discussed.

Chapter seven describes continuative to chapter six a protocol for a possible technical route of data sampling. To realize the suggested broad taxon sampling DNA-probes were amplified and the large scale probe detection experiments are described in chapter seven: *Realizing broad taxon sampling with large datasets: Probe detection of ribosomal proteins from cDNA libraries*. This approach outlines the technical opportunities to expand the taxon sampling far beyond the current sampling, hopefully shedding light on some evolutionary questions regarding mollusc phylogeny.

Chapter eight: *General Discussion* summarizes the most important results of this thesis and discusses their impact on the conceptions of lophotrochozoan phylogeny and the suggested future molecular data sampling for the Mollusca.

2. Spiralian phylogenomics supports the resurrection of the Bryozoa comprising Ectoprocta and Entoprocta

Abstract

Phylogenetic analyses based on 79 ribosomal proteins of 38 metazoans, partly derived from six new EST projects for Ectoprocta, Entoprocta, Sipuncula, Annelida and Acanthocephala, indicate the monophyly of Bryozoa comprising Ectoprocta and Entoprocta, two taxa which have been separated for more than a century based on seemingly profound morphological differences. Our results also show that bryozoans are more closely related to Neotrochozoa including molluscs and annelids than to Syndermata, the latter comprising Rotifera and Acanthocephala. Furthermore, we find evidence for the position of Sipuncula within Annelida. These findings suggest that classical developmental and morphological key characters such as cleavage pattern, coelomic cavities, gut architecture and body segmentation are subject to greater evolutionary plasticity than traditionally assumed.

Introduction

With the establishment of Lophotrochozoa and Ecdysozoa (Halanych et al., 1995; Aguinaldo et al., 1997), molecular data have substantially changed our view of animal evolution. Recent phylogenomic approaches have generally sustained these hypotheses (Philippe et al., 2005; Philippe and Telford, 2006; Baurain et al., 2007) but adequate genomic data are still lacking for many minor phyla whose affinities are still in dispute (Giribet et al., 2000; Halanych, 2004) Two of the most enigmatic minor animal phyla are the moss animals, i.e., Ectoprocta and Entoprocta. When first discovered, entoprocts (Kamptozoa) were treated together with the ectoproct bryozoans because of their sessile life style and ciliated tentacles. Nitsche (1869) pointed to the differences between the position of the anus and the retractability of the tentacle crowns and proposed the names Entoprocta and Ectoprocta for the two main groups of bryozoans. Subsequently, the two groups have almost unanimously been treated as separate higher taxa, mainly based on the differences in cleavage patterns and body cavities. So far, all analyses of rDNA sequences have supported the assumption that they do not constitute sister taxa (Mackey et al., 1996; Littlewood et al., 1998; Zrzavy et al., 1998; Giribet et al., 2000; Peterson and Eernisse, 2001; Passamaneck and Halanych, 2006). However, Nielsen (1971; 1985; 2001) and Cavalier-Smith (1998) maintained the monophyly of Bryozoa in the broader sense.

To acquire molecular data sufficient for a resolution of the phylogenetic relationships of ectoprocts and entoprocts, we generated 2,000 to 4,000 expressed sequence tags (ESTs) from representatives of Ectoprocta, Entoprocta, Sipuncula, Annelida, and Acanthocephala (Table 2.1). The comparison of the six analyzed transcriptomes revealed a broad coverage of ribosomal proteins, which are valuable markers for phylogenomic analyses (Veuthey and Bittar, 1998; Philippe et al., 2004; Hughes et al., 2006; Marletaz et al., 2006) because of the rarity of known gene duplications resulting in paralogs and their conservation among eukaryotes. We compiled from our EST projects a dataset comprising 79 ribosomal proteins, which we complemented by orthologous sequences of 32 additional taxa obtained from public databases.

Materials and Methods

Isolation of RNA and library construction

Total RNA of the organisms specified in Table 2.1 was extracted from living or frozen tissue employing TRIzol (Invitrogen) or column-based methods (Qiagen RNeasy Plant Mini Kit). *Flustra* RNA was additionally purified by the RNeasy Mini Kit cleanup procedure (Qiagen), while for the purification of Barentsia RNA we applied the NucleoSpin RNA II kit (Macherey-Nagel). Quality of total RNA was visually checked on agarose gel and mRNA was subsequently captured by using the polyATract mRNA Isolation System

Table 2.1: List of investigated taxa and data used in phylogenetic analyses

Species	Taxon	Origin	# EST	# RP
Flustra foliacea (Linnaeus, 1758)	Ectoprocta	Helgoland, North Sea	4.074	77
Barentsia elongata Jullien & Calvet, 1903 [a]	Entoprocta	lab culture	2.154	47
Arenicola marina (Linnaeus, 1758)	Annelida	Sylt, North Sea	2.199	61
Eurythoe complanata (Pallas, 1776)	Annelida	lab culture	2.257	41
Sipunculus nudus Linnaeus, 1766	Sipuncula	Roscoff, France	2.329	48
Pomphorhynchus laevis (Müller, 1776)	Acanthocephala	Gimbsheim, Germany (from host *Barbus fluviatilis*)	2.207	65

Note: # EST: number of sequenced EST clones; # RP: number of ribosomal proteins retrieved at least partially from the EST datasets. Voucher specimens were deposited at the Zoological Museum Hamburg.

[a] The dataset of *Barentsia elongata* was complemented by two sequences derived from 95 ESTs of *B. benedeni* (Foettinger, 1886).

III (Promega) or Dynabead (Invitrogen) for *Sipunculus nudus*. All cDNA libraries were constructed at the Max Planck Institute for Molecular Genetics in Berlin by primer extension, size fractioning and directional cloning applying the Creator SMART cDNA Libraries Kit (Clontech) or Invitrogen's CloneMiner technology (Arenicola only), using the respective vectors pDNR-LIB or pDONR222. Clones containing cDNA inserts were sequenced from the 5' end on the automated capillary sequencer systems ABI 3730 XL (Applied Biosystems) and MegaBace 4500 (GE Healthcare) using BigDye chemistry (Applied Biosystems). If possible, clones containing ribosomal proteins from the libraries of *Barentsia elongata* and *Sipunculus nudus* were completed by reverse sequencing with polyT- and vector specific reverse primer to maximize sequence coverage.

EST processing

EST processing was accomplished at the Center for Integrative Bioinformatics in Vienna. Sequencing chromatograms were first base-called and evaluated using the Phred application (Ewing et al., 1998). Vector, adaptor, poly-A and bacterial sequences were removed employing the software tools Lucy (www.tigr.org), SeqClean (compbio.dfci.harvard.edu/tgi/software), and CrossMatch (www.phrap.org), respectively. Repetitive elements were subsequently masked with RepeatMasker. Clustering and assembly of the clipped sequences was performed using the TGICL program package (compbio.dfci.harvard.edu/tgi/software) by first performing pairwise comparisons (MGIBlast) and a subsequent clustering step (CAP3). Low quality regions were then removed by Lucy. Finally, contigs were tentatively annotated by aligning them pairwise with the 25 best hits retrieved from NCBI's non-redundant protein database using the BlastX algorithm (www.ncbi.nlm.nih.gov). Alignment and computation of the resulting match scores on which annotation was based were conducted by GeneWise (Birney et al., 2004) in order to account for frameshift errors. The EST data used in our analyses have been deposited in GenBank under the accession numbers (*Flustra*), EU116892-EU116936 (*Barentsia*), EU116844-EU116891 (*Sipunculus*), EU124931-EU124992 (*Arenicola*), EU124993-EU125033 (*Eurythoe*) and AM849482 bis AM849546 (*Pomphorhynchus*).

Sequence analyses and ribosomal proteins alignment

Ribosomal protein sequences were extracted from the newly obtained EST data by their annotation, or by using the human ribosomal proteome retrieved from the Ribosomal Protein Gene Database (ribosome.med.miyazaki-u.ac.jp) as search template during local BLAST searches (using the tblastn algorithm and an e-value < e-10 as match criterion). The observed sequences were checked for assembly errors by visual inspection and by comparison with corresponding sequences of related taxa, and translated into amino acid sequences. Orthologous sequences of *Priapulus caudatus, Ascaris suum, Aplysia californica, Idiosepius paradoxus, Macrostomum lignano, Philodina roseola, Flaccisagitta enflata,* and *Strongylocentrotus purpuratus* were obtained from public EST databases using tblastn searches also employing human sequences as query. Additional ribosomal protein data were retrieved from the alignments compiled by Baurain, Brinkmann, and Philippe (2007) and provided by H. Philippe (Université de Montréal) and complemented for missing genes. Ribosomal proteins of *Ciona intestinalis, Takifugu rubripes, Anopheles gambiae* and, in part, *Apis mellifera* were acquired directly from the Ribosomal Protein Gene Database. Sequences of *Spadella cephaloptera* were provided by F. Marlétaz (Station Marine d'Endoume, Marseille).

All ribosomal protein sequences obtained were aligned by the ClustalW algorithm (Thompson et al., 1994). The resulting 79 ribosomal protein alignments were inspected and adjusted manually. Questionably aligned positions were eliminated with Gblocks (Castresana, 2000), applying all less stringent block selection parameters available and thereafter concatenated to a single multiple sequence alignment. The concatenated alignment has been deposited in TreeBASE (http://www.treebase.org; accession S1884).

Phylogenetic analysis

ML analyses were conducted with Treefinder (Jobb et al., 2004; Jobb, 2007). The rtRev+G+F model of protein evolution was used for the ML analyses, because it was superior to other uniform models for the concatenated dataset as well as a mixed model combining separate models as determined by ProtTest (Abascal et al., 2005) for each of the 79 gene partitions according to the AICc criterion. Confidence values for the edges of the ML tree were computed by applying expected-likelihood weights (Strimmer and Rambaut, 2002) to all local rearrangements of tree topology around an edge (LR-ELW; 1,000 replications).

To test predefined phylogenetic hypotheses we used constrained trees and the 'resolve multifurcations' option of Treefinder to obtain the ML tree for a specified hypothesis. Then we investigated whether the ML trees for these hypotheses are part of the confidence set of trees applying the expected likelihood weights method (Strimmer and Rambaut, 2002).

BI analyses based on the site-heterogeneous CAT model (Lartillot and Philippe, 2004) were performed using PhyloBayes v2.1c (Blanquart and Lartillot, 2006) Two independent chains were run simultaneously for 10,000 points each. Chain equilibrium was estimated by plotting the log-likelihood and the alpha-parameter as a function of the generation number. The first 1,000 points were consequently discarded as burn-in. According to the divergence of bipartition frequencies, both chains reached convergence (maximal difference < 0.3, mean difference < 0.005), supported by the fact that both chains produced the same consensus tree topology. Taking every 10th sampled tree, a 50%-majority rule consensus tree was finally computed using both chains.

Results and Discussion

Bryozoa sensu lato: a century-old hypothesis resurrected

Phylogenetic analyses of the concatenated sequences of 79 ribosomal proteins encompassing 11,428 amino acid positions show for the first time Bryozoa as a monophyletic clade comprising Entoprocta and Ectoprocta. The monophyly is supported by strong nodal support values (Figure 2.1). Therefore, the century-old hypothesis of Bryozoa in the broader sense has to be resurrected.

Ectoprocts have been included in Lophophorata based on similarities of the tentacular apparatus and the radial cleavage they share with phoronids and brachiopods. Lophophorata was traditionally considered the sister or paraphyletic stem group of Deuterostomia (Hennig, 1979; Schram, 1991; Ax, 1995; Brusca and Brusca, 2003). However, studies employing rDNA (Halanych et al., 1995; Mackey et al., 1996; Littlewood et al., 1998; Peterson and Eernisse, 2001; Mallatt and Winchell, 2002; Halanych, 2004; Passamaneck and Halanych, 2006), Hox genes (Passamaneck and Halanych, 2004), multiple nuclear genes (Helmkampf et al., 2008) and mitochondrial protein sequences (Stechmann and Schlegel, 1999; Helfenbein and Boore, 2004; Waeschenbach et al., 2006) showed that Ectoprocta as well as Phoronida and Brachiopoda are more closely related to Annelida, Mollusca and allies than to Deuterostomia or Ecdysozoa. Therefore, Halanych et al. (1995) united them under the name Lophotrochozoa. Some of these studies further demonstrated that Lophophorata is polyphyletic (Halanych et al., 1995; Mackey et al., 1996; Littlewood et al., 1998; Giribet et al., 2000; Halanych, 2004; Passamaneck and Halanych, 2006; Helmkampf et al., 2008). On the basis of our data, the hypotheses that ectoprocts are related to Deuterostomia, that they are sister to all remaining Spiralia (Halanych et al., 1995; Littlewood et al.,

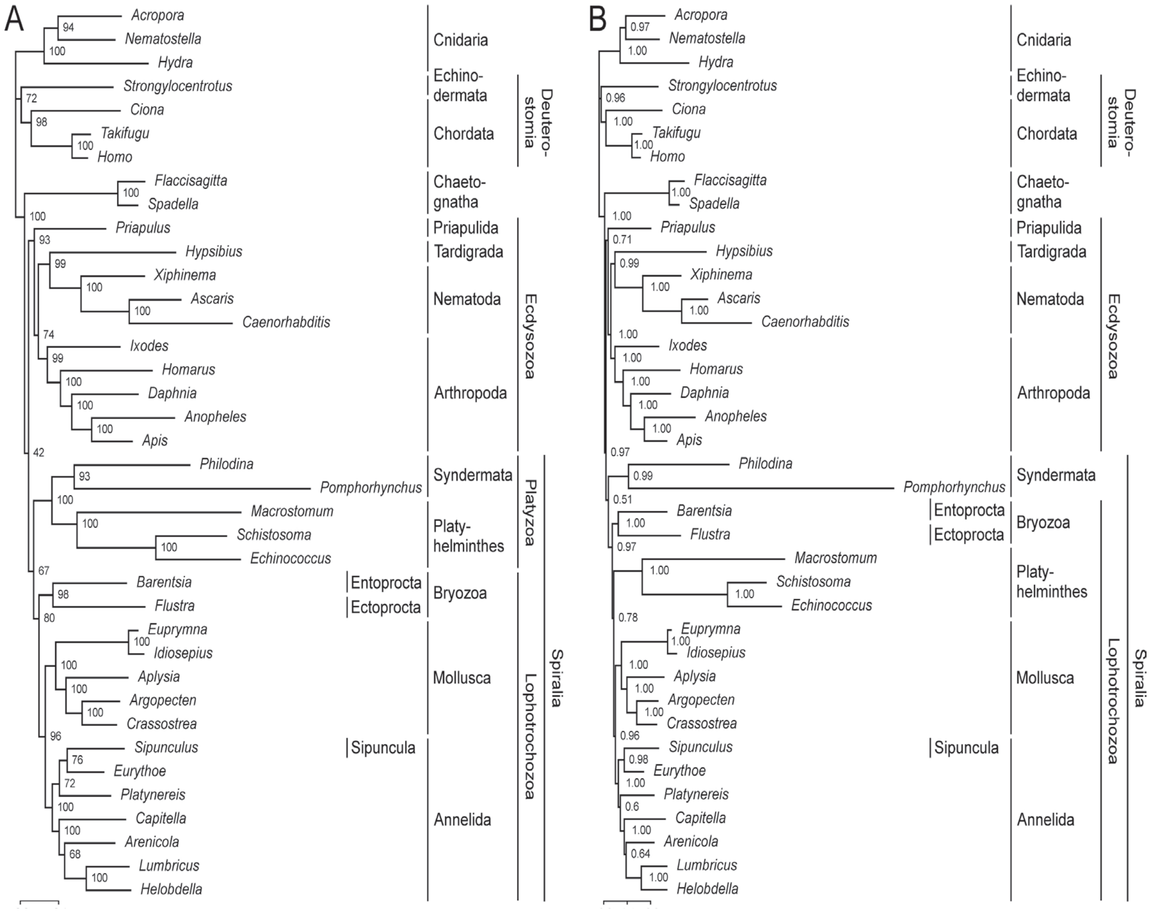


Figure 2.1: Spiralian phylogenomics unites ectoprocts with entoprocts, resurrecting Bryozoa sensu lato. Phylogenetic analyses were performed on the basis of 11,428 amino acid positions derived from 79 concatenated ribosomal proteins. (A) Maximum likelihood tree. Approximate bootstrap support values (LR-ELW) are shown to the right of the nodes. (B) Bayesian inference reconstruction. Bayesian posterior probabilities are shown to the right of the nodes.

1998; Halanych, 2004; Passamaneck and Halanych, 2006) and that they are sister to all other protostomes except chaetognaths (Giribet et al., 2000) could be rejected by topology tests (Table 2.2, hypotheses 1-3).

Entoprocts exhibit spiral cleavage and trochophora-type larvae, leading to the assumption of closer connections to taxa also possessing these features (Ax, 1995; Zrzavy et al., 1998; Giribet et al., 2000; Peterson and Eernisse, 2001). Molecular phylogenetic analyses of 18S rDNA generally confirmed the affiliation of entoprocts with taxa having trochophora larvae, but their exact relationships remained controversial (Mackey et al., 1996; Littlewood et al., 1998; Zrzavy et al., 1998; Giribet et al., 2000; Peterson

Table 2.2: Topology test results. Numbers refer to the order of appearance in the text. ELW: expected-likelihood weights test. Values for the topologies included in the 0.95 confidence set are indicated by an asterisk (i.e., expected-likelihood weights of the trees with the highest confidence levels that add up to 0.95).

Number	Phylogenetic hypothesis	References	ELW-Test
	ML tree (Figure 2.1A)		0.3452*
1	Lophophorata + Deuterostomia	Hennig (1979), Schram (1991), Ax (1995), Sørensen et al. (2000), Brusca and Brusca (2002)	0.0000
2	Ectoprocta sister to other Spiralia	Halanych et al. (1995), Halanych (2004), Passamaneck and Halanych (2006)	0.0007
3	Ectoprocta sister to other Spiralia + Ecdysozoa	Giribet et al. (2000)	0.0006
4	Lacunifera (= Entoprocta + Mollusca)	Bartolomaeus (1993), Haszprunar (1996, 2000), Ax (1999)	0.0037
5	Entoprocta + Annelida (+ Sipuncula)	Emschermann (1982)	0.0094
6	Entoprocta + Platyzoa	Halanych (2004), Passamaneck and Halanych (2006)	0.0262
7	Entoprocta + Neotrochozoa	Zrzavý et al. (1998), Giribet et al. (2000), Peterson and Eernisse (2001)	0.0804*
8	Articulata (= Annelida + Arthropoda)	Hennig (1979), Schram (1991), Ax (1999), Sørensen et al. (2000), Nielsen (2001), Brusca and Brusca (2002)	0.0000
9	Annelida monophyly (exclusive Sipuncula)	Schram (1991), Zrzavý et al. (1998), Ax (1999), Giribet et al. (2000), Sørensen et al. (2000), Nielsen (2001), Brusca and Brusca (2002), Passamaneck and Halanych (2006)	0.1000*
10	Sipuncula + Mollusca	Scheltema (1993), Zrzavý et al. (1998)	0.0000
11	Sipuncula sister to (Annelida + Mollusca)	Giribet et al. (2000)	0.0000
12	Eubilateria	Hennig (1979), Ax (1985)	0.0000
13	Chaetognatha + Deuterostomia	Ghirardelli (1981), Sørensen et al. (2000), Brusca and Brusca (2002)	0.0271*
14	Chaetognatha sister to Spiralia	Matus et al. (2006)	0.2221*
15	Chaetognatha + Ecdysozoa	Littlewood et al. (1998), Zrzavý et al. (1998), Peterson and Eernisse (2001)	0.1847*

Note: Numbers refer to the order of appearance in the text. ELW: expected-likelihood weights test. Values for the topologies included in the 0.95 confidence set are indicated by an asterisk (i.e., expected-likelihood weights of the trees with the highest confidence levels that add up to 0.95).

and Eernisse, 2001). Combined analyses of 18S and 28S rDNA data resulted in a placement within Platyzoa, but also without significant support (Passamaneck and Halanych, 2006).

With our data, most alternative hypotheses concerning the phylogenetic position of Entoprocta, in particular a sister group relationship between Entoprocta and Mollusca (Bartolomaeus, 1993; Haszprunar, 1996, 2000; Ax, 1999), a neotenic origin of entoprocts from annelids (Emschermann, 1982) and their placement within Platyzoa (Halanych, 2004; Passamaneck and Halanych, 2006) could be ruled out according to the expected-likelihood weights test (Table 2.2, hypotheses 4-6). However, a sister group relationship between Entoprocta and Neotrochozoa, which comprises Mollusca, Sipuncula and Annelida (Zrzavy et al., 1998; Giribet et al., 2000; Peterson and

Eernisse, 2001), could not be significantly rejected (Table 2.2, hypothesis 7). Nonetheless, our analyses strongly support the monophyly of Bryozoa in the broader sense including Ectoprocta and Entoprocta, and thus confirm the morphology-based argumentation of Nielsen (1971; 1985; 2001) and Cavalier-Smith (1998). Morphological data (Funch and Kristensen, 1995; Zrzavy et al., 1998; Sorensen et al., 2000) and rDNA sequences (Passamaneck and Halanych, 2006) indicate that Entoprocta and Cycliophora are sister groups. Although genomic data for Cycliophora are unfortunately still missing, we suggest to also include Cycliophora in Bryozoa sensu lato as has been done by Cavalier-Smith (1998).

Sipuncula as an annelid taxon

Both maximum likelihood (ML) (Figure 2.1A) and Bayesian inference (BI) analyses (Figure 2.1B) recovered Neotrochozoa, which comprises Mollusca, Sipuncula and Annelida, thus confirming studies using morphological and molecular data (Zrzavy et al., 1998; Giribet et al., 2000; Peterson and Eernisse, 2001). Based on segmentation, Annelida has traditionally been regarded as sister to Arthropoda (Hennig, 1979; Schram, 1991; Sorensen et al., 2000; Nielsen, 2001; Brusca and Brusca, 2003), but this so-called Articulata hypothesis is significantly rejected by topology testing (Table 2.2, hypothesis 8).

In accordance with mitochondrial amino acid sequences and gene order data (Boore and Staton, 2002; Staton, 2003; Jennings and Halanych, 2005; Bleidorn et al., 2006), our analyses indicate with strong support that Sipuncula are more closely related to Annelida than to Mollusca (Figure 2.1). More precisely, these unsegmented worms appear as a subtaxon of Annelida, which has also been suggested in some previous analyses (Peterson and Eernisse, 2001; Bleidorn et al., 2006; Struck et al., 2007). However, the monophyly of Annelida excluding Sipuncula (Schram, 1991; Zrzavy et al., 1998; Ax, 1999; Giribet et al., 2000; Sorensen et al., 2000; Nielsen, 2001; Brusca and Brusca, 2003; Passamaneck and Halanych, 2006) could not be ruled out by topology testing (Table 2.2, hypothesis 9). On the other hand, the alternative hypotheses that Sipuncula forms a monophyletic group with Mollusca (Scheltema, 1993; Zrzavy et al., 1998), and that Sipuncula is sister to Annelida plus Mollusca (Giribet et al., 2000) were rejected (Table 2.2, hypotheses 10-11).

Spiralia – Syndermata, Platyhelminthes and Lophotrochozoa

Our analyses strongly support the clade Syndermata, formed by Rotifera and Acanthocephala (Figure 2.1). This taxon has been established on the basis of morphological evidence (Ahlrichs, 1995a; 1995b; 1997) and has been further supported by analyses of 18S rDNA sequences (Garey et al., 1996; Garey et al., 1998; Littlewood et al., 1998; Zrzavy et al., 1998; Giribet et al., 2000; Herlyn et al., 2003).

The position of Platyhelminthes differs in our analyses as either being sister to Syndermata (Figure 2.1A) or to Neotrochozoa (Figure 2.1B). The former confirms the Platyzoa hypothesis. Platyzoa comprise Platyhelminthes, Syndermata, Gastrotricha and Gnathostomulida (Cavalier-Smith, 1998; Garey and Schmidt-Rhaesa, 1998; Giribet et al., 2000) and has first been hypothesized by Ahlrichs (1995a) based on sperm morphology. Platyzoa was either corroborated (Giribet et al., 2000; Passamaneck and Halanych, 2006) or contradicted (Zrzavy et al., 1998; Peterson and Eernisse, 2001) by rDNA and total evidence analyses. The lack of a robust resolution of the phylogenetic relationships of Platyhelminthes within Spiralia despite the large available dataset is probably due to increased substitution rates in Platyhelminthes and Syndermata causing long-branch attraction artifacts. However, the Eubilateria hypothesis (Hennig, 1979; Ax, 1985) can clearly be rejected by topology testing (Table 2.2, hypothesis 12). According to this hypothesis, Platyhelminthes, which do not have an anus, are considered to be the sister group of all other Bilateria possessing a one-way gut and an anus.

Lophotrochozoa is defined as including the last common ancestor of lophophorates, molluscs and annelids, and its descendants (Halanych et al., 1995). Because Bryozoa is more closely related to Neotrochozoa than to Syndermata in our analyses (Figure 2.1), syndermatans (and according to the ML analysis also platyhelminths) are not lophotrochozoans, even though to further substantiate this conclusion genomic data of Phoronida and Brachiopoda are necessary.

For the clade including Lophotrochozoa, Platyhelminthes and Syndermata, some authors have used the name Spiralia (Garey and Schmidt-Rhaesa, 1998; Giribet et al., 2000; Helmkampf et al., 2008). We follow this usage, because spiral quartet cleavage might be an autapomorphy of that taxon (see below).

Chaetognatha remain enigmatic

Chaetognatha, or arrow worms, represents the sister group of Spiralia and Ecdysozoa in our analyses (Figure 2.1). This confirms previous findings based on analyses of 18S rDNA (Giribet et al., 2000), mitochondrial DNA (Helfenbein et al., 2004), and an EST dataset (Marletaz et al., 2006). However, alternative hypotheses, namely a common ancestry with Deuterostomia (Ghirardelli, 1981; Brusca and Brusca, 2003) or Ecdysozoa (Littlewood et al., 1998; Zrzavy et al., 1998; Peterson and Eernisse, 2001) or a sister group relationship to Spiralia (Matus et al., 2006) could not be excluded (Table 2.2, hypotheses 13-15). The phylogenetic position of chaetognaths thus remains elusive.

Implications for character evolution

Cleavage pattern was often considered a key character for the reconstruction of metazoan phylogeny. Typical spiral quartet cleavage with mesoderm formation by the 4d mesoteloblast or one of its daughter cells (Sorensen et al., 2000; Nielsen, 2001) is known from several lophotrochozoan groups (Mollusca, Annelida, Nemertea, Entoprocta), Platyhelminthes, and Gnathostomulida. If we map this character state on our tree (Figure 2.1) considering the close relationship of Syndermata to Gnathostomulida (Ahlrichs, 1995a; Ahlrichs, 1995b; Ahlrichs, 1997; Cavalier-Smith, 1998; Garey and Schmidt-Rhaesa, 1998; Giribet et al., 2000; Sorensen et al., 2000; Nielsen, 2001) it turns out to be a possible autapomorphy of the clade including Syndermata, Plathyhelminthes and Lophotrochozoa, for which we accepted the name Spiralia, although it has been secondarily modified several times within this clade (e.g., in Syndermata, Neoophora, Ectoprocta, Brachiopoda, Cephalopoda). The sister group relation of ectoprocts and entoprocts demonstrates that the transition from spiral to radial cleavage can happen within a clade without any transitional stages being preserved. After all, the different cleavage types were one of the main reasons that the two taxa were classified in different major groups for more than a century.

Often coelomic cavities were considered an autapomorphy of a clade Coelomata (Hennig, 1979; Blair et al., 2002; Philip et al., 2005). If the coelomic cavities of lophotrochozoans are considered homologous to those of deuterostomes and to the small coelomic cavities present in some ecdysozoans, our trees would indicate a frequent reduction of coelomic cavities in several bilaterian lineages (e.g., in chaetognaths, priapulids, nematodes, platyzoans, entoprocts). However, the differing developmental origin of coelomic cavities in the different bilaterian lineages cast doubts on the homology of the coelom across bilaterians (Nielsen, 2001).

The significant rejection of the Eubilateria hypothesis, and the derived position of platyhelminths within Spiralia indicates that the anus has been secondarily reduced in platyhelminths, in which the mouth is the only opening to the intestinal system.

Finally, the significant rejection of Articulata as well as the derived position of Annelida within Spiralia supports the hypothesis that segmentation originated convergently in annelids and arthropods. The placement of unsegmented worms within Annelida, namely Sipuncula (this study, Peterson and Eernisse, 2001; Bleidorn et al., 2006; Struck et al., 2007) and Echiura (McHugh, 1997; Bleidorn et al., 2003; Struck et al., 2007), further reveals that segmentation has been secondarily lost in annelid subtaxa. Sipunculans possess a U-shaped gut, a feature already established in Cambrian fossils (Huang et al., 2004). The movement of the anus in the anterior direction requires the disorganisation of segmentation, a factor that may have eased inhabiting holes in solid substrates.

The results presented herein therefore indicate that several of the supposed key characters of animal phylogeny such as cleavage pattern, coelomic cavities, body segmentation and gut architecture are much more variable during evolution than previously thought.

Acknowledgments

We would like to thank [redacted] (University Freiburg) for providing us with samples of the entoprocts *Barentsia elongata* and *B. benedeni*, and C. Nielsen (University Copenhagen) and an unknown referee for helpful comments on the manuscript. We are also grateful to [redacted] and [redacted] (Max Planck Institute for Molecular Genetics, Berlin) for the construction and sequencing of cDNA libraries, and to [redacted] (Max F. Perutz Laboratories, Center for Integrative Bioinformatics, Vienna) for the processing of our ESTs. This study was funded by the priority program "Deep Metazoan Phylogeny" of the Deutsche Forschungsgemeinschaft (grants HA 2103/4-1 to TH; HA 2763/5-11 to BH and IB; LI 338/3-1 to BL; STR 683/3-1 to THS).

3. Mitochondrial genome sequence and gene order of *Sipunculus nudus* give additional support for an inclusion of Sipuncula into Annelida

Abstract

Background: Mitochondrial genomes are a valuable source of data for analysing phylogenetic relationships. Besides sequence information, mitochondrial gene order may add phylogenetically useful information, too. Sipuncula are unsegmented marine worms, traditionally placed in their own phylum. Recent molecular and morphological findings suggest a close affinity to the segmented Annelida.

Results: The first complete mitochondrial genome of a member of Sipuncula, *Sipunculus nudus*, is presented. All 37 genes characteristic for metazoan mtDNA were detected and are encoded on the same strand. The mitochondrial gene order (protein-coding and ribosomal RNA genes) resembles that of annelids, but shows several derivations so far found only in Sipuncula. Sequence based phylogenetic analysis of mitochondrial protein-coding genes results in significant bootstrap support for Annelida sensu lato, combining Annelida together with Sipuncula, Echiura, Pogonophora and Myzostomida.

Conclusion: The mitochondrial sequence data support a close relationship of Annelida and Sipuncula. Also the most parsimonious explanation of changes in gene order favours a derivation from the annelid gene order. These results complement findings from recent phylogenetic analyses of nuclear encoded genes as well as a report of a segmental neural patterning in Sipuncula.

Background

Molecular sequence analysis has become the method of choice to address phylogenetic questions. The applied techniques improve continually and the rapidly growing amount of available data helps to broaden our knowledge of phylogenetic relationships within the animal kingdom. Nevertheless, different molecular datasets often show conflicting phylogenetic signals, so that results relying on just one dataset may be interpreted with caution (Rokas et al., 2003).

Unlike nuclear DNA, the mt-genome of animals is normally rather small and simply structured: haploid, without or only few non-coding segments, repetitive regions and transposable elements. Derived from endosymbiotic bacteria only a few genes are retained in the mitochondrial genomes of Bilateria: 13 protein subunits (nad1-6, nad4L, cox1-3, cob, atp6/8), 2 ribosomal RNAs (rrnL, rrnS) and 22 tRNAs are found encoded on a circular doublestranded DNA molecule sized about 15 kb (Wolstenholme, 1992; Boore, 1999). As such sequencing and annotation of mt-genomes is much easier and faster than analysing nuclear genomes, making mt-genomes one of the commonly used sources of sequence data for phylogenetic analyses. Apart from sequence data other features of the genome may contain phylogenetic information, too. Taxon-specific gene order often remains identical over long periods of time (Boore et al., 1995; Shao et al., 2004; Valles and Boore, 2006). Simultaneously, the intra-taxonomic variances of these characteristic orders are quite distinctive and convergent changes in the positioning of single genes are rather unlikely, due to the vast number of possible combinations (Dowton et al., 2002). Thus changes in the mitochondrial gene order have proved to be valuable tools in phylogenetic analyses (Boore et al., 1998; Lavrov et al., 2004; Bleidorn et al., 2007). Less often secondary structures of tRNAs or rRNAs show distinct differences between taxa (e.g. loss of a stem/ loop region) and hence may also contribute to a phylogenetic analysis (Haen et al., 2007).

The taxon Sipuncula (peanut worms) comprises about 150 species, being found in all water depths of different marine habitats. The hemisessile organisms dwell in mud and sand, but settle also in empty mollusc shells or coral reef clefts for instance. Their body shows no segmentation, but a subdivision into a posterior trunk and an anterior introvert that can be fully retracted into the trunk is observeable (Cutler, 1994b). Fossils that date back into the later Cambrian (Huang et al., 2004) suggest that sipunculans have undergone little morphologically change over the past 520 Myr. The monophyly of this morphologically uniform taxon is well founded by morphological (Ax, 2000) and molecular data (Schulze et al., 2007). However, the phylogenetic position within Bilateria was highly disputed. Based on morphological characters, very different phylogenetic positions of Sipuncula were discussed. Early in history an affinity to Echinodermata, especially holothurians was mentioned and later again propagated by Nichols (1967), but with little acceptance from other authors. Scheltema (1993) proposed a close relationship to molluscs based on the presence of the so calles “molluscan cross” organization of micromeres during spiral cleavage. The usefulness of this character for phylogenetic inference was neglected by Malaskova (2004). Other analyses found Sipuncula to be sister group of Mollusca, Annelida and Arthropoda (Nielsen, 2001), Articulata (Annelida and Arthropoda) (Ax, 2000), Echiura (Meglitsch and

Schram, 1991), Mollusca (Brusca and Brusca, 2003), Annelida (Erber et al., 1998) or Annelida+Echiura (Eernisse et al., 1992). More details about the different hypotheses of sipunculid relationships are reviewed in (Schulze et al., 2005a).

In contrast to all these studies, molecular analyses of large datasets from 18S/28S data (Struck et al., 2007), ESTs (Hausdorf et al., 2007; Dunn et al., 2008) or mitochondrial genome data (Boore and Staton, 2002; Bleidorn et al., 2006) favour an inclusion of Sipuncula into annelids. An implication of this hypothesis is that we have to assume that segmentation has been reduced within Sipuncula (Bleidorn, 2007). A derivation from segmented ancestors of Sipuncula was recently also supported by a segmental mode of neural patterning in ontogeny (Kristof et al., 2008).

Relationhips within Sipuncula are well investigated (Cutler and Gibbs, 1985; Staton, 2003; Schulze et al., 2005b; Schulze et al., 2007). An analysis using combined molecular and morphological data recovered five major clades and supports that *Sipunculus* is the sister group to all other sipunculids (Schulze et al., 2007).

Up to now mt-genome data from Sipuncula was restricted to a partial mtDNA sequence from Phascolosoma gouldii (Boore and Staton, 2002), comprising only about half of the complete genome. Here we describe the first complete mitochondrial genome for another representative of the Sipuncula, *Sipunculus nudus*. We analyse sequence data in comparison with mitochondrial genomes of various Bilateria to evaluate the phylogenetic position of Sipuncula. In addition we compare gene order among Lophotrochozoa and evaluate the most parsimonious explanation for gene order changes.

Results and discussion

Genome organisation

The complete mt-genome of *S. nudus* is a circular DNA doublestrand of 15502 bp length. As usual in bilateria, 13 genes coding for different protein subunits and two encoding ribosomal RNA genes were identified. In addition 22 tRNA genes were detected and thus all 37 genes typically present in bilaterian mt genomes, were found (Figure 3.1, Table 3.1). All of these genes are located on the (+)- strand, as is the case in annelid and echiurid mt-genomes. There are two small gene overlaps: one between nad4L and nad4 (7 bp), the other one between trnS (AGN) and nad2 (1 bp). The putative control region is 441 bp in length and flanked by trnF and trnT. Besides the control

Table 3.1: Genome organisation of *Sipunculus nudus*. Complete circular mtDNA has a lenght of 15502 bp.

Gene	Strand	Position (start – end)	Length (nuc.)	GC-/AT-skew	Start-codon	Stop-codon	Intergenic bp
trnN	+	1556 – 1624	69	-0.24/-0.07			0
cox2	+	1625 – 2319	695		ATG	TA	0
trnD	+	2320 – 2385	66	-0.26/-0.07			0
atp8	+	2386 – 2544	159		ATG	TAG	2
trnY	+	2547 – 2609	63	-0.38/0.12			36
trnE	+	2646 – 2712	67				1
trnG	+	2714 – 2780	67				0
cox3	+	2781 – 3560	780		ATG	TAA	4
trnQ	+	3565 – 3632	68	-0.27/-0.07			0
nad6	+	3633 – 4106	474		ATG	TAG	1
cob	+	4108 – 5247	1140	-0.34/-0.18	ATG	TAA	7
trnP	+	5255 – 5322	68	-0.30/-0.06			0
trnS-UCN	+	5323 – 5389	67				5
trnC	+	5395 – 5455	61				5
trnM	+	5461 – 5527	67				0
rrnS (12S)	+	5528 – 6373	846				0
trnV	+	6374 – 6442	69	-0.23/0.18			0
rrnL (16S)	+	6443 – 7929	1487				0
trnL-CUN	+	7930 – 7995	66	-0.26/0.08			7
trnA	+	8003 – 8070	68				0
trnI	+	8071 – 8139	69				0
trnK	+	8140 – 8207	68				0
nad3	+	8208 – 8565	358		ATG	T	2
trnF	+	8568 – 8631	64	-0.35/-0.05			0
Major NCR	+	8632 – 9072	441				0
trnT	+	9073 – 9142	70	-0.14/0.09			0
nad4L	+	9143 – 9424	282		ATG	TAA	-7
nad4	+	9418 – 10774	1357	-0.43/-0.06	ATG	T	7
trnL-UUR	+	10782 – 10846	65	-0.37/-0.02			0
nad1	+	10847 – 11789	943	-0.32/-0.09	ATG	T	0
trnW	+	11790 – 11855	66				7
atp6	+	11863 – 12550	688	-0.41/-0.14	ATG	T	0
trnR	+	12551 – 12619	69				1
trnH	+	12689 – 12688	68				39
nad5	+	12728 – 12727	1698	-0.37/0.01	ATA	TAA	21
trnS-AGN	+	14447 – 14518	72				-1
nad2	+	14518 – 15502	985	-0.45/-0.13	ATG	T	0

* start and stop position of ribosomal RNA and NCR according to adjacent gene boundaries

region 15 other non-coding regions are dispersed over the whole genome, ranging from one to 39 base pairs. The three largest of these are located between trnY and trnE (35 bp), trnH and nad5 (39 bp) and nad5 and trnS (AGN) (21 bp).

The GC-skew [(G-C)/(G+C)] reflects the relative number of cytosin to guanine and is often used to describe the strand-specific bias of the nucleotide composition (Perna and Kocher, 1995). In *S.nudus* the complete (+)-strand genome sequence has a clear bias toward Cytosine (GC-skew -0.296). As all genes are coded on (+)-strand, all single gene sequences exhibit a negative GC-skew, too (Table 1), ranging from -0.23 (rrnS) to -0.45 (nad2). A negative GC-skew is also found in most of the mitochondrial genomes known from annelids, pogonophorans, and myzostomids, with the exception of the annelid Eclysippe vanelli (Zhong et al., 2008). AT-skew of the complete (+)-strand is close to evenness (-0.013) and single gene AT-skews are distributed around evenness with a range between 0.18 (rrnS) and -0.18 (nad6), see also Table 1. AT content of the complete genome is 54.2%, AT contents of protein-coding and rRNA genes are not much derived from this value, between a minimum of 50,3% (nad3) and a maximum of 59,8% (atp8).

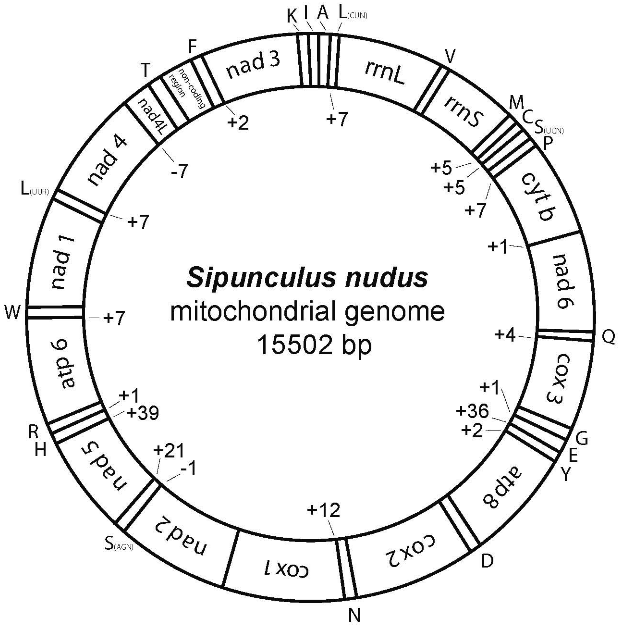


Figure 3.1: Circular map of the mitochondrial genome of *Sipunculus nudus.*

Protein coding genes

All but one of the protein subunits begin with start codon ATG, only nad5 starts with ATA. Both are prevalent in mitochondrial genomes. The commonly found stop codons TAA and TAG are present, as well as the abbreviated forms TA (cox2) and T (nad1-4, atp6). Putative shortened stop codons were already found in other species and are thought to be complemented via post-transcriptional polyadenylation (Ojala et al., 1981).

Ribosomal RNA genes and control region

The sizes of the ribosomal RNAs (rrnS: 846 bp; rrnL: 1487 bp) are within the range of their sizes in other animals including molluscs and annelids. The two genes only separated by trnV, a feature often found in animals from vertebrates to arthropods, so therefore this represent an ancestral condition. Among annelids and their kin only echiurans (*Urechis caupo*) and myzostomids (*Myzostoma seymourcollegiorum*) differ from that condition in that there is no tRNA gene separating the two ribosomal genes. AT content of ribosomal genes is 50.8% (rrnS) and 53.1% (rrnL), so well within the range of AT content of proteincoding genes.

Noncoding regions, putative control region

The putative control region is found between nad3/trnF on one side and trnT/nad4L/nad4 on the other side. While gene order (or protein-coding and rRNA genes) in Annelida is more or less conserved there is a great variation in the position of the control region: (a) Species from Clitellata, Maldanidae and Terebellidae have a major noncoding region between atp6/trnR and trnH/nad5; (b) in *Orbinia* it is located between nad4/trnC/trnL2 and trnL1/ trnM/rrnS; (c) in *Platynereis* it is found between cox2/trnG and trnV/atp8 (Boore and Brown, 2000; Bleidorn et al., 2006; Bleidorn et al., 2007; Zhong et al., 2008). Such great variability is not found in other taxa like Arthropoda or Vertebrata, where also the control region is found in the same position in different species, when gene order of the rest of the mtgenome is conserved. In *Sipunculus nudus* the major non-coding region has a size of 441 bp and is clearly more AT rich (66.1%) than the rest of the genome (53.9%). Structural elements know from arthropod mitochondrial control regions (Zhang and Hewitt, 1997) are present also in *S. nudus*: (1) a poly-TA(A) stretch of 50 bp including a tenfold TA repeat; (2) a poly-T stretch flanked by purin bases; (3) a GA-rich block of 16 bases length. Although we examined the complete non-coding region intensively by software and by eye, no large stem-loop structure was identified. Such a structure is normally found between the poly-T stretch and the GA rich region in arthropods.

Transfer RNAs

All typical 22 tRNAs were detected in the mitochondrial genome of *S. nudus*, their putative secondary structures are depicted in Figure 3.2. All but three tRNA genes are capable to be folded in the usual cloverleaf structure, consisting of TψC stem and loop, anticodon stem and loop, DHU stem and loop, and the acceptor stem – tRNA-Ser(AGN) and tRNA-Ser(UCN) have no DHU stem. While tRNA-Ser(AGN) shows this feature in many bilaterian mt-genomes, the other one must have changed its secondary structure in the lineage leading to Sipuncula and after the split of its sister group. The putative secondary structure of tRNA-Cys shows no TψC, in addition there are two mismatches in the anticodon stem and an unusual anticodon (ACA), weakening this secondary structure hypothesis. But intensive search for an alternative sequence of

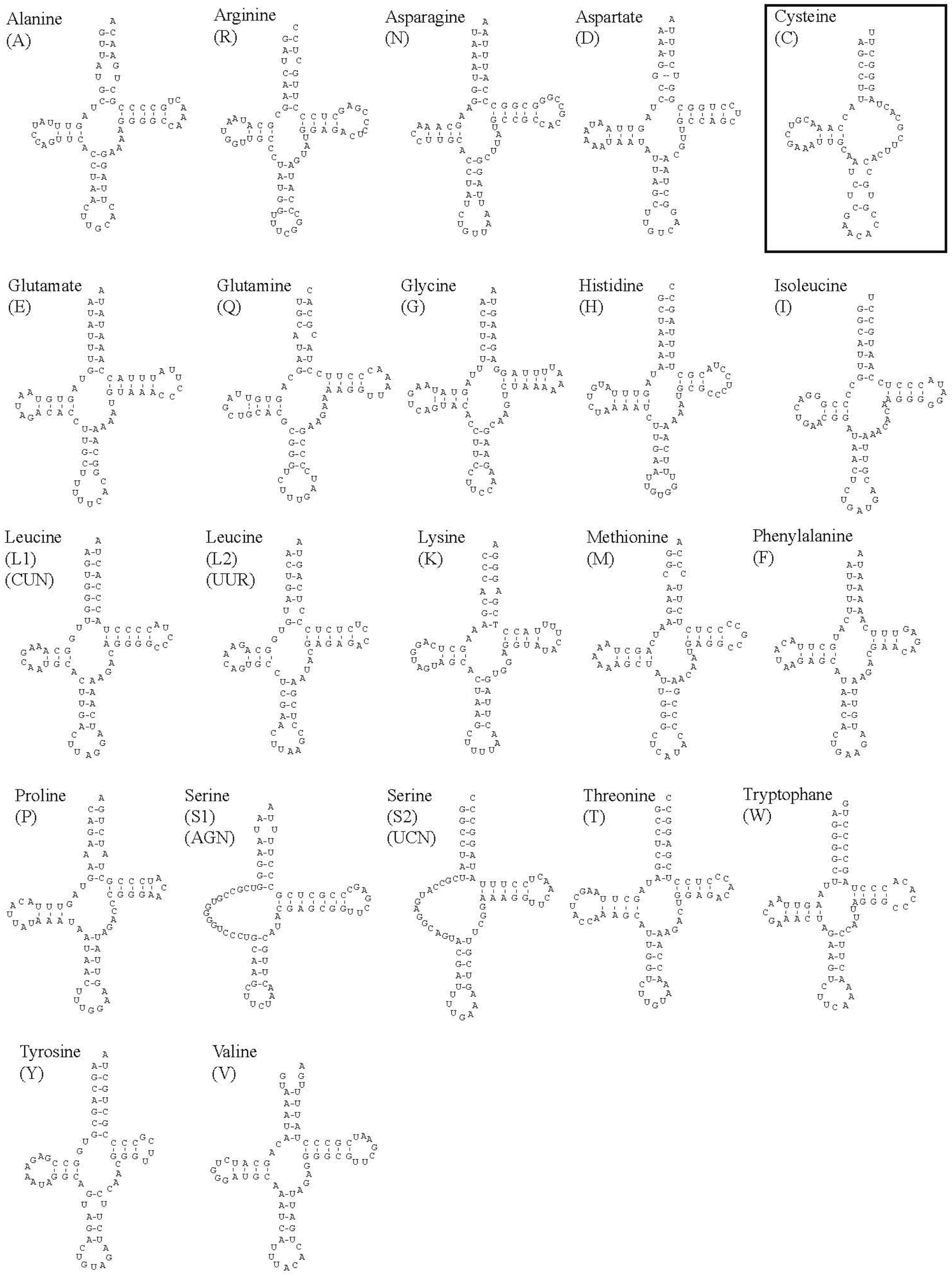


Figure 3.2: Secondary structure of tRNAs identified in the mitochondrial genome of *S. nudus*. The best found putative secondary structure of tRNA-Cys (box) seems to be strongly derived, probably non-functional or subject to gene editing.

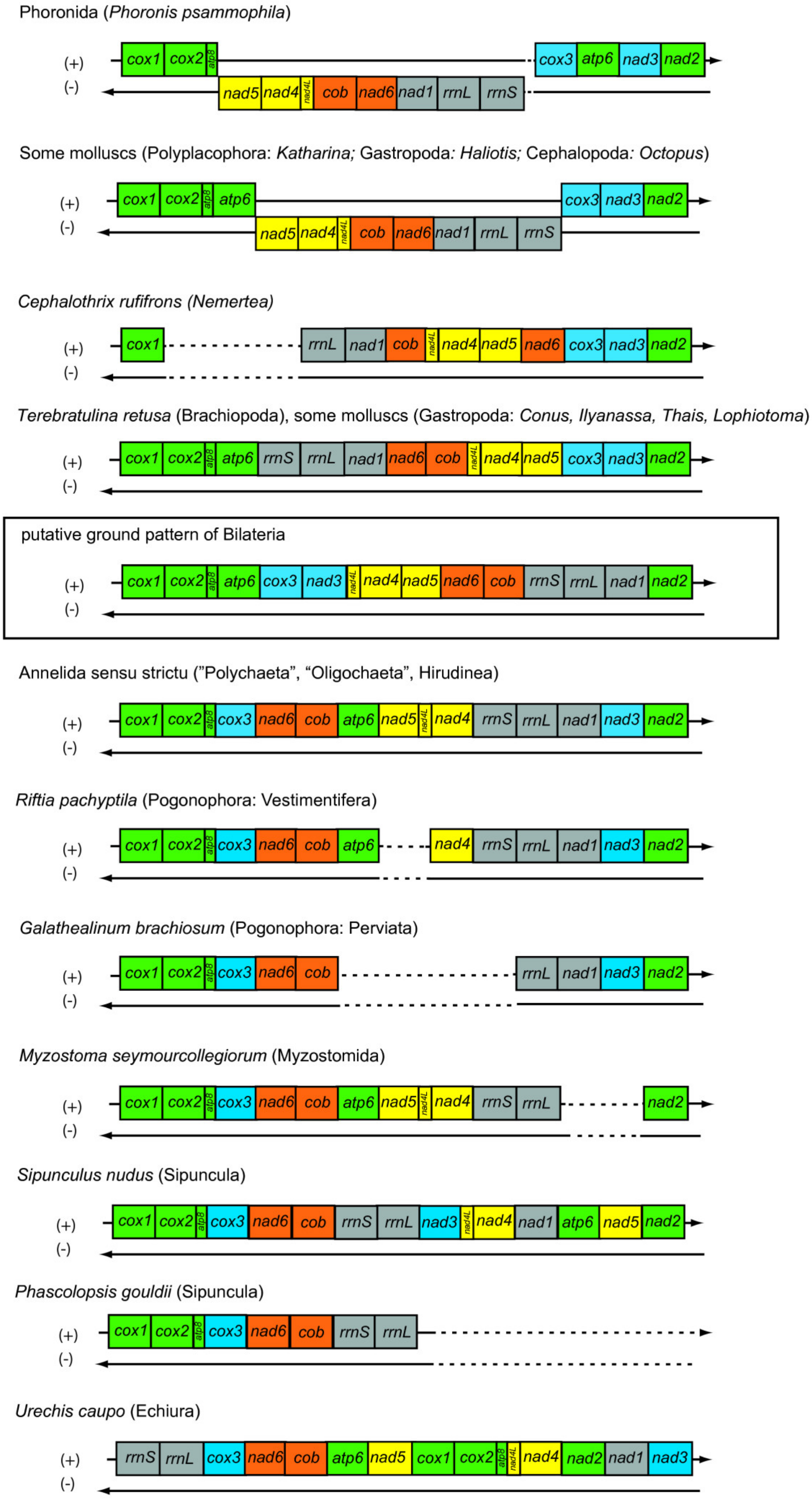
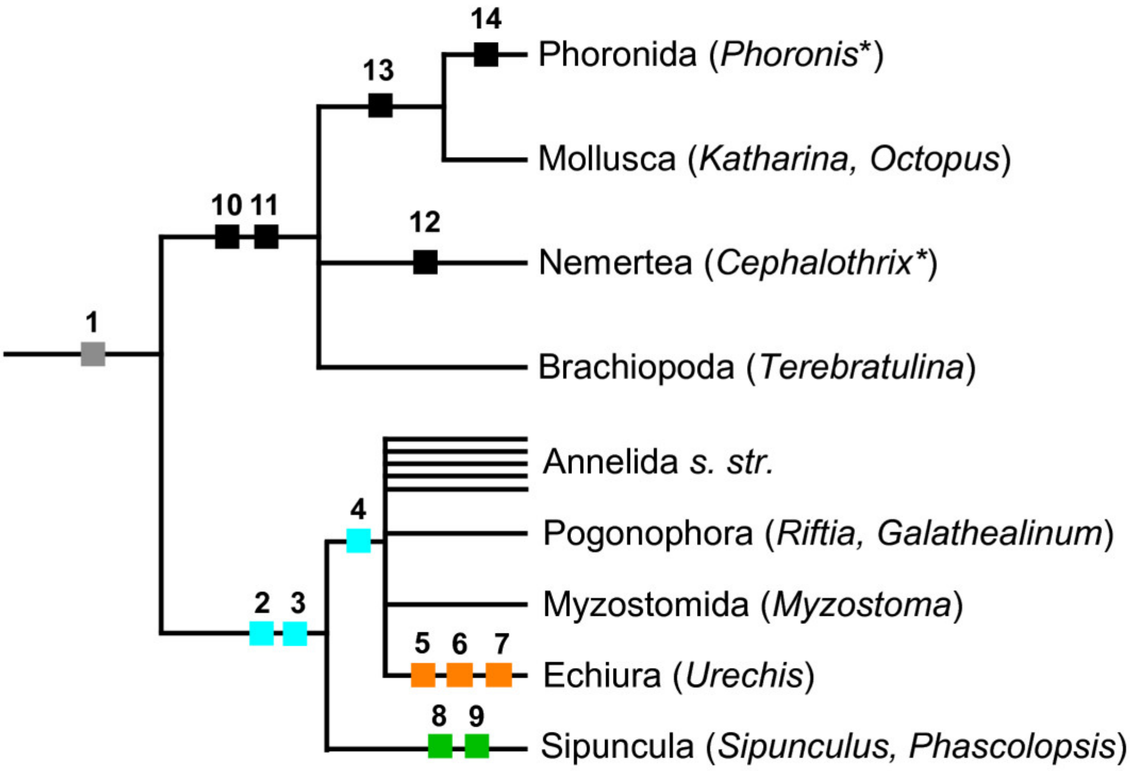

Phoronida (Phoronis psammophila)
(+)
(-)
cox1 cox2 atp8 cox3 atp6 nad3 nad2
nad5 nad4 nad4L cob nad6 nad1 rrnL rrnS
Some molluscs (Polyplacophora: Katharina; Gastropoda: Haliotis; Cephalopoda: Octopus)
(+)
(-)
cox1 cox2 atp8 atp6 cox3 nad3 nad2
nad5 nad4 nad4L cob nad6 nad1 rrnL rrnS
Cephalothrix rufifrons (Nemertea)
(+)
(-)
cox1 rrnL nad1 cob nad4L nad4 nad5 nad6 cox3 nad3 nad2
Terebratulina retusa (Brachiopoda), some molluscs (Gastropoda: Conus, Ilyanassa, Thais, Lophiotoma)
(+)
(-)
cox1 cox2 atp8 atp6 rrnS rrnL nad1 nad6 cob nad4L nad4 nad5 cox3 nad3 nad2
putative ground pattern of Bilateria
(+)
(-)
cox1 cox2 atp8 atp6 cox3 nad3 nad4L nad4 nad5 nad6 cob rrnS rrnL nad1 nad2
Annelida sensu strictu ("Polychaeta", "Oligochaeta", Hirudinea)
(+)
(-)
cox1 cox2 atp8 cox3 nad6 cob atp6 nad5 nad4L nad4 rrnS rrnL nad1 nad3 nad2
Riftia pachyptila (Pogonophora: Vestimentifera)
(+)
(-)
cox1 cox2 atp8 cox3 nad6 cob atp6 nad4 rrnS rrnL nad1 nad3 nad2
Galathealinum brachiosum (Pogonophora: Perviata)
(+)
(-)
cox1 cox2 atp8 cox3 nad6 cob rrnL nad1 nad3 nad2
Myzostoma seymourcollegiorum (Myzostomida)
(+)
(-)
cox1 cox2 atp8 cox3 nad6 cob atp6 nad5 nad4L nad4 rrnS rrnL nad2
Sipunculus nudus (Sipuncula)
(+)
(-)
cox1 cox2 atp8 cox3 nad6 cob rrnS rrnL nad3 nad4L nad4 nad1 atp6 nad5 nad2
Phascolopsis gouldii (Sipuncula)
(+)
(-)
cox1 cox2 atp8 cox3 nad6 cob rrnS rrnL
Urechis caupo (Echiura)
(+)
(-)
rrnS rrnL cox3 nad6 cob atp6 nad5 cox1 cox2 atp8 nad4L nad4 nad2 nad1 nad3

Figure 3.3: Comparison of mitochondrial gene order (protein-coding genes and ribosomal RNAs only) of several lophotrochozoan taxa compared and the putative bilaterian ground pattern (according to Lavrov and Lang 2005). Genome segments from the bilaterian ground pattern are colour coded for a better visualization of differences between gene orders. For complete species names and accession numbers see table 3.3.


14
13
Phoronida (*Phoronis**)
Mollusca (*Katharina, Octopus*)
10 11
12
Nemertea (*Cephalothrix**)
1
Brachiopoda (*Terebratulina*)
Annelida *s. str.*
4
Pogonophora (*Riftia, Galathealinum*)
2 3
Myzostomida (*Myzostoma*)
5 6 7
Echiura (*Urechis*)
8 9
Sipuncula (*Sipunculus, Phascolopsis*)

Species		1	2	3	4	5	6	7	8	9
		nad6+ cob	*atp6*	*nad5*	*nad3*	*cox1-atp8*	*nad2*	*rrnS-rrnL*	*rrnS+rrnL*	*atp6+ nad5*
Platynereis dumerilii	„Polychaeta“	x	x	x	x	o	o	o	o	o
Orbinia latreilii	„Polychaeta“	x	x	x	x	o	o	o	o	o
Clymenella torquata	„Polychaeta“	x	x	x	x	o	o	o	o	o
Lumbricus terrestris	„Oligochaeta“	x	x	x	x	o	o	o	o	o
*Helobdella robusta**	Hirudinea	x	x	?	x	o	o	o	?	o
*Galathealinum br.**	Perviata	x	?	?	x	o	o	o	?	?
*Riftia pachyptila**	Vestimentifera	x	x	?	x	o	o	o	o	?
Urechis caupo	Echiura	x	x	x	x	x	x	x	o	o
*Phascolopsis gouldii**	Sipuncula	x	(x)	?	(x)	o	o	o	?	x
Sipunculus nudus	Sipuncula	x	(x)	(x)	o	o	o	o	x	x
*Myzostomum seym.**	Myzostomida	x	x	x	?	o	o	o	o	o

*: partial genome data

Figure 3.4: Cladogram for changes in gene order of lophotrochozoan taxa (only changes in protein-coding and rRNA genes were analysed). The translocation of a gene or a gene block is treated as an apomorphic feature (small box) with numbers according to translocated genes in the table below. “x” indicates derived gene positions, circles stand for an unvaried order. “(x)” symbolizes that although the position of the gene is now different there is evidence that it allows no definite conclusion so far. Questionmarks indicate missing sequence data or putative secondary events complicating the interpretation. Changes not mentioned in the table: (10) translocation of cox3/ nad3; (11) translocation of rrnS/rrnL/nad1; (12) translocation of nad6; (13) large inversion of a segment spanning from rrnS to nad5; (14) translocation of atp6. See text for further details.

tRNA-Cys was not successful, so we stuck with this hypothesis although we cannot rule out that this is a non-functional sequence or subject to gene editing. In several other tRNAs there are mismatches in the acceptor or anticodon stem.

Mitochondrial gene order

Figure 3.3 shows a comparison of lophotrochozoan mitochondrial gene orders and the ground pattern of Bilateria (as mentioned in (Lavrov and Lang, 2005)). We restrict the discussion of gene order to the protein-coding and rRNA genes, as tRNA genes change their relative position much faster than the former, as seen in gene order comparisons of e.g. annelids (Bleidorn et al., 2007) or crustaceans (Kilpert and Podsiadlowski, 2006). The annelids, pogonophorans and myzostomids do not differ from each other in the relative positions of protein-coding and rRNA genes.

Compared to the ground pattern of Bilateria several genes have a different relative position: (1) nad6/cob are found right after cox3, (2) atp6 is found between cob and nad5, (3) nad5 and nad4L/nad4 have interchanged positions, and (4) nad3 is found between nad1 and nad2 (numbers refer also to hypothesized events in Figure 3.4). Mollusca (*Conus textile* (Bandyopadhyay et al., 2008), *Ilyanassa obsoleta* (Simison et al., 2006)) and Brachiopoda (*Terebratulina retusa* (Stechmann and Schlegel, 1999)) show a different pattern, with derived positions for three gene blocks: rrnS/rrnL/nad1, cox3/nad3 and nad6/cob. The translocation of nad6/cob may be explained as a commonly derived feature of Lophotrochozoa, or a subtaxon of it including Mollusca, Phoronida, Brachiopoda, Nemertea, Annelida s. l. (including Pogonophora, Echiura and Myzostomida) and Sipuncula (compare Figure 3.4). The other translocation events found in annelids and their kin (2.–4.) seem to be restricted to that group. The gene order so far known from Nemertea (*Cephalothrix rufifrons*, partial genome (Turbeville and Smith, 2007)) can be easily derived with one change (translocation of nad6) from the pattern of the brachiopod *Terebratulina* and the gene order of Phoronida (*Phoronis psammophila*, partial genome (Helfenbein and Boore, 2004b)) from that of the mollusc *Katharina tunicata* with only one event (translocation of atp6). Much more variation is seen within Mollusca (Valles and Boore, 2006; Yokobori et al., 2007) and Brachiopoda (Noguchi et al., 2000; Helfenbein et al., 2001; Endo et al., 2005) (not shown). Compared to the Annelida and their kin, the mitochondrial gene order of *Sipunculus nudus* differs clearly: (a) atp6 and nad5 are found between nad1 and nad2. This may be interpreted as two events restricted to the sipunculid lineage and independently achieved from the bilaterian or lophotrochozoan ground pattern. But another explanation would be a singular event translocating the block atp6/nad5 compared to the annelid ground pattern (No. 8 in Figure 3.4); (b) rrnS/rrnL found a different position, between cob and nad3 – this is as well different from the situation in Brachiopoda and Mollusca, so probably another event in the lineage leading to Sipuncula (No. 9 in Figure 3.4); (c) nad3 is found right after rrnL and adjacent to nad4L/nad4. This is different from its position in annelids, pogonophorans, myzostomids and echiuran taxa and is more similar to the bilaterian ground pattern. Visualized in Figure 3.4 the most parsimonious explanation of sipunculid gene order is that Sipuncula share two events with annelids, but lack the translocation of nad3. In addition two events have to be assumed in the lineage of Sipunula (rrnS/rrnL and atp6/nad5, corresponding to 8 and 9 in Figure 3.4). Derivation of the *Sipunculus* gene order directly from the bilaterian ground pattern would demand four translocation events (nad6/cob, rrnS/rrnL, atp6, nad5) from which only one is shared with other lophotrochozoan taxa (nad6/cob). So this hypothesis is in demand of three additional events instead of two for the "annelid" hypothesis. Derivation of the sipunculid gene order from the brachiopod/ mollusc pattern is in demand of five additional events. Therefore the most parsimonious explanation of gene order changes would be that Sipuncula is sister group to a group comprising Annelida s.str., Myzostomida, Echiura and Pogonophora.

At first sight gene order of the echiurid *Urechis caupo* (Boore, 2004) is completely different from that of annelids and *Sipunculus*, but the position of atp6 between cob and nad5 and that of nad3 adjacent to nad1 clearly hint to the derived features postulated for the annelid ground pattern (see b and c in the discussion of annelid gene order above). As well adjacency of nad6 to cox3 is found in all annelids and *Sipunuculus.* So the gene order of *Urechis* may be derived from the annelid ground pattern, with additional translocations of three genome segments: (a) cox1/cox2/atp8, (b) rrnS/rrnL and (c) nad2.

Phylogenetic analysis of mitochondrial sequences

The phylogenetic analysis was performed with a concatenated amino acid alignment of 11 protein-coding genes (exept atp8 and nad4L) from 74 species. Figure 3.5 shows the best tree of the Maximum Likelihood analysis with RAxML (mtREV+G+I). A close relationship of *Sipunculus* and *Phascolopsis* and thus monophyletic Sipuncula is well supported (ML bt: 100%). Sipuncula appears to be close related to the classic "Annelida", Echiura and Pogonophora – this assemblage has a bootstrap support of 93%. This assemblage is also other recovered in recent molecular analyses of 18S/28S rRNA and EF1α (Struck et al., 2007) or EST data (Dunn et al., 2008). The internal relationships of these taxa are not well resolved by our analysis. With high bootstrap

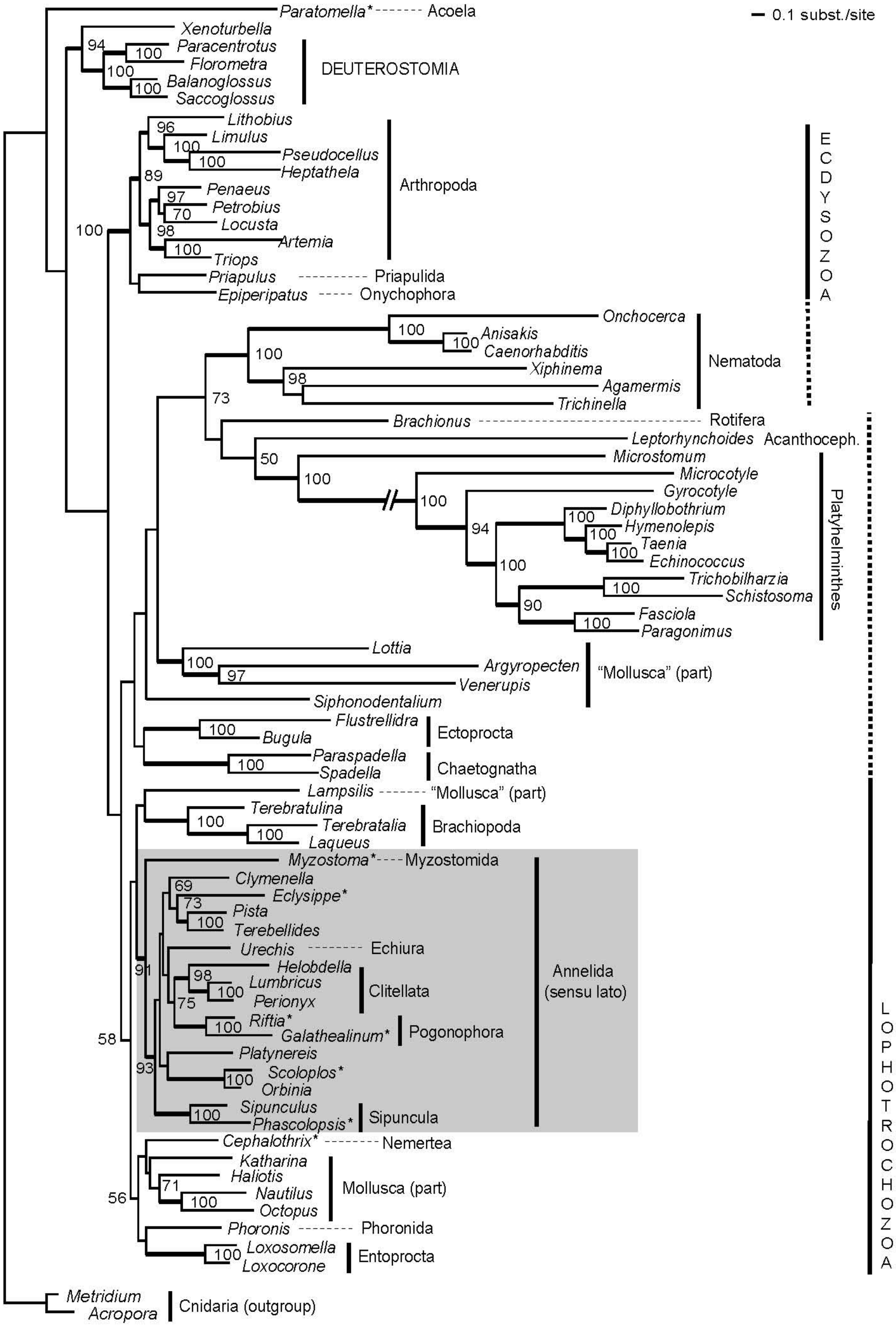


Figure 3.5: Best tree from the Maximum Likelihood analysis, inferred from the mitochondrial amino acid data set of 11 protein coding genes (RAxML 7.00, mtREV, G+I, single gene partitions). Numbers beneath nodes are ML bootstrap percentages, bold branches indicate bootstrap percentages >85%. See table 3.3 for complete species names and accession numbers. Asterisks indicate taxa with incomplete mt-genome information.

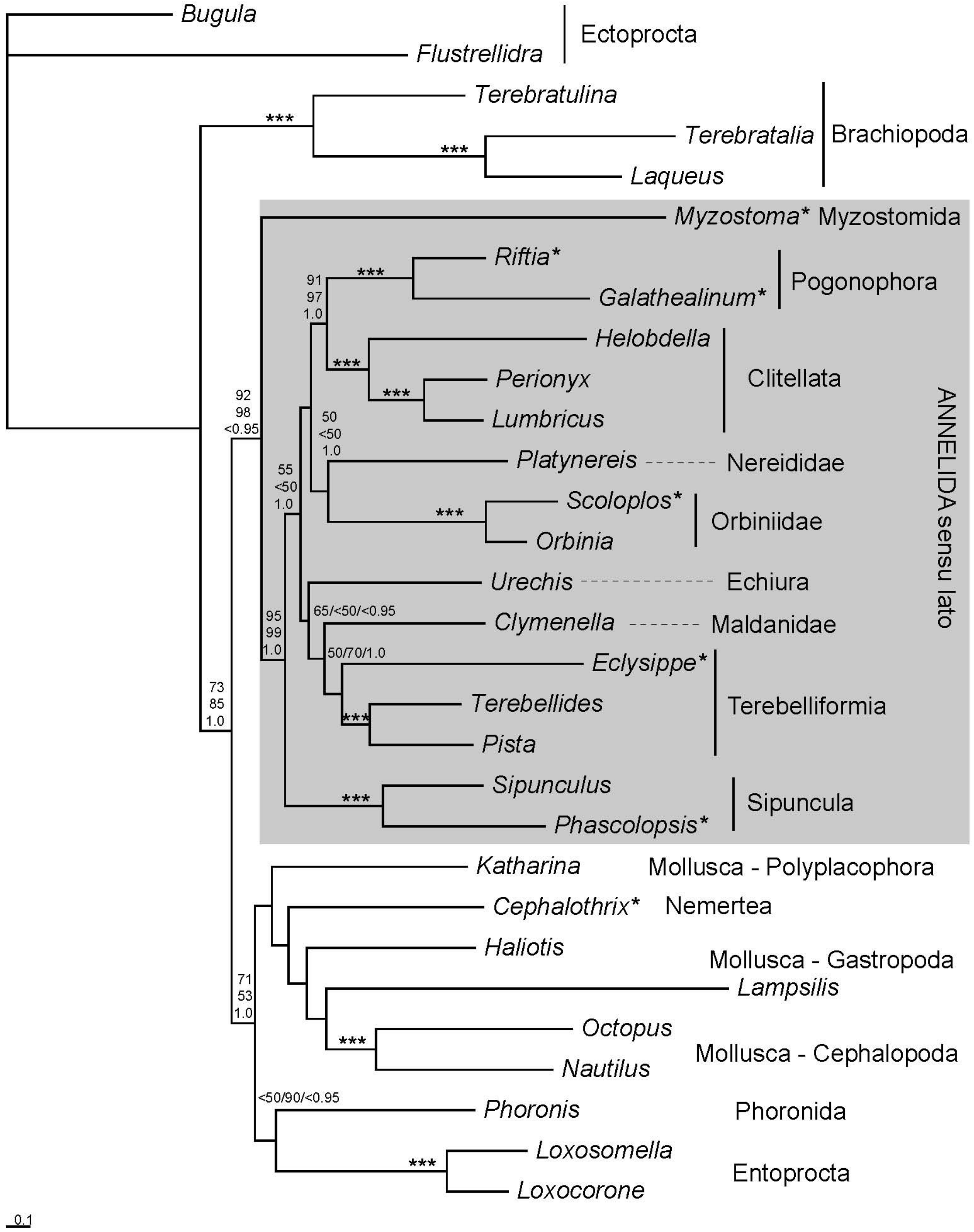


Figure 3.6: Best tree from the Maximum Likelihood analysis (RAxML 7.00, mtREV, G+I, single gene partitions) of the reduced taxon set (30 lophotrochozoan species). Numbers beneath nodes indicate support (from left to right or up to down, respectively): (1) through RaxML bootstrapping (1000 pseudoreplicates) (2) ML analysis with Treefinder (1000 pseudoreplicates), model mtART+G+I, (3) Bayesian posterior probabilities (model mtREV+G+I). Triple asterisks indicate maximum support from all three analyses (100/100/1.0). See table 3.3 for complete species names and accession numbers. Single asterisks indicate taxa with incomplete mt-genome information. Scalebar depicts substitutions per site in the best RAxML tree.

support Clitellata (98%) and Pogonophora (100%) appear monophyletic, while their sister group relationship found only weak support (bootstrap: 75%). Sister group to the Sipuncula/ Annelida/Echiura/Pogonophora taxon is Myzostomida (ML bt: 91%), this relationship is also supported by morphological characters and mitochondrial gene order as recently detailed elsewhere (Bleidorn et al., 2007). The position of this "Annelida sensu lato" among other Lophotrochozan subtaxa is not well resolved in our analysis.

Probably due to long branch effects, Ecdysozoa and Lophotrochoza appear not to be monophyletic in our analysis. While the former miss Nematoda, the latter miss Platyhelminthes, Ectoprocta, Rotifera, Acanthocephala and some molluscs. All these taxa are associated with long branches and form a probably artificial clade, which was never recovered in analyses with molecular data from nuclear genes or morphological data. Apart from this the most "problematic" taxon are Mollusca, with some taxa (*Lottia, Argopecten, Venerupis, Siphonodentalium*) found clustering with the above mentioned nematode-platyhelminth assemblage, others (*Katharina, Haliotis, Nautilus, Octopus*) clustering with Nemertea, Phoronida and Entoprocta, while *Lampsilis* appears as sister taxon to Brachiopoda.

For further evaluation the interrelationships of Annelida sensu lato, we performed additional phylogenetic analyses with a smaller taxon set comprising 30 species (all species from the lophotrochozoan branch of the larger taxon set). ML analyses were done comparing mtREV (RaxML) and mtART (Treefinder) models; in addition a Bayesian analysis was performed with mtREV model (MrBayes). Myzostomida, Sipuncula and other Annelida formed a monophyletic group (Figure 3.6) supported by ML bootstrapping (mtREV: 92%, mtART: 98%), but not by BI, where support is below 0.95 (Bayesian posterior probabilities). Sipuncula and Annelida together form a clade well supported by all three analyses, while Annelida without Sipuncula found best support only in BI, while the ML analyses do not significantly support this group, leaving open if there is a basal split between Sipuncula and the rest of the annelids. In the best ML-mtART tree *Platynereis* is found as sister to Sipuncula tree, but with bootstrap support below 50%. Well supported subtaxa of annelids are Pogonophora (s.lato), Clitellata, Pogonophora+Clitellata, Orbiniidae (*Scoloplos*+*Orbinia*). Topologies obtained in the three analyses differ in the position of *Urechis* (Echiura), which is found as sister to Maldanidae+Terebelliformia in the best ML tree with mtREV model (bootstrap support 65%), as sister to Orbiniidae in the best tree with mtART model (bootstrap support below 50%) and as sister to Pogonophora+Clitellata in BI (BPP below 0.95).

In addition we performed an AU test as implemented in CONSEL to statistically test the hypothesis of a sister group relationship between Sipuncula and Mollusca. We were able to significantly reject ($p < 0.001$) this hypothesis compared to the best ML-tree (mtREV).

Conclusion

Annelida, in traditional phylogenetic systems the sister group to Arthropoda, are nowadays included in the taxon Lophotrochozoa by almost all large scale analyses (Halanych et al., 1995; Mallatt and Winchell, 2002; Halanych, 2004; Hausdorf et al., 2007; Dunn et al., 2008). In this view more and more molecular studies no longer support the monophyly of the classical Annelida ("polychaetes" and clitellates). As well as the unsegmented Pogonophora, Echiura, and Myzostomida the Sipuncula have also been under suspect to be included in what was called Annelida sensu lato (Bleidorn et al., 2006; Bleidorn et al., 2007; Struck et al., 2007; Dunn et al., 2008). The complete mitochondrial genomic sequence of *Sipunculus nudus* presented in this paper, adds an important piece of evidence to answer the question of sipunculid position in the metazoan tree of life. Our sequence data and gene order analysis clearly support an affinity of Sipuncula to Annelida s. l. (including Pogonophora, Echiura and Myzostomida) rather than to Mollusca or any other phylum. It still remains an open question if Sipuncula and the whole Annelida s. l. are sister groups (as the most parsimonious explanation of gene order data suggests), or if Myzostomids form the sister group to Sipuncula and the remaining Annelida (as sequence based analyses favour). In sequence-based analyses the myzostomid is the annelid taxon with the longest branch, suggesting a more rapid evolution of mitochondrial sequence in this taxon. Therefore analyses placing Myzostomids outside the Annelida are probably misleading due to higher substitution rates in myzostomids.

Methods

Animals, DNA purification

A specimen of *S. nudus* was collected in Concarneau, France and conserved in 100% ethanol. Using the DNeasyR Blood & Tissue kit (Qiagen, Hilden, Germany) we followed the instructions given to extract DNA from animal tissues and used approximately 1 x 1 cm of the body wall from one individual.

PCR and purification of DNA fragments

EST sequence fragments for the genes nad1, nad3, rrnL, cob, cox1, cox2 and cox3 were used to design the first species specific primer pairs (Hausdorf et al., 2007). The complete mitochondrial genome of *S. nudus* was amplified in PCR fragments generated with species specific primer pairs from EST information (see Table 2). All PCRs were done with Eppendorf Mastercycler or Eppendorf Mastercycler Gradient thermocyclers. PCRs were carried out in 50 µl volumes (41.75 µl water, 5 µl 10× buffer, 0.25 µl Taq polymerase (5 U/µl), 1 ml dNTP mixture, 1 µl template DNA, 1 µl primer mixture (10 µM each)) using the Eppendorf 5- prime kit (Eppendorf, Germany). The cycling conditions were as follows: 94°C for 2 min (initial denaturation); 40 cycles of 94°C for 30 sec (denaturation); primer-specific temperature (see table 3.2) for 1 min (annealing), 68°C for 1 min (elongation), was followed by 68°C for 2 min (final elongation). After 40 cycles the samples were stored at 4°C and visualised on a 1% ethidium bromide-stained TBE agarose gel, respectively. DNA fragments expected to be larger than 3 kb, were amplified in 25 µl volumes (16.75 µl water, 2.5 µl buffer, 0.25 µl Takara LA Taq polymerase, 4 µl dNTP mixture, 1 µl template DNA, 0.5 µl primer mixture (10 µM each)) under the following long PCR conditions (Takara LA kit): 94°C for 2 min (initial denaturation); 40 cycles of 94°C for 30 sec (denaturation), primer-specific temperature for 1 min (annealing) and 72°C for 10 min (elongation). After the final elongation step (68°C for 2 min), samples were treated as described above. PCR products were purified with minispin columns provided in the Nucleo Spin Extract II kit (Macherey & Nagel) and the Blue Matrix PCR/DNA clean up DNA Purification kit (EurX, Gdansk, Poland). Dependent on the band intensity on the agarose gel, DNA was eluted in 30–60 µl elution buffer and stored at -20°C. Slightly contaminated samples

Table 3.2: Primer pairs and corresponding annealing temperatures used for successful amplification of mitochondrial genome fragments from *Sipunculus nudus.*

Primer names	Primer sequence (5'-3')	Annealing temperature	Approxim. size of PCR product
Sn-cox1-f	CTCCCACTTAGCACCCTC	48°C	400 bp
Sn-cox2-r	TAAGAGAATAATGGCGGG	50°C	700 bp
Sn-cox2-f	CCAACCACTCTTTTATGCC		
Sn-cox3-r	CCAGGATTAGGGCGGT	48°C	700 bp
Sn-cox3-f	TTTTCCTATACCTCTGCATC		
Sn-cob-r	TTGAATGACAAGGCAGAGA	48°C	2500 bp
Sn-cob-f	CTCCTCGCCCCCAA		
Sn-16S-r	GATTTATTGAAGAGTGGTTAGTGA	48°C	600 bp
Sn-16S-f	TCATACCCCGCACTCC		
Sn-nd3-r	CAAACCCGCACTCAAAC	50°C	2500 bp
Sn-nd3-f	GGTAAGTGAACGGGGAAC		
Sn-nd1-r	AAAAGGTTGGGGGAGG	50°C	4000 bp
Sn-nd1-f	CGCTATCACCTCCACCTT		
Sn-cox1-r	ATTCGGCACGGATAAGA		

were cut from a 1% ethidium bromide-stained TAE agarose gel and purified with the QIAquick Gel Extraction kit (Qiagen) afterwards.

Cloning

If the DNA amount, obtained by PCR, turned out to be insufficient for sequencing, the respective fragment was cloned in a pGEM-T Easy Vector (Promega). Ligation was carried out in 5 µl volumes instead of the double amount, proposed in the protocol. In each case 2 µl of the sample were used for transformation in 50 µl competent E. coli XL Gold (Stratagene) cells. Colonies, containing recombinant plasmids, were detected via blue-white screen on LB selection plates, charged with IPTG, ampicillin and Xgal. To check whether the desired insert had been really transferred to the picked out colonies, a minimum amount of each colony (approximately half of it) was utilized as DNA template in a colony PCR. PCRs were run in 50 µl volumes (ingredients, amounts and conditions as above named), using M13F and M13R vector primers. Products were checked on 1% TBE agarose gels and – if they contained an insert of the anticipated size – transferred to LB/ampicillin medium. After proliferation over night, samples were purified according to the guidelines of the Quantum Prep-Kit (Bio Rad) and finally stored at - 20 °C.

Sequencing and gene annotation

The amplified fragments were set up in 10 µl reaction volumes (2.5 µl DNA, 2.5 µl water, 1 µl primer (10 µM), 4 µl DCTS master mix) and sequencing PCR reactions were carried out according to the following procedure: 96 °C for 20 sec (denaturation); primer-specific temperature for 20 sec (annealing); 60 °C for 2 min (elongation). After 30 cycles the samples were sequenced with a CEQ™8000 capillary sequencer (Beckmann-Coulter) and the appropriate CEQ DCTS Quick Start kit (Beckmann-Coulter).

While the first checking of the sequences was carried out with the CEQ 8000 software (Beckman-Coulter), the actual sequence assemblage was done with BioEdit, version 7.0.5 (Hall, 1999). Protein coding and ribosomal RNA genes, encoded in the mtDNA, were identified by BLAST (blastn, tblastx) searches on NCBI databases and by aligning the different sipunculid fragments with the mt genome of the echiurid *Urechis caupo*. To revise the final consensus sequence of *S. nudus*, further mt-genome data of relatively closely related taxa were retrieved from the OGRe database (Jameson et al., 2003). The species used for sequence comparison were: *Platynereis dumerilii* (Annelida), *Clymenella torquata* (Annelida), *Orbinia latreillii* (Annelida), *Lumbricus terrestris*

(Annelida), *Terebratalia transversa* (Brachiopoda), *Terebratulina retusa* (Brachiopoda), *Laqueus rubellus* (Brachiopoda), *Urechis caupo* (echiura), *Epiperipatus biolleyi* (Onychophora), and *Flustrellidra hispida* (Bryozoa), see table 3.3 for accession numbers. Transfer RNA genes and their putative secondary structures, were determined with the tRNAscan-SE (Lowe and Eddy, 1997) and ARWEN (Laslett and Canback, 2008) and for the missing ones by eye inspection of candidate regions. The genome sequence was deposited in NCBI database [Gen- Bank: FJ422961].

Table 3.3: Species, systematic position and accession number of mitochondrial genome sequences used in the phylogenetic analysis and/or for of gene order comparisons.

Species	Taxonomic position	Accession no.
Sipunculus nudus	Sipuncula	FJ422961
*Phascolopsis gouldii**	Sipuncula	AF374337
Urechis caupo	Echiura	NC_006379
*Myzostoma seymourcollegiorum**	Myzostomida	EF506562
Lumbricus terrestris	Annelida – Clitellata	NC_001677
Perionyx excavatus	Annelida – Clitellata	NC_009631
*Helobdella robusta**	Annelida – Clitellata	AF178678
Platynereis dumerilii	Annelida – "Polychaeta"	NC_000931
Orbinia latreillii	Annelida – "Polychaeta"	NC_007933
*Eclysippe vanelli**	Annelida – "Polychaeta"	EU239687
Clymenella torquata	Annelida – "Polychaeta"	NC_006321
Pista cristata	Annelida – "Polychaeta"	NC_011011
Terebellides stroemi	Annelida – "Polychaeta"	NC_011014
*Scoloplos armiger**	Annelida – "Polychaeta"	DQ517436
*Galathealinum brachiosum**	Annelida – Pogonophora	AF178679
*Riftia pachyptila**	Annelida – Pogonophora	AY741662
Epiperipatus biolleyi	Onychophora	NC_009082
Limulus polyphemus	Chelicerata – Xiphosura	NC_003057
Heptathela hangzhouensis	Chelicerata – Araneae	NC_005924
Pseudocellus pearsei	Chelicerata – Ricinulei	NC_009985
Lithobius forficatus	Myriapoda – Chilopoda	NC_002629
Petrobius brevistylis	Hexapoda – Archaeognatha	NC_007689
Locusta migratoria	Hexapoda – Orthoptera	NC_001712
Artemia franciscana	Crustacea – Anostraca	NC_001620
Triops cancriformis	Crustacea – Phyllopoda	NC_004465
Penaeus monodon	Crustacea – Decapoda	NC_002184
Priapulus caudatus	Priapulida	NC_008557
*Cephalothrix rufifrons**	Nemertea	EF140788
*Phoronis psammophila**	Phoronida	AY368231
Terebratulina retusa	Brachiopoda	NC_000941
Laqueus rubellus	Brachiopoda	NC_002322
Terebratalia transversa	Brachiopoda	NC_003086
Katharina tunicata	Mollusca – Polyplacophora	NC_001636
Lottia digitalis	Mollusca – Gastropoda	NC_007782
Haliotis rubra	Mollusca – Gastropoda	NC_005940
Conus textile	Mollusca – Gastropoda	NC_008797

*: incomplete genome sequence

Table 3.3: Species, systematic position and accession number of mitochondrial genome sequences used in the phylogenetic analysis and/or for of gene order comparisons. (*Continued*)

Species	Taxonomic position	Accession no.
Ilyanassa obsoloeta	Mollusca – Gastropoda	NC_007781
Thais clavigera	Mollusca – Gastropoda	NC_010090
Lophiotoma cerithiformis	Mollusca – Gastropoda	NC_008098
Nautilus macromphalus	Mollusca – Cephalopoda	NC_007980
Octopus ocellatus	Mollusca – Cephalopoda	NC_007896
Venerupis phllippinarum	Mollusca – Bivalvia	NC_003354
Argopecten irradians	Mollusca – Bivalvia	NC_009687
Lampsilis ornata	Mollusca – Bivalvia	NC_005335
Siphonodentalium lobatum	Mollusca – Scaphopoda	NC_005840
Loxocorone allax	Entoprocta	NC_010431
Loxosomella aloxiata	Entoprocta	NC_010432
Flustrellidra hispida	Bryozoa/Ectoprocta	NC_008192
Paraspadella gotoi	Chaetognatha	NC_006083
Spadella cephaloptera	Chaetognatha	NC_006386
Brachionus plicatilis	Rotifera	NC_010484
Leptorhynchoides thecatus	Acanthocephala	NC_006892
Anisakis simplex	Nematoda	NC_007934
Agamermis sp.	Nematoda	NC_008231
Onchocercus volvulus	Nematoda	NC_001861
Caenorhabditis elegans	Nematoda	NC_001328
Trichinella spiralis	Nematoda	NC_002681
Xiphinema americanum	Nematoda	NC_005928
*Paratomella rubra**	Acoela	AY228758
*Microstomum lineare**	Platyhelminthes – "Turbellaria"	AY228756
Fasciola hepatica	Platyhelminthes – "Trematoda"	NC_002546
Paragoniums westermanni	Platyhelminthes – "Trematoda"	NC_002354
Gyrodactylus salaris	Platyhelminthes – "Trematoda"	NC_008815
Microcotyle sebastis	Platyhelminthes – "Trematoda"	NC_009055
Schistosoma haematobium	Platyhelminthes – "Trematoda"	NC_008074
Trichobilharzia regenti	Platyhelminthes – “Trematoda”	NC_009680
Hymenolepis diminuta	Platyhelminthes – Cestoda	NC_002767
Taenia asiatica	Platyhelminthes – Cestoda	NC_004826
Echinococcus granulosus	Platyhelminthes – Cestoda	NC_008075
Balanoglossus carnosus	Enteropneusta	NC_001887
Saccoglossus kowalevskii	Enteropneusta	NC_007438
Florometra serratissima	Echinodermata – Crinoidea	NC_001878
Paracentrotus lividus	Echinodermata – Echinoidea	NC_001572
Xenoturbella bocki	Xenoturbellida	NC_008556
Acropora tenuis	Cnidaria – Anthozoa	NC_003522
Metridium senile	Cnidaria – Anthozoa	NC_000933

*: incomplete genome sequence

Phylogenetic analysis

The amino acid alignments of the protein-coding genes (except the two short and highly variable genes atp8 and nad4L) were concatenated. Sequence data from 74 species were included in the large analyses (see table 3.3 for all species names and accession numbers). The tree was rooted with two representatives of Cnidaria. Maximum likelihood analysis was performed with RAxML, ver. 7.00 (Stamatakis, 2006;

Stamatakis et al., 2008). mtREV+G+I was chosen as model for aminoacid substitutions. The complete dataset was partitioned, so that model parameters and amino acid frequencies were optimized for each single gene alignment. 100 bootstrap replicates were performed to infer the support of clades from the best tree. A second set of analyses were done with a reduced dataset of 30 species. This dataset was analyzed with RAxML as described above (model mtREV+G+I, partitioned according to the 12 single gene sequences), with 1000 bootstrap replicates. Secondly we did a Bayesian analysis with MrBayes ver. 3.1.2 (Huelsenbeck and Ronquist, 2001). In BI the mtREV+G+I model was used and 1.000.000 generations were run with 8 chains in parallel. Trees were sampled every 1000 generations, while the first 200 trees were discarded as burn-in (according to the likelihood plot). In addition we performed a ML analysis using the mtART+G+I model with Treefinder (Jobb, 2007) and "edge support" analysis, again with a partitioned dataset (= independently optimizing model parameters for the 12 genes).

For comparison of the hypothesis that sipunculids might be closely related with molluscs and our best tree, we used a constraint for a ML-analysis (Sipuncula + Mollusca) of the sequence dataset using RaxML (Stamatakis, 2006) with parameters described above. We computed per-site log-likelihoods with RAxML for both topologies (best tree and constrained topology) and conducted an au-test as implemented in CONSEL (Shimodaira and Hasegawa, 2001).

Abbreviations

atp6 and 8: genes encoding ATPase subunit 6 and 8; bp: base pairs; bt: bootstrap; cox 1–3: genes encoding cytochrome oxidase subunits I-III; cob: gene encoding cytochrome b; BI: Bayesian Inference; ML: Maximum Likelihood; mtDNA: mitochondrial DNA; mt-genome: mitochondrial genome; nad1-6 and nad4L: genes encoding NADH dehydroenase subunits 1–6 and 4L; PCR: polymerase chain reaction; rRNA: ribosomal RNA; rrnL: large rRNA subunit (16S); rrnS: small rRNA subunit (12S); tRNA: transfer RNA; trnX: tRNA gene (X is replaced by one letter amino acid code).

Acknowledgements

We are grateful to [redacted] for providing us with a *Sipunculus nudus* specimen. Financial support for this study came from Deutsche Forschungsgemeinschaft (DFG): DFG Ba 1520/10-1,2 to TB and LP; DFG BL 787/2-1 to CB, both from priority programm "Deep Metazoan Phylogeny" SPP 1174.

4. Respiratory proteins in *Sipunculus nudus* Linnaeus 1766 – implications for phylogeny and evolution of the hemerythrin family

Abstract

Three major classes of respiratory proteins are known, hemoglobin, molluscan and arthropod hemocyanin, and hemerythrin (Hr). Similar to hemoglobin, respiratory Hr is packed into erythrocytes floating in the coelomic fluid and is only known from sipunculids, brachiopods, and priapulids. Owing to this scattered distribution, the presence of Hr is generally assumed to be the plesiomorphic condition without phylogenetic importance. By sequencing 2,000 Expressed Sequence Tags (ESTs) from *Sipunculus nudus*, we found 75 Hr-coding ESTs assembled to 20 cDNA contigs classified as four distinct Hr isoforms: three polymeric Hrs (subunit A, A', and B) and the monomeric myo-hemerythrin (myoHr). Phylogenetic analyses revealed a clade of annelid and sipunculan monomeric Hrs, distinct from polymeric Hrs. Monomeric Hrs from annelids and sipunculids can be clustered together using Maximum Likelihood tree-building and network analyses, as well as applying Bayesian methods. Three distinct Hr clusters were found for *S. nudus*, suggesting a new monomeric Hr isoform.

Introduction

Respiratory proteins are among the first proteins used in molecular systematics (Zuckerkandl and Pauling 1965), and methods like molecular clock approaches (Wilson and Sarich 1969) or the problem of comparing paralogous genes (Goodman et al. 1979) were initially introduced using hemoglobins. Many of the phylogenetic conclusions inferred from these early analyses (mostly distance- and parsimony-based) have been proven to be incorrect, but their impact on the development of new methods has been outstanding. The benefit of analyzing respiratory proteins in recent phylogenetic studies is the knowledge about the drawbacks and opportunities based on inferences using large hemoglobin datasets. For example Archosauria (i.e., crocodilians + birds) were not recovered using hemoglobin data (Gorr et al. 1998) but on the other hand orthologues haemoglobin isoforms have constrained positions and phylogenetic value in vertebrates (Gribaldo et al. 2003).

The Hr gene family has been identified in a variety of organisms: archaea, bacteria, several fungi, priapulids, brachiopods, a cnidarian (*Nematostella vectensis*), several annelids, and all sipunculids (Bailly et al. 2008). Owing to the scattered distribution across the phyla, the presence of Hrs is considered as the plesiomorphic condition in eukaryotes without phylogenetic importance. Sipunculids carry out reversible oxygen binding and transport using Hr as the respiratory pigment.

Hr binds a dioxygen molecule via two Fe^{2+} ions which are bound by direct coordination to the polypeptide chain. The monomeric molecule has a molecular mass of 13.0–13.5 kDa. Respiratory Hr is often found as an octamer, thus having molecular masses of 100–110 kDa, but further oligomeric molecules also exist, depending on the different species or as the result of reversible disassociation (Kurtz 1986). The respiratory Hr of Sipuncula is carried by nucleated erythrocytes floating in the coelomic cavity and vascular system, both showing different oxygen affinities (Cutler 1994, pp. 256-265). The vascular Hr is found in the tentacular contractile vessel system, and interacts with the second coelomic Hr as donor or acceptor of oxygen. The gradient depends on the involvement of the tentacular crown in gas exchange: Sipunculids adapt to different respiration strategies depending on the habitat and morphology of a certain species, and the tentacular crown does not have a respiratory function in every species (Cutler 1994, pp. 256-265).

A monomeric isoform is found in sipunculids and annelids, characterized as myoHr (Takagi and Cox 1991), neuroHr (Vergote et al. 2004), or Metalloprotein II (Demuynck et al. 1991). NeuroHr has been described in the leech *Hirudo medicinalis*, and this initially iron-free protein is found to be upregulated in response to septic injury. The potential role of neuroHr is associated with its ability to bind oxygen and iron, as an immune response of the leech nervous system to bacterial invasion. A similar protein from *Helobdella robusta* has a signal peptide sequence mediating extracellular localization (Bailly et al. 2008). Metalloprotein II is an additionally described Hr and is highly similar to sipunculan myoHr, but binds Cadmium instead of iron (Demuynck et al. 1993). It has been shown experimentally that the diiron in Hr can easily be substituted without affecting its quaternary structure (Zhang and Kurtz 1992).

The only investigated Hr from the two brachiopods of the genus *Lingula* is octameric, but has distinct differences in cooperativity and oxygen affinity, and in contrast to octameric sipunculan Hrs, exhibits a Bohr effect (Manwell 1960; Negri et al. 1994). Ludt (1995) sequenced a short Hr peptide fragment (36AA) of the priapulid, *Halicryptus spinulosus*, and determined a tetrameric structure. Klippenstein (1980) assumed a tetrameric quaternary structure for *Priapulus caudatus*. The prerequisite for phylogenetic reconstructions are inferences using orthologous genes, but the delineated knowledge about the Hr protein family is fragmentary.

Peanut worms (Sipuncula) are unsegmented marine worms comprising about 150 species (Cutler 1994) and are known since 500 Myr (Huang et al. 2004). Their systematic position within the Lophotrochozoa was controversially discussed in the past, e.g., a sister group relationship to the molluscs was repeatedly suggested (Scheltema 1993; 1996). However, no traces of Hrs within Mollusca could be observed. Most recently, a close relationship between the Annelida and Sipuncula has been reported based on phylogenomic, mitochondrial, and morphological data (Dunn et al. 2008; Hausdorf et al. 2007; Kristof et al. 2008; Mwinyi et al. 2009; Shen et al. 2009).

Sipunculus nudus has been reported to be the most basal offshoot of extant sipunculids (Maxmen et al. 2003; Schulze et al. 2007); but see Cutler (1994, p. 376). We selected one individual to screen for new simultaneously expressed Hr isoforms by EST sequencing, and analyzed the data in a phylogenetic context. The combination of paralogous Hr isoforms in phylogenetic analyses can identify orthologs and shed light on protein and species evolution simultaneously.

Material and Methods

Isolation of RNA and Library Construction

Sipunculus nudus was collected near Roscoff (France), frozen in liquid N_2, and stored at −80 °C. The total RNA was extracted from 1 g of frozen tissue gained by cross section of the anterior part of the trunk from an adult specimen. The sample contained body wall musculature and coelomic fluid. We used Trizol (Invitrogen, Karlsruhe, Germany) according to the manufacturer’s recommendations, but adding 100 µl of H_2O/ml of Trizol and homogenized with an Ultra-Turrax. The RNA was visually checked by agarose gel electrophoresis, and mRNA was subsequently captured using Dynabeads (Invitrogen). The cDNA library was constructed by primer extension, size fractioning, and directional cloning, by applying the Creator SMART cDNA Libraries Kit (Clontech, Heidelberg, Germany). A total of 2,000 ESTs were sequenced from the 5’ end with the automated capillary sequencer systems ABI 3730 XL (Applied Biosystems, Darmstadt, Germany) using the BigDye chemistry (Applied Biosystems).

EST processing

EST processing was accomplished at the Center for Integrative Bioinformatics in Vienna. The sequencing chromatograms were first base-called and evaluated using the Phred application (Ewing et al. 1998). Vector, adapter, poly-A, and bacterial sequences were removed by employing the software tools Lucy (www.tigr.org), SeqClean (compbio.dfci.harvard.edu/tgi/software), and CrossMatch (www.phrap.org). The repetitive elements were subsequently masked with RepeatMasker (http://www.repeatmasker.org). Clustering and assembly of the clipped sequences were performed with EST data for *S. nudus, Themiste lageniformis,* and *Priapulus caudatus* from the trace archives using the TGICL program package (compbio.dfci.harvard.edu/tgi/software) by first performing pairwise comparisons (MGIBlast) and a subsequent clustering step (CAP3). Low-quality regions were then removed by Lucy.

Alignment construction and phylogenetic analyses

The Hr sequences were assigned and collected using tBlastn and Blastp searches (Altschul et al. 1997), and are presented in Table 4.1. Additionally, up to 40 ESTs for each Hr contig of the Clitellata were assembled using the CAP function implemented in BioEdit (Hall 1999). Mean genetic distance was calculated using Mega version 4 (Tamura et al. 2007) excluding all sequences surpassing 5% missing data. Screen-

Table 4.1: Species with the Genbank accession numbers of the Hr sequences used in this study, except the assembled EST contigs. AA = amino acid data gained by peptide sequencing, NT = nucleotide data.

Species and Hr isoform/lable	Higher taxon	Data-type	Accession Number
Golfingia vulgaris subunit 1	Sipuncula	AA	AJ632200
Golfingia vulgaris subunit 2	Sipuncula	NT	AJ632199
Golfingia vulgaris subunit 3	Sipuncula	AA	AJ632202
Golfingia vulgaris subunit 4	Sipuncula	AA	AJ632201
Hediste diversicolor MPII 1	Nereididae, Polychaeta, Annelida	NT	AAB26091
Hediste diversicolor MPII 2	Nereididae, Polychaeta, Annelida	AA	P80255.2
Hediste diversicolor MPII frag.	Nereididae, Polychaeta, Annelida	NT	S57799
Hediste diversicolor	Nereididae, Polychaeta, Annelida	AA	P22761
Hirudo medicinalis neuroHr	Hirudinea, Clitellata, Annelida	NT	AY521548
Lingula anatina subunit A	Inarticulata, Brachiopoda	AA	P22764.3
Lingula anatina subunit B	Inarticulata, Brachiopoda	AA	P22765
Lingula reevii subunit A 1	Inarticulata, Brachiopoda	AA	P23543.2
Lingula reevii subunit A 2	Inarticulata, Brachiopoda	AA	2022172A
Lingula reevii subunit B	Inarticulata, Brachiopoda	AA	P23544.2
Perinereis aibuhitensis	Nereididae, Polychaeta, Annelida	NT	FJ212325
Periserrula leucophryna	Nereididae, Polychaeta, Annelida	NT	AY312845

Table 4.1: Species with the Genbank accession numbers of the Hr sequences used in this study, except the assembled EST contigs. AA = amino acid data gained by peptide sequencing, NT = nucleotide data. (*Continued*)

Species and Hr isoform/lable	Higher taxon	Data-type	Accession Number
Phascolopsis gouldii	Sipuncula	AA	AF220529
Phascolopsis gouldii 1	Sipuncula	NT	AF220529
Phascolopsis gouldii 2	Sipuncula	AA	1I4Y
Phascolopsis gouldii 3	Sipuncula	AA	1I4Z
Phascolopsis gouldii myoHr	Sipuncula	AA	P27686.2
Phascolosoma arcuatum	Sipuncula	NT	EU368753
Ridgeia piscesae	Siboglinidae, Annelida	NT	EV802527
Riftia pachyptila 1	Siboglinidae, Annelida	NT	EF648563
Riftia pachyptila 2	Siboglinidae, Annelida	NT	AM886446
Scoloplos armiger myoHr	Orbiniidae, Polychaeta, Annelida	NT	AM886447
Siphonosoma cumanense	Sipuncula	AA	P22766
Sipunculus nudus myoHr	Sipuncula	NT	AM886444
Sipunculus nudus myoHr2	Sipuncula	NT	AM886445
Sipunculus nudus subunitA	Sipuncula	NT	AJ632021
Sipunculus nudus subunitB	Sipuncula	NT	AJ632197
Sipunculus nudus subunitB2	Sipuncula	NT	AJ632198
Themiste dyscritum 1	Sipuncula	AA	P02246
Themiste dyscritum 2	Sipuncula	AA	1HMD_A
Themiste dyscritum 3	Sipuncula	AA	2HMQ_A
Themiste zostericola myoHr 1	Sipuncula	AA	1A7D
Themiste zostericola myoHr 2	Sipuncula	AA	1A7E
Themiste zostericola myoHr 3	Sipuncula	AA	P02247.2
Theromyzon tessulatum	Hirudinea, Clitellata, Annelida	NT	AF279333

ing for signal peptides was carried out at the SignalP 3.0 webserver (Emanuelsson et al. 2007) to identify a possible extracellular target location for monomeric Hrs. All the nucleotide sequences were translated using GeneWise (Birney et al. 2004) and aligned using muscle (Edgar 2004). Back translation of the nucleotide sequences post alignment was performed using TranslatorX (Telford, unpublished). Alisore (Misof and Misof 2009) was used to trim the resulting amino acid alignment. The protein substitution model was estimated by ProtTest 1.3 (Abascal et al. 2005), by applying the AICc criterion with a correction for small sample size. Neighbor-net analysis using ML distances were inferred from an amino acid alignment under the WAG + Γ model with SplitsTree4 (Huson 1998). Maximum Likelihood using RAxML (Stamatakis 2006) was performed using the phylobench.vital-it.ch (Stamatakis et al. 2008) under the GTR+ Γ model. RAxML carried out 100 bootstrap replicates and an ML search using every tenth tree as the starting point in a thorough ML analysis. The Bayesian inference using MrBayes 3.1.2 (Huelsenbeck and Ronquist 2001) at the CIPRES-Portal 2.0 (http://www.phylo.org/) accomplished two runs, four chains each, with 10,000,000 generations discarding 25% of the sampled trees as burnin.

Results

A total of 20 *S. nudus* Hr contigs were recovered (Acession Nr.: FN393830 - FN393849) and were clustered to four distinct isoforms (Table 4.2). The previously published *S. nudus* Hr subunit alpha (AJ632021) and contig FN393837 from the specimen used in this study differs only in two nucleotide substitutions. The Hr beta subunit AJ632197 is identical to FN393843. The within mean genetic distance of the newly generated Hr sequences from this study is: HrB = 0.014 (five sequences) and HrA = 0.004 (four sequences). Two additional distinct Hr isoforms have been discovered: (i) FN393838, stated as Hr A' and (ii) a cluster of three distinct monomeric Hr sequences (Sinu FN393847-9, mean genetic distance 0.107) lacking any extracellular signal peptide contrary to signal peptides found in monomeric MPII proteins from leeches (Bailly et al. 2008). Extensive database searches revealed 60 Hr protein sequences from Annelida, Sipuncula, Brachiopoda, and one Hr fragment from Priapulida, but no entries were found for other Lophotrochozoa, such as molluscs. Similarly, the polychaete model organism, *Capitella capitata*, lacks Hr such as most annelids rather exhibit erythrocruorin or hemoglobin. cDNA Hemerythrin sequences were only found in some species belonging to Clitellata, Nereididae, Orbiniidae and Siboglinidae. All the sequences found were used to construct an alignment and the final amino acid alignment (all alignments available from the authors) comprised 75 sequences with 123 amino acid positions from 23 taxa. The AICc criterion with a correction for small sample size suggested the WAG+Γ substitution model. The Maximum Likelihood network analysis (Figure 4.1) distinguishes Hrs packed in erythrocytes (Figure 4.1; Sipuncula I, II and Brachiopoda) to all other Hr sequences. *Lumbricus rubellus* exhibits one monomeric Hr and two additional proteins with unknown quaternary structure and function. These two *L. rubel-*

Table 4.2: Coverage of Hr isoforms found in 2,000 ESTs from *S. nudus*. FN 393847-9 are in all probability representing three myoHr isoforms (see discussion). A = alpha subunit, B = beta subunit, A' = previously unknown polymeric Hr isoform related to the alpha subunit.

Accession number	No. of ESTs	Hr
FN393830	2	A
FN393831	1	A
FN393832	1	A
FN393833	1	A
FN393834	1	A
FN393835	1	A
FN393836	30	A
FN393837	1	A
FN393839	1	A
FN393838	1	A'
FN393840	26	B
FN393841	1	B
FN393842	1	B
FN393843	1	B
FN393844	1	B
FN393845	1	B
FN393846	1	B
FN393847	1	myo
FN393848	1	myo
FN393849	1	myo

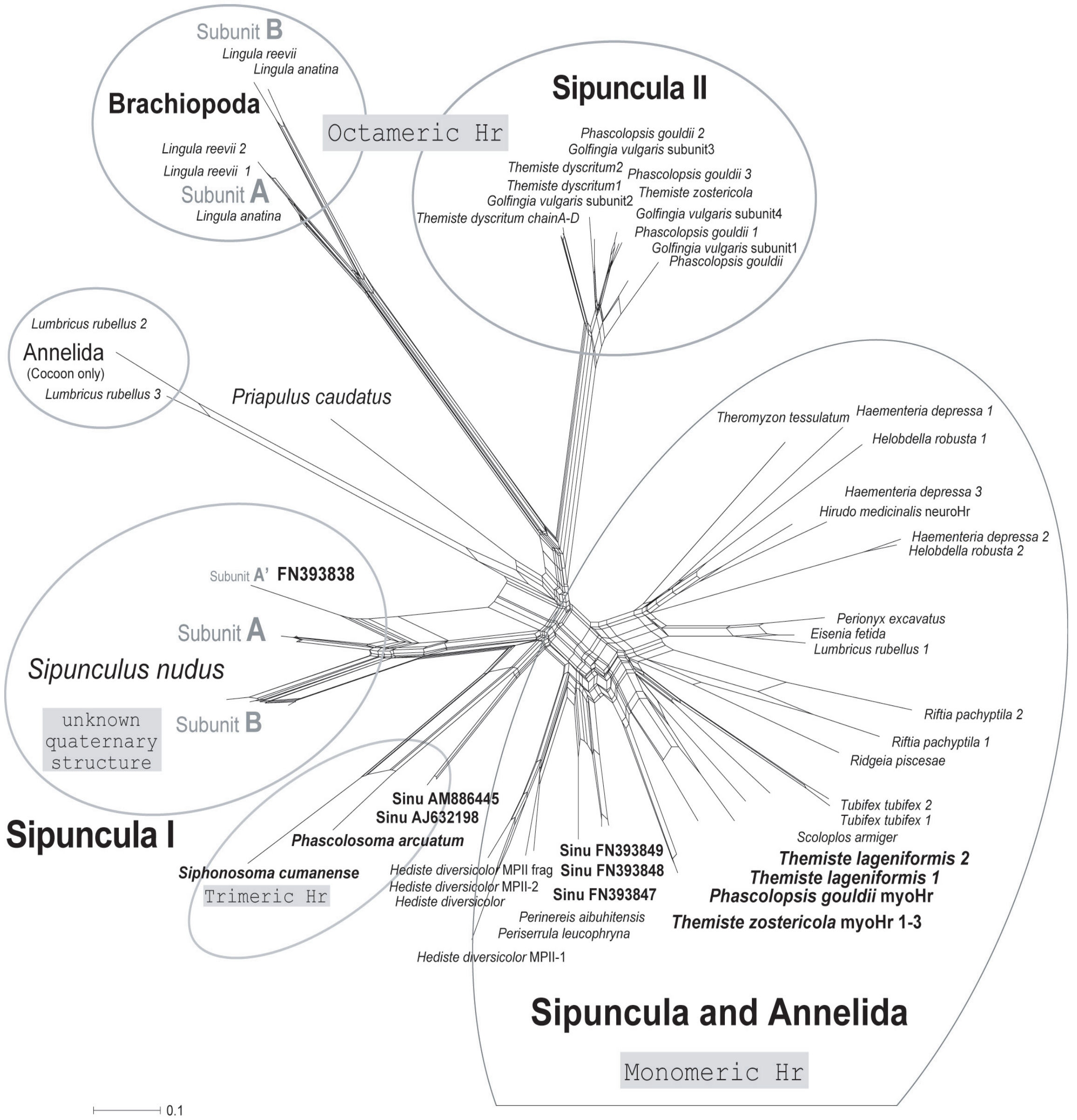


Figure 4.1: Neighbor network using Maximum Likelihood distances and 123 amino acid positions. Despite the overall short alignment length, some well-defined clusters were recognized (encircled). Groups were either separated according to their different quaternary structures (e.g., trimeric vs. octameric sipunculan Hr) or phylogenetic distances (Brachiopoda vs. Sipuncula II). The *Lumbricus rubellus* 2 and 3 sequences are exclusively expressed in late cocoons with unknown function and structure, but the expected myoHr is also found (*Lumbricus rubellus* 1).

lus Hrs (2 and 3), each covered by more than 25 ESTs, are exclusively found in late cocoon cDNA libraries, except one single EST (DR076331) reported from head tissue. The overall coverage for *L. rubellus* in Genbank is 19,934 ESTs.

The network analysis confirmed the large distance of the Brachiopoda to all remaining sequences, and recovered similarities of the monomeric Hrs from the Annelida and Sipuncula. Additionally, ML and Baysian analyses were performed using all 43 available cDNA sequences from 18 taxa (Figure 4.2). The underlying nucleotide alignment comprised 351 positions after discarding four columns owing to random signal from the amino acid alignment (position 39–41, 51) using Aliscore. Monomeric Hrs from Annelida and Sipuncula were observed to intermingle in all analyses, but their polymeric Hrs were distinct depending on their quaternary structure (Figure 4.1). Within the Hr clades of different quaternary structure, the resolution was very low and paral-

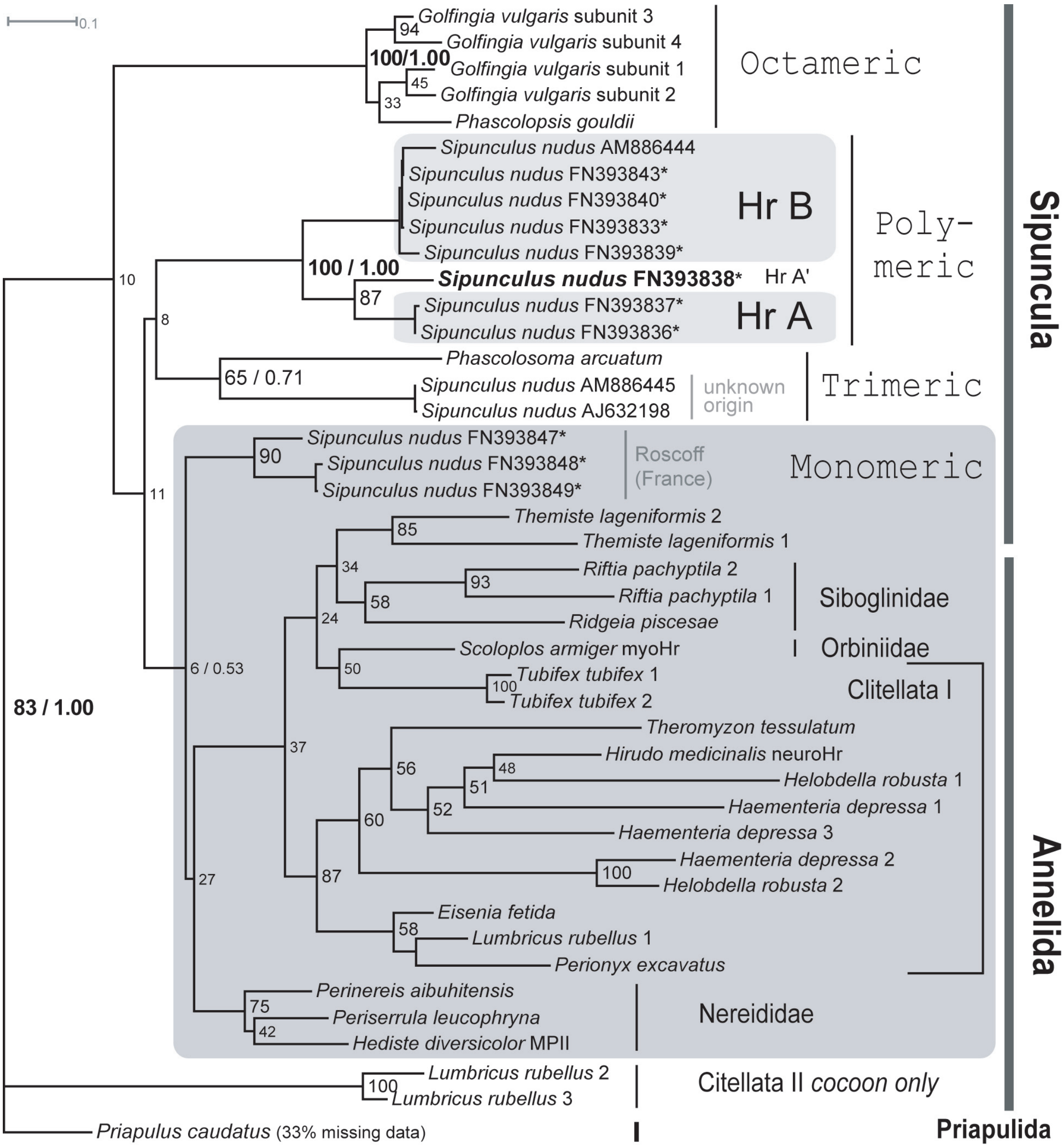


Figure 4.2: Only a subset of the amino acid sequences is covered with cDNA data. The Maximum Likelihood tree based on the nucleotide alignment was inferred using RAxML. *Themiste lageniformes* is deeply nested in the monomeric hemerythrin clade, and distinct polymeric isoforms (HrA and HrB) of *S. nudus* are highly supported. *S. nudu*s sequences generated for this study are marked with an asterisk. Only Hr sequences with complete coding regions were included, reducing the subset of new Hr sequences to ten. Some full-supported nodes from the Bayesian analysis are plotted on the ML tree.

ogs were observed. Despite the high proteome coverage leading to 75 Hr ESTs, no hit was observed to correspond to the suggested myoHr (AM886445, AJ632198) similar to coelomic trimeric Hr from *Siphonosoma cumanense* (Figure 4.1). Instead, we found a monomeric Hr clustering with previously suggested myoHrs from the *Themiste* spp. and *Phascolopsis gouldii*.

Discussion

For the first time, we have shown that the mRNA variability of the polymeric Hr family simultaneously expressed within a single individual of *S. nudus*. Klippenstein (1972) isolated one distinct and one “broad protein band” from the coelomic fluid of *Phascolosoma gouldii*, both representing Hr, and concluded one conserved and one more variable Hr subunit. The same pattern can be observed within the EST data of *S. nudus* mirroring the homogeneity of the subunit A sequences (mean genetic distance 0.004) and a more variable set of subunit B sequences (mean genetic distance 0.014), possibly subsequently enlarged by posttranslational protein modifications. Furthermore, we discovered a new isoform in *S. nudus*, stated as subunit Hr A’ (see also Figures 4.1 and 4.2, FN393838). According to these findings, the different Hr isoforms are simultaneously expressed in the erythrocytes; it is obvious that the experimental difficulties impede the determination of the still unknown quaternary structure of *S. nudus* Hr (Bates et al. 1968).

Quaternary structure and phylogenetic implications

Our analyses support a clade of monomeric Hrs restricted to Annelida and Sipuncula (Figure 4.1). The evolutionary data often contain a number of different and sometimes conflicting signals (homoplasy), but which can be visualized using network methods (Huson 1998). The ML neighbor network demonstrates the high amount of “noisy signal” within the Hr dataset, but also infers unambiguous support for distinct Hr groups.

The clades are formed either owing to their quaternary structure or phylogenetic distances: Octameric Hrs from evolutionary distinct Sipuncula and Brachiopoda are well discerned, but monomeric Hrs from closer related taxa like annelids and sipunculids intermingle. It has been shown that intersubunit surfaces of hemoglobins co-evolve under selective pressure and obscure historical relationships if different quaternary structures are compared (Gribaldo et al. 2003, Lieb et al. 2001). The separate position of polymeric *S. nudus* Hr subunit A and B in the network analysis suggest a non-octameric structure for the coelomic Hr (Figure 4.1). The functional constrains adopting different quaternary structures may then impede phylogenetic conclusions, in particular, if combined with the overall short sequences in Hr. However, monomeric Hrs form a well-recognized group in the network analysis, and also analyzing the coding cDNA revealed monophyletic monomeric Hrs (Figure 4.2). These monomeric Hrs found in the Sipuncula and Annelida might be plesiomorphic. The inferred relationship of *Themiste lageniformis* Hrs to Hrs from Siboglinidae (Figure 4.2) has no statistical support and is artifactual because the monomeric Hrs from Sipuncula did not cluster together. On the other hand the placement of Siboglinidae (Pogonophora) as derived polychaetes with sabellid affinity is in consensus with many investigations (McHugh 1997).

Unfortunately, only peptide sequences were available for brachiopods and some of the other Hrs. Some polychaete Hrs were biochemically detected without providing any sequence data like *Ophelia bicornis* (Scolecida) (Sanna et al. 2005) or *Magelonia papillicornis* (Spionida); the latter species has respiratory Hr packed in coelomic cells, similar to the situation seen in sipunculan Hrs (Wells and Dales 1974).

The structure and function of the two *L. rubellus* Hr isoforms 2 and 3 are unknown. Both are nearly exclusively expressed in cocoons, and might not account for monomeric Hrs, as indicated by their sequence divergence and the presence of the expected monomeric *L. rubellus* Hr (isoform 1, Figures 4.1 and 4.2). Further developmental EST data are available for Clitellata, *Tubifex tubifex* and three Hirudinea species, but the distinct cocoon-Hr seem to be unique for *L. rubellus*. This apparently demonstrates that Hrs have evolved a variety of functional potentials and features analogous to hemocyanins (Lieb 2003).

Hemerythrins in *Sipunculus nudus*: new isoform or cryptic species?

Siphonosoma cumanense Hr has a trimeric quaternary structure, and is notably different from octameric Hrs (Figure 4.1). Similarly, the Hr from *Phascolosoma arcuatum* has a trimeric quaternary structure (Addison and Bruce 1977). For the nomenclature of the *Phascolosoma* species, we followed Cutler and Cutler (1990). They recognized *P. lurco* and *P. esculenta* as invalid synonyms of *P. arcuatum*. The *P. arcuatum* Hr clusters together with *Siphonosoma cumanense* and the previously published *S. nudus* Hr isoforms Sinu AM886445 and Sinu AJ632198, respectively (Figure 4.1). Within our EST data, we could not find any Hr sequence similar to this previously published *S.*

nudus Hr (AM886445, AJ632198). In contrast, we found a second previously unknown isoform (FN393847–FN393849) clustering with other myoHrs from Sipuncula (Figure 4.1). Vanin et al. (2006) stated two separate sources of the utilized *S. nudus* tissue leading to the published *S. nudus* Hrs: One specimen from Roscoff (France) is the same collection side as the specimen used in this study, and as a second source, a specimen provided by the fishermen in Venice (Italy) is used. The different Hr isoforms are not assigned to a distinct specimen locality. *Sipunculus nudus* is extensively shipped as bait from China and Vietnam to Europe, and therefore, the given origin unfortunately did not allow conclusions about the collection side. Costa et al. (2006) reported that every "live sea worms" shipped to Europe from Vietnam was *S. nudus*, reaching, e.g., more than 5,000,000 specimen at Lisbon airport a year. Notably, the amino acid sequence of the previously published *S. nudus* Hrs AM886445 and AJ632198 cluster together with the trimeric Hr of *Siphonosoma cumanense* and the *Phascolosoma arcuatum* Hr, which are the two tropical sipunculids. The trimeric quaternary structure might be an adaptation to the different temperature and oxygen level of the tropical habitats. We cannot conclude that all *S. nudus* Hr sequences account, in fact, for a single *S. nudus* species; however, Bates et al. (1968) suggested hexameric or octameric Hr for *S. nudus*, excluding trimeric Hr. Coexistent and fundamentally different quaternary structures of coelomic Hrs are unknown and unlikely, owing to the physiological constrains, and thus, the observed clustering of myoHrs (Figure 4.1) might hint to the existence of a cryptic *S. nudus* species with trimeric Hr. Polymeric Hrs are highly expressed and therefore, easy to recover from mRNA extractions, in contrast to rare myoHrs (Table 4.2). The use of different source specimens (cryptic species?) combined with the expected existence of two coelomic and one myoHr renders possible misspecifications. Further studies are needed to screen for cryptic species in the cosmopolitan *S. nudus*, especially using molecular methods, owing to limited diagnostic character available. Additionally, protein characterization of the polymeric *S. nudus* Hr would be helpful to determine the exact quaternary structure.

Acknowledgments

This study was supported by grants from the DFG within the priority program Deep Metazoan Phylogeny (Li 998/3-1). We are thankful to [redacted] (Friedrich-Wilhelms-Universität, Bonn) for providing *S. nudus* specimen, and [redacted] [redacted] (Max F. Perutz Laboratories, Center for Integrative Bioinformatics, Vienna) for processing of the EST data. We thank [redacted] for providing excellent working facilities.

5. Fast evolving 18S rRNA sequences from Solenogastres (Mollusca) resist standard PCR amplification and give new insights into mollusk substitution rate heterogeneity

Abstract

Background

The 18S rRNA gene is one of the most important molecular markers, used in diverse applications such as molecular phylogenetic analyses and biodiversity screening. The Mollusca is the second largest phylum within the animal kingdom and mollusks show an outstanding high diversity in body plans and ecological adaptations. Although an enormous amount of 18S data is available for higher mollusks, data on the early branching taxa are still limited. Despite of some partial success in obtaining these data from Solenogastres, by some regarded to be the most basal mollusks, this taxon still remained problematic due to contamination with food organisms and general amplification difficulties.

Results

We report here the first authentic 18S genes of three Solenogastres species (Mollusca), each possessing a unique sequence composition with regions conspicuously rich in guanine and cytosine. For these GC-rich regions we calculated strong secondary structures. The observed high intramolecular forces hamper standard amplification and appear to increase formation of chimerical sequences caused by contaminating foreign DNAs from potential prey organisms. In our analyses, contamination was avoided by using RNA as a template. Indication for contamination of previously published Solenogastres sequences is presented. In addition to our new sequences, we collected 831 mollusk 18S sequences from Genbank to infer taxon specific substitution rates.

Conclusions

The extreme morphological diversity of mollusks is mirrored in the molecular 18S data and shows elevated substitution rates mainly in three higher taxa: true limpets (Patellogastropoda), Cephalopoda and Solenogastres. Our phylogenetic tree based on 122 species, including representatives of all mollusk classes but Monoplacophora, shows limited resolution at the class level but illustrates the pitfalls of artificial groupings formed due to shared biased sequence composition.

Background

The small subunit (SSU) 18S rRNA is one of the most frequently used genes in phylogenetic studies (see below) and an important marker for random target PCR in environmental biodiversity screening (Chenuil 2006). In general, rRNA gene sequences are easy to access due to highly conserved flanking regions allowing for the use of universal primers (Hillis and Dixon 1991). Their repetitive arrangement within the genome provides excessive amounts of template DNA for PCR, even in smallest organisms. The 18S gene is part of the ribosomal functional core and is exposed to similar selective forces in all living beings (Moore and Steitz 2002). Thus, when the first large-scale phylogenetic studies based on 18S sequences were published - first and foremost Field et al.'s (1988) phylogeny of the animal kingdom - the gene was celebrated as the prime candidate for reconstructing the metazoan tree of life. And in fact, 18S sequences later provided evidence for the splitting of Ecdysozoa and Lophotrochozoa (Aguinaldo et al. 1997; see also Halanych 2004), thus contributing to the most recent revolutionary change in our understanding of metazoan relationships.

During recent years and with increased numbers of taxa included into molecular phylogenies, however, two problems became apparent. First, there are prevailing sequencing impediments in representatives of certain taxa, such as the mollusk classes Solenogastres and Tryblidia (Giribet et al. 2006; Okusu and Giribet 2003), selected bivalve taxa (pers. comment H. Dreyer), and the enigmatic crustacean class Remipedia (pers. comment H. Glenner). Failure to obtain 18S sequences of single taxa is considered a common phenomenon but is rarely ever reported. Secondly, in contrast to initially high hopes, 18S cannot resolve nodes at all taxonomic levels and its efficacy varies considerably among clades. This has been discussed as an effect of rapid ancient radiation within short periods (Abouheif et al. 1998). Multigene analyses are currently thought to give more reliable results for tracing deep branching events in Metazoa but 18S still is extensively used in phylogenetic analyses.

Considering the wide range of studies based on the 18S gene as a molecular marker, both sequencing problems and the applicability of 18S for phylogenetic inferences need to be scrutinized. To address these questions, we focus on the Mollusca, the second largest animal phylum. There are eight higher taxa (classes) defined within Mollusca: the aplacophoran Solenogastres (= Neomeniomorpha) and Caudofoveata (= Chaetodermomorpha), two small clades with about 250 and 150 currently described species; Polyplacophora (ca. 920 species); and the conchiferan clades Tryblidia (= Monoplacophora; 29 species described), Scaphopoda (ca. 520 species), Cephalopoda (ca. 1,000 species), Bivalvia (ca. 30,000 species), and Gastropoda (40,000-150,000 species). The monophyly of the phylum is well established based on morphological characters, but 18S phylogenies often show the Mollusca as polyphyletic or paraphyletic with low resolution of the deeper nodes (e.g: Adoutte et al. 2000; Passamaneck and Halanych 2006; Passamaneck et al. 2004; Peterson and Eernisse 2001). One of the main deficiencies of all published studies on mollusk phylogeny is the underrepresentation of the minor taxa Solenogastres, Caudofoveata, and Monoplacophora. The aplacophoran Solenogastres and Caudofoveata together with Polyplacophora (chitons) are traditionally thought to represent basal clades within Mollusca, but their relative position is still controversially discussed (for review see Haszprunar et al. 2008, pp. 19-32; Todt et al. 2008, pp 71-96). Despite of their key position in mollusk phylogeny and evolution, the number of molecular phylogenetic studies including any aplacophoran mollusk species is extremely low: Winnipeninckx et al. (1996); Okusu et al. (2003); Lindgren et al. (2004); Passamaneck et al (2004); Giribet et al. (2006); Dunn et al. (2008). To date there are no more than four caudofoveate and five solenogaster (partial) 18S sequences published in Genbank. In solenogasters, Okusu and Giribet (2003) described severe contamination issues caused by cnidarian prey. Here, we specifically address contamination issues and point out technical problems hampering amplification during PCR. This is important not only for prompting representative taxonomic sampling for phylogenetic analyses, but also for avoiding under- or overestimation of biodiversity in environmental screening programs. Moreover, we evaluate the usefulness of the 18S gene for phylogenetic inferences by combining our new authentic solenogaster 18S data with published molluscan sequences and – based on an extended taxon sampling - analyze sequence divergence and substitution rates within the phylum.

Results

Solenogastres sequences

Initial experiments using standard PCR protocols and genomic DNA from starved specimens resulted in sequences from prey organisms and epibionts or in chimerical PCR products. At best, cDNA templates led to shortened 18S sequences. Finally, authentic solenogaster 18S sequences for three species were obtained via isolation of total RNA from starved specimens, followed by reverse transcription and utilizing additives for GC-rich templates. 10% DMSO was applied in sequencing reactions.

Strong secondary structures and GC clamps were observed analyzing the *Wirenia argentea* (2161bp, GC content = 63.12%), *Simrothiella margaritacea* (2149bp, GC content = 61.42%), and *Micromenia fodiens* (2087bp, GC content = 62.72%) 18S sequences. Stitch Profiles determined three to four helical sections in dsDNA (Figure 5.1, blue bars), which additionally bear prominent stem regions in ssDNA as inferred with the secondary structure probability plot in MFold at 72°C. Elevated GC contents above 60% were observed in the helical regions, with a maximum of 82% in the second helix of *S. margaritacea*.

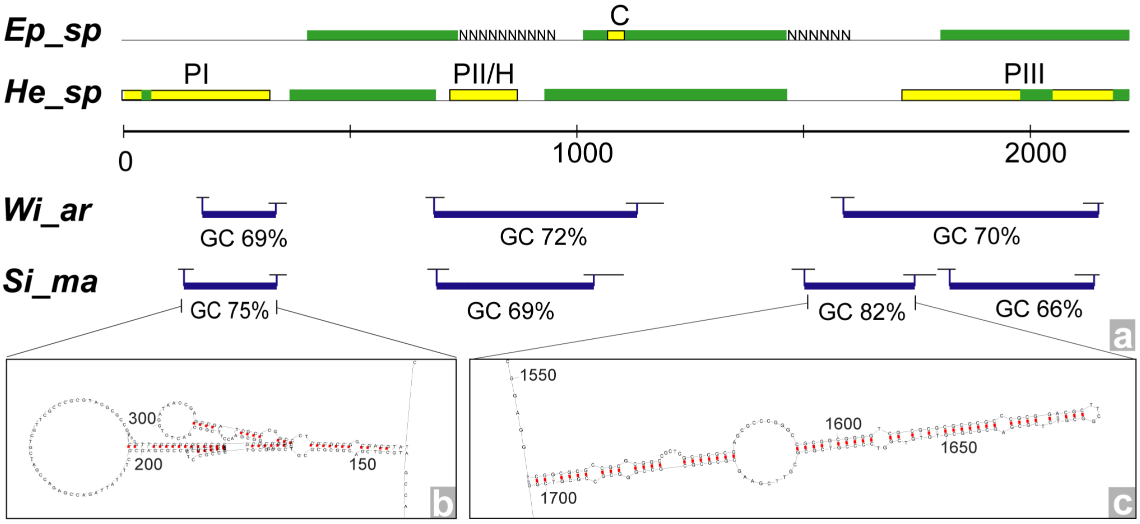


Figure 5.1 - Helical regions correspond with missing data or possible foreign DNA

5.1a: Possible chimerical patterns in previously published sequences (yellow bars) match with regions of double stranded DNA forming helical regions (blue bars) in newly generated sequences. *Epimenia* sp. (Ep_sp) and *Helicoradomenia* sp. (He_sp) sequences have been aligned to the nucleotide positions (represented by the scale bar) of the sequences from this study *Wirenia argentea* (Wi_ar) and *Simrothiella margaritacea* (Si_ma). The simplified schematic alignments (green bars) show high similarity (>90%) whereas other regions (yellow bars) possess lower similarity and point to contamination issues. BLASTn searches of yellow domains indicate high similarity with polychaetes (PI to III) cnidarians (C) or any significant hit, indicated as *Helicoradomenia* only (H). Detailed BLASTn results are given in table 5.1. Below the scale bar, the double stranded helical regions of 18S sequences are indicated containing 66-82% GC islands. **5.1b and c**: Close-up views of single stranded regions of helix 1 (1b) and helix 2 (1c) of *S. margaritacea* 18S. These stems have strong adhesive forces probably hampering PCR. Secondary structures were calculated using Mfold, applying 72°C. G-C hydrogen bonds are indicated in red.

Table 5.1: BLASTn results from previously published Solenogastres sequences using strings within regions of strong secondary structure (helices 1-3; see figure 5.1a). The short fragments of *Epimenia babai* 18S lie outside of helix regions and do not contain alien DNA. For the two published helix regions of *Epimenia* sp. the query results in Cnidaria (27 matches within Octocorallia). In all *Helicoradomenia* fragments the 18S of the polychaete *Amphisamytha galapagensis* was found within the best six BLASTn hits (excluding *Helicoradomenia*; mean=3), while the first hit for Mollusca was on rank 16 (mean=35).

Species:	*Epimenia babai*	*Epimenia babai*	*Epimenia* sp.	*Helicoradomenia* sp.	*Helicoradomenia* sp.
Accession No.	AY212107	AY212106	AY377657	AY212108	AY145377
Total length	396bp	248bp	1389bp	1902bp	1833bp
internal NNNs (position)	-	-	Helix2: 350-528, Helix3: 932-1006	-	-
Helix 1 (position)	-	-	-	1-348	1-369
BLASTn similarity				90%	91%
with				*A. galapagensis*	*A. galapagensis*
Helix 2 (position)	-	-	634-675	642-750	650-768
BLASTn similarity			100%	100%	94%
with			Cnidaria (and see above)	*Helicoradomenia* only	*A. galapagensis*
Helix 3 (position)	-	-	internal Ns (see above)	1597-1902	1490-1822
BLASTn similarity				95%	95%
with				*A. galapagensis*	*A. galapagensis*

A previously published complete caudofoveate (*Scutopus ventrolineatus*) 18S sequence showed only one short helical region (~100bp, GC content 63%) that could be aligned to the second helix of our solenogaster sequences (Figure 5.2). All previously published solenogaster 18S sequences lack exceptional secondary structures and the GC contents are below 60%. The *Epimenia* sp. sequence includes a 46bp fragment with 100% cnidarian sequence identity (Figure 5.1; yellow bar, C). The two nearly complete *Helicoradomenia* sp. sequences show 79.3% identity to each other. Analyses of selected sections from these sequences using BLASTn searches showed significant identities of >95% with the polychaete *Amphisamytha galapagensis* and other polychaetes (Table 5.1). Sections with missing data or exogenous DNA in previously published sequences match the regions of strong secondary structures determined within the sequences of *W. argentea, S. margaritacea* and *M. fodiens* (Figure 5.1, yellow bars: PI, PII/H and PIII and Figure 5.2).

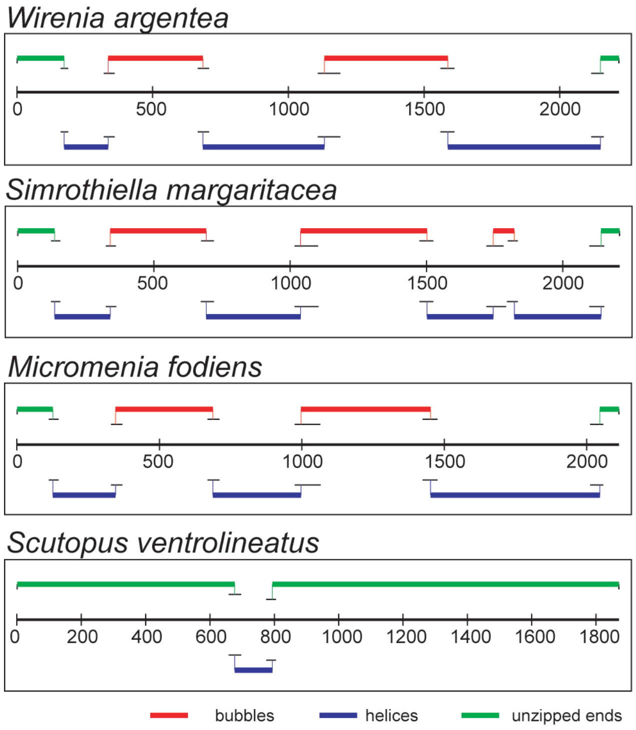


Figure 5.2: Secondary structures of double stranded 18S sequences. Stich Profiles calculated at 90.7°C showing melted and double stranded regions for the three new solenogaster 18S sequences and the caudofoveate *Scutopus ventrolineatus*. The probability to observe the calculated secondary structure is given. One short helical region in the *S. ventrolineatus* sequence of approximately 100bp length and 63% GC content is at the same position as in the solenogaster sequences.

Phylogenetic analyses

The 122 taxa alignment comprises 1967 aligned positions past trimming and contains 16% gaps or completely undetermined characters (N). The χ^2 test of homogeneity of base frequencies across taxa hints to significant rate heterogeneity caused by 36 taxa, including the Solenogastres species (χ^2=1247.144434 (df=363), P = 0.00000000). The recoded alignment passed without significant F values (χ^2= 43.588336 (df=121), P = 1.00000000). The maximum likelihood (ML) tree inferred on the recoded RY alignment has a log-likelihood value of -ln=18925.157694 and a gamma shape parameter of alpha=0.339513 (Figure 5.3 and 5.4). Bootstrap support is given for clades representing the mollusk classes and some major bivalve and gastropod groups, if detected. Posterior probabilities (Figure 5.5) of the NT and the RY coded alignment are plotted on the depicted ML tree of the recoded alignment (Figure 5.3 and 5.4). An additional ML analysis (-ln=17199.145422, alpha= 0.323075) of the recoded alignment excluding all cephalopods resulted in a fundamental rearrangement of class level groupings (Figure 5.6).

Path length estimation and substitution rates

The refined muscle (Edgar 2004) alignment of 873 species comprises 7,338 aligned positions and was used without trimming but retaining 75% undetermined positions (gaps). The resulting ML tree gave 872 branch length estimates starting from the basal node that were summarized for 17 higher taxa and depicted with Box plots in figure 5.3. Patellogastropoda, Cephalopoda, and Solenogastres show the highest substitution rates, exceeding two substitutions per site for more than 50% of their

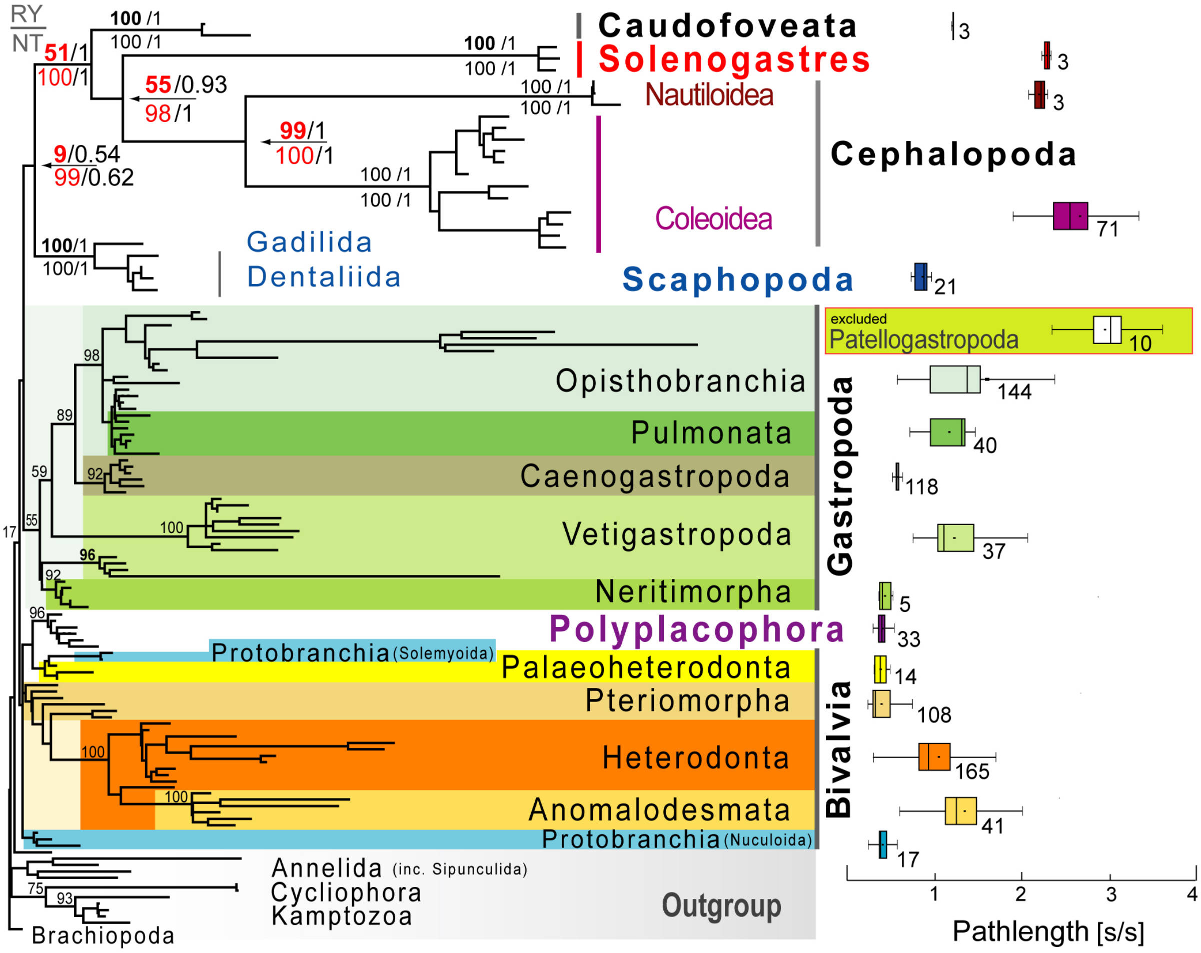


Figure 5.3: ML tree and taxon specific substitution rates. Solenogastres show a derived 18S sequence composition: Maximum Likelihood (ML) tree inferred using a recoded and trimmed alignment constructed with Prank$_{+F}$. Bootstrap support values of 1,000 replicates are given for labelled (sub)clades. Three additional phylogenetic analyses were run. Above branch support values correspond to the RY coded alignment, below branch values to the original NT coded alignment. The first value indicates ML bootstrap support and the second one posterior probabilities of the Bayesian inference. In particular the ML bootstrap support values show considerable decrease for the class level branching pattern in this part of the tree. To the right of the phylogenetic tree, sequence divergence is depicted as box plots of path length estimates in corresponding colours. The number of species pooled in each box plot is given alongside. Underlying substitutions per site have been estimated using a broader taxon sampling of 874 taxa and, in contrast to the phylogenetic analyses, an untrimmed, NT-coded muscle alignment. Path length estimates (substitutions per site) start at the most basal node separating *Priapulus caudatus* from Lophotrochozoa.

species. The overall fastest evolving sequences are found as outliers in the opistobranchs: *Dondice banyulensis* (6.0643521), *Facelina bostoniensis* (6.046491) and *Phidiana lynceus* (5.9970576), all Facelinidae (Gastropoda).

All trees and alignments are available at Treebase (http://www.treebase.org).

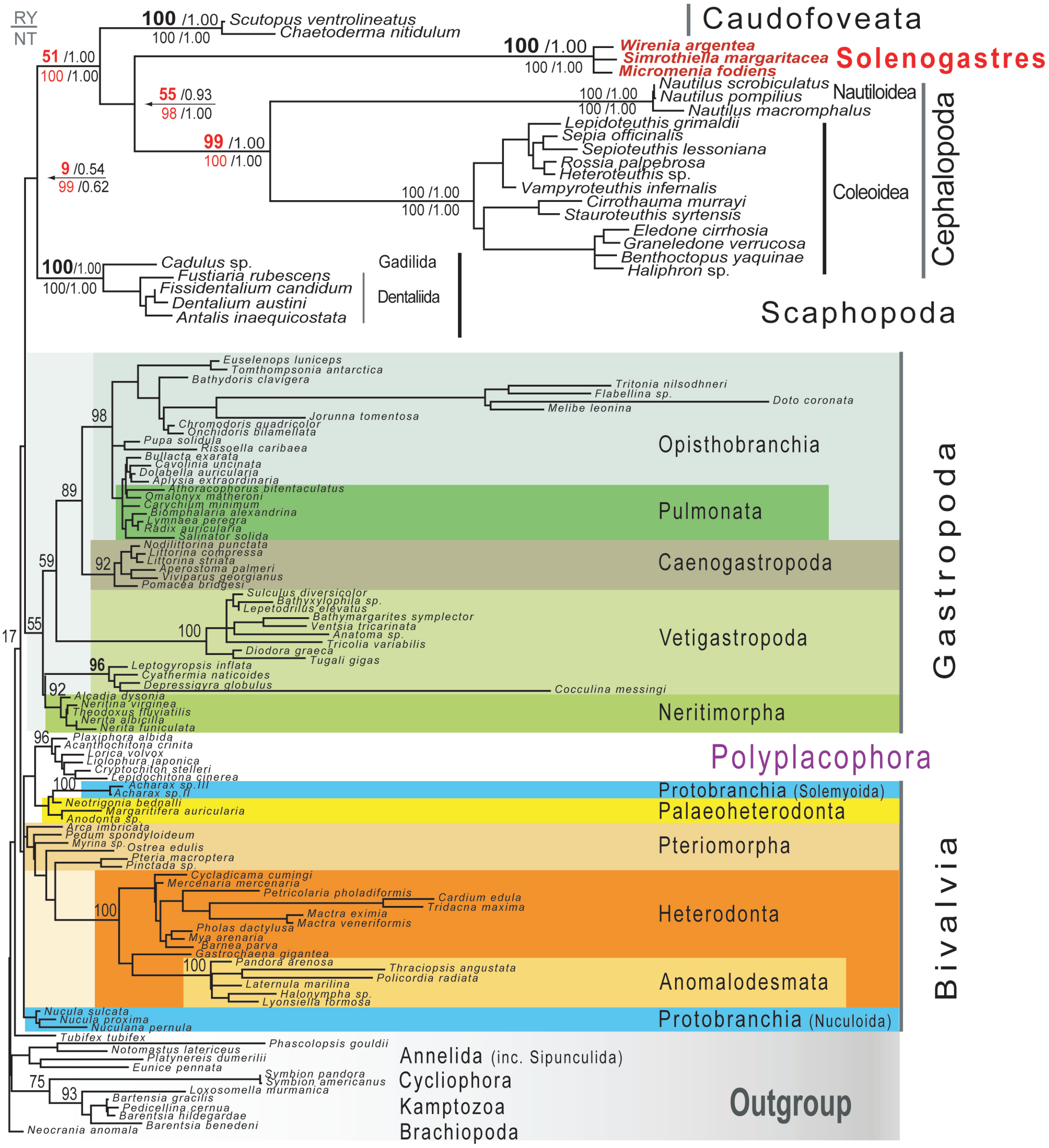


Figure 5.4: ML tree from figure 5.3 with complete species names. All branch length and support values are identical to figure 5.3.

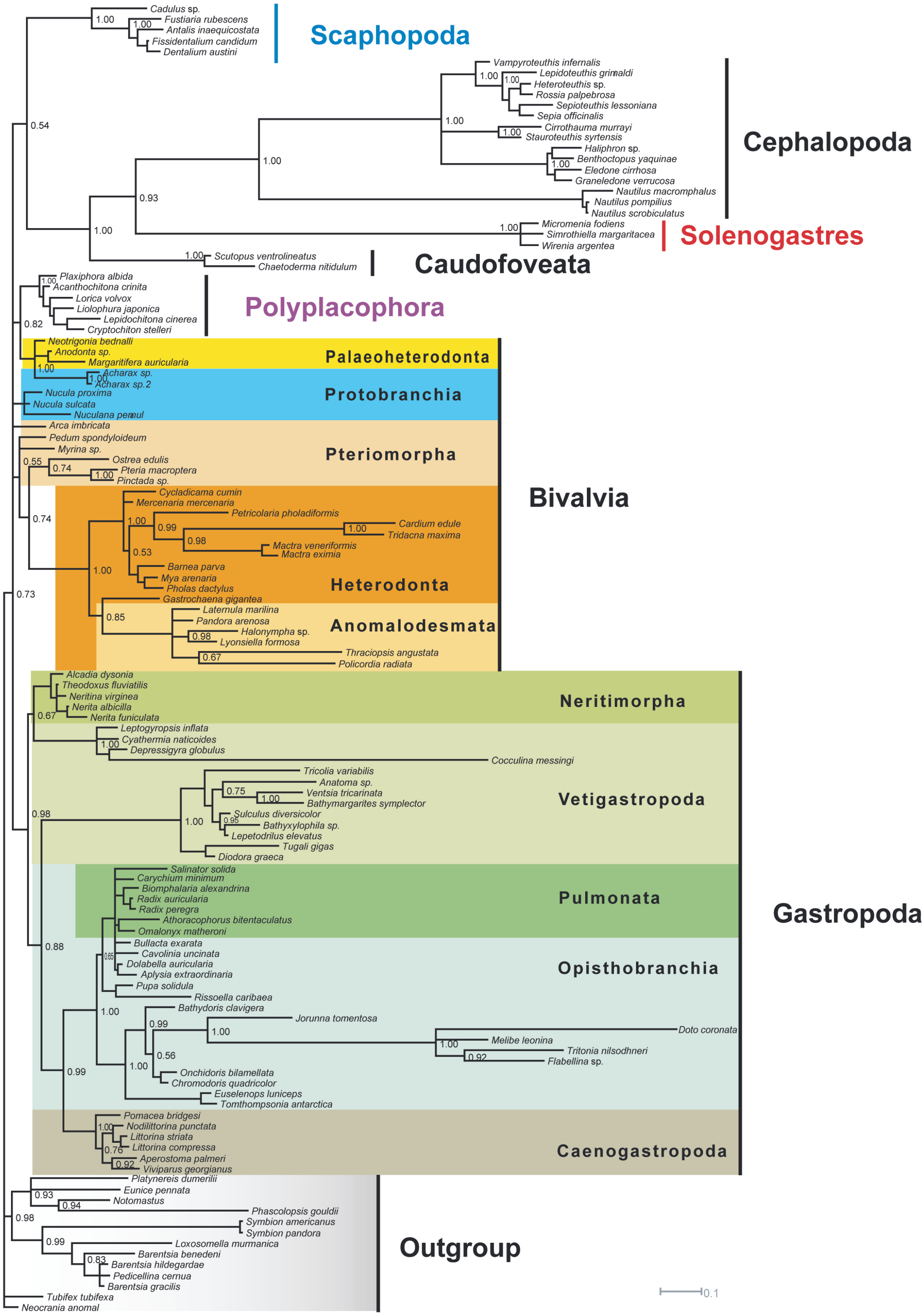


Figure 5.5: Baysian inference. Consensus tree from posterior distribution of the RY-coded alignment widely agrees with the overall topology of the ML analysis.

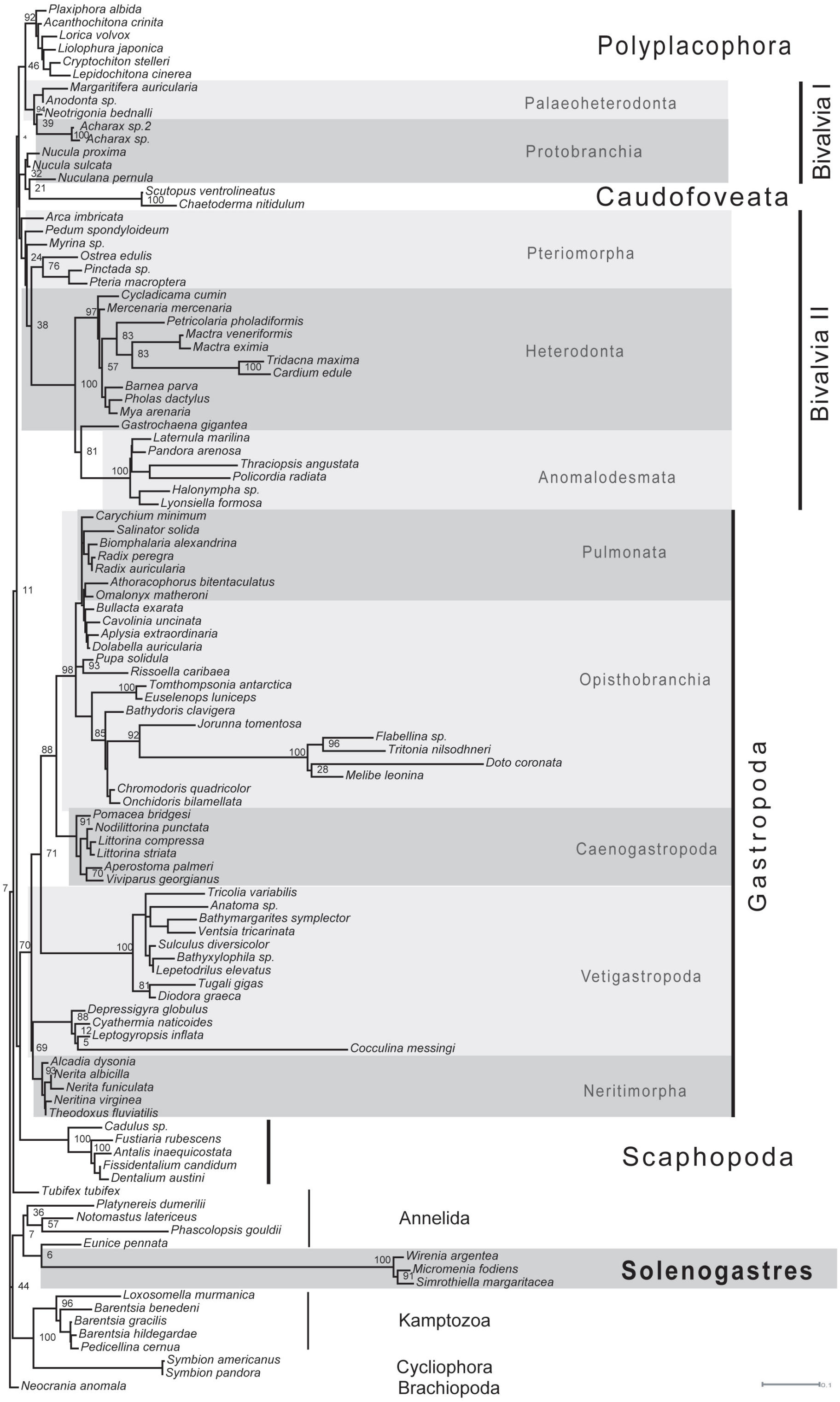


Figure 5.6: ML tree excluding cephalopods. ML analysis of the RY-coded alignment excluding all cephalopod 18S sequences (-ln= -17199.145422, alpha: 0.323075). Solenogastres cluster with Annelida (sensu lato) rendering Mollusca paraphyletic. The positions of Caudofoveata and Scaphopoda are in strong conflict with the inference that includes cephalopods, indicating long-branch attraction effects. Here, Caudofoveata is nested within Protobranchia and Scaphopoda is sistergroup to Gastropoda, although with weak support.

Discussion

GC rich 18S and the chimera problem

Amplification problems are a frequent phenomenon, even if using standard markers such as the 18S rRNA gene. Secondary structures and GC-rich sections comparably pronounced as for the solenogaster 18S RNAs shown here have not been reported earlier. May GC-rich sequences cause hampered PCR reactions in other taxa, too? Based on our results, we assume this to be likely, but equally startling as the failure to obtain gene sequences is the danger to produce chimerical products.

GC-rich sequences demand higher melting temperatures to be converted into and kept as single stranded DNA molecules. Under standard amplification conditions, sequence sections exclusively composed of GC residues will terminate elongation steps in PCR by forming highly stable stem loops or GC clamps causing amplification to be refractory. Incompletely extended primers can anneal to heterologous 18S sequences and, despite of some degree of nucleotide mismatching, they will often be completed in the subsequent polymerization step resulting in chimeras (Wang and Wang 1997). Alternatively, compatible priming sequences from genes with lower GC contents are favored and successfully amplified (Meyerhans et al. 1990). Single stranded sequences with pronounced GC-stem regions also may lead the Taq polymerase to skip the ‘locked’ sections, which leads to shortened PCR products lacking the stem-loop regions (e.g.: Musso et al. 2006). This potentially explains the lack of data for the GC-rich regions in the previously published solenogaster sequences, in contrast to the *Wirenia argentea*, *Simrothiella margaritacea*, and *Micromenia fodiens* sequences described herein. This interpretation is corroborated by the severe amplification problems we experienced when using recombinant plasmids as templates under standard PCR conditions (denaturing step at 94 °C). The cloned *Wirenia argentea* 18S inserts could not be amplified using conventional PCR protocols but resulted in blank agarose gels (not documented) although the number of templates accessible from clone amplification usually exceeds the gene copy number of genomic DNA extractions by far. Failure to amplify solenogaster 18S under such conditions demonstrates that clean starting material is just the first step. In addition to ours, a number of protocols for difficult or GC rich templates are published (Frey et al. 2008; Hube et al. 2005), supplemented by several commercial kits.

Contamination

The solenogaster midgut is extremely voluminous and can hold undigested food that may provide considerable amounts of DNA templates leading to non-target or chimerical amplification products (see above). To avoid contamination with prey organisms, Okusu and Giribet (Okusu and Giribet 2003) suggested the use of gonad tissue, larval material or of species that are not predators of metazoan organisms. The first two options are feasible where material is available, but the third option can be misleading when the diet is unknown. A prey-predator relationship between *Helicoradomenia* and non-cnidarian metazoan organisms, probably polychaetes, has been proposed earlier based on transmission electron microscopy data (Todt and Salvini-Plawen 2005).

Starving animals in the laboratory for several days can reduce the amount of contaminating prey tissue considerably (Todt and Salvini-Plawen 2004), but exogenous DNA was amplified even after starvation times of up to three months when using standard PCR protocols. This may be due to residues of prey cnidocysts held back in the midgut epithelial cells. Thus, we found isolation of total RNA from starved specimens, followed by reverse transcription, to be most effective to avoid contamination. Isolation of total RNA followed by DNase digestion makes up to 95% of rRNA available for RT-PCR and destroys contaminating non-target DNA templates. Due to the ubiquitous presence of RNases causing instability of RNA outside of living cells, the presence of considerable amounts of exogenous RNA within an isolate is highly unusual, but the possibility of contamination with parasites, epibionts, and undigested prey tissue always has to be considered.

Phylogenetic inferences

Our analyses across the Mollusca resulted in short internal branches combined with single long terminal branches. This scenario is known to be critical in phylogenetic tree reconstructions (Felsenstein 1978). Both, the failure to detect Bivalvia as monophyletic and to unambiguously align Patellogastropoda sequences demonstrate the limits of 18S analyses in Mollusca using contemporary methods (see also Passamaneck et al. 2004; Steiner and Müller 1996; Winnepenninckx et al. 1996), even though our new and more representative dataset of 122 taxa covered for the first time all mollusk classes but Monoplacophora. Considering these pitfalls and taking into account well-established knowledge on mollusk relationships, a number of groupings in our tree are ambiguous. This concerns the position of Polyplacophora grouping with Bivalvia and the clade composed of Cephalopoda and Solenogastres, where molecular inferences may reflect a shared bias in base composition rather than a sistergroup relationship: The recoded alignment shows a considerable decrease in support values for the Caudofoveata and Solenogastres node as compared to the NT alignment. In contrast, there is no such decrease for the equally distant Coleoidea-Nautiloidea grouping that is well established also from morphological data. Hence recoded alignments have previously been shown to be more reliable for reconstructing deeper nodes, in using a coding scheme allowing reduction of both substitution saturation and nucleotide compositional bias (Delsuc et al. 2003; Phillips et al. 2006), the branching pattern of Caudofoveata and Solenogastres is probably artificial. Within Solenogastres, the trimming of the alignment adapted to the broad taxon sampling diminished the phylogenetic signal at the species level.

The fundamental modification of the tree after removing Cephalopoda from the taxon sampling is a further indication of long-branch attraction not representing a true phylogenetic signal. The support for a Scaphopoda-Cephalopoda clade is weak in our analyses with the recoded dataset, but is once more (see Haszprunar et al. 2008; Steiner and Dreyer 2003) questioning the Diasoma hypothesis (Runnegar and Pojeta 1974; Salvini-Plawen 1980). Within our outgroup we recovered a previously proposed clade composed of Cycliophora and Kamptozoa (Baguñà et al. 2008; Giribet et al. 2000; Paps et al. 2009; Sørensen and Giribet 2006).

Sequence divergence in Mollusca

Elevated substitution rates are known to be gene specific but also characteristic for certain lineages across different genes (Hedges and Kumar 2003; Welch and Bromham 2005). High substitution rates in the generally well-conserved 18S gene thus may point to fast evolving taxa. To allow for general conclusions across the Mollusca, we inferred substitution rates for all published molluscan 18S sequences longer than 1600bp. Using an untrimmed alignment in the ML inference inflates path length estimates (tree expansion), but our aim was to document diverging lineage specific rates of evolution by including all available information rather than to determine exact 18S substitution rates per site. An estimate of absolute substitution rates in 18S sequences is not feasible due to frequent indel events that cannot be scored confidently at deep time scales.

The solenogaster 18S sequences obtained in this study are among the fastest evolving 18S sequences within the Mollusca. Solenogastres bear a number of assumed ancestral mollusk features, but substitution rates in the 18S gene are nearly doubled in our selected species compared to other presumably “basal” mollusks, such as Caudofoveata and Polyplacophora. If the assumed plesiomorphic morphological characters in Solenogastres are in fact conserved ancestral features, then molecular and morphological rates of evolution are unlinked, at least for the 18S gene. Similar cases of possibly ancestral morphology combined with exceptionally high substitution rates as shown in Figure 5.3 are Nautilus, the “living fossil” cephalopods, and Patellogastropoda, the true limpets (Harasewych and McArthur 2000; Ponder and Lindberg 1997). A number of factors, for example functionality of proteins and RNAs, generation time, metabolic rate, population size, and life histories, are thought to influence substitution rates (Smith and Donoghue 2008; Thomas et al. 2006).

Outlook and Conclusions

We show that the solenogaster 18S gene has an exceptional base composition, thus resulting in a number of technical difficulties. Amplification problems are a common phenomenon but can be overcome by combining known methods in a new framework and by employing alternative strategies. We suggest to include assays with modified PCR methods and to creatively vary PCR conditions if amplification fails or if it leads to ‘peculiar’ results (see also Wintzingerode et al. 1997).

We show that the practical amplification issues of 18S are conquerable, whereas the future of mollusk class level phylogeny appears not to lie in this gene. Multigene analyses are more effective to resolve such ancient splits, as recently demonstrated by Dunn et al. (2008). Within the Mollusca, where evolutionary rates are highly variable between clades, both alignment methods and models of evolution used in phylogenetic analyses at date still remain the main bottlenecks in tracing deep phylogeny.

Methods

Animals and PCR conditions

All Solenogastres specimens were collected off Bergen, Norway and starved up to three months at 4 °C in natural seawater. RNA was extracted from six to twenty animals of each species using Trizol (Invitrogen, Germany) and followed by DNAse digestion on NucleoSpin II columns (Macherey-Nagel, Germany). RT was performed for 45min at

55 °C using 100ng of RNA as template. Either random or the gene specific R1843 primer (Elwood et al. 1985) were applied using Superscript III (Invitrogen, Germany). The GC-rich PCR system (Roche) was used to amplify cDNA adding the supplied GC-rich resolution solution to a final concentration of 0.5M. Three non overlapping fragments were amplified in all species. The thermal profile for the primers 500F (5'GCGGCGCGACGAT CGAAATGAGTCGG3') and 2000R (5'GCCTTATCCCGAGCACGCGCGGGGTTCG3'), annealing at 58 °C was setup as time incremental PCR with 1'35'' + 5''/cycle at 72 °C for 25 times, and a final elongation at 68 °C for 7'. The two smaller, neighboring 18S regions were annealed at 53 °C [primer F19 (Turbeville et al. 1994) and 500R (5'CCGACTCAT TTCGATCGTCGCGCCGC3')] or 56 °C [primer 2000F (5'CGAACCCCGCGCGTGCTCGG3') / R1843 (Elwood et al. 1985)]. All PCR products were excised from 0.8% TAE agarose gels and isolated using the Agarose-Out kit (EURx, Poland). All large central 18S PCR fragments (500F/2000R) were cloned using the TOPO-TA vector (Invitrogen, Germany) and electrocompetent cells DH10B pulsed at 1800 V. Recombinant plasmids from over night grown liquid cultures (1.5 ml) were isolated using the GeneMatrix Miniprep purification kit (EURx, Poland). Sequencing was performed using a M13 primer, gene specific primer or Wiar900F (CCGCGGCCGCCTCG) and R427 (Bleidorn 2005). Cycle sequencing was done applying the Taq Dye Terminator system (Big Dye 3.1, Applied Biosystems) including 10% DMSO in each reaction. All sequences were submitted to Genbank: *Wirenia argentea* FJ649599, *Simrothiella margaritacea* FJ649600 and *Micromenia fodiens* FJ649601.

Sequence conformation, alignment and phylogenetic inferences

Melting curves of double stranded DNA products were analyzed with Stitch Profiles (Tostesen 2008). We determined the critical temperature to melt half of the DNA (helicity=0.5) in *Wirenia argentea* and than applied the determined temperature of 90.7° for all sequences applying the Blossey and Carlon (2003) parameter set. The profiles were aligned to the public available sequences for *Epimenia babai* (AY212107, AY212106), *Epimenia* sp. (AY377657) and *Helicoramenia* sp. (AY145377, AY212108). Secondary structures were calculated at 72°C and 65°C with MFold (Zuker et al. 1999) at the Institut Pasteur webserver (http://bioweb2.pasteur.fr). 18S sequences with a minimum length of 1600bp were collected from GenBank. Prank$_{+F}$ (Loytynoja and Goldman 2008) run twice on a 145 taxon alignment, and a guide tree for a third run was constructed from this alignment using RAxML vers. 7.0.4 (Stamatakis 2006) constraining the molluscan classes as monophyletic except unconstrained Solenogastres. Patellogastropoda was excluded due to adjusting problems leading to 122 aligned species in the final alignment. Columns bearing >90% missing data were discarded. A χ^2 test of homogeneity of base frequencies across taxa was performed using PAUP (Swofford 2000). The resulting alignment was analyzed with RAxML-HPC version 7.0.4 (Stamatakis 2006) and MrBayes (Huelsenbeck and Ronquist 2001) at the CIPRES portal (Stamatakis et al. 2008) under GTR+G substitution model with standard NT-coding (ACTG) as well as an recoded purine (A,G=R) and pyrimidine (C,T=Y) coded alignment. The RY coded alignment was additionally analysed without cephalopods. Node support was calculated from 1000 bootstrap replicates. MrBayes (Huelsenbeck et al. 2001) run 108 generations under the GTR+G model sampling every 250th generation.

Substitution rates

18S data (>1.6kb) from 834 mollusks and 40 non-mollusk outgroup species including all sequence data available for Sipunculida (11 species) and Kamptozoa (5 species) were collected, aligned using muscle (Edgar 2004) and improved after discarding identical or doubtful sequences (option –refine). The resulting alignment was analyzed untrimmed using RAxML with the CAT approximation of rate heterogeneity, conducting a thorough ML search under the GTR+G substitution model after inferring 200 bootstrap replicates. The resulting tree was used to infer the substitutions per site from the most basal node (LCA of *Priapulus caudatus* and all remaining Lophotrochozoans) for all mollusks.

Acknowledgements

We gratefully acknowledge [redacted] for the access to the computer cluster at the FH-Bingen (Bioinformatik) and [redacted] for general support. Funding was provided by Bergens Forskningsstiftelse, Norway (project nr. 481103) to AM and by the University of Bergen free researcher-initiated program (project nr. 226270 Friforsk_08_Todt) to CT.

6. Selecting ribosomal protein genes for invertebrate phylogenetic inferrences - How many genes to resolve the Mollusca?

Abstract

Phylogenomic projects are currently the gold standard in phylogenetics, with sequencing costs rapidly decreasing. Yet, taxon sampling for multigene analyses is in general limited. To facilitate broad taxon representation on an economically tolerable level we suggest restricting datasets to a selected subset of ribosomal protein (RP) genes. Here, we optimized such a subset for the Mollusca, the second most diverse animal phylum. The phylogenetic position of Mollusca is still uncertain and internal relationships are to a high degree unresolved. Starting with a small EST project for *Lepidochitona cinerea* (Polyplacophora, Mollusca), we screened 1,000 ESTs for RP genes. The obtained 32 RP gene sequences were integrated into a data matrix of 79 RP genes (11,963 aa) covering 16 mollusc taxa (four classes). The resulting Maximum Likelihood (ML) tree was used to evaluate each single RP-ML tree according to its fit. RP genes were sorted in ascending order of their subtree agreement metric values. All alignments were successively concatenated and evaluated by ML per site optimisation and the Shimodaira–Hasegawa test. In this manner we determined an economically efficient set of 18 molecular markers to be used for phylogenetic inferences in Mollusca, striving to maximize the informative signal. To further test the influence of increased taxon sampling, we amplified five of these genes via RT-PCR from six species with significant phylogenetic status, including *Micromenia fodiens* (Solenogastres), *Nautilus pompilius* (Cephalopoda) and *Nucula nucleus* (Bivalvia). Phylogenetic analyses based on these five RP genes including 36 molluscan species from six classes recover for the first time the Conchifera and all classes of Mollusca as monophyletic, albeit with weak support. Future analyses will benefit of an increased taxon sampling for the herein specified 18 RPs by using techniques alternative to large EST projects, for example RT-PCR or probe detection of specifically targeted clones.

Introduction

Phylogenomic datasets using Expressed Sequence Tags (EST) data provide high resolution when inferring relationships of major metazoan taxa (e.g.: Steinke, Salzburger, and Meyer 2006; Roeding et al. 2007; Dunn et al. 2008; Philippe et al. 2009). Despite the outstanding advantages of this method are incongruent single gene trees a well known issue. Ebersberger et al. (2007) documented a rate of 23% incongruent trees to the species tree when comparing the genomes of humans and great apes. These results are considered to result from rapid speciation and they are disturbing considering the erosion of phylogenetic signal over time. Especially when looking at events much further back in time, such as the Cambrian explosion when molluscs arose. A closer look at subsets of molecular data thus is necessary, particularly when topologies conflicting with general knowledge arise.

The ribosome is a delineated macromolecular complex and the core of the translation machinery of any organism. Ribosomal protein (RP) genes are highly expressed and have already been successfully applied in a number of phylogenetic studies (Hausdorf et al. 2007; Roeding et al. 2007; Helmkampf, Bruchhaus, and Hausdorf 2008; Struck and Fisse 2008; Timmermans et al. 2008; Witek et al. 2008). Eighty RPs are known in eukaryotes. By definition RPs are present in the ribosome in stoichiometric amounts, unlike e.g. transcription factors and other proteins with less than one copy per ribosome. RPs were originally defined according their arrangement on a two-dimensional polyacrylamid gel, resulting in large acidic proteins to have small numbers and small alkaline proteins to have large numbers (Kaltschmidt and Wittmann 1970). A high degree of sequence conservation across phyla is demonstrated by the overall similarity between sponge and rat RP genes of 79% (Perina et al. 2006). Another important advantage of RPs for phylogenomics is the scattered distribution across the genome, which indicates unlinked loci (Uechi, Tanaka, and Kenmochi 2001; Marygold et al. 2007). As a disadvantage, coevolving sites have been postulated for amino acid residues near tRNA binding sites of four procaryote RPs (Yeang and Haussler 2007).

Metazoan RPs are best studied in mammals. Human RPs have an average number of 5.6 exons per protein (Yoshihama, Nguyen, and Kenmochi 2007) and hardly any splicing variants have been reported so far (Kenmochi et al. 1998). As a rule, each mammalian RP is encoded by a single functional gene (Dudov and Perry 1984; Kuzumaki et al. 1987; Wool, Chan, and Gluck 1995; Kenmochi et al. 1998) but duplicated functional (paralogous) genes may exist. Comparing the RPs from nine metazoan model organisms deposited in the Ribosomal Protein Gene Database on average three out of eighty duplicated RPs per species (2.4%) were detected (Nakao, Yoshihama, and Kenmochi 2004). The maximum of nine paralogous RPs (10%) is found in *Drosophila melanogaster*. All nine duplications appear to have arisen within the Drosophilidae and in eight of the nine duplication events the younger copy has a lower expression level (Marygold et al. 2007). The report of more than 2,000 RP pseudogenes within the human genome also seems worrying. RP pseudogenes are reverse-transcribed mRNAs integrated into the genome, but directly exposed to selection (Zhang, Harrison, and Gerstein 2002). Any significant preservation of processed pseudogenes between human and rodents is missing (Balasubramanian et al. 2009). Due to the recent origin of duplicated RPs and RP pseudogenes and low expression levels of RP paralogues, the paralogy problem appears to be minor when using RP genes for deep metazoan phylogenetics. This encouraged us to sample and apply all 79 RPs available for molluscs, but see Dunn et al. (2008) for a different approach. Even if RPs are frequently recovered by random sequencing of cDNA libraries (ESTs) may taxon rich invertebrate groups not be sampled adequately due to limited funds.

Here we choose the phylum Mollusca as a model to employ gene selection strategies linked with different alignment sizes for RPs. The Mollusca is the second largest metazoan phylum but still very sparsely sampled for EST data. Knowledge about molluscan evolutionary history is surprisingly incomplete despite of a pronounced fossil record and numerous morphological, molecular and developmental data published. Extant molluscs are diagnosed into seven to eight separate lineages, usually referred to as classes (e.g.: Salvini-Plawen and Steiner 1996; Brusca and Brusca 2002; Ponder and Lindberg 2008). Agreement can be found for the monophyletic shell bearing Conchifera separated from a more basal aculiferan grade or clade (molluscs with spicules). The latter comprises three higher taxa with yet unknown phylogenetic status, the vermiform Solenogastres and Caudofoveata (“aplacophoran” molluscs) and the Polyplacophora (chitons). Today’s Conchifera include monoplacophorans (Tryblidia), limpets, snails and slugs (Gastropoda), clams and mussels (Bivalvia), tusk shells (Scaphopoda) as well as octopuses and squids (Cephalopoda). Nearly all the molluscan class level relationships are under dispute.

The sistergroup of the Molluca is likewise a matter of debate. A common annelid and mollusc ancestor has been repeatedly proposed (e.g.: Pelseneer 1899; Pojeta and Runnegar 1976) and a close relationship with sipunculids has been discussed (Scheltema 1993; 1996). Kamptozoans (entoprocts) and molluscs show possible synapomorphies, such as a dorsal chitinous cuticle, a nervous system with four longitudinal nerve cords, and a sinusoid circulatory system (Bartolomaeus 1993; Ax 2000; Haszprunar and Wanninger 2008; Wanninger 2009). Molecular phylogenetic analyses of molluscan class level relationships are to date hampered by limited character and taxon sampling. Using a combination of 18S and 28S genes brought about no convincing support for a certain branching pattern of the molluscan classes (Passamaneck, Schander, and Halanych 2004) and phylogenomic datasets in general suffer from sparse taxon sampling.

Here, we enlarged the available data on mollusc ribosomal proteins and evaluated the phylogenetic value of individual RP genes and of combined RP datasets. To get a more robust phylogenetic signal at the molluscan root we generated 1,000 EST's for the chiton *Lepidochitona cinerea* and additionally enlarged our taxon sampling by amplifying five selected genes from six additional mollusc species. The economically efficient size of a selected dataset was determined by maximizing the fit between tree, substitution model, and alignment per site. We discuss, how the restriction to subsets deduced from large EST-based datasets and the application of selective medium scale data sampling may improve phylogenetic analyses by optimizing taxon and character coverage.

Material and Methods

Generation and processing of EST's

A *Lepidochitona cinerea* specimen was collected at Helgoland island (Germany) and stored at -80°C. RNA was extracted using the Trizol reagent (Invitrogen, Karlsruhe, Germany). mRNA was purified applying the Dynabeads mRNA Purification Kit (Invitrogen, Karlsruhe, Germany) and transcribed by primer extension using SuperScript II (Invitrogen, Karlsruhe, Germany). Size fractioning and directional cloning was done by the CloneMiner cDNA Library Kit (Invitrogen, Karlsruhe, Germany) and the pDONR222 vector. Clones containing cDNA inserts were sequenced from the 5' end on ABI 3730 capillary sequencer systems using the BigDye chemistry (Applied Biosystems, Darmstadt, Germany). EST sequencing of 1,000 reads led to 32 RP contigs from the chiton *Lepidochiton cinerea*: 'positive' clones were additionally sequenced from 3' end or using an oligoT primer.

EST processing was accomplished at the Center for Integrative Bioinformatics in Vienna: Sequence chromatograms were first base-called and evaluated using the Phred application (Ewing et al. 1998). Vector, adaptor, poly-A tract and bacterial sequences were removed employing the software tools Lucy (Chou and Holmes 2001), SeqClean (http://compbio.dfci.harvard.edu/tgi/software, webcite), and CrossMatch (Ewing and Green 1998), respectively. Clustering and assembly of the clipped sequences was performed using the TGICL program package (http://compbio.dfci.harvard.edu/tgi/software, webcite) by performing pairwise comparisons (MGIBlast) and a subsequent clustering step (CAP3). Low quality regions were then removed by Lucy. Contig assembly for taxa found exclusively in trace archives was done analogues in Vienna and retrieved via tBlastn (Altschul et al. 1997) searches (see also Acknowledgements).

Alignment, phylogenetic reconstructions and tree evaluation

Amino acid alignments from the Hausdorf et al. (2007) alignment were completed via tBlastn searches using Human RPs as queries, resulting in a data matrix of 79 RP genes for 16 mollusc and 46 outgroup species. Contigs from up to 50 EST's were assembled in BioEdit (Hall 1999) and translated using GeneWise (Birney, Clamp, and Durbin 2004). *Capitella* sp. sequences not found with a second divergent hit in *Helobdella robusta* ESTs or had 100% identity with *H. robusta* were excluded. We assume a miss specified batch of *Capitella* sp. sequences in genbank. Single genes were aligned using muscle (Edgar 2004) and trimmed with GBlock (Castresana 2000) using all options for a less stringent excision. Substitution models for each RP were determined with MultiPhyl (Keane, Naughton, and McInerney 2007) applying the AICc criterion (Sugiura 1978) before concatenation in BioEdit. Single gene ML phylogenies were calculated using the determined models with Treefinder (Jobb, von Haeseler, and Strimmer 2004) version October 2008. We performed 100 inferences on the original 79 gene alignment and 1,000 bootstrap replicates with RAxML (Stamatakis 2006). We adapted the program settings to this alignment with an initial rearrangement setting of 10, using 25 rate categories and the rtREV model with empirical frequencies. Two Phylobayes (Blanquart and Lartillot 2006) chains sampled 23,000 points under the CAT model (Lartillot and Philippe 2004). 3,000 points were discarded before summarizing over every 10th of the remaining trees (maxdiff. < 0.09; meandiff. < 0.002). Agreement metric and symmetric distance between single gene trees to the 79 gene ML-tree were estimated with PAUP 4.0 (Swofford 2002). The agreement metric results were used to concatenate RPs ordered to their fit with the 79 gene tree. Additionally two alignments comprising half of the genes with poorly matching tree topologies were assembled (labeled 'worst39' and 'worst 40'). The proposal of the substitution model under the AICc criterion and inference of ML trees from the sorted and concatenated genes were conducted using Treefinder. RAxML inferences of selected alignments were additionally run to directly compare the ML topologies with the 79 gene tree applying the SH-test (Shimodaira and Hasegawa 1999) implemented in RAxML.

Amplification of single genes

Animals used for mRNA extraction are listed in table 6.1. Trizol (Invitrogen, Karlsruhe, Germany) or NucleoSpin RNA II (Machery-Nagel, Düren, Germany) was used in total RNA extractions. Frozen tissue (-80°C) from single individuals was used except for Micromenia fodians where multiple living specimen were homogenized after several weeks of starvation in natural seawater. RT-PCR was performed using 200ng total RNA as template applying Superscript III after manufacturer's specifications at 50°C for 45min with an anchored oligoT primer (GAGAGAGGATCCAAGTACTAATACGACTCACTATAGGGAGAT$_{25}$V). Recombinant Taq DNA polymerase, (Invitrogen, Karlsruhe, Germany) was used in subsequent PCR reactions using 1-4 µl cDNA as template. A gene specific forward primer was combined with GAGAGAGGATCCAAGTACTAATACGACTCACTATAGG (Schramm, Bruchhaus, and Roeder 2000) or a gene specific reverse primer with 45s to 1.30min elongation time. PCR products were purified on a 1%TAE agarose gel using the Agarose - Out DNA Purification Kit (EURx, Gdansk, Poland) after excision and sequenced directly applying the BigDye chemistry. Heterogeneous products were cloned with the Topo TA cloning kit (Invitrogen, Karlsruhe, Germany). Whitish clones were selected for clone PCR with redTaq (Genaxxon, Biberach, Germany) and one strand was sequenced using vector specific primer M13 for/rev past isolation of plasmids using the Plasmid Miniprep DNA Purification Kit (EURx, Gdansk, Poland).

Table 6.1: Mollusks collected for this study and RPs amplified via RT-PCR.

Species	Origin	RPs
Gibbula varia	Georgioupolis (Greece)	L10a
Hanleya nagelfar	Bergen (Norway)	L10a, P0
Nucula nucleus	Roscoff (France)	L10a, P0, S10
Mya truncata	Tasiilaq (Greenland)	L4
Micromenia fodiens	Bergen (Norway)	L4, L10a, L17, P0, S10
Nautilus pompilius	Pet store	L4, L10a, S10

EST sequencing of 98 clones from a cDNA library of *Littorina saxatilis* (set up analog to the *Lepidochitona cinerea* cDNA library) led to the sequence of RPL10a (Accession No.: GQ122192).

Alignment and phylogenetic inferences of five RPs

Nucleotide sequences were retrieved via tBlastn searches using human amino acid sequences as queries and possible frame shifts were corrected with GeneWise (see above). Amino acid sequences were aligned using muscle (L10a, L17 and P0). Prank$_{+F}$ (Loytynoja and Goldman 2008) was applied if highly variable 3' regions were observed (S10 and L4). All alignments were visually inspected and corrected for alignment errors. AA alignments were trimmed using Aliscore (Misof and Misof 2009) or first retranslated using translatorX (Telford, unpublished) before fed into Aliscore for the NT alignments. Additionally one cohesive region in P0 contained large amount of missing data and was manually discarded. The five trimmed genes were concatenated with Phyutility (Smith and Dunn 2008) and analysed with RAxML (Stamatakis, Hoover, and Rougemont 2008) conducting 500 bootstrap replicates and using every tenth tree as starting point in a ML search.

Results

Thirty-two RPs were obtained from the EST dataset for *Lepidochitona cinerea* (Accession no. FJ429199-FJ429230). The final concatenated alignment spanned 79 RPs, covered 62 taxa, and comprised 11,911 amino acid positions after the exclusion of ambiguous sites using GBlock. All alignments are available from the authors. Percentages of missing data for the Mollusca and Kamptozoa and species condensed to chimerical taxonomic units are given in table 6.2. The likelihood of the best known ML tree shown in figure 6.1 is -lnL=416928.853285 (Γ shape parameter of a=0.599528). The clade Mollusca got full support (100% bs, 1.00 pp) with gastropods and bivalves forming a robustly supported clade in all analyses. The topologies of the inferred ML and Bayesian trees are largely congruent except the unstable positions of the caudofoveate *Chaetoderma nitidulum*, rotifers, and chaetognaths. *C. nitidulum* got 78% bootstrap support (bs) for a grouping with cephalopods in the ML inference but 1.00 posterior probability as sistertaxon to chitons in the Bayesian analysis. Minor changes between the two inferences were additionally observed for the branching pattern of bivalves. The molluscan classes with more than one representative species in the alignment were recovered with high support values despite the presence of long branching taxa such as the true limpet *Lottia gigantea*. The tree topology inferred with a constraint on Cyrtosoma (Gastropoda+Cephalopoda) is significantly rejected by SH-test (Likelihood: -lnL= 417052.973151 D(LH): -124.119865, SD: 40.294081).

Table 6.2: Chimerical **O**perational **T**axonomic **U**nits (OTU) and the data coverage for Mollusca and Kamptozoa in the concatenated amino acid alignment. *Helix aspersa* and *Mandarina ponderosa* point out more than 80% missing data.

OTU	Species	Missing data
Aplysia califaornica	-	1.75%
Biomphalaria glabrata	-	2.45%
Chaetoderma nitidulum	-	21.45%
Chaetopleura apiculata	-	42.41%
Crassostrea spp.	*Crassostrea gigas, Crassostrea virginica*	2.94%
Dreissena spp.	*Dreissena rostriformis, Dreissena polymorpha*	26.87%
Euprymna scolopes	-	21.22%
Haliotis spp.	*Haliotis discus, Haliotis asanina*	38.01%
Helix aspersa	-	83.33%
Idiosepius paradoxus	-	40.95%
Lepidochitona cinerea	-	57.86%
Lottia gigantea	-	0.68%
Lymnaea stagnalis	-	37.70%
Mandarina ponderosa	-	81.63%
Mytilus spp.	*Mytilus galloprovincialis, Mytilus edulis*	22.87%
Pectinoidea spp.	*Argopecten irradians, Pecten maximus*	2.10%
Barentsia spp.	*Barentsia elongata, Barentsia benedeni*	46.14%
Pedicellina cernua	-	10.87%
Acropora spp.	*Acropora millepora, Acropora palmata*	0.11%
Dugesia spp.	*Dugesia ryukyuensis, Dugesia japonica*	23.40%
Hirudinea spp.	*Helobdella robusta, Haementeria depressa*	33.80%
Lumbricidae spp.	*Lumbricus rubellus, Eisenia andrei*	0.52%

The alternating position of Kamptozoa among the 1,000 bootstrap replicates lies next to Bryozoa, Platyzoa or at the base of Trochozoa. Among the trees sampled from the posterior distribution Kamptozoa settles at the base of Lophotrochozoa and Platyhelminthes (Figure 6.2).

Single gene maximum likelihood analyses of all ribosomal genes were compared with the ML tree of the concatenated alignment using the agreement metric (Goddard et al. 1994) and symmetric distance metric (Penny and Hendy 1985) (Table 6.3; p76). The single gene topologies are to varying extent in conflict with the result from the concatenated dataset and may even disagree with well settled evolutionary hypotheses, with for example the RPL7 tree rejecting monophyletic Mollusca (data not shown). Single gene trees were sorted due to their agreement metric results and concatenated for subsequent analyses. A plot of alignment length against agreement metric results shows the approximate logarithmic correlation of amino acid positions and resolution (Figure 6.3). The likelihood per site is maximised with 18 sorted and concatenated genes comprising 3,348 amino acid (aa) positions and can be increased if L9 is shifted to position 19 (diamond, Figure 6.4A). The Likelihood per site results calculated using Treefinder (Jobb, von Haeseler, and Strimmer 2004) were confirmed in the critical region between 10 and 30 genes with the RAxML (Stamatakis 2006) inferences.

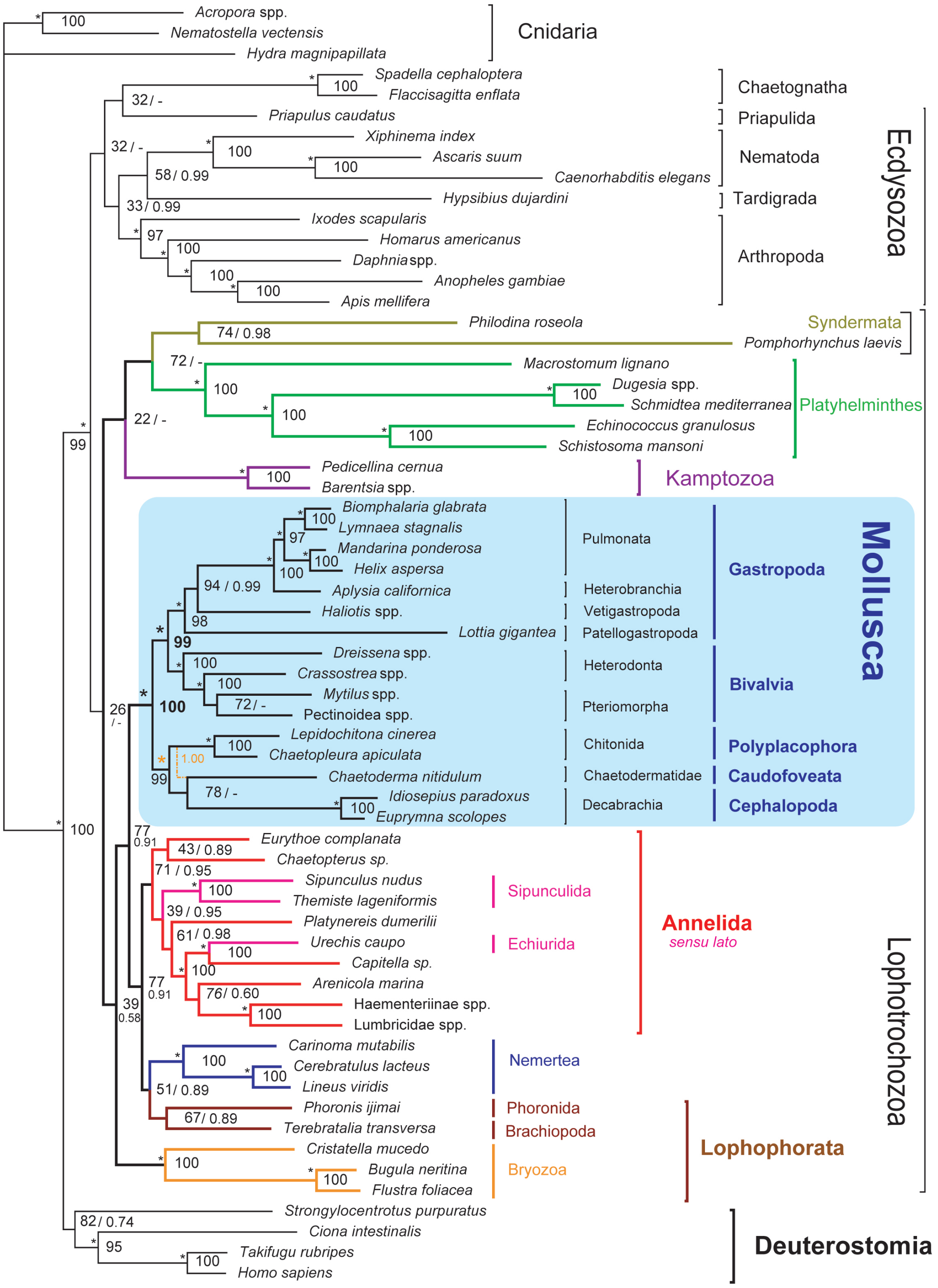


Figure 6.1: ML tree inferred from all 79 RP genes. Topology resulted from a thorough search strategy (100 searches from different starting trees on the original alignment and 1,000 bootstrap replicates). Posterior probabilities from two phylobayes runs are given if the node was conform (pp1.00 = *). The alternative position of *Chaetoderma nitidilum* in the Bayesian analysis is indicated in yellow.

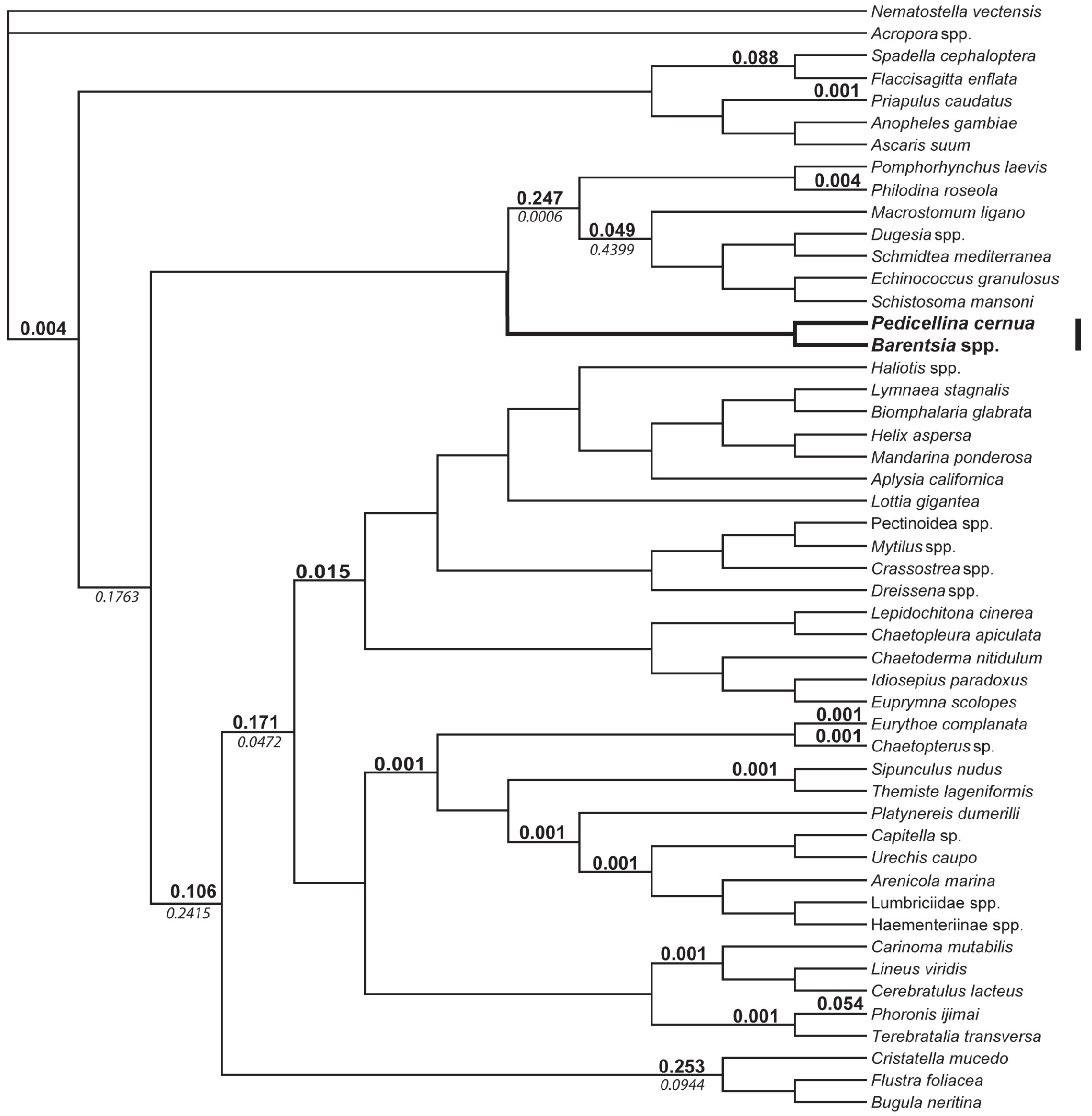


Figure 6.2: Lineage movement for the Kamptozoa mapped on the ML-tree from figure 6.1. Ecdysozoa were pruned and branch lengths are not given. Distribution within the bootstrap replicates above and in the posterior distribution below branches. Frequently a sistergroup relationship to the Bryozoa was recovered (253 out of 1000 replicates), whereas 15 bootstrap replicates recovered the Mollusca as their sistergroup. Within the sampled trees of the posterior distribution the Kamptozoa prevalently oscillate between flatworms and the base of the Lophotrochozoa.

The topological changes using 18 or 79 RPs affect three nodes and changed the position of Priapulida, Polyplacophora and Nemertea (arrows, figure 6.5). The SH test did not reject the 18 gene (3,348 aa) ML-tree if testing against the 79 RP alignment and ML tree (Figure.6.4B). The RAxML tree calculated from the 39 RP genes (4,739 aa) with the lowest agreement metric values was significantly worse and suggests a number of doubtful clades such as (Platyhelminthes+Nematoda) or (Chaetognatha+ Kamptozoa) and ((Gastropoda)+(Haliotis+Bivalvia)).

Table 6.3: Comparison of single gene ML topologies with the ML-tree from the concatenated alignment using all 79 RPs. RPs are sorted due to their agreement metric results (Goddard et al. 1994). The 18 best scoring RPs (see also figure 6.5) are shown on white background. Additionally the results from the partition metric (Symmetric Difference) are given (Penny and Hendy 1985).

Gene	**L10a**	**L17**	**P0**	**SA**	**L4**	**S10**	**L18a**
Agreement-subtree metric	33.078	33.093	34.089	34.108	35.093	35.094	36.082
*Symetric-Difference metric * rank	66 \ 3	63 \ 2	62 \ 1	66 \ 3	69 \ 6	76 \ 18	72 \ 8
Taxon coverage [%], n=62	87	89	87	81	73	87	90
Positions past Gblock [aa]	213	173	286	212	290	112	170
Substitutionmodel [AICc]	rtREV+G	WAG+G	rtREV+G+F	rtREV+G	rtREV+I+G	WAG+G	rtREV+G

Gene	**S4**	**L14**	**S3a**	**L8**	**L13**	**L31**	**L13a**	**L24**	**L23a**
Agree.	36.086	36.086	36.086	36.091	36.095	37.090	37.091	37.098	38.088
Sym. Diff.	82 \ 34	81 \ 31	76 \18	84 \ 44	72 \ 8	81 \ 31	75 \ 14	74 \ 13	79 \ 26
Taxa %	94	79	87	90	84	85	79	84	79
Size aa	253	126	244	248	201	112	194	117	134
Model	rtREV+G	WAG+G	rtREV+G	rtREV+I+G	WAG+I+G	rtREV+G	WAG+G	rtREV+G	WAG+G

Gene	**L21**	**S30**	**S8**	**S3**	**S6**	**S23**	**S17**	**L19**	**L9**
Agree.	38.089	38.090	38.091	39.083	39.085	38.090	39.091	40.080	40.083
Sym. Diff.	82 \ 34	82 \ 34	93 \ 73	72 \ 8	68 \ 3	76 \18	78 \ 24	72 \ 8	80 \ 29
Taxa %	90	77	89	87	81	84	85	84	85
Size aa	152	111	189	222	232	143	173	191	176
Model	rtREV+G	WAG+G	rtREV+G	rtREV+G	WAG+G	rtREV+G	rtREV+I+G	rtREV+G	rtREV+I+G

Gene	**L6**	**S27**	**L18**	**L3**	**S7**	**S2**	**L15**	**L10**	**S14**
Agree.	40.084	40.096	40.102	41.087	41.087	41.091	41.096	41.098	41.102
Sym. Diff.	79 \ 26	82 \ 34	88 \ 65	70 \ 7	73 \ 12	83 \ 39	75 \ 14	76 \ 18	82 \ 34
Taxa %	79	84	84	74	79	79	79	94	87
Size aa	182	84	184	391	188	242	204	206	148
Model	WAG+G	JTT+G	rtREV+G+F	rtREV+I+G	rtREV+G	WAG+I+G	rtREV+G	rtREV+I+G	rtREV+G

Gene	**L7**	**L27a**	**L40**	**S15**	**S12**	**L30**	**S27a**	**S19**	**L12**
Agree.	42.079	42.083	42.083	42.084	42.085	42.088	42.090	42.091	42.093
Sym. Diff.	91 \ 70	88 \ 65	95 \ 78	77 \ 23	79 \ 26	76 \ 18	88 \ 65	75 \ 14	84 \ 44
Taxa %	81	85	85	82	85	73	77	85	77
Size aa	218	137	52	141	118	109	77	131	163
Model	rtREV+G	rtREV+G	rtREV+G	rtREV+G	rtREV+G	JTT+G	rtREV+G	WAG+I+G	rtREV+I+G+F

Gene	**L22**	**L35**	**L35a**	**S9**	**S15a**	**L27**	**L28**	**L11**	**L7a**
Agree.	42.093	42.094	42.094	42.094	42.096	42.108	43.083	43.083	43.086
Sym. Diff.	88 \ 65	83 \ 39	83 \ 39	83 \ 39	87 \ 60	92 \ 72	83 \ 39	86 \ 57	75 \ 14
Taxa %	87	79	84	74	73	87	76	77	84
Size aa	108	122	108	183	130	135	107	169	244
Model	rtREV+G	rtREV+I+G	rtREV+I+G	JTT+G	cpREV+G	rtREV+G	rtREV+G+F	rtREV+G	rtREV+G

Gene	**L32**	**L29**	**S5**	**S18**	**S11**	**S13**	**S24**	**S26**	**L26**
Agree.	43.087	43.087	43.088	43.091	43.092	43.092	43.095	43.098	43.098
Sym. Diff.	85 \ 51	89 \ 69	78 \ 24	84 \ 44	84 \ 44	85 \ 51	86 \ 57	87 \ 60	91 \ 70
Taxa %	89	69	81	84	77	81	81	79	85
Size aa	131	52	189	152	143	201	120	96	132
Model	JTT+G	Dayhoff+I+G	JTT+G	rtREV+G	rtREV+I+G	JTT+G	rtREV+G	JTT+G	rtREV+G

Gene	**L34**	**L37a**	**P1**	**L39**	**L36**	**L5**	**L23**	**S29**	**S21**
Agree.	43.099	43.114	44.072	44.081	44.085	44.086	44.087	44.092	44.104
Sym. Diff.	85 \ 51	87 \ 60	85 \ 51	93 \ 73	84 \ 44	80 \ 29	85 \ 51	86 \ 57	84 \ 44
Taxa %	76	84	81	82	68	84	82	77	69
Size aa	113	91	104	51	95	271	140	56	83
Model	WAG+G+F	rtREV+G	Dayhoff+G	rtREV+G	rtREV+G	WAG+I+G	WAG+G	JTT+G	WAG+G

Gene	**L36a**	**S20**	**S25**	**S16**	**P2**	**L37**	**S28**	**L38**	**L41**
Agree.	45.084	45.090	46.076	46.086	46.093	46.098	46.109	47.100	51.076
Sym. Diff.	81 \ 31	94 \ 76	84 \ 44	85 \ 51	98 \ 79	94 \ 76	93 \ 73	87 \ 60	87 \ 60
Taxa %	71	87	74	81	84	81	76	76	53
Size aa	105	108	93	142	104	86	59	66	25
Model	rtREV+G	JTT+G	rtREV+G	rtREV+G	WAG+G+F	rtREV+G	rtREV+G	rtREV+G	WAG+G

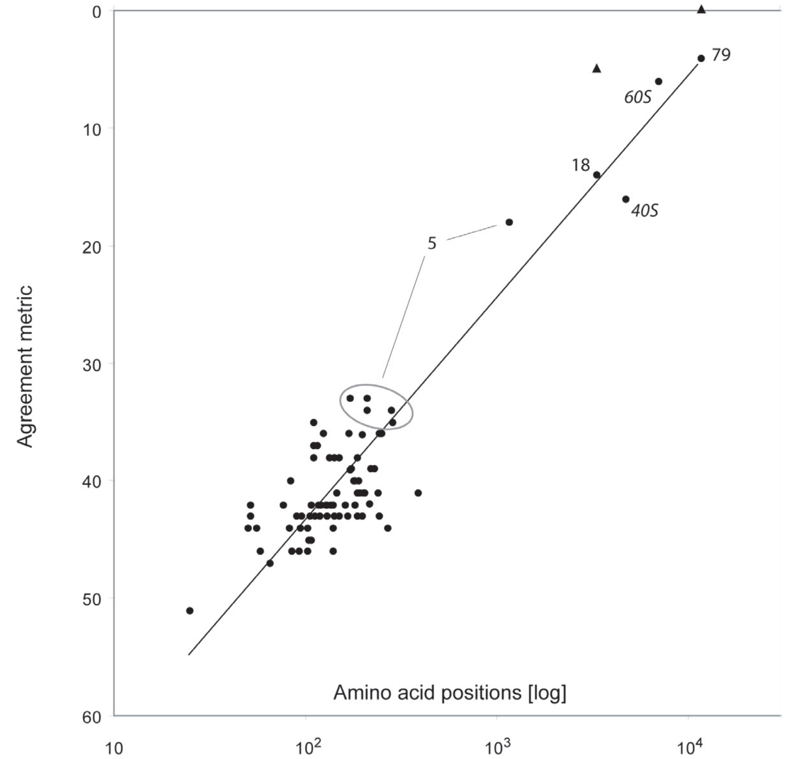
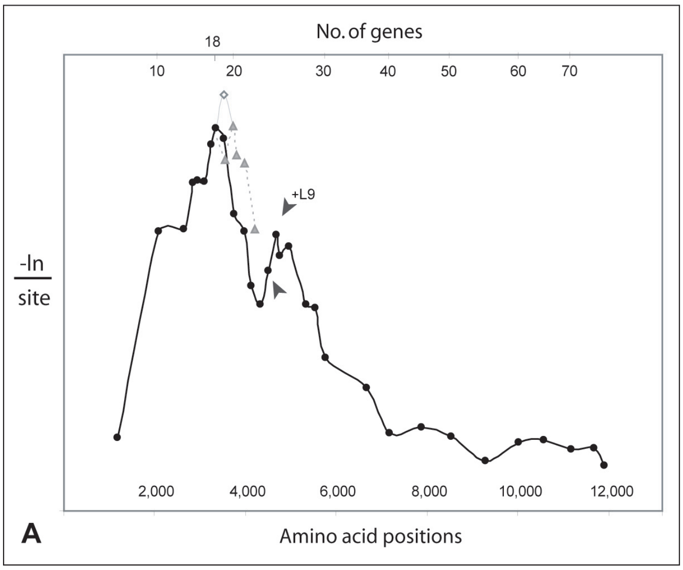


Figure 6.3: Recovery of the tree topology proportional to alignment length. Agreement metric results from the 79RP-ML tree for all single gene analyses and five selected concatenated alignments against alignment length. The logarithmic trend line is given. 40S = small ribosomal subunit (4,723AA), 60S = large ribosomal subunit (7,188AA), Circles = Treefinder topologies, triangle = RAxML topologies. Treefinder and RAxML differ in substitution models and search strategies.



Tree	D(LH)	SD	Significantly worse
best 14	-436.334425	61.537245	Yes
best 15	-367.196877	46.514255	Yes
best 16	-302.791333	42.988815	Yes
best 17	-182.770114	41.455881	Yes
best 18	-36.620631	44.219097	No
best 18+L9	-45.335341	45.625691	No
best 19	-103.619831	46.507915	Yes
best 20	-150.522305	58.108814	Yes
best 21	-254.013939	61.736517	Yes
best 22	-103.619831	46.507941	Yes
best 23	-103.619831	46.507919	Yes
best 24	-112.247924	47.856604	Yes
best 25	-103.619831	46.507873	Yes
best 26	-51.386375	34.82665	No
best 27	-103.619831	46.50795	Yes
best 28	-51.386375	34.826653	No
best 29	-8.883769	23.619968	No
best 30	-8.883767	23.620015	No
best 35	-0.000001	0.000877	No
best 40	-8.883768	23.619955	No
best 50	-19.335846	26.153765	No
worst 39	-418.058212	95.936027	Yes
worst 40	-326.886095	92.650269	Yes

B

Figure 6.4: Alignment size optimisation. A: Genes were sorted due to their agreement metric results as shown in table 6.2. Likelihood per site of ML-trees from concatenated alignments is given against alignment length. The global maximum is found with 18 genes. The addition of L19 and L9 from position 24 and 25 resulted in increasing –ln/site values. Triangles indicate the gradient when these genes are pulled at position 19 and 20. The diamond indicates a global maximum with only one change in agreement order: L9 is shifted to position 19 (best18+L9). B: SH test (Shimodaira and Hasegawa 1999) results for topologies inferred with RAxML from sorted and concatenated alignments of different size tested against the result using all 79 genes.

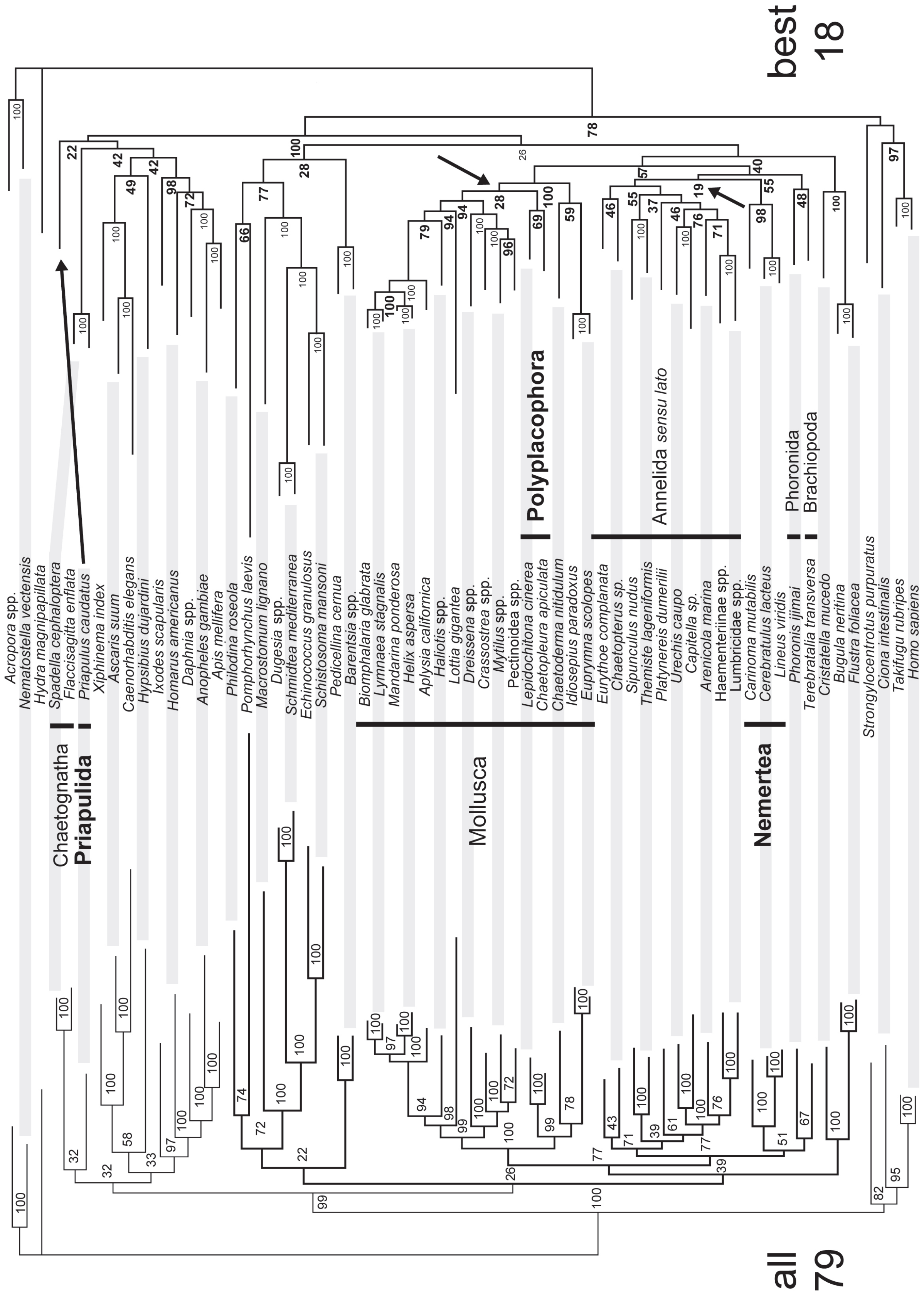


Figure 6.5: Comparison of the ML-tree using all 79 RPs with a subset of 18 selected genes. The RPs with best agreement metric results are listed on blank background in table 6.3 (best 18). Arrows indicate the three rearrangements. The taxa affected by rearrangements are highlighted with bold font (Priapulida, Polyplacophora and Nemertea).

Fifteen RPs were amplified from six additional mollusc species (Table 6.1) and deposited in Genbank with accession No. GQ122191 - GQ122205. The base composition of the concatenated and trimmed alignment is heterogeneous: chi-square = 1462.181904 (df=207), P = 0.00000000. The resulting ML tree inferred from the trimmed nucleotide (NT) and recoded (G,A=R;CT=Y) alignment differs from a tree inferred based on the amino acid alignment regarding the position of cephalopods (see discussion). Aculifera and Conchifera were detected on the DNA alignment (RY and NT) albeit with low support values (Figure 6.6).

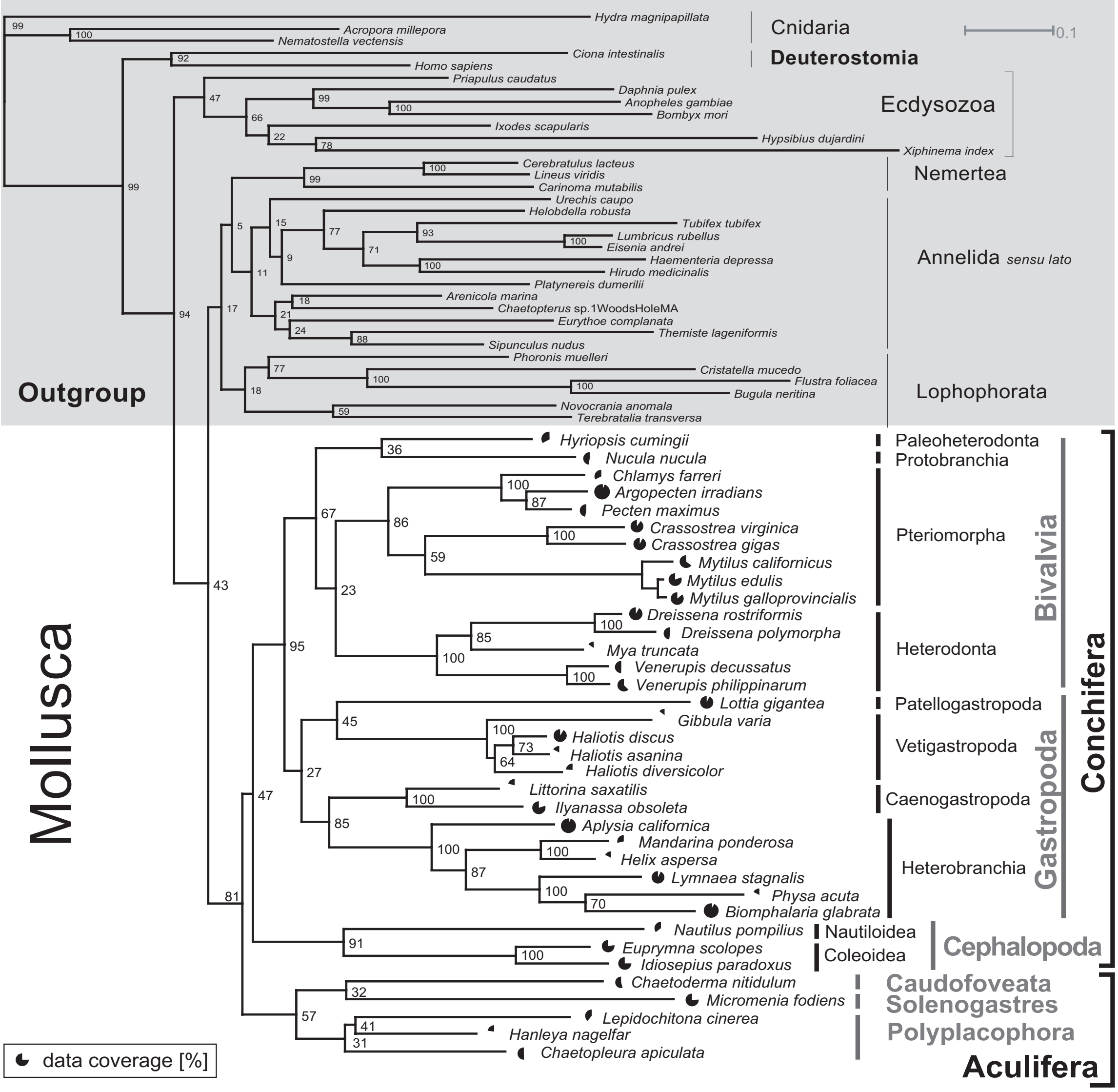


Figure 6.6: RAxML tree inferred from a nucleotide alignment of five RPs. Conchifera and Aculifera are recovered. Bootstrap support of 500 replicates is given. Black pie charts show the data coverage. Fully covered taxa such as *Aplysia californica* or *Lottia gigantea* still have missing data due to indel events.

Discussion

Ribosomal protein genes are highly valuable phylogenetic markers. The inferred topology of the analyses using all 79 RP genes (Figure 6.1) largely agrees with molecular trees inferred from phylogenomic datasets (Bourlat et al. 2006; Philippe and Telford 2006; Dunn et al. 2008). Mollusca and the molluscan classes are monophyletic in our analyses using 18 and 79 RPs, in clear contrast to previously published results based on single or a few genes (Winnepenninckx, Backeljau, and De Wachter 1996; Giribet et al. 2000; Peterson and Eernisse 2001). Generally, phylogenomic approaches remove parts of the raw data that contain high level of noisy signal by gene selection and amino acid coding prior to analysis. For Mollusca, we suggest to sample a small subset of 18 selected ribosomal genes for maximum efficiency, knowing well that the data needed is dependent on the particular question, respectively the taxon, in focus (Gatesy, DeSalle, and Wahlberg 2007).

Gene selection

The selection of adequate genes has been done by comparison of each single gene ML topology with the tree inferred on the concatenated 79 gene alignment. Single gene trees usually display a subset of species reducing the available methods for tree comparison. The symmetric difference or RF-distance (Robinson and Foulds 1981) measures the distance to transfer one tree into another. This leads to a better score of unresolved comb-like trees compared to classification using the agreement metric, as seen with the small L41 gene (Table 6.3). In contrast, the agreement metric prunes leaves from the subtree until the tree topologies fit each other. The drawback here is that infrequently sampled genes are slightly favoured and branch length and node support are not considered. This might cause some misplacing of single genes. Multiple local maxima in the likelihood per site estimation of classified concatenated genes (Figure 6.4A) may point to such misplacing. If, for example, the two genes (L19, L9) inducing the first prominent local maximum (arrows, figure 6.4A) are shifted to position 19 and 20 (triangle, dotted line), the likelihood scores are not improved and a bimodal curve is formed. In contrast, when L9 alone is pulled to position 19 (diamond) the global maximum is elevated (diamond) and the previously bimodal curve (triangle, dotted line) smoothed. The Shimodaira- Hasegawa (SH) test estimated a slightly worse likelihood for this tree (best18+L9) using the 79RP tree and alignment for estimation (Figure 6.4B). Altogether the exact determination of the phylogenetic value of single RP genes is difficult to define, but the 18 selected genes resulted in a tree favoured by the SH test and showing only minor topological changes to the 79 gene tree (arrows, figure 6.5).

Tree inference is on average improved with gene length underlying an approximately logarithmic relation between amino acid positions and resolution (Figure 6.3). Therefore the conducted gene selection favours longer genes, which here is seen an advantage. Larger genes are beneficial when applying directed sampling strategies, such as PCR or probe detection because the effort of sampling one large gene is far less costly than sampling two small genes offering the same amount of sequence data.

Trees inferred using the “worst 39” RP dataset show typical long branch attraction effect issues, thus suggesting that analyses using our complete dataset may be flawed. Obviously such “bad” genes should be excluded from analyses, but it is impossible to select them a priori. Concaterpillar (Leigh et al. 2008) and supertree bootstrapping (Burleigh, Driskell, and Sanderson 2006) are able to identify congruent sets of markers, but the method of gene selection we present here reduces the number of markers simultaneously with quality evaluation. Our approach is optimized for direct sampling strategies and offers the opportunity to incorporate patchy EST datasets. In this manner broader taxon sampling can complement the steadily growing EST datasets.

Phylogenetic inferences based on 79 RP genes

Our new data from the chiton *Lepidochitona cinerea*, representing an assumed basal molluscan lineage, increases markedly the overall gene coverage for Polyplacophora, but nevertheless the position of the clade remains unsettled. Looking at the chiton branch, the Baysian and ML results are inconsistent regarding the placement of Caudofoveata and Cephalopoda (Figure 6.1). Using the subset of 18 RPs revealed three topological changes (Figure 6.5), again affecting the position of Polyplacophora. To evaluate these results, in particular considering the unexpected placement of Cephalopoda, clearly a broader taxon sampling is needed. The SH test significantly rejects a Gastropoda-Cephalopoda clade (Cyrtosoma) but instead we found a close relationship of Gastropoda to Bivalvia. Due to the sessile lifestyle bivalves have many modifications that hamper cladistic analyses inside the group as well as comparison with other molluscs. Beside our molecular data, no possible synapomorphy for gastropods and bivalves is known today, but all characters listed in favour of a gastropod-cephalopod clade might be due to the similar lifestyle of their adult members: concentration of ganglia, free head and relatively small mantle cavity lacking filtering gills. Meyer et al. (Chapter 5) showed that cephalopods have high substitution rates in 18S sequences and that thus artefacts in molecular phylogenetic reconstructions are to be expected. This may also explain the unstable position of *Chaetoderma nitidulum*, changing from sistergroup-relationship to Cephalopoda or Polyplacophora between the two inferences from the 79 gene alignment (Figure 6.1), and the low support for Conchifera in analyses using five RPs.

None of our analyses convincingly recovers the sistergroup of Mollusca within Lophotrochozoa (Figure 6.1, 6.5 and 6.6). Lacunifera – a clade comprising Mollusca and Kamptozoa – has been proposed earlier based on morphological data (Bartolomaeus 1993; Ax 2000; Haszprunar and Wanninger 2008; Wanninger 2009) but is not suggested in our analyses. Alternative positions among bootstrap samples and the posterior distribution detect strong affinity of Kamptozoa to Bryozoa but with a similar frequency a sistergroup relationship of Kamptozoa to Platyzoa is revealed, indicating long-branch attraction problems. The previously proposed Bryozoa sensu lato (Hausdorf et al. 2007; Witek et al. 2008), uniting Kamptozoa and Bryozoa, are not consistent with our results where additional species, in particular the basal bryozoan *Cristatella mucedo*, are included.

Inferences based on the 18 RP gene subset

Topological changes in comparison to the phylogeny derived from the the 79 gene alignment affect only unsettled, weak supported nodes: The grouping (Nemertea,(Annelida, Mollusca)), bs support 19%, versus a (Nemertea,(Brachiopoda, Phoronida) clade, bs support 51%, is not disturbing. The inferred flatworm-nematode clade using the 39 worst scoring genes, in contrast, suggests LBA problems. The dubious groupings inferred with the ‘worst 39’ alignment underline the advantage of discarding such data. We assume that the gain of adding more taxa will outperform the gain of sampling more sequence data.

Inferences based on five RP genes

The concern to add more taxa has often been emphasized (e.g.: Cummings and Meyer 2005) and this is particularly important for phylogenetic reconstructions of diverse and ancient groups, such as the Mollusca (Bergsten 2005). The enlarged number of species included in our five RP gene dataset (Figure 6.6) shows the potential of some taxa, e.g. *Nautilus pompilius,* to break up long branches. Conchifera thus emerge as monophyletic and Solenogastres, Caudofoveata, and Polyplacophora form a monophyletic clade, Aculifera. Additionally, the dataset confirms the close relationship of species condensed to chimerical operational taxonomic units in the large 79 gene alignment, such as *Mytilus* spp. (Table 6.2). The five gene tree, however, is strongly dependent on the exclusion of fast evolving taxa and on the coding of data. For example, if Kamptozoa is added it emerges within or at the base of Gastropoda (NT, RY) or within Annelida (AA). The five-gene tree in some parts obviously suffers from limited sequence data, but the individual and taxon specific sequence composition is crucial. *Gibbula varia* got full support for a plausible placement (Vetigastropoda) near the *Haliotis* species despite 85.51% missing data, whereas three cephalopod species (altogether all genes more than once covered) are not sufficient for a robust placement. Conchifera was not recovered using all RP genes or phylogenomic datasets (Dunn et al. 2008), but emerge as monophyletic with the five RP gene dataset. The weak statistical support might be caused by the limited sequence data sampled. The five gene tree thus clearly demonstrates the importance of broadened taxon sampling (see below).

Conclusion

Phylogenomic multigene datasets generally are restricted in taxon sampling, despite decreasing sequencing costs. In particular species-poor taxa that are not in the centre of public interest are underrepresented, leading to biased species representation. This is also worrisome because a broad taxon representation is important for the identification of saturated positions. This is not possible with only one or a few taxa per gene on hand. The bootstrap support is high if only a few taxa are analysed, but there are examples of fully supported but wrong nodes from pylogenomic datasets (see Soltis et al. 2004). Furthermore, randomly assembled data (ESTs) usually suffer from incomplete (‘gappy’) alignments and poor quality of single reads. Our results suggest to focus on a subset of gene data and to increase efforts to sample additional species for high quality data. We show that assembling 18 selected genes gives promising results (Figure 6.5). Ribosomal proteins were recently suspected to be affected by long branch attraction effects (Bleidorn et al. 2009). This assumption is confirmed here when analyzing the 39 and 40 worst scoring ribosomal proteins (from a total of 79 genes), underlining the importance to carefully select marker genes. Long branch issues can be met with several strategies first and foremost with an enlarged taxon sampling of high quality data focusing on short branching taxa (reviewed in Bergsten 2005).

The Mollusca comprises about 200,000 species (Heywood 1995) that display diverse morphologies. Trying to analyse mollusc phylogenetic relationships using only 0.01% of the clade's extant diversity appears somewhat naive. Collecting larger datasets can be done by RT-PCR methods but especially promising appears the use of alternative strategies, such as probe detection of RP clones from cDNA libraries (Chapter 7). The high degree of sequence conservation within RPs facilitates the application of amplified probes at a broad range of species. Our subset of 18 selected RP genes already now resolves the Mollusca and their classes as monophyletic clades and in future it can be sampled on a broader taxonomic scale to increase resolution and support.

Acknowledgements

We thank the DFG for funding this study within the priority program "Deep Metazoan Phylogeny" project (Li 998/3-1) as well as [redacted] (Max F. Perutz Laboratories, Center for Integrative Bioinformatics, Vienna) for processing of EST data.

7. Realizing broad taxon sampling with large datasets: Probe detection of ribosomal proteins from cDNA libraries

Abstract

The sequencing of 10 to 20 ribosomal protein genes of a specimen is time consuming and labor intensive if applying conventional RT-PCR techniques for data collection. Alternative strategies have to be developed for an economically effective data mining. Five cloned ribosomal protein genes from the two phylogenetically distant species *Sipunculus nudus* (Sipuncula) and *Barentsia elongata* (Kamptozoa) were selected for cross-hybridisation experiments detecting ortholgous target DNAs. Four of the five genes were successfully stained using DIG labled probes. Additionally the plasmid isolation protocol to screen large number of clones had to be adapted and is tested for the simultaneous plasmid isolation from 192 overnight cultures.

Introduction

The previous chapter 6 (Selecting ribosomal protein genes for invertebrate phylogenetic inferences - How many genes to resolve the Mollusca?) described the selection of an economically efficient dataset of ribosomal protein (RP) genes to infer the phylogeny of molluscs. It is possible to collect such data using gene specific primer after the reverse transcription of mRNA, but to reduce the efforts alternative methods to screen for RPs are in need. The staining of orthologous cDNA clones by cross hybridisation of probes and known targets from two phylogenetically distant organisms will show if the preselecting of specifically targeted clones is practicable. Two Expressed Sequence Tag (EST) datasets led to a number of clones for RPs from *Sipunculus nudus* (Sipuncula) and *Barentsia elongata* (Kamptozoa). RPs are highly conserved (Perina et al., 2006) and highly expressed (e.g.: Philippe and Telford, 2006) offering excellent conditions for probe detection. The DIG labelling system is a nonradioactive technology with sensitive and reusable probes for filter hybridization. The DIG detection of labelled DNA is based on antibody conjugates with DIG-11-dUTP, a steroid isolated from digitalis plants. DIG is incorporated during PCR using gene specific primer and sensitivity of this probe is among other things coupled with the amount of incorporated DIG-dUTP. The stringency of hybridisation can be influenced by hybridisation temperature or/and salt concentration. Probes can be reused often and even screened membranes can be stripped and incubated with different probes, repeatedly (Kreike et al., 1990). Five RPs with clones available for both species (*S. nudus* and *B. elongata*) were chosen for a single cross hybridisation experiment. Additionally, a shortened plasmid isolation protocol was tested to skip costly column based DNA purification steps and simplifying large scale dot blot analyses.

Procedure

PCR labelling of probes

Isolated plasmids (Plasmid Miniprep DNA Purification Kit, EURx, Poland) of overnight grown bacterial liquid cultures (LB/ampicillin) from six clones were selected as PCR template: L10 (dmp011tP0001J14) and L27 (dmp011tP0003M04) from *Barentsia elongata* and L8 (dmp010P0005J13), L27a (dmp010P0005E15, dmp010P0002E11) and L29 (dmp010P0005P12) from *Sipunculus nudus*. The recombinant taq polymerase (Invitrogen, Germany) was used to amplify and label six RP probes using 1ng plasmid DNA as template and the following cycling conditions: Initial denaturation at 94°C for 3 min', (95°C for 20'', 55°C for 30'', 72°C for 3') x 32 cycles, and final elongation at 72°C for additional 5 min supplying DIG labelling dNTPs (Roche, Germany). Gene specific primer sequences are given in table 7.1. Dig labelled PCR products (probes) were excised from a 0.8% TAE agarose gel. Bands were purified using Agarose-out (EURx, Poland) and eluted with 50µl H_2O. Staining intensity of the probes was tested applying 1 µl of the eluted probes with dilution series (1:1; 1:10; 1:100; 1:1,000) on a nylon membrane. The detection was performed analogue to the protocol given for the filter hybridisation but substituting the Roche blocking solution with 2% milk powder (Milupa, Germany) in TBS.

Table 7.1: Primer used in probe amplifications

Name	Sequence
L8_for	CCWGGGTCGAGTAATCAGGAGC
L8_rev	GATGGTTCCCTCAGGCATGGT
L10_for	AGGARRSMGYCRAGATGGGKCGC
L10_rev	GGYTTGTAYTGNACRSTNACRCCRTCNGG
L27_for	GGCCGTTATGCTGGCAGAAAAGG
L27_rev	GAAGCGCAACTTTGAGAAGAACCACC
L27a_for	CAGCCATGGACACGGC
L27a_rev	GGAGAATGTCTGAATGGAGG
L29_for	ATGGCCAAGTCCAAGAACCACAC
L29_rev	TCTTCTTGTTGTGCCTCTTGGCAA

Filter hybridisation

The dot blot protocol was adapted from the DIG labelling manual applying all components listed therein (Roche, Germany). Cross linking of target DNA (0.12 Joule/ cm^2, Biolink BLX 254, Peqlab, Germany) was done after denaturating the DNA, and neutralisation and equilibration of the membrane. The membrane was blocked for 45min using 1% blocking solution (Roche, Germany) and hybridized for 1h with diluted 1:7,500 Anti-DIG-alkaline phosphatase coupled antibody. Maleic acid buffer was substituted by TBS (0.1M Tris/Cl, 0.15M NaCl, pH 7.5). All washing and hybridisation solutions were incubated in a shaking bath at 60°C. Detection of target DNA was perfomed NBT and BCIP as recommended by the manufacturer's instruction. Staining was performed in complete darkness to reduce background staining.

Preparation of plasmid DNA for large scale dot blots

192 clones from a *Lepidochitona cinerea* cDNA library (chapter 6) were transferred into two 2ml 96 well plates (Eppendorf, Germany) containing 1.5ml LB/ampicillin for overnight culturing.

Preparation of plasmid DNA by alkaline lysis with SDS was adapted from Sambrook and Russell Protocol 1 (2001). Deep well plates were sealed with parafilm and locked with an in-house production padded cap to enable inverting of the solutions during alkaline lysis. The ethanol purification step was skipped but applying 0.5µl supernatant directly on the nylon membrane. Two probes from ribosomal proteins were kindly provided by M. Heller and used as stated above (Filter hybridisation).

Results

Four of the five RP probes successfully detected the foreign RP clones by cross hybridisation. Plasmid concentration (*Barentsia elongata*, L10) and Dig dUTP incorporation (*Sipunculus nudus*, L29) directly affected the staining intensity achieved during incubation. The staining was documented after two and 12 hours (Figure 7.1). Two hours of incubation with BCIP and NBT led to incomplete RP detection. After 12 h of incubation all orthologous RPs except L27 were stained and negative controls started to appear on the blot (2 central rows, figure 7.1). The plasmid preparation in two (96x) deep well plates from the *Lepidochitona cinerea* cDNA library was successful. The

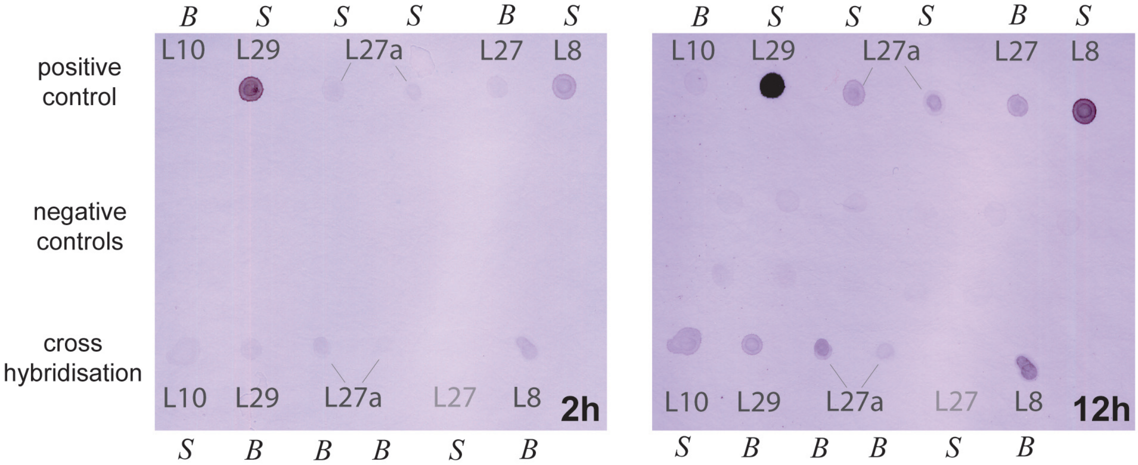


Figure 7.1: Dot Blot with cross hybridisation of ribosomal proteins. Four rows with plasmid preparations from *Sipunculus nudus* and *Barentsia elongata* after two and twelve hours of incubation. The twelve central negative controls applied in two rows include plasmid preparations of different housekeeping genes from Mollusca and Sipuncula. All orthologues genes were detected except RPL27. The L27 probe (app. 100bp) fall below the estimated critical size needed to hybridize at 60°C.

direct application of 0.5µl supernatant from 192 clones on a nylon membrane led to six stained dots using two RP probes during overnight incubation (not documented).

Discussion

The application of DIG labelled probes to detect RPs is applicable for a huge number of clones from cDNA libraries. Four of the five genes were successfully detected. Uprising staining of negative control plasmids during long time incubation hint to prospective false positives if screening unknown clones, but the preselection reduce costs noticeable compared to a random EST sequencing. The L27 probe (app. 100bp) may fell below the estimated critical size needed to hybridise at 60 °C. False positives as well as multiple detections of the same RP has to be taken into account if performing large scale screening with RP-probes, but the separate incubation of less sensitive probes such as L27 at lower temperatures still offers some opportunities for recovery. The expected multiple sequencing of the same RP this is likewise a negligible issue due to steadily decreasing sequencing costs but even enhance data quality. A practicable number of probes simultaneously used for hybridisation still have to be adapted, considering expected sequence similarity and the known amount of incorporated dUTP.

Results from ongoing experiments will allow assembling compatible sets of probes. Even the reuse of stripped membranes may be considered if different probes interfere during hybridisation. The manual of the employed DIG labelling kit (Roche, Germany) delimitate the alkali stripping and reprobing procedure to a maximum of 20 recurrences on the same membrane.

However, the suggested 18 genes (Chapter 6) might not all be accessible using probe detection, but this method can easily be combined with RT-PCR protocols to fully complete the gene sampling. If conserved sequence pattern of a desired RP are below the usual primer length of 15 to 20 nucleotides, peptide nucleic acid primer binding with 6mer to 10mer may enable access to such genes (Ray and Norden, 2000). On the other hand single missing genes are usually not hampering phylogenetic analyses (Philippe et al., 2004).

The construction of a cDNA library from the target organism is costly. In particular, the size fractioning of cDNA inserts remains important to enhance the quality of the cDNA library. The effort to conduct over night cultures and plasmid isolation is minimized here, but still laborious and recommend the use of high quality cDNA libraries. The gains using the outlined protocol are high quality data, compared to low quality reads from EST projects and concomitant a reduction of fund. The cost estimate of this approach is about 25,000.00 EUR for consumables if screening 15 species, but the generation of 1,000 EST’s starting from tissue is tendered about 7,000.00 EUR per species (Aug. 2008, MPI-Berlin) summing up to 100,000.00 EUR total.

Resulting alignments will be much denser compared to ‘gappy’ EST alignments, which than allow to detect saturated positions and improve phylogenetic reconstructions. The drawback is an increased lab work, especially during the initial development of the method, in particular adjusting optimal hybridisation conditions. Once established, a well settled protocol will allow for easy upscale of target species.

Acknowledgement

I thank [redacted]ller for the support during the development of the large scale plasmid isolation procedure.

8. General Discussion

A well-considered data sampling by the means of genes and species is always a trade-off and depends on the available funds. The amount and selection of the required molecular data is again closely associated to the question asked. Here Maximum Likelihood optimization and the Shimodaira–Hasegawa (SH) test (Figure 6.4; p77) suggest to sample the 18 selected marker genes listed in table 6.2 to infer mollusc high level relationships. Auxiliary 18S analyses combined with knowledge from morphological classification can guide the prospective molluscan taxon sampling (see below). The suggested strategy to sample a medium sized dataset is without doubt the most important result of consequence from my thesis and has to be applied to a large number of species. An adapted sampling protocol to collect the eighteen selected ribosomal protein genes is outlined in chapter 7.

Before judging results from large scale analyses and discussing these results in greater detail, it might be helpful to recall some of the most important findings from the recent literature. Jermiin et al. (2005) demonstrated that bootstrap values will increase considerably if longer sequences are used. Likewise the number of species in a phylogenetic inference directly affects node support values by showing decreasing support for intensively sampled clades (Sanderson and Wojciechowski, 2000). Both observations are simple statistical properties of the used methods and thus node support values can not be transferred directly to the confidence in a clade or the underlying phylogenetic hypothesis. Analyzing many characters from a few taxa will always result in well resolved and fully supported trees (e.g.: Rokas et al., 2003). The problem of overestimated confidence for certain nodes is seen in all phylogenetic large scale inferences and needs careful analyses of the data. The choice of the outgroup can affect the result of an inference (Gatesy et al., 2007; Philippe et al., 2009), as well as the individual taxon specific substitution rate (Baurain et al., 2007). Support values and even tree topologies might be incorrect despite using phylogenomic data and sophisticated analytical methods (Soltis et al., 2004); yet some errors can be recognized by applying different methods, using sophisticated models (e.g.: CAT), coding schemes (AA,NT) or simply comparing independent data. Independent data might be different genes and is shown here by comparing single gene analyses in chapter 6. The phylogenetic inference on sipunculids gain from the comparison of mitochondrial, nuclear ribosomal and housekeeping genes. Additionally morphological data are the most important independent data source.

The Mollusca

Molluscs are an outstanding evolutionary model taxon because they offer a comprehensive morphological data record which is a precondition for well-founded testing of molecular phylogenetic hypotheses. The inferences performed within the framework of my thesis include the currently (June 2009) largest molecular datasets (Chapter 5 and 6) published for the Mollusca. Unfortunately the molecular data sampling for mollusks suffers either from limited character (Chapter 5) or relatively sparse taxon sampling (Chapter 6). In general the molluscan classes are morphologically well-defined and have to be recovered in a reliable molecular analysis. Using the ribosomal protein genes, the Mollusca get full support from Maximum Likelihood and Baysian inferences and furthermore the bivalves are found to be monophyletic (Chapter 2 and 6) - both results are difficult to prove using single or few gene datasets (Chapter 5; Winnepenninckx et al., 1996). The limited resolution of major lophotrochozoan lineages is usually assumed to be a result of ancient rapid radiation (Passamaneck et al., 2004). Nevertheless morphological data undoubtedly support the shell bearing Conchifera as a monophyletic higher-ranking taxon within the Mollusca including the cephalopods. Within molecular analyses the position of cephalopods is problematic. Cephalopods render Conchifera paraphyletic in all so far published phylogenomic trees if any representative of ‘Aculifera’ is included (Chapter 2, 5 and 6; Dunn et al., 2008). The five gene tree (Figure 6.6; p79) is the first molecular tree detecting the Conchifera. The overall support of that tree is weak though and furthermore depends on the exclusion of long branching taxa from the analysis, hence clearly demonstrating the future data sampling needed: carefully selected genes with a broad taxon sampling and representing - as far as possible - the observed morphological and evolutionary diversity of recent mollusks.

The results from the 18S analysis are well suited to trace this diversity by molecular methods and have to be discussed in greater detail. The conclusions for the phylogeny of the Mollusca are completed by results based on the protein coding datasets and are discussed afterwards. The phylogenetic inference using the 18S gene identified molecular well-supported clades, as well as clades or taxa with limited or no resolution such as bivalves or Vetigastropoda (Figure 5.3, 5.4 and 5.5; pp 56-58). Gastropods might illustrate how this inference can guide prospective data sampling: within Gastropoda the Caenogastropoda 18S is currently sampled with 118 species offering extremely low molecular variability. Caenogastropoda are the species-richest gastropod lineage with an enormous adaptive radiation and comprising 60% of the described extant 60,000 gastropod species. As a result this taxon needs to be sampled with only one representative in ‘deep node’ inferences due to the low genetic distances within this lineage, whereas the Vetigastropoda would benefit from a larger taxon sampling to trace their diversity. Within Vetigastropoda two distinct clades are observed in the 18S analysis. These clades suggest sampling the fast evolving true limpets with relatively short branching Acmaeidae and moreover with species from the hot vent taxa Neomphalina and Cocculinoidea. Thus beyond a pre-selection grounded on morphological differences, 18S analyses can guide future taxon sampling. The 18S ribosomal RNA and ribosomal protein genes form a macromolecular complex and are expected to be closely linked and resolve different taxonomic levels. According to that the combination of 18S sequences with ribosomal protein data may even suggest supertree approaches for future analyses, without constructing heterogeneous datasets.

Phylogenetic implications of the 18S analyses suffer from short internal branches and limited support but nevertheless give some new insights. To stay close to the above outlined example, within Gastropoda, the Neomphalina and Cocculinoidea branch off before Vetigastropoda sensu stricto. This supports the traditional view (e.g., Ponder and Lindberg, 1997) but conflicts to recent morphological (Sasaki, 1998; Geiger et al., 2008) and also molecular (Geiger and Thacker, 2005) findings. Vetigastropoda are known since the Cambrian/Ordovician boundary and Patellogastropoda are thought to be the most basal gastropod clade (Geiger et al., 2008). Patellogastropoda show exceptionally high substitution rates in the 18S gene and are difficult to align, but ribosomal proteins can be generally aligned using the coding amino acid sequences. Using ribosomal protein data the true limpets (*Lottia gigantea*) are found at a reasonable position at the base of the gastropods.

Similar cases of possibly ancestral morphology combined with exceptionally high substitution rates in the 18S gene are also shown for cephalopods including *Nautilus*, the “living fossil”, and the derived sequences of Solenogastres. A number of factors, such as generation time, metabolic rate, population size, and life histories, are thought to influence substitution rates (Thomas et al., 2006; Smith and Donoghue, 2008). Coleoid cephalopods, for example, have short generation times and high metabolic rates that may cause elevated evolution rates, while correlations of high substitution rates with internal or external factors in nautilids and patellogastropods are more difficult to explain and remain speculative. *Nautilus* spp. have life spans of ten to twenty years (Saunders, 1984) and their high substitution rates may be caused by their low population sizes. Increased substitution rates are also assumed to reflect environmental stress (Pawlowski et al. 1997, Tinn and Oakley 2008). As Davies et al. (2004) showed, molecular evolutionary rates in flowering plants have been faster in high-energy habitats, and this correlation could be a possible reason for long branches seen in true limpets too. The majority of Patellogastropoda inhabit temperate oceans and the temperature profile which limpets are exposed to on intertidal rocks is extreme: they have to resist dramatic temperature shifts if exposed to the sun on dark rocks in summer, getting abruptly cooled down with temperate ocean water during tide or even endure freezing in winter. Furthermore salinity can vary over a broad range and they are exposed to strong UV radiation, highlighting an extreme habitat with dramatic environmental stress, and potentially causing accelerated substitution rates in true limpets.

High substitution rates are also characteristic for parasitic organisms and within sequenced Mollusca representatives of the nudibranch family Facelinidae (Gastropoda, Opistobranchia) show maximum rates. These slugs are food specialists that live as epizoans/parasites on cnidarians and use cnidocysts taken from their host for defense. Some solenogasters have a similar specialized lifestyle, but without the ability to recycle cnidocysts. Not all solenogasters, however, feed on cnidarians for example *Helicoradomenia* spp. predate on polychaets (Todt and Salvini-Plawen, 2005), and also *Simrothiella margaritacea* has a non cnidarian diet (Todt, pers. comm.). The new data for Solenogastres are important due to their exceptional sequence composition and their critical taxonomic position. The previously published sequences are supposed to be contaminated with foreign DNA, but were nonetheless used in a number of phylogenetic analyses (e.g.: Giribet et al., 2000). Unfortunately the high substitution rates of the 18S sequences generated in this study hamper unambiguous alignment construction as well as inferences of class level relationships. The five gene analysis already confirms the close relationship to Caudofoveata although long branches and exceptional base frequencies seen in the 18S analyses of Solenogastres warn of a generally unsettled behavior of Solenogastres in phylogenetic inferences. Likewise, Cephalopods have to be suspected to create respectively suffer from long branch effects. Here morphological data, the 18S gene (Chapter 5) and the five ribosomal gene tree (Chapter 6) strongly suggest sampling more data for Nautiloidea.

The cephalopods are an outstanding taxon to test the plausibility of phylogenomic analyses. Their position within the Conchifera is well founded by synapomorphies like the presence of a single shell and paired statocysts (Haszprunar, 2000). The position of cephalopods inferred using ribosomal genes is contradicting the Conchifera and thus seems unlikely. Scaphopoda and Monoplacophora are still missing in large scale analyses but are essential to finally evaluate the Conchifera conclusively. The Cyrtosoma are a taxon proposed to comprise Gastropoda and Cephalopoda (Salvini-Plawen, 1980; Haszprunar, 2000). The SH test using ribosomal protein genes (Chapter 6) significantly rejects the Cyrtosoma hypothesis: Instead, all large scale analyses result in a close and robustly supported relationship of gastropods with bivalves. Due to their ecologically extreme lifestyle, bivalves display many modifications that hamper both, the cladistic analyses inside the group and comparisons with other mollusks. Bivalved shells evolved convergent within gastropods (Juliidae) and in Brachiopoda, and might not be an obstacle to consider a close relationship between bivalves and gastropods. Apart from molecular data, no possible synapomorphy for gastropods and bivalves is known up to know, but all characters listed in favour of the conflicting Cyrtosoma hypothesis might be shaped by the diverging lifestyle of their adult members: concentration of ganglia, a free head and reduced mantle cavity. The conchiferan trend to spiralisation of the shell is less pronounced in bivalves, although the individual valves of a bivalve shell are coiled (e.g.: Clarkson, 1998). Hence coiled shells are not a possible synapomorphic character uniting Gastropoda and Cephalopoda. Torsion is only known for gastropods. Unfortunately the direct development without larvae in cephalopods impedes the comparison of larval characters found in trochophora and veliger larvae, which are both present in bivalves and gastropods. For this reason larval character can not be consulted if the mode of development differs. Trochophora larvae are the plesiomorphic condition and allow to compare developmental data from the outgroup.

The Kamptozoa

The kamptozoan trochophora larvae possesses a foot with creeping sole similar to the foot of the Mollusca. The adult animal shows a number of possible autapomorphies which support a clade of Mollusca and Kamptozoa (reviewed in Haszprunar and Wanninger, 2008). In this regard Ax (2000) proposed the taxon Lacunifera, but his nomenclature didn’t obtain broad approval. Recently Wanninger (2009) suggested Tetraneuralia to unite both taxa and thereby highlighted another possible important synapomorphie: “...one pair of ventral (pedal) nerve cords with associated serotonergic perikarya and commissures, as well as one pair of lateral (visceral) nerve cords that do not run in the same plane as the pedal cords but more dorsal to them”. Thus morphological data clearly support kamptozoans as sistergroup of the Mollusca (but see Nielsen, 1971; and Emschermann, 1982 for a different view).

The position of Kamptozoa using ribosomal protein analyses is unsettled. The inference described in chapter two includes one kamptozoan genus and the bryozoan *Flustra foliacea* resulting in high support for a unifying clade Bryozoa + Kamptozoa as proposed by Nielsen (2001). Although this clade was not found in the Maximum Likelihood tree after adding more species for both lineages, it still obtained some support from the resampled data as seen with the lineage movement (Figure 6.2; p75).

The 18S data clearly favours a sistergroup relationship of Cycliophora and Kamptozoa. The bootstrap support values for this clade range from 75% to 100%, depending on the included taxa (cephalopods) but irrespective of the coding of the data (NT/RY). This grouping has already been suggested earlier using combined 18S-28S analyses (Passamaneck and Halanych, 2006) and due to the similar possession of protonephridia with multiciliated terminal cells and asexual reproduction by budding (Funch and Kristensen, 1995). Phylogenomic data for Cycliophora are currently on the way and likewise support a grouping of Kamptozoa with Cycliophora (Obst pers. com.: Hejnol et al., submitted). On the other hand there are several proposals for alternative hypotheses: Conflict arise with the typical tetraneural nerve system which is modified in Cycliophora. Cycliophorans have four longitudinal nerve cords but both pairs show different immunocytochemical properties, whereas Kamptozoa and Mollusca own similar staining properties for all four nerve cords (Wanninger, 2005). 18S analyses including rotifers and acanthocephalans revealed a clade of Platyzoa + Cycliophora, but this dataset lacks any basal solitary kamptozoan (Winnepenninckx et al., 1998). Sörensen and Giribet (2006) suggested Gnathostomulida + Cycliophora, but this dataset excluded Kamptozoa.

Cycliophora are filter feeders with a downstream collecting system (similar to Kamptozoa) and an anus straight behind the ciliated funnel comparable to its position in Bryozoa (Funch and Kristensen, 1995). Funch and Kristensen (1995) thus suggested affinities to Kamptozoa and Bryozoa. The close relationship of these three taxa is also suggested summarizing the molecular results presented in this study, though they need some cautious remarks. All molecular analyses suffer from unusual base frequencies in Kamptozoa and high substitution rates in Bryozoa, leading to unstable results. The five gene analyses in chapter 6 excluded Kamptozoa because of their unsettled placing: while amino acid data placed them within annelids, nucleotide data recovered a sistergroup relationship to gastropods, probably indicating unidentified methodological problems. Hopefully multi gene data sampling for the solitary and assumed basal Loxosomatidae accompanied with intensive sampling for further lophotrochozoan taxa will lead to non-ambiguous molecular results in the future, as seen here with sipunculids.

The Sipuncula

The most important phylogenetic implication gained from my thesis concerns the position of peanut worms. The results presented in chapter two show the first statistically high supported subsumption of annelids and sipunculids, evidenced by hypotheses testing. Previous studies likewise assumed an annelid affinity of sipunculids but these results were generally not robustly supported (e.g.: Boore and Staton, 2002). Ribosomal protein data suggest an annelid ingroup position of sipunculids, thus voting for a loss of segmentation. Such possible reduction is further corroborated by trace amounts of neuronal repetitive patterns in *Phascolosoma agassizii* (Kristof et al., 2008), again confirming a loss of segmentation (Wanninger, 2009). Similar results were described earlier for Echiura (e.g.: Bleidorn et al., 2003a), where adults lack of segmentation, but the serial repetition of groups of neuronal perikarya correspond to typical metameric ganglia of annelids (Hessling, 2002). Echiura additionally resemble annelids in embryology and anatomy (Nielsen, 2001). In this study the sistergroup relationship of Capitellidae and Echiura, as proposed by Bleidorn (2003b), is again supported by the analyses on ribosomal proteins (Chapter 6). Considering the loss of segmentation within annelids and any mollusks, the ancestral trochozoan was segmented.

Despite of the complete lack of segmentation in sipunculids, there are similarities to annelid features like the double nerve cord observed in the early development of some species (Rice, 1985; Wanninger et al., 2005) or the nuchal tentacles resembling the polychaete nuchal organ (Brusca and Brusca, 2003). Hazprunar (1996) proposed a metanephridial duct from proper anlage as synapomorphie of Annelida including Echiura, and Sipuncula. The new data from *Sipunculus nudus* presented here can supplement additional molecular characters: the high support from phylogenetic inferences using ribosomal protein genes (Chapter 2 and 6), the translocation of the mitochondrial genes atp6 and nad5 (Chapter 3) and the presence and high similarity of monomeric hemerythrin (Chapter 4). The respiratory hemerythrin protein family is characterized by short sequence lengths causing limited resolution; consequentially these data do not distinguish between an ingroup or sistergroup relationship to annelids.

The cosmopolitan *S. nudus* has only a few taxonomic characters available and as a spin-off from the phylogenetic hemerythrin inference, the existence of a cryptic species is proposed. Each assumed species possesses distinct coelomic hemerythrins characterized by considerably different quaternary structures. The existence of a cryptic species must be proven immediately since the current practice to disperse large quantities of this species as bait in foreign habitats is startling.

The sipunculid hemerythrin data suggest annelid affinity and the Maximum Likelihood inference using eleven protein coding genes from the mitochondrial genome suggests a sistergroup relationship to annelids, concordant to conclusions based on mitochondrial gene order. Other studies based on mitochondrial data gained from the sipunculid genus *Phascolopsis* support an ingroup position of sipunculids (Bleidorn et al., 2006; Struck et al., 2007), in line with studies on ribosomal genes (Chapter 2 and 6) and phylogenomic data (Dunn et al., 2008). The phylogeny of annelids is only poorly understood, which impedes robust hypotheses on the origin of Sipuncula. However, most of the outlined results point to include Sipuncula as annelid subtaxon without peculiar molluscan affinities. Character loss is a frequent phenomenon in annelids and can even affect metamerism (Bleidorn, 2007). Modern sipunculids have a helicoidal digestive tract with an anterior position of the anus (e.g.: Cutler, 1994), whereas that of Cambrian forms is a simple U-shaped tube (Huang et al., 2004). The relocation of the anus at the anterior end of the trunk is an ancient feature which preclude segmentation. The shift of the anus towards the anterior direction from an assumed segmented annelid-like ancestor may has enforced the loss of segmentation. The crossing and reintegration of the intestine thru multiple segments without dissolving the metameric organization is hardly imaginable. An anterior anus must have eased inhabiting holes in solid substrates or empty gastropod shells, habitats occupied by many recent forms (Cuttler, 1994).

Outlook

The molecular data on hand suggest to focus future analyses on (Annelida + Sipuncula) and (Kamptozoa + Cycliophora) to reveal the molluscan sistergroup. Molluscs itself have to be sampled with many more species to get a better understanding of their internal phylogeny. The reduced set of ribosomal protein genes will allow the sampling of a large number of species with high quality data. EST sequences, especially if mainly based on singletons, are error-prone. Additionally the random sequencing of genes leads to varying degrees of missing data in phylogenomic datasets and concatenating multiple genes can mix different phylogenetic signals (Sanderson and Driskell, 2003). Gene families have complex histories of duplication, recombination and silencing. The separate inference of gene families allow to incorporate additional knowledge, such as functional constrains. High quality nucleotide sequences from 18 well characterized ribosomal proteins sampled with large taxon coverage will then allow for careful inspection of the small number of alignments. This is seen as advantageous to highly automated alignment construction and concatenation of EST data. Accessory methods are applicable on smaller datasets, such as spectral analyses to detect alternative splits in the data (Waegele and Mayer, 2007).

Ribosomal proteins were recently suspected to be affected by long branch attraction effects, which were likewise assumed to affect phylogenomic analyses (Bleidorn et al., 2009). This assumption is confirmed in this study after having analyzed the 39 and 40 worst scoring ribosomal proteins (from a total of 79 genes) sorted using the subtree agreement metric (Chapter 6). These results should not completely abandon ribosomal protein genes from phylogenetic inferences though. In fact they rather remind to carefully select marker genes and to compare different and independent datasets. Long branch issues can be met with several strategies first and foremost with an enlarged taxon sampling of high quality data focusing on short branching taxa (see Bergsten, 2005). Ribosomal proteins are very conserved, highly expressed, generally single copy genes, unlinked and scattered over the whole genome, thus ideal molecular markers (reviewed in chapter 6).

The mitochondrial genome demonstrates the high value of the gene position as a valuable phylogenetic character in annelids, but mollusks show highly rearranged mtDNAs, difficult to interpret (Boore et al., 2004). Whole nuclear genome data are currently assembled amongst others for *Lottia gigantea, Octopus vulgaris* and *Capitella* sp. and may prepare future analyses of rare genomic changes to infer phylum level relationships. Identified possible markers have then to be screened for Kamptozoa, Annelida and basal molluscan lineages. These rare genomic changes and presence or absence of genes respectively might be helpful to detect the evolution of taxa with a long and independent history. This is particularly important if species that break up long branches are not available, as it is the case for molecular analyses of kamptozoans. In the end morphological and molecular data must agree the same phylogeny, as all these data own the same genealogical history (Lüter and Bartolomaeus, 1997). Conflicting results can only arise by undetected systematic error within one dataset, as seen with the conflicting position of cephalopods. The Conchifera are morphologically well founded and generally the Mollusca is morphologically very well defined so that they might serve to sharpen molecular methods.

9. Summary

The position of the Mollusca in the phylogenetic system is uncertain and their internal relationships are only scarcely known. To infer mollusk phylogeny three EST-datasets were generated: *Sipunculus nudus* (Sipuncula), *Barentsia elongata* (Kamptozoa) and *Lepidochitona cinerea*, (Polyplacophora, Mollusca). These data were supplemented by single gene amplification experiments.

Kamptozoa and Sipuncula were both discussed as possible molluscan sistergroups, but all data presented here favor a close relationship of Sipuncula with annelids: (i) similar monomeric hemerythrins occur in annelids and sipunculids; (ii) mitochondrial gene order support Annelida as their sistergroup, and (iii) phylogenetic analyses of mitochondrial protein-coding genes and nuclear ribosomal proteins suggest annelid sensu lato comprising Annelida, Echiura and Sipuncula. As mollusks, Sipuncula lack segmentation and these findings have important consequences to interpret the evolution of metamerism, because here molecular phylogenetic analyses suggest that segmentation can get lost. Morphological characters robustly support a clade comprising Mollusca and Kamptozoa but molecular inferences show an unsettled position of Kamptozoa. Within the Mollusca, Solenogastres were previously proposed as possible sistergroup to all remaining extant mollusks. For the first time, valid 18S sequences from the Solenogastres were amplified here and are included in molecular analyses demonstrating huge substitution rate heterogeneity within Mollusca. A possible route to infer their phylogenetic position is suggested using ribosomal protein genes. Subsequently selection of marker genes for concatenation and number of genes included, respectively alignment size is optimized performing Maximum Likelihood per site inferences. Finally eighteen carefully selected ribosomal protein genes are suggested to be economically efficient recovering the major molluscan clades. Medium scale datasets will allow extended taxon sampling, because the reduction in gene number, compared to phylogenomic approaches, enables the application of alternative sampling strategies, other than EST-sequencing. Probe detection of cDNA library clones carrying these genes is introduced here as feasible strategy to assemble species rich datasets.

10. Zusammenfassung

Die phylogenetische Position der Mollusken innerhalb der Trochozoa sowie die interne Evolution der Klassen der Mollusca sind weitgehend unbekannt und werden in meiner Arbeit anhand molekularer Merkmale untersucht. Phylogenomische Analysen zeigten in der Vergangenheit eine gute Auflösung für ursprüngliche Speziationsereignisse. Daher wurden hier drei neue EST Datensätze generiert: für *Sipunculus nudus* (Sipuncula), *Barentsia elongata* (Kamptozoa) und *Lepidochitona cinerea*, (Polyplacophora, Mollusca). Zusätzlich wurden gezielt Gene verschiedener Mollusken mittels RT-PCR amplifiziert.

Sowohl Kamptozoen als auch Sipunculiden wurden aufgrund morphologischer Kriterien bisher als mögliche Schwestergruppe der Mollusken gehandelt, aber die hier erzielten Ergebnisse zur Evolution der Hämerythrine, Gen-Anordnungen der mitochondrialen Genome und phylogenetische Analysen der ribosomalen und der mitochondriellen Proteine stützen diese Hypothese nicht. Die Position der Kamptozoa erwies sich generell als unbeständig; phylogenomische Analysen deuten eine Nähe zu den Bryozoen an, aber diese Position wird stark durch die Auswahl der Taxa beeinflusst. Dagegen weisen meine Analysen klar auf eine nähere Beziehung zwischen Annelida und Sipuncula hin. Die ribosomalen Proteine zeigen Sipuncula (und Echiura) sogar als Subtaxa der Anneliden. Wie den Mollusken fehlt den Sipunculiden jegliche Segmentierung und meine Ergebnisse legen hier die Möglichkeit des Verlusts dieses Merkmals innerhalb der Anneliden bei den Sipunculiden nahe. Innerhalb der Mollusken wurden die Solenogastren bereits als Schwestergruppe aller rezenten Mollusken vorgeschlagen. Im Rahmen meiner Arbeit wurden von drei verschiedenen Solenogastren-Arten die ersten zuverlässigen 18S rRNA-Sequenzen ermittelt, und es zeigte sich, dass alle bisher veröffentlichten 18S-Sequenzen dieser Molluskenklasse höchst unvollständig oder fehlerhaft sind.

Ribosomale Proteine sind gute phylogenetische Marker und hier wurden die Auswahl und Anzahl dieser Gene für phylogenetische Analysen optimiert. Über Sonden-basierte Detektion wurde eine sampling-Strategie getestet, die im Vergleich mit standard-phylogenomischen Ansätzen zukünftige molekulare Stammbaumrekonstruktionen mit größerem Taxonsampling ermöglicht.

11. Collected References

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

Abouheif, E., Zardoya, R., Meyer, A. 1998. Limitations of Metazoan 18S rRNA Sequence Data: Implications for Reconstructing a Phylogeny of the Animal Kingdom and Inferring the Reality of the Cambrian Explosion. *J Mol Evol* **47**, 394-405.

Addison, A. W., Bruce, R. E. 1977. Chemistry of *Phascolosoma lurco* hemerythrin. *Arch Biochem Biophys* **183**, 328-332.

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

Aguinaldo, A. M. A., Turbeville, J. M., Linford, L. S., Rivera, M. C., Garey, J. R., Raff, R. A., Lake, J. A. 1997. Evidence for a clade of nematodes, arthropods and other moulting animals. Nature **387**, 489-493.

Ahlrichs, W. H. 1995a. *Seison annulatus* and *Seison nebaliae*—Ultrastruktur und Phylogenie. *Verh Dtsch Zool Ges* **88**, 115.

Ahlrichs, W. H. 1995b. Zur Ultrastruktur und Phylogenie von *Seison nebaliae* Grube, 1859 und *Seison annulatus* Claus, 1876 Hypothesen zu phylogenetischen Verwandtschaftsverhältnissen innerhalb der Bilateria. Cuvillier, Göttingen.

Ahlrichs, W. H. 1997. Epidermal ultrastructure of *Seison nebaliae* and *Seison annulatus*, and a comparison of epidermal structures within the Gnathifera. Zoomorphology **117**, 41-48.

Aktipis, S. W., Giribet, G., Lindberg, D. R., Ponder, W. F. 2008. Gastropoda. In: Ponder, W. F., Lindberg, D. R. (Eds.), Phylogeny and Evolution of the Mollusca. University of California Press, Berkeley and Los Angeles CA.

Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., Lipman, D. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res*. **25**, 3389-3402.

Ax, P. 1985. The position of the Gnathostomulida and Platyhelminthes in the phylogenetic system of the Bilateria. In: Conway Morris, S., George, J. D., Gibson, R., Platt, H. M. (Eds.), The origins and relationships of lower invertebrates. Clarendon Press, Oxford.

Ax, P. 1995. Das System der Metazoa I. G. Fischer, Stuttgart.

Ax, P. 1999. Das System der Metazoa II. G. Fischer, Stuttgart.

Ax, P. 2000. Multicellular Animals: The phylogenetic system of the metazoa II. Springer Verlag, Berlin.

Baguñà, J., Martinez, P., Paps, J., Riutort, M. 2008. Back in time: a new systematic proposal for the Bilateria. *Philos Trans R Soc Lond B Biol Sci* **363**, 1481-1491.

Bailly, X., Vanin, S., Chabasse, C., Mizuguchi, K., Vinogradov, S. 2008. A phylogenomic profile of hemerythrins, the nonheme diiron binding respiratory proteins. *BMC Evol Biol* **8**, 244.

Balasubramanian, S., Zheng, D., Liu, Y.-J., Fang, G., Frankish, A., Carriero, N., Robilotto, R., Cayting, P., Gerstein, M. 2009. Comparative analysis of processed ribosomal protein pseudogenes in four mammalian genomes. Genome Biology **10**, R2.

Balfour, F. M. 1880. A treatise on comparative embyology. MacMillan, London.

Bandyopadhyay, P., Stevenson, B., Ownby, J., Cady, M., Watkins, M., Olivera, B. 2008. The mitochondrial genome of *Conus textile*, coxI-coxII intergenic sequences and Conoidean evolution. *Mol Phylogenet Evol* **46**, 215 - 223.

Bartolomaeus, T. 1993. Die Leibeshöhlenverhältnisse und Verwandtschaftsbeziehungen der Spiralia. Verhandlungen der Deutschen Zoologischen Gesellschaft 86.

Bartolomaeus, T. 1997. Ultrastructure of the renopericardial complex of the interstitial gastropod *Philinoglossa helgolandica* Hertling, 1932 (Mollusca: Opisthobranchia). Zoologischer Anzeiger **235**, 165-176.

Bates, G., Brunori, M., Amiconi, G., Antonini, E., Wyman, J. 1968. Hemerythrin. I. Thermodynamic and kinetic aspects of oxygen binding. Biochemistry **7**, 3016-&.

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.

Beesley, P. L., Ross, G. J. B., Wells, A. 1998. Mollusca: The Southern Synthesis. CSIRO Publishing, Melbourne.

Bergsten, J. 2005. A review of long-branch attraction. Cladistics **21**, 163 - 193.

Berner, R. A., VandenBrooks, J. M., Ward, P. D. 2007. Evolution - Oxygen and evolution. Science **316**, 557-558.

Birney, E., Clamp, M., Durbin, R. 2004. GeneWise and Genomewise. *Genome Res* **14**, 988 - 995.

Blair, J., Ikeo, K., Gojobori, T., Hedges, S. B. 2002. The evolutionary position of nematodes. *BMC Evol Biol* **2**, 7.

Blanquart, S., Lartillot, N. 2006. A Bayesian Compound Stochastic Process for Modeling Nonstationary and Nonhomogeneous Sequence Evolution. *Mol Biol Evol* **23**, 2058-2071.

Bleidorn, C. 2005. Phylogenetic relationships and evolution of Orbiniidae (Annelida, Polychaeta) based on molecular data. *Zool J Linnean Soc* **144**, 59-73.

Bleidorn, C. 2007. The role of character loss in phylogenetic reconstruction as exemplified for the Annelida. *J Zool Sys Evol Res* **45**, 299-307.

Bleidorn, C., Eeckhaut, I., Podsiadlowski, L., Schult, N., McHugh, D., Halanych, K., Milinkovitch, M., Tiedemann, R. 2007. Mitochondrial genome and nuclear sequence data support myzostomida as part of the annelid radiation. *Mol Biol Evol* **24**, 1690 - 1701.

Bleidorn, C., Podsiadlowski, L., Bartolomaeus, T. 2006. The complete mitochondrial genome of the orbiniid polychaete *Orbinia latreillii* (Annelida, Orbiniidae) - A novel gene order for Annelida and implications for annelid phylogeny. Gene **370**, 96 - 103.

Bleidorn, C., Podsiadlowski, L., Zhong, M., Eeckhaut, I., Hartmann, S., Halanych, K., Tiedemann, R. 2009. On the phylogenetic position of Myzostomida: Can 77 genes get it wrong? *BMC Evol Biol* **9**, 150.

Bleidorn, C., Vogt, L., Bartolomaeus, T. 2003a. A contribution to sedentary polychaete phylogeny using 18S rRNA sequence data. *J Zool Sys Evol Res* **41**, 186-195.

Bleidorn, C., Vogt, L., Bartolomaeus, T. 2003b. New insights into polychaete phylogeny (Annelida) inferred from 18S rDNA sequences. *Mol Phylogenet Evol* **29**, 279-288.

Blossey, R., Carlon, E. 2003. Reparametrizing the loop entropy weights: effect on DNA melting curves. *Phys Rev E Stat Nonlin Soft Matter Phys* **68**, 061911.

Boore, J. 1999. Animal mitochondrial genomes. *Nucleic Acids Res* **27**, 1767 - 1780.

Boore, J. 2004. Complete mitochondrial genome sequence of *Urechis caupo*, a representative of the phylum Echiura. BMC Genomics **5**, 67.

Boore, J., Brown, W. 2000. Mitochondrial genomes of *Galathealinum, Helobdella*, and *Platynereis*: sequence and gene arrangement comparisons indicate that Pogonophora is not a phylum and Annelida and Arthropoda are not sister taxa. *Mol Biol Evol* **17**, 87 - 106.

Boore, J., Collins, T., Stanton, D., Daehler, L., Brown, W. 1995. Deducing the pattern of arthropod phylogeny from mitochondrial DNA rearrangements. Nature **376**, 163 - 165.

Boore, J., Lavrov, D., Brown, W. 1998. Gene translocation links insects and crustaceans. Nature **392**, 667 - 668.

Boore, J. L., Medina, M., Rosenberg, L. A. 2004. Complete Sequences of the Highly Rearranged Molluscan Mitochondrial Genomes of the Scaphopod *Graptacme eborea* and the Bivalve *Mytilus edulis*. *Mol Biol Evol* **21**, 1492-1503.

Boore, J. L., Staton, J. L. 2002. The Mitochondrial Genome of the Sipunculid *Phascolopsis gouldii* Supports Its Association with Annelida Rather than Mollusca. *Mol Biol Evol* **19**, 127-137.

Bourlat, S. J., Juliusdottir, T., Lowe, C. J., Freeman, R., Aronowicz, J., Kirschner, M., Lander, E. S., Thorndyke, M., Nakano, H., Kohn, A. B., Heyland, A., Moroz, L. L., Copley, R. R., Telford, M. J. 2006. Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida. Nature **444**, 85-88.

Brusca, R. C., Brusca, G. J. 2003. Invertebrates. Sinauer Associates, Inc, Sunderland, MA.

Burleigh, J. G., Driskell, A. C., Sanderson, M. J. 2006. Supertree Bootstrapping Methods for Assessing Phylogenetic Variation among Genes in Genome-Scale Data Sets. *Syst Biol* **55**, 426 - 440.

Butterfield, N. J. 2008. An Early Cambrian Radula. Journal of Paleontology **82**, 543-554.

Caldwell, W. H. 1882. Preliminary note on the structure, development and affinities of *Phoronis*. *Proc R Soc Lond B* 371-383.

Caron, J.-B., Scheltema, A., Schander, C., Rudkin, D. 2007. Reply to Butterfield on stem-group ldquowormsrdquo: fossil lophotrochozoans in the Burgess Shale. Bioessays **29**, 200-202.

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17**, 540 - 552.

Cavalier-Smith, T. 1998. A revised six-kingdom system of life. *Biol Rev Camb Philos Soc* **73**, 203-266.

Chenuil, A. 2006. Choosing the right molecular genetic markers for studying biodiversity: from molecular evolution to practical aspects. Genetica **127**, 101-120.

Chou, H. H., Holmes, M. H. 2001. DNA sequence quality trimming and vector removal. Bioinformatics **17**, 1093-1104.

Clark, A. 1921. A new classification of animals. Bulletin de l'Institut Oceanographique.

Clarkson, E. N. K. 1998. Invertebrate palaeontology and evolution. Chapman & Hall, London.

Costa, P. F. E., Gil, J., Passos, A. M., Pereira, P., Melo, P., Batista, F., Da Fonseca, L. C. 2006. The market features of imported non-indigenous polychactes in Portugal and consequent ecological concerns. Scientia Marina **70**, 287-292.

Cummings, M. P., Meyer, A. 2005. Magic bullets and golden rules: Data sampling in molecular phylogenetics. Zoology **108**, 329-336.

Cutler, E. B. 1994. THE SIPUNCULA. Their Systematics, Biology, and Evolution. Cornell University Press, Ithaca and London.

Cutler, E. B., Gibbs, P. E. 1985. A Phylogenetic Analysis of Higher Taxa in the Phylum Sipuncula. *Syst Zool* **34**, 162 - 173.

Cutler, N. J., Cutler, E. B. 1990. A revision of the subgenus *Phascolosoma* (Sipuncula: Phascolosoma). *Proc Biol Soc Wash* **103**, 691-730.

Delsuc, F., Phillips, M. J., Penny, D. 2003. Comment on “Hexapod Origins: Monophyletic or Paraphyletic?”. Science **301**, 1482d-.

Demuynck, S., Li, K. W., Schors, R., Dhainaut-Coutois, N. 1993. Amino acid sequence of the small cadmium-binding protein (MP II) from *Nereis diversicolor* (Annelida, Polychaeta). *Eur J Biochem* **217**, 151-156.

Demuynck, S., Sautiere, P., Vanbeeumen, J., Dhainautcourtois, N. 1991. Homologies between hemerythrins of sipunculids and a cadmium-binding protein (MP II) from a polychaete annelid, *Nereis diversicolor*. *C R Seances Acad Sci III* **312**, 317-322.

Dowton, M., Castro, L., Austin, A. 2002. Mitochondrial gene rearrangements as phylogenetic characters in the invertebrates: The examination of genome ‘morphology’. Invertebrate Systematics **16**, 345 - 356.

Dudov, K. P., Perry, R. P. 1984. The gene family encoding the mouse ribosomal protein L32 contains a uniquely expressed intron-containing gene and an unmutated processed gene. Cell **37**, 457-468.

Dunn, C. W., Hejnol, A., Matus, D. Q., Pang, K., Browne, W. E., Smith, S. A., Seaver, E., Rouse, G. W., Obst, M., Edgecombe, G. D., Sorensen, M. V., Haddock, S. H., Schmidt-Rhaesa, A., Okusu, A., Kristensen, R. M., Wheeler, W. C., Martindale, M. Q., Giribet, G. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature **452**, 745-749.

Ebersberger, I., Galgoczy, P., Taudien, S., Taenzer, S., Platzer, M., von Haeseler, A. 2007. Mapping Human Genetic Ancestry. *Mol Biol Evol* **24**, 2266-2276.

Edgar, R. C. 2004. MUSCLE: Multiple sequence alignment with improved accuracy and speed. Computational Systems Bioinformatics Conference, Proceedings. 2004IEEE, 728-729.

Edwards, S. V., Liu, L., Pearl, D. K. 2007. High-resolution species trees without concatenation. Proceedings of the National Academy of Sciences **104**, 5936-5941.

Eernisse, D. J. 2007. Chitons. University of California Press, Berkley, California.

Eernisse, D. J., Albert, J. S., Anderson, F. E. 1992. Annelida and Arthropoda are Not Sister Taxa: A Phylogenetic Analysis of Spiralian Metazoan Morphology. *Syst Biol* **41**, 305-330.

Ehrenberg, C. G. 1834. Beitrage zur physiologischen Kenntniss der Corallenthiere im allgemeinen, und besonders des rothen Meeres, nebst einem Versuche zur physiologischen Systematik derselben. *Phys. Math. Abh. K. Akad. Wiss*., Berlin.

Elwood, H., Olsen, G., Sogin, M. 1985. The small-subunit ribosomal RNA gene sequences from the hypotrichous ciliates *Oxytricha nova* and *Stylonychia pustulata*. *Mol Biol Evol* **2**, 399-410.

Emanuelsson, O., Brunak, S., von Heijne, G., Nielsen, H. 2007. Locating proteins in the cell using TargetP, SignalP and related tools. *Nat. Protocols* **2**, 953-971.

Emschermann, P. 1972. Loxokalypus socialis gen. et sp. nov. (Kamptozoa, Loxokalypodidae fam. nov.), ein neuer Kamptozoentyp aus dem nördlichen Pazifischen Ozean. Ein Vorschlag zur Neufassung der Kamptozoensystematik. Marine Biology **12**, 237-254.

Emschermann, P. 1982. Les Kamptozoaires. État actuel de nos connaissances sur leur anatomie, leur développement, leur biologie et leur position phylogénétique. *Bull Soc Zool Fr* 317-344.

Emschermann, P. 1995. Kamptozoa (Entoprocta), Kelchwürmer. In: Westheide, W., Rieger, R. (Eds.), Spezielle Zoologie. Gustav Fischer Verlag, Stuttgart, pp. 337-344.

Endo, K., Noguchi, Y., Ueshima, R., Jacobs, H. 2005. Novel repetitive structures, deviant protein-encoding sequences and unidentified ORFs in the mitochondrial genome of the brachiopod *Lingula anatina*. *J Mol Evol* **61**, 36 - 53.

Erber, A., Riemer, D., Bovenschulte, M., Weber, K. 1998. Molecular phylogeny of metazoan intermediate filament proteins. *J Mol Evol* **47**, 751 - 762.

Ewing, B., Green, P. 1998. Base-Calling of Automated Sequencer Traces UsingPhred. II. Error Probabilities. Genome Research **8**, 186-194.

Ewing, B., Hillier, L., Wendl, M., Green, P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Research **8**, 175 - 185.

Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Syst Zool* **27**, 401 - 410.

Field, K. G., Olsen, G. J., Lane, D. J., Giovannoni, S. J., Ghiselin, M. T., Raff, E. C., Pace, N. R., Raff, R. A. 1988. Molecular phylogeny of the animal kingdom. Science **239**, 748-753.

Frey, U. H., Bachmann, H. S., Peters, J., Siffert, W. 2008. PCR-amplification of GC-rich regions: ‘slowdown PCR’. *Nat. Protocols* **3**, 1312-1317.

Friedrich, S., Wanninger, A., Brückner, M., Haszprunar, G. 2002. Neurogenesis in the mossy chiton, *Mopalia muscosa* (Gould) (Polyplacophora): Evidence against molluscan metamerism. *J Morphol* **253**, 109-117.

Funch, P., Kristensen, R. M. 1995. Cycliophora is a new phylum with affinities to Entoprocta and Ectoprocta. Nature **378**, 711-714.

Garey, J., Near, T., Nonnemacher, M., Nadler, S. 1996. Molecular evidence for Acanthocephala as a subtaxon of Rotifera. *J Mol Evol* **43**, 287 - 292.

Garey, J., Schmidt-Rhaesa, A. 1998. The essential role of “minor” phyla in molecular studies of animal evolution. *Amer Zool* **38**, 907-917.

Garey, J., Schmidt-Rhaesa, A., Near, T., Nadler, S. 1998. The evolutionary relationships of rotifers and acanthocephalans. Hydrobiologia 388, 83 - 91.

Gatesy, J., DeSalle, R., Wahlberg, N. 2007. How Many Genes Should a Systematist Sample? Conflicting Insights from a Phylogenomic Matrix Characterized by Replicated Incongruence. *Syst Biol* **56**, 355 - 363.

Geiger, D. L., Nützel, A., Sasaki, T. 2008. Vetigastropoda. In: Ponder, W. F., Lindberg, D. R. (Eds.), Phylogeny and Evolution of the Mollusca. University of California Press, Berkley and Los Angeles, pp. 297-330.

Geiger, D. L., Thacker, C. E. 2005. Molecular phylogeny of Vetigastropoda reveals non-monophyletic Scissurellidae, Trochoidea,and Fissurelloidea. Molluscan Research **25**, 47-55.

Ghirardelli, E. 1981. Chaetognati: posizione sistematica, affinità ed evoluzione del phylum. Origine dei grande phyla dei metazoi. Accademia dei Lincei, Rome, pp. 191-233.

Giribet, G., Distel, D. L., Polz, M., Sterrer, W., Wheeler, W. C. 2000. Triploblastic relationships with emphasis on the acoelomates and the position of Gnathostomulida, Cycliophora, Plathelminthes, and Chaetognatha: a combined approach of 18S rDNA sequences and morphology. *Syst Biol* **49**, 539-562.

Giribet, G., Okusu, A., Lindgren, A. R., Huff, S. W., Schroedl, M., Nishiguchi, M. K. 2006. Evidence for a clade composed of molluscs with serially repeated structures: Monoplacophorans are related to chitons. *Proc Natl Acad Sci U S A* **103**, 7723-7728.

Giribet, G., Wheeler, W. C. 2002. On bivalve phylogeny: a high-level analysis of the Bivalvia (Mollusca) based on combined morphology and DNA sequence data. Invertebrate Biology **121**, 271-324.

Goddard, W., Kubicka, E., Kubicki, G., McMorris, F. R. 1994. The agreement metric for labeled binary-trees. Mathematical Biosciences 123, 215-226.

Goodman, M., Czelusniak, J., Moore, G. W., Romero-Herrera, A. E., Matsuda, G. 1979. Fitting the Gene Lineage into its Species Lineage, a Parsimony Strategy Illustrated by Cladograms Constructed from Globin Sequences. *Syst Zool* **28**, 132-163.

Gorr, T. A., Mable, B. K., Kleinschmidt, T. 1998. Phylogenetic Analysis of Reptilian Hemoglobins: Trees, Rates, and Divergences. *J Mol Evol* **47**, 471-485.

Gribaldo, S., Casane, D., Lopez, P., Philippe, H. 2003. Functional Divergence Prediction from Evolutionary Analysis: A Case Study of Vertebrate Hemoglobin. *Mol Biol Evol* **20**, 1754-1759.

Haen, K., Lang, B., Pomponi, S., Lavrov, D. 2007. Glass sponges and bilaterian animals share derived mitochondrial genomic features: a common ancestry or parallel evolution? *Mol Biol Evol* **24**, 1518 - 1527.

Halanych, K., Bacheller, J., Aguinaldo, A., Liva, S., Hillis, D., Lake, J. 1995. Evidence from 18S ribosomal DNA that the lophophorates are protostome animals. Science **267**, 1641-1643.

Halanych, K. M. 2004. The new view of animal phylogeny. *Annu Rev Ecol Evol Syst* **35**, 229-256.

Hall, T. A. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp Ser* **41**, 95-98.

Harasewych, M. G., McArthur, A. G. 2000. A molecular phylogeny of the Patellogastropoda (Mollusca : Gastropoda). Marine Biology **137**, 183-194.

Haszprunar, G. 1996. The Mollusca: coelomate turbellarians or mesenchymate annelids? Oxford University Press, Oxford.

Haszprunar, G. 2000. Is the Aplacophora monophyletic? A cladistic point of view. American Malacological Bulletin **15**, 115-130.

Haszprunar, G., Schander, C., Halanych, K. M. 2008. Relationships of higher molluscan taxa. University of California Press, Berkley, Los Angeles, London.

Haszprunar, G., Wanninger, A. 2008a. Mollusca and Entoprocta Are Sister Groups: Implications for the Origin of Both Phyla. *J Morphol* **269**, 1473-1473.

Haszprunar, G., Wanninger, A. 2008b. On the fine structure of the creeping larva of *Loxosomella murmanica*: additional evidence for a clade of Kamptozoa (Entoprocta) and Mollusca. *Acta Zool* **89**, 137-148.

Hausdorf, B., Helmkampf, M., Meyer, A., Witek, A., Herlyn, H., Bruchhaus, I., Hankeln, T., Struck, T. H., Lieb, B. 2007. Spiralian Phylogenomics Supports the Resurrection of Bryozoa Comprising Ectoprocta and Entoprocta. *Mol Biol Evol* **24**, 2723-2729.

Hedges, B. S., Kumar, S. 2003. Genomic clocks and evolutionary timescales. *Trends Genet* **19**, 200-206.

Hejnol, A., Obst, M., Stamatakis, A. *submitted*. Further Insight into Animal Relationships from Scalable Phylogenomic and Supercomputing Tools.

Helfenbein, K., Brown, W., Boore, J. 2001. The complete mitochondrial genome of the articulate brachiopod *Terebratalia transversa*. *Mol Biol Evol* **18**, 1734 - 1744.

Helfenbein, K. G., Boore, J. L. 2004. The mitochondrial genome of *Phoronis architecta* - Comparisons demonstrate that phoronids are lophotrochozoan Protostomes. *Mol Biol Evol* **21**, 153-157.

Helfenbein, K. G., Fourcade, H. M., Vanjani, R. G., Boore, J. L. 2004. The mitochondrial genome of *Paraspadella gotoi* is highly reduced and reveals that chaetognaths are a sister group to protostomes. *Proc Natl Acad Sci U S A* **101**, 10639-10643.

Helmkampf, M., Bruchhaus, I., Hausdorf, B. 2008a. Multigene analysis of lophophorate and chaetognath phylogenetic relationships. *Mol Phylogenet Evol* **46**, 206-214.

Helmkampf, M., Bruchhaus, I., Hausdorf, B. 2008b. Phylogenomic analyses of lophophorates (brachiopods, phoronids and bryozoans) confirm the Lophotrochozoa concept. *Proc R Soc Lond [Biol]* **275**, 1927-1933.

Hennig, W. 1979. Wirbellose I (ausgenommen Gliedertiere). G. Fischer, Jena.

Herlyn, H., Piskurek, O., Schmitz, J., Ehlers, U., Zischler, H. 2003. The syndermatan phylogeny and the evolution of acanthocephalan endoparasitism as inferred from 18S rDNA sequences. *Mol Phylogenet Evol* **26**, 155 - 164.

Hessling, R. 2002. Metameric organisation of the nervous system in developmental stages of *Urechis caupo* (Echiura) and its phylogenetic implications. Zoomorphology **121**, 221 - 234.

Heywood, V. H. 1995. Global biodiversity assessment : published for the United Nations Environment Programme. Cambridge Univ. Press, Cambridge.

Hillis, D. M., Dixon, M. T. 1991. Ribosomal DNA: molecular evolution and phylogenetic inference. *Q Rev Biol* **66**, 411-453.

Huang, D., Chen, J., Vannier, J., Salinas, J. 2004. Early Cambrian sipunculan worms from southwest China. *Proc R Soc Lond [Biol]* **271**, 1671 - 1676.

Hube, F., Reverdiau, P., Iochmann, S., Gruel, Y. 2005. Improved PCR method for amplification of GC-Rich DNA sequences. Molecular Biotechnology **31**, 81-84.

Huelsenbeck, J., Ronquist, F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics **17**, 754 - 755.

Huelsenbeck, J. P., Ronquist, F., Nielsen, R., Bollback, J. P. 2001. Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology. Science **294**, 2310-2314.

Hughes, J., Longhorn, S. J., Papadopoulou, A., Theodorides, K., de Riva, A., Mejia-Chang, M., Foster, P. G., Vogler, A. P. 2006. Dense Taxonomic EST Sampling and Its Applications for Molecular Systematics of the Coleoptera (Beetles). *Mol Biol Evol* **23**, 268-278.

Huson, D. 1998. SplitsTree: analyzing and visualizing evolutionary data. Bioinformatics **14**, 68-73.

Hyman, L. 1967. The invertebrates: Mollusca. McGraw-Hill Book Co., New York.

Ivanov, d. L. 1996. Origin of Aculifera and problems of monophyly of higher taxa in molluscs. Oxford University Press, Oxford.

Jacobs, D. K., Wray, C. G., Wedeen, C. J., Kostriken, R., DeSalle, R., Staton, J. L., Gates, R. D., Lindberg, D. R. 2000. Molluscan engrailed expression, serial organization, and shell evolution. Evolution & Development 2, 340-347.

Jägersten, G. 1964. On the morphology and reproduction of entoproct larvae. Zoologiska Bidrag fran Uppsala, 295-315.

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.

Jeffroy, O., Brinkmann, H., Delsuc, F., Philippe, H. 2006. Phylogenomics: the beginning of incongruence? *Trends Genet* **22**, 225-231.

Jennings, R. M., Halanych, K. M. 2005. Mitochondrial Genomes of *Clymenella torquata* (Maldanidae) and *Riftia pachyptila* (Siboglinidae): Evidence for Conserved Gene Order in Annelida. *Mol Biol Evol* **22**, 210-222.

Jermiin, L. S., Poladian, L., Charleston, M. A. 2005. EVOLUTION: Is the “Big Bang” in Animal Evolution Real? Science **310**, 1910-1911.

Jobb, G. 2007. Treefinder. Version of Feb. 2007. Munich, Germany.

Jobb, G., von Haeseler, A., Strimmer, K. 2004. TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. *BMC Evol Biol* **4**, 9.

Jonston, J. 1657. An History of the Wonderful Things of Nature. John Streater, London.

Kaltschmidt, E., Wittmann, H. G. 1970. Ribosomal proteins. XII. Number of proteins in small and large ribosomal subunits of *Escherichia coli* as determined by two-dimensional gel electrophoresis. *Proc Natl Acad Sci U S A* **67**, 1276-1282.

Keane, T. M., Naughton, T. J., McInerney, J. O. 2007. MultiPhyl: a high-throughput phylogenomics webserver using distributed computing. *Nucl Acids Res* **35**, W33-37.

Kenmochi, N., Kawaguchi, T., Rozen, S., Davis, E., Goodman, N., Hudson, T., Tanaka, T., Page, D. 1998. A map of 75 human ribosomal protein genes. *Genome Res* **8**, 509 - 523.

Kilpert, F., Podsiadlowski, L. 2006. The complete mitochondrial genome of the common sea slater, *Ligia oceanica* (Crustacea, Isopoda) bears a novel gene order and unusual control region features. BMC Genomics **7**, 241.

Klippenstein, G. L. 1972. Molecular variants of *Golfingia gouldii* hemerythrin. Primary structure of the variants arising from five amino acid interchanges. Biochemistry **11**, 372-380.

Klippenstein, G. L. 1980. Structural Aspects of Hemerythrin and Myohemerythrin. *Amer Zool* **20**, 39-51.

Kreike, C., de Koning, J., Krens, F. 1990. Non-radioactive detection of single-copy DNA-DNA hybrids. *Plant Mol Biol Reptr* **8**, 172-179.

Kristof, A., Wollesen, T., Wanninger, A. 2008. Segmental mode of neural patterning in sipuncula. *Curr Biol* **18**, 1129-1132.

Kurtz, D. M. 1986. Structure, Function and Oxidation Levels of Hemerythrin. In: Linzen, B. (Ed.) Invertebrate oxygen carriers: International conference., Berlin, pp. 8-22.

Kuzumaki, T., Tanaka, T., Ishikawa, K., Ogata, K. 1987. Rat ribosomal protein L35a multigene family: molecular structure and characterization of three L35a-related pseudogenes. *Biochim Biophys Acta* - Gene Structure and Expression **909**, 99-106.

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., Canback, B. 2008. ARWEN: a program to detect tRNA genes in metazoan mitochondrial nucleotide sequences. Bioinformatics **24**, 172 - 175.

Lauterbach, K. E. 1983. Considerations on the phylogeny of the Mollusca with special reference to the Conchifera. *Z Zool Syst Evol* **21**, 201-216.

Lavrov, D., Brown, W., Boore, J. 2004. Phylogenetic position of the Pentastomida and (pan)crustacean relationships. *Proc R Soc Lond B Biol Sci* **271**, 537 - 544.

Lavrov, D., Lang, B. 2005. Poriferan mtDNA and animal phylogeny based on mitochondrial gene arrangements. *Syst Biol* **54**, 651 - 659.

Leigh, J. W., Susko, E., Baumgartner, M., Roger, A. J. 2008. Testing congruence in phylogenomic analysis. *Syst Biol* **57**, 104-115.

Lieb, B., Altenhein, B,, Markl Jr., Vincent A., van Olden E., van Holde K. E., Miller K. I. 2001. Structures of two molluscan hemocyanin genes: Significance for gene evolution. *PNAS* 98:4546-4551.

Lieb, B. 2003. A new metallothionein gene from the giant keyhole limpet *Megathura crenulata*. *Comp Biochem Physiol C* 134, 131-137.

Lieb, B., Todt, C. 2008. Hemocyanin in mollusks-A molecular survey and new data on hemocyanin genes in Solenogastres and Caudofoveata. *Mol Phylogenet Evol* 49:382-385.

Lindberg, D. R., Ponder, W. F., Haszprunar, G. 2004. The Mollusca: relationships and patterns from their first half-billion years. Oxford University Press, Oxford.

Lindgren, A. R., Giribet, G., Nishiguchi, M. K. 2004. A combined approach to the phylogeny of Cephalopoda (Mollusca). Cladistics **20**, 454-486.

Littlewood, D. T. J., Telford, M. J., Clough, K. A., Rohde, K. 1998. Gnathostomulida--An Enigmatic Metazoan Phylum from both Morphological and Molecular Perspectives. *Mol Phylogenet Evol* **9**, 72-79.

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

Loytynoja, A., Goldman, N. 2008. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science **320**, 1632-1635.

Ludt, C. 1995. Charakterisierung des Hämerythrins von *Halicryptus spinulosus* (Priapulida). University of Heidelberg, Heidelberg.

Lüter, C., Bartolomaeus, T. 1997. The phylogenetic position of Brachiopoda – a comparison of morphological and molecular data. Zoologica Scripta **26**, 245-253.

Mackey, L. Y., Winnepenninckx, B., DeWachter, R., Backeljau, T., Emschermann, P., Garey, J. R. 1996. 18S rRNA suggests that Entoprocta are protostomes, unrelated to Ectoprocta. *J Mol Evol* **42**, 552-559.

Mallatt, J., Winchell, C. J. 2002. Testing the New Animal Phylogeny: First Use of Combined Large-Subunit and Small-Subunit rRNA Gene Sequences to Classify the Protostomes. *Mol Biol Evol* 19, 289-301.

Manwell, C. 1960. Oxygen Equilibrium of Brachiopod *Lingula* Hemerythrin. Science **132**, 550-551.

Marletaz, F., Martin, E., Perez, Y., Papillon, D., Caubit, X., Lowe, C. J., Freeman, B., Fasano, L., Dossat, C., Wincker, P., Weissenbach, J., Le Parco, Y. 2006. Chaetognath phylogenomics: a protostome with deuterostome-like development. Current Biology **16**, R577-R578.

Marygold, S., Roote, J., Reuter, G., Lambertsson, A., Ashburner, M., Millburn, G., Harrison, P., Yu, Z., Kenmochi, N., Kaufman, T., Leevers, S., Cook, K. 2007. The ribosomal protein genes and Minute loci of *Drosophila melanogaster*. Genome Biology **8**, R216.

Maslakova, S., Martindale, M., Norenburg, J. 2004. Fundamental properties of the spiralian developmental program are displayed by the basal nemertean C*arinoma tremaphoros* (Palaeonemertea, Nemertea). *Dev Biol* **267**, 342 - 360.

Matus, D. Q., Copley, R. R., Dunn, C. W., Hejnol, A., Eccleston, H., Halanych, K. M., Martindale, M. Q., Telford, M. J. 2006. Broad taxon and gene sampling indicate that chaetognaths are protostomes. Current Biology **16**, R575-R576.

Maxmen, A. B., King, B. F., Cutler, E. B., Giribet, G. 2003. Evolutionary relationships within the protostome phylum Sipuncula: a molecular analysis of ribosomal genes and histone H3 sequence data. *Mol Phylogenet Evol* **27**, 489-503.

McHugh, D. 1997. Molecular evidence that echiurans and pogonophorans are derived annelids. *Proc Natl Acad Sci U S A* **94**, 8006 - 8009.

Meglitsch, P., Schram, F. 1991. Invertebrate zoology. Oxford University Press, New York.

Meyerhans, A., Vartanian, J.-P., Wain-Hobson, S. 1990. DNA recombination during PCR. *Nucl Acids Res* **18**, 1687-1691.

Misof, B., Misof, K. 2009. A Monte Carlo Approach Successfully Identifies Randomness in Multiple Sequence Alignments : A More Objective Means of Data Exclusion. *Syst Biol* **58**, 21-34.

Moore, P. B., Steitz, T. A. 2002. The involvement of RNA in ribosome function. Nature **418**, 229-235.

Musso, M., Bocciardi, R., Parodi, S., Ravazzolo, R., Ceccherini, I. 2006. Betaine, Dimethyl Sulfoxide, and 7-Deaza-dGTP, a Powerful Mixture for Amplification of GC-Rich DNA Sequences. *J Mol Diagn* **8**, 544-550.

Mwinyi, A., Meyer, A., Bleidorn, C., Lieb, B., Bartolomaeus, T., Podsiadlowski, L. 2009. Mitochondrial genome sequence and gene order of *Sipunculus nudus* give additional support for an inclusion of Sipuncula into Annelida. BMC Genomics **10**, 27.

Nakao, A., Yoshihama, M., Kenmochi, N. 2004. RPG: the Ribosomal Protein Gene database. *Nucl Acids Res* **32**, D168-170.

Negri, A., Tedeschi, G., Bonomi, F., Zhang, J.-H., Kurtz Jr, D. M. 1994. Amino-acid sequences of the alpha- and beta-subunits of hemerythrin from *Lingula reevii*. *Biochim Biophys Acta* - Protein Structure and Molecular Enzymology 1208, 277-285.

Nichols, D. 1967. The origin of echinoderms. Symposia of the Zoological Society London **20**, 209 - 229.

Nielsen, C. 1971. Entoproct life-cycles and the entoproct/ectoproct relationship. Ophelia, 209-341.

Nielsen, C. 1985. Animal phylogeny in the light of the trochaea theory. *Biol J Linnean Soc* **25**, 243-299.

Nielsen, C. 2001. Animal Evolution. Interrelationships of the living phyla. Oxford University Press, Oxford.

Nitsche, H. 1869. Beitraege zur Kenntnis der Bryozoen. *Zeitsch für wiss Zool* **20**, 1-36.

Noguchi, Y., Endo, K., Tajima, F., Ueshima, R. 2000. The mitochondrial genome of the brachiopod *Laqueus rubellus*. Genetics **155**, 245 - 259.

Obst, M., Funch, P., Kristensen, R. M. 2006. A new species of Cycliophora from the mouthparts of the American lobster, *Homarus americanus* (Nephropidae, Decapoda). Organisms Diversity & Evolution **6**, 83-97.

Ojala, D., Montoya, J., Attardi, G. 1981. tRNA punctuation model of RNA processing in human mitochondria. Nature **290**, 470 - 474.

Okusu, A., Giribet, G. 2003. New 18S rRNA sequences from neomenioid aplacophorans and the possible origin of persistent exogenous contamination. *J Mollus Stud* **69**, 385-387.

Okusu, A., Schwabe, E., Eernisse, D. J., Giribet, G. 2003. Towards a phylogeny of chitons (Mollusca, Polyplacophora) based on combined analysis of five molecular loci. Organisms Diversity & Evolution **3**, 281-302.

Paps, J., Baguñà, J., Riutort, M. 2009. Lophotrochozoa internal phylogeny: new insights from an up-to-date analysis of nuclear ribosomal genes. *Proc R Soc Lond [Biol]* **276**,1245-1254

Passamaneck, Y., Halanych, K. 2006. Lophotrochozoan phylogeny assessed with LSU and SSU data: Evidence of lophophorate polyphyly. *Mol Phylogenet Evol* **40**, 20 - 28.

Passamaneck, Y. J., Halanych, K. M. 2004. Evidence from Hox genes that bryozoans are lophotrochozoans. Evolution & Development **6**, 275-281.

Passamaneck, Y. J., Schander, C., Halanych, K. M. 2004. Investigation of molluscan phylogeny using large-subunit and small-subunit nuclear rRNA sequences. *Mol Phylogenet Evol* **32**, 25-38.

Pelseneer, P. 1899. Recherches morphologiques et phylogénétiques sur les mollusques archaïques. Mém Curonnées et Mém Savants Étrangers Acad Bélgique, 1-113.

Penny, D., Hendy, M. D. 1985. The Use of Tree Comparison Metrics. *Syst Zool* **34**, 75-82.

Perina, D., Cetkovic, H., Harcet, M., Premzl, M., Lukic-Bilela, L., Muller, W. E., Gamulin, V. 2006. The complete set of ribosomal proteins from the marine sponge *Suberites domuncula*. Gene **366**, 275-284.

Perna, N., Kocher, T. 1995. Patterns of nucleotide composition at fourfold degenerate sites of animal mitochondrial genomes. *J Mol Evol* **41**, 353 - 358.

Peterson, K. J., Eernisse, D. J. 2001. Animal phylogeny and the ancestry of bilaterians: inferences from morphology and 18S rDNA gene sequences. Evol Dev **3**, 170-205.

Philip, G. K., Creevey, C. J., McInerney, J. O. 2005. The Opisthokonta and the Ecdysozoa May Not Be Clades: Stronger Support for the Grouping of Plant and Animal than for Animal and Fungi and Stronger Support for the Coelomata than Ecdysozoa. *Mol Biol Evol* **22**, 1175-1184.

Philippe, H., Delsuc, F. d. r., Brinkmann, H., Lartillot, N. 2005a. PHYLOGENOMICS. *Annu Rev Ecol Evol Syst* **36**, 541-562.

Philippe, H., Derelle, R., Lopez, P., Pick, K., Borchiellini, C., Boury-Esnault, N., Vacelet, J., Renard, E., Houliston, E., Quéinnec, E., Da Silva, C., Wincker, P., Le Guyader, H., Leys, S., Jackson, D. J., Schreiber, F., Erpenbeck, D., Morgenstern, B., Wörheide, G., Manuel, M. 2009. Phylogenomics Revives Traditional Views on Deep Animal Relationships. Current Biology **19**, 706-712.

Philippe, H., Lartillot, N., Brinkmann, H. 2005b. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. *Mol Biol Evol* **22**, 1246 - 1253.

Philippe, H., Snell, E., Bapteste, E., Lopez, P., Holland, P., Casane, D. 2004. Phylogenomics of eukaryotes: impact of missing data on large alignments. *Mol Biol Evol* **21**, 1740 - 1752.

Philippe, H., Telford, M. J. 2006. Large-scale sequencing and the new animal phylogeny. Trends in Ecology & Evolution **21**, 614-620.

Phillips, M., McLenachan, P., Down, C., Gibb, G., Penny, D. 2006. Combined Mitochondrial and Nuclear DNA Sequences Resolve the Interrelations of the Major Australasian Marsupial Radiations. *Syst Biol* **55**, 122-137.

Plate, L. H. 1897-1901. Anatomie und Phylogenie der Chitonen 1-3. Gustav Fischer, Jena.

Pojeta, J., Jr., Runnegar, B. 1976. The paleontology of rostroconch mollusks and early history of the phylum Mollusca. *U S Geol Surv Prof Paper*, 1-88.

Ponder, W., Lindberg, D. (Eds.) 2008. Molluscan Evolution and Phylogeny. University of California Press, Berkley and Los Angeles.

Ponder, W. F., Lindberg, D. R. 1997. Towards a phylogeny of gastropod molluscs: an analysis using morphological characters. *Zool J Linnean Soc* **119**, 83-265.

Ray, A., Norden, B. 2000. Peptide nucleic acid (PNA): its medical and biotechnical applications and promise for the future. FASEB J. **14**, 1041-1060.

Rice, M. E. 1985. Sipuncula: Developmental Evidence for Phylogenetic Inference. Oxford University Press, New York.

Robinson, D. F., Foulds, L. R. 1981. Comparison of phylogenetic trees. *Math Biosci* **53**, 131-147.

Roeding, F., Hagner-Holler, S., Ruhberg, H., Ebersberger, I., von Haeseler, A., Kube, M., Reinhardt, R., Burmester, T. 2007. EST sequencing of Onychophora and phylogenomic analysis of Metazoa. *Mol Phylogenet Evol* **45**, 942-951.

Rokas, A., Williams, B. L., King, N., Carroll, S. B. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature **425**, 798-804.

Runnegar, B., Pojeta, J., Jr. 1974. Molluscan Phylogeny: The Paleontological Viewpoint. Science **186**, 311-317.

Salvini-Plawen, L., Steiner, G. 1996a. Synapomorphies and plesiomorphies in higher classification of Mollusca. Oxford Univ. Press.

Salvini-Plawen, L. v. 1980. A reconsideration of systematics in the Mollusca (Phylogeny and higher classification). Malacologia **19**, 249-278.

Salvini-Plawen, L. v. 1990. Origin, phylogeny and classification of the phylum Mollusca. Iberus.

Salvini-Plawen, L. v. 2003. On the phylogenetic significance of the aplacophoran Mollusca. Iberus, 67-97.

Salvini-Plawen, L. v., Steiner, G. 1996b. Synapomorphies and plesiomorphies in higher classification of Mollusca. Oxford University Press, Oxford.

Salvini-Plawen, v. L. 2008. Three new species of Simrothiellidae (Solenogastres) associated with the hot-vent biotope. *J Mollusc Stud* **74**, 223-238.

Sambrook, J., Russell, D. W. 2001. Molecular Cloning: A laboratory manual. Cold Spring Harbor Laboratory Press, New York.

Sanderson, M. J., Driskell, A. C. 2003. The challenge of constructing large phylogenetic trees. *Trends Plant Sci* **8**, 374-379.

Sanderson, M. J., Wojciechowski, M. F. 2000. Improved bootstrap confidence limits in large-scale phylogenies, with an example from Neo-Astragalus (Leguminosae). *Syst Biol* **49**, 671-685.

Sanna, M. T., Manconi, B., Castagnola, M., Giardina, B., Masia, D., Messana, I., Olianas, A., Patamia, M., Petruzzelli, R., Pellegrini, M. 2005. Functional and structural characterization of the myoglobin from the polychaete *Ophelia bicornis*. *Biochem J* **389**, 497-505.

Sasaki, T. 1998. Comparative anatomy and phylogeny of the recent Archaeogastropoda (Mollusca: Gastropoda). University Museum, Tokjo.

Saunders, W. B. 1984. Nautilus Growth and Longevity: Evidence from Marked and Recaptured Animals. Science **224**, 990-992.

Schaefer, K., Haszprunar, G. 1996. Anatomy of *Laevipilina antarctica*, a monoplacophoran limpet (Mollusca) from Antarctic waters. *Acta Zool* 77, 295-314.

Scheltema, A. H. 1993. Aplacophora as progenetic aculiferans and the coelomate origin of mollusks as the sister taxon of Sipuncula. *Biol Bull* **184**, 57-78.

Scheltema, A. H. (Ed. 1996. Phylogenetic position of the Sipuncula, Mollusca and the progenetic Aplacophora. Oxford University Press, Oxford.

Schmidt-Rhaesa, A. 2007. The evolution of organ systems. Oxford University Press, Oxford.

Schöne, B. R., Fiebig, J., Pfeiffer, M., Gle, R., Hickson, J., Johnson, A. L. A., Dreyer, W., Oschmann, W. 2005. Climate records from a bivalved Methuselah (*Arctica islandica*, Mollusca; Iceland). Palaeogeography, Palaeoclimatology, Palaeoecology **228**, 130-148.

Schram, F. R. 1991. Cladistic analysis of metazoan phyla and the placement of fossil problematica. In: Conway-Morris, S. (Ed.) The early evolution of Metazoa and the significance of problematic taxa. Cambridge University Press., Cambridge, pp. 35-46.

Schramm, G., Bruchhaus, I., Roeder, T. 2000. A simple and reliable 5'-RACE approach. *Nucl Acids Res* **28**, e96-.

Schulze, A., Cutler, E., Giribet, G. 2005a. Reconstructing the phylogeny of the Sipuncula. Hydrobiologia **535**, 277 - 296.

Schulze, A., Cutler, E. B., Giribet, G. 2005b. Molecular and morphological evolution in sipunculan worms. Integrative and Comparative Biology **45**, 1070.

Schulze, A., Cutler, E. B., Giribet, G. 2007. Phylogeny of sipunculan worms: A combined analysis of four gene regions and morphology. *Mol Phylogenet Evol* **42**, 171-192.

Shao, R., Aoki, Y., Mitani, H., Tabuchi, N., Barker, S., Fukunaga, M. 2004. The mitochondrial genomes of soft ticks have an arrangement of genes that has remained unchanged for over 400 million years. *Insect Mol Biol* **13**, 219 - 224.

Shen, X., Ma, X., Ren, J., Zhao, F. 2009. A close phylogenetic relationship between Sipuncula and Annelida evidenced from the complete mitochondrial genome sequence of *Phascolosoma esculenta*. BMC Genomics **10**, 136.

Shimodaira, H., Hasegawa, M. 1999. Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. *Mol Biol Evol* **16**, 1114-1116.

Shimodaira, H., Hasegawa, M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics **17**, 1246 - 1247.

Sigwart, J. D., Sutton, M. D. 2007. Deep molluscan phylogeny: synthesis of palaeontological and neontological data. *Proc R Soc Lond [Biol]* **274**, 2413-2419.

Simison, W., Lindberg, D., Boore, J. 2006. Rolling circle amplification of metazoan mitochondrial genomes. *Mol Phylogenet Evol* **39**, 562 - 567.

Smith, S. A., Donoghue, M. J. 2008. Rates of Molecular Evolution Are Linked to Life History in Flowering Plants. Science **322**, 86-89.

Smith, S. A., Dunn, C. W. 2008. Phyutility: a phyloinformatics tool for trees, alignments and molecular data. Bioinformatics **24**, 715-716.

Soltis, D. E., Albert, V. A., Savolainen, V., Hilu, K., Qiu, Y.-L., Chase, M. W., Farris, J. S., Stefanovic, S., Rice, D. W., Palmer, J. D., Soltis, P. S. 2004. Genome-scale data, angiosperm relationships, and `ending incongruence': a cautionary tale in phylogenetics. *Trends Plant Sci* **9**, 477-483.

Sorensen, M., Funch, P., Willerslev, E., Hansen, A., Olesen, J. 2000. On the phylogeny of the Metazoa in the light of Cycliophora and Micrognathozoa. Zoologischer Anzeiger **239**, 297 - 318.

Sørensen, M. V., Giribet, G. 2006. A modern approach to rotiferan phylogeny: combining morphological and molecular data. Mol Phylogenet Evol 40, 585-608.

Stamatakis, A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics **22**, 2688-2690.

Stamatakis, A., Hoover, P., Rougemont, J. 2008. A Rapid Bootstrap Algorithm for the RAxML Web Servers. *Syst Biol* **57**, 758 - 771.

Stamatakis, A., Ott, M. 2008. Efficient computation of the phylogenetic likelihood function on multi-gene alignments and multi-core architectures. *Proc R Soc Lond [Biol]* **363**, 3977-3984.

Staton, J. 2003. Phylogenetic analysis of the mitochondrial cytochrome c oxidase subunit 1 gene from 13 sipunculan genera: intra- and interphylum relationships. Invertebrate Biology **122**, 252 - 264.

Stechmann, A., Schlegel, M. 1999. Analysis of the complete mitochondrial DNA sequence of the brachiopod *Terebratulina retusa* places Brachiopoda within the protostomes. *Proc R Soc Lond [Biol]* **266**, 2043-2052.

Steiner, G., Dreyer, H. 2003. Molecular phylogeny of Scaphopoda (Mollusca) inferred from 18S rDNA sequences: support for a Scaphopoda-Cephalopoda clade. *Zool Scripta* **32**, 343-356.

Steiner, G., Müller, M. 1996. What can 18S rDNA do for bivalve phylogeny? *J Mol Evol* **43**, 58-70.

Steinke, D., Salzburger, W., Meyer, A. 2006. Novel Relationships Among Ten Fish Model Species Revealed Based on a Phylogenomic Analysis Using ESTs. *J Mol Evol* **62**, 772-784.

Strimmer, K., Rambaut, A. 2002. Inferring confidence sets of possibly misspecified gene trees. *Proc Biol Sci* **269**, 137 - 142.

Struck, T. H., Fisse, F. 2008. Phylogenetic Position of Nemertea Derived from Phylogenomic Data. *Mol Biol Evol* **25**, 728-736.

Struck, T. H., Schult, N., Kusen, T., Hickman, E., Bleidorn, C., McHugh, D., Halanych, K. M. 2007. Annelid phylogeny and the status of Sipuncula and Echiura. BMC Evol Biol **7**, 57

Strugnell, J., Jackson, J., Drummond, A. J., Cooper, A. 2006. Divergence time estimates for major cephalopod groups: evidence from multiple genes. Cladistics **22**, 89-96.

Sugiura, N. 1978. Further analysts of the data by Akaike' s Information Criterion and the finite corrections. Communications in Statistics - Theory and Methods **7**, 13 - 26.

Swofford, D. 2002. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Sinauer Associates, Sunderland Massachusetts.

Takagi, T., Cox, J. A. 1991. Primary structure of myohemerythrin from the annelid *Nereis diversicolor*. FEBS Letters **285**, 25-27.

Tamura, K., Dudley, J., Nei, M., Kumar, S. 2007. MEGA 4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol Biol Evol* 24, 1596-1599.

Thiele, J. 1902. Die systematische Stellung der Solenogastren und die Phylogenie der Mollusken. Zeitschrift für Wissenschaftliche Zoologie, 249-466.

Thomas, J. A., Welch, J. J., Woolfit, M., Bromham, L. 2006. There is no universal molecular clock for invertebrates, but rate variation does not scale with body size. **103**, 7366-7371.

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

Timmermans, M., Roelofs, D., Marien, J., van Straalen, N. 2008. Revealing pancrustacean relationships: Phylogenetic analysis of ribosomal protein genes places Collembola (springtails) in a monophyletic Hexapoda and reinforces the discrepancy between mitochondrial and nuclear DNA markers. *BMC Evol Biol* **8**, 83.

Todt, C., Okusu, A., Schander, C., Schwabe, E. 2008. Phylogeny and Evolution of the Mollusca. University of California Press, Berkley, Los Angeles, London.

Todt, C., Salvini-Plawen, L. v. 2004. Ultrastructure of the midgut epithelium of *Wirenia argentea* (Mollusca: Solenogastres). *J Mollus Stud* **70**, 213-224.

Todt, C., Salvini-Plawen, L. v. 2005. The digestive tract of *Helicoradomenia* (Solenogastres, Mollusca), aplacophoran molluscs from the hydrothermal vents of the East Pacific Rise. Invertebrate Biology **124**, 230-253.

Tostesen, E. 2008. A stitch in time: Efficient computation of genomic DNA melting bubbles. Algorithms for Molecular Biology **3**.

Turbeville, J., Schulz, J., Raff, R. 1994. Deuterostome phylogeny and the sister group of the chordates: evidence from molecules and morphology. *Mol Biol Evol* **11**, 648-655.

Turbeville, J., Smith, D. 2007. The partial mitochondrial genome of the Cephalothrix rufifrons (Nemertea, Palaeonemertea): Characterization and implications for the phylogenetic position of Nemertea. *Mol Phylogenet Evol* **43**, 1056 - 1065.

Uechi, T., Tanaka, T., Kenmochi, N. 2001. A complete map of the human ribosomal protein genes: assignment of 80 genes to the cytogenetic map and implications for human disorders. Genomics **72**, 223 - 230.

Valles, Y., Boore, J. 2006. Lophotrochozoan mitochondrial genomes. Integrative and Comparative Biology **46**, 544 - 557.

Vanin, S., Negrisolo, E., Bailly, X., Bubacco, L., Beltramini, M., Salvato, B. 2006. Molecular Evolution and Phylogeny of Sipunculan Hemerythrins. *J Mol Evol* 62, 32-41.

Vergote, D., Sautiere, P.-E., Vandenbulcke, F., Vieau, D., Mitta, G., Macagno, E. R., Salzet, M. 2004. Up-regulation of Neurohemerythrin Expression in the Central Nervous System of the Medicinal Leech, *Hirudo medicinalis*, following Septic Injury. *J Biol Chem* **279**, 43828-43837.

Veuthey, A. L., Bittar, G. 1998. Phylogenetic relationships of fungi, plantae, and animalia inferred from homologous comparison of ribosomal proteins. *J Mol Evol* **47**, 81-92.

Waegele, J. W., Mayer, C. 2007. Visualizing differences in phylogenetic information content of alignments and distinction of three classes of long-branch effects. *BMC Evol Biol* **7**, 147.

Waeschenbach, A., Telford, M. J., Porter, J. S., Littlewood, D. T. J. 2006. The complete mitochondrial genome of *Flustrellidra hispida* and the phylogenetic position of Bryozoa among the Metazoa. *Mol Phylogenet Evol* **40**, 195-207.

Waller, T. R. 1998. Origin of the molluscan class Bivalvia and the phylogeny of major groups. University of Calgary Press, Calgary.

Wang, G., Wang, Y. 1997. Frequency of formation of chimeric molecules as a consequence of PCR coamplification of 16S rRNA genes from mixed bacterial genomes. *Appl Environ Microbiol* **63**, 4645-4650.

Wanninger, A. 2005. Immunocytochemistry of the nervous system and the musculature of the chordoid larva of *Symbion pandora* (Cycliophora). J Morphol **265**, 237-243.

Wanninger, A. 2009. Shaping the Things to Come: Ontogeny of Lophotrochozoan Neuromuscular Systems and the Tetraneuralia Concept. *Biol Bull* **216**, 293-306.

Wanninger, A., Fuchs, J., Haszprunar, G. 2007. Anatomy of the serotonergic nervous system of an entoproct creeping-type larva and its phylogenetic implications. Invertebrate Biology **126**, 268-278.

Wanninger, A., Koop, D., Bromham, L., Noonan, E., Degnan, B. 2005. Nervous and muscle system development in *Phascolion strombus* (Sipuncula). *Dev Genes Evol* **215**, 509-518.

Wehe, A., Bansal, M. S., Burleigh, J. G., Eulenstein, O. 2008. DupTree: a program for large-scale phylogenetic analyses using gene tree parsimony. Bioinformatics **24**, 1540-1541.

Welch, J. J., Bromham, L. 2005. Molecular dating when rates vary. Trends in Ecology & Evolution 20, 320-327.

Wells, R. M. G., Dales, R. P. 1974. Oxygenational properties of haemerythrin in blood of *Magelonia papillicornis* Mueller (Polychaeta-Magelonidae). *Comp Biochem Physiol* **49**, 57-64.

Wilson, A. C., Sarich, V. M. 1969. A molecular time scale for human evolution. Proceedings of the National Academy of Sciences of the United States of America 63, 1088-1093.

Wilson, N. G., Huang, D., Goldstein, M. C., Cha, H., Giribet, G., Rouse, G. W. 2009. Field collection of *Laevipilina hyalina* McLean, 1979 from southern California, the most accessible living monoplacophoran. *J Mollus Stud* **75**, 195-197.

Winnepenninckx, B., Backeljau, T., De Wachter, R. 1996. Investigation of molluscan phylogeny on the basis of 18S rRNA sequences. *Mol Biol Evol* **13**, 1306-1317.

Winnepenninckx, B. M. H., Backeljau, T., Kristensen, R. M. 1998. Relations of the new phylum Cycliophora. Nature **393**, 636-638.

Wintzingerode, F. v., Goebel, U. B., Stackebrandt, E. 1997. Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. FEMS Microbiology Reviews **21**, 213-229.

Witek, A., Herlyn, H., Meyer, A., Boell, L., Bucher, G., Hankeln, T. 2008. EST based phylogenomics of Syndermata questions monophyly of Eurotatoria. *BMC Evol Biol* **8**, 345.

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

Wool, I., Chan, Y., Gluck, A. 1995. Structure and evolution of mammalian ribosomal proteins. *Biochem Cell Biol* **73**, 933 - 947.

Yeang, C.-H., Haussler, D. 2007. Detecting Coevolution in and among Protein Domains. *PLoS Comput Biol* **3**, e211.

Yokobori, S.-i., Lindsay, D. J., Yoshida, M., Tsuchiya, K., Yamagishi, A., Maruyama, T., Oshima, T. 2007. Mitochondrial genome structure and evolution in the living fossil vampire squid, *Vampyroteuthis infernalis*, and extant cephalopods. *Mol Phylogenet Evol* **44**, 898-910.

Yoshihama, M., Nguyen, H. D., Kenmochi, N. 2007. Intron dynamics in ribosomal protein genes. PLoS ONE **2**, e141.

Zhang, D., Hewitt, G. 1997. Insect mitochondrial control region: A review of its structure, evolution and usefulness in evolutionary studies. *Biochem Sys Ecol* **25**, 99 - 120.

Zhang, J. H., Kurtz, D. M. 1992. Metal substitutions at the diiron sites of hemerythrin and myohemerythrin: contributions of divalent metals to stability of a four-helix bundle protein. *Proc Natl Acad Sci U S A* **89**, 7065-7069.

Zhang, Z., Harrison, P., Gerstein, M. 2002. Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. *Genome Res* **12**, 1466-1482.

Zhong, M., Struck, T., Halanych, K. 2008. Phylogenetic information from three mitochondrial genomes of Terebelliformia (Annelida) worms and duplication of the methionine tRNA. Gene **416**, 11 - 21.

Zrzavy, J., Mihulka, S., Kepka, P., Bezdek, A., Tietz, D. 1998. Phylogeny of the Metazoa based on morphological and 18S ribosomal DNA evidence. Cladistics **14**, 249-285.

Zuckerkandl, E., Pauling, L. 1965. Molecules as documents of evolutionary history. *J Theor Biol* **8**, 357-366.

Zuker, M., Mathews, D. H., Turner, D. H. (Eds.) 1999. Turner Algorithms and Thermodynamics for RNA Secondary Structure Prediction: A Practical Guide in RNA Biochemistry and Biotechnology. Kluwer Academic Publishers.

12. Curriculum Vitae

Name	Achim Meyer
Addresse	[redacted]
	Frankfurt am Main
Email	[redacted]
Geboren	22.10.1965 in Essen
Nationalität	deutsch
[redacted]	[redacted]

1984	Abitur am Otto Hahn Gymnasium in Göttingen.
1984 - 1986	Zivildienst an der Sozialstation Bovenden.
1987 - 1988	Studium Chemie/Biologie an der Georg August Universität Göttingen.
1988 - 1989	Geographiestudium an der FU Berlin.
1989 - 1992	Erzieherausbildung am Oberlin Seminar in Berlin.
1992 - 1993	Anerkennungsjahr am Kinderbauernhof „Görlitzer Park e.V.“ (Berlin).
1994 - 2003	Erzieher im Schülerladen der Elterninitiative „Krümelkinder e.V.“
1999 - 2005	Studium der Biologie. Populationsgenetische Diplomarbeit in der AG Systematik und Evolution der Tiere, FU-Berlin.
seit 12/ 2005	Wissenschaftlicher Mitarbeiter und Promotionsstudent an der Johannes Gutenberg-Universität Mainz im Rahmen des Schwerpunktprogamms „Deep Metazoan Phylogeny“ der DFG betreut von [redacted].

Publikationen

Gross W & **Meyer A** (2003) Distribution of myo-inositol dehydrogenase in algae. *Eur J Phycol* 38:191-194

Meyer A, Bleidorn C, Rouse GW, Hausen H (2007) Morphological and molecular data suggest a cosmopolitan distribution of the polychaete *Proscoloplos cygnochaetus* Day, 1954 (Annelida, Orbiniidae). Marine Biology 153:879-889

Hausdorf B, Helmkampf M, **Meyer A**, Witek A, Herlyn H, Bruchhaus I, Hankeln T, Struck TH, Lieb B (2007) Spiralian Phylogenomics Supports the Resurrection of Bryozoa Comprising Ectoprocta and Entoprocta. Mol Biol Evol 24:2723-2729

Witek A, Herlyn H, **Meyer A**, Boell L, Bucher G, Hankeln T (2008) EST based phylogenomics of Syndermata questions monophyly of Eurotatoria. BMC Evolutionary Biology 8:345

Mwinyi A, **Meyer A**, Bleidorn C, Lieb B, Bartolomaeus T, Podsiadlowski L (2009) Mitochondrial genome sequence and gene order of *Sipunculus nudus* give additional support for an inclusion of Sipuncula into Annelida. BMC Genomics 10:27

Meyer A & Lieb B (submitted) Respiratory proteins in *Sipunculus nudus* Linnaeus 1766 – implications for phylogeny and evolution of the hemerythrin family***.***

Meyer A, Todt C, Mikkelsen NT, Lieb B (submitted) Fast evolving 18S rRNA sequences from Solenogastres (Mollusca) resist standard PCR amplification and give new insights into mollusk substitution rate heterogeneity.

Meyer A, Todt C, Lachnit H, Lieb B (submitted) Selecting ribosomal protein genes for invertebrate phylogenetic inferences - How many genes to resolve the Mollusca?

Nicht-wissenschaftliche Artikel:

Meyer A (2003). Einfach schön: *Palaemon elegans*. Arbeitsmaterial der ZAG: 11-12

Meyer A (2005). Stumm wie ein Fisch? Netz-Preußenfische sind es nicht. Der Meerwasseraquarianer, 2: 19-21.

Poster und Vorträge

Meyer, A., Hausen, H., Bleidorn, C., Rouse, G. (2005). The *Proscoloplos* species complex (Annelida: Orbiniidae) is a single disjunctively distributed species: support from molecular and morphological data. GfBS - Annual meeting. Basel, Switzerland. (Talk)

Meyer A., Lieb B. (2006). The phylogeny of mollusks and their relatives within the Lophotrochozoans. Molluscan Forum 2006. Natural History Museum. London, UK. (Poster)

Meyer A., Streit K., Kelly R.P., Eernisse D.J. &Lieb B. (2007). Hemocyanin as a promising molecular marker for phylogenetic analyses in chitons (Polyplacophora). 9. Jahrestagung der Gesellschaft für Biologische Systematik. Wien, Austria. (Poster)

Meyer, A Struck, T. H., Lieb B., (2007). Ribosomal protein sequence data identify *Sipunculus nudus* as an annelid. - Evolution of the animals – a Linnean tercentenary celebration / Royal Society, London, UK. (Poster)

Meyer, A., Struck, T. H., Lieb, B. (2007). Sipuncula are annelids: EST analyses strongly support an ingroup relationship. - 100th Annual Meeting of the German Zoological Society, Köln, Germany. (Talk)

Helmkampf M., **Meyer A.**, Lieb B, Bruchhaus I., Hausdorf B. (2007). Phylogenomic analyses attest the resurrection of Bryozoa sensu lato. 100. Jahrestagung der Deutschen Zoologischen Gesellschaft. Köln, Germany. (Talk)

Meyer, A., Lieb, B. (2008). Sipunculan hemerythrin protein family with annelid affinity. GfBS-Annual Meeting and Systematics 2008 Göttingen, Germany. (Poster)

Mwinyi, A., **Meyer, A.**, Bleidorn, C., Lieb, B., Bartolomaeus, T., Podsiadlowski, L. (2008). Mitochondrial genome sequence and gene order of *Sipunculus nudus* give additional support for an inclusion of Sipuncula into annelida. 10th Young Systematists Forum, Natural History Museum London, UK. (Poster)

Meyer A., Todt, C., Lachnit, H. and Lieb, H. (2009). Identification of new target genes and the application of ribosomal gene data in molecular phylogenetics of molluscs. Celebrating Darwin: From The Origin of Species to Deep Metazoan Phylogeny. Berlin, Germany. (Talk)

Meyer A., Todt, C., Mikkelsen, NT., Lieb, B (2009). 18S rRNA sequences from Solenogastres (Mollusca) exhibit strong secondary structures and elevated substitution rates. Systematics, Natural History Museum Leiden, NL. (Talk)

Warnke KM., **Meyer A**., Ebener B., Lieb B.(2009). Phylogenetic inferences on the taxonomic position of *Spirula spirula* (Cephalopoda) based on hemocyanin sequence data. CIAC Conference. Vigo, Spain. (Talk)

13. Danksagung

Liebe Claudia, danke für alles!

Mein besonderer Dank gilt [redacted] für die freundschaftliche Betreuung und die Überlassung des Themas. Ich danke ihm auch für die freie Entfaltungsmöglichkeit und beständige Unterstützung die er mir in meiner Arbeit gab. [redacted] danke ich dass er freundlicherweise die Mühen eines Zweitgutachtens auf sich genommen hat. [redacted] danke ich für die wohlwollende Duldung eines Systematikers in seiner physiologisch ausgerichteten AG und die hervorragende Laborausstattung. Der gesamten Arbeitsgruppe „Molekulare Tierphysiologie" der Uni-Mainz danke ich für die gute Arbeitsatmosphäre, sowie die beständige Hilfsbereitschaft und die vielen schönen Momente während der gesamten Promotionszeit. Dank gilt auch den Wissenschaftlern die mir lebende Tiere für molekulare Analysen zur Verfügung gestellt haben: [redacted]. [redacted] danke ich zusätzlich für die herzliche Aufnahme in Norwegen, die Co-Autorenschaft und die damit verbundenen "English lessons". [redacted] und [redacted] danke ich für hilfreiche Diskussionen rund um die Stammbaumanalysen. [redacted] und [redacted] danke ich für ihre Korrektur-Vorschläge. Und schließlich gilt ganz besonderer Dank meinen Eltern [redacted] und [redacted] die stets mein biologisches Interesse gefördert haben.