Molecular Phylogenetics and Evolution 52 (2009) 268–272

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Short Communication

MtZoa: A general mitochondrial amino acid substitutions model for animal evolutionary studies

Omar Rota-Stabelli, Ziheng Yang, Maximilian J. Telford *

Department of Genetics, Evolution and Environment, University College London, Darwin Building, Gower Street, London WC1E 6BT, UK

ARTICLE INFO

Article history:
Received 10 October 2008
Revised 22 December 2008
Accepted 23 January 2009
Available online 30 January 2009

Keywords:
Mitochondrial genome
Mitogenomics
Phylogeny
Metazoa
Model selection
Empirical
Mechanistic
Maximum likelihood
Bayesian inference

1. Introduction

Mitochondrial genome coded proteins are widely used as markers for the inference of phylogeny (mitogenomics). Their main advantages are unambiguous orthology, the richness of available sampling among eukaryotes and the relative ease of generating new data. On the other hand, mitochondrial sequences have been reported to suffer from an accelerated substitution rate and among-lineages compositional heterogeneity (Foster et al., 1997; Rota-Stabelli and Telford, 2008). These characteristics, if shared by phylogenetically unrelated species, may be responsible for convergent evolution (homoplasy) and promote the dilution of the true phylogenetic signal. Furthermore, the mitochondrial genetic code varies to different degrees between different metazoan lineages. In the light of this, mitogenomic studies are in need of realistic models of evolution that best represent the evolutionary process and reduce systematic bias.

The majority of deep level mitogenomic analyses are carried out at the amino acid level as nucleotide sequences are more susceptible to substitutional saturation and codon-based phylogenies may be complicated by differences in the genetic code. Amino acid datasets can be analyzed using the mechanistic GTR (general time reversible) model (Yang et al., 1998), allowing all the parameters of the model to be estimated from the dataset during the inference of phylogeny. A clear problem in this procedure is the large size of the amino acid alphabet, which makes the estimation of all the parameters a demanding computational task. Additionally, reliable estimation of the amino acid replacement rates needs a significant amount of substitutional information from the dataset and the small datasets typically used in phylogenetic analyses may not contain sufficient information. Consequently, amino acid alignments are generally analyzed using empirical amino acid replacement matrices, which have been pre-estimated from a large dataset and are represented in fixed matrices.

A current problem with existing empirical models is that they are based on the comparison of restricted datasets; MtREV (Adachi and Hasegawa, 1996) or MtMamm (Yang et al., 1998) are dominated by mammalian sequences and the recently released MtArt (Abascal et al., 2007) and MtPan (Carapelli et al., 2007) are both based on the analysis of arthropod-only datasets (Fig. 1). These matrices consequently reflect the substitution processes of either mammals or arthropods only and may be not appropriate for the analysis of other metazoan lineages, in particular lophotrochozoans and non-mammalian deuterostomes, for which many mitogenomic datasets are available, but few analyses have been conducted (Waeschenbach et al., 2006).

* Corresponding author. Fax: +44 2076797096.
E-mail address: m.telford@ucl.ac.uk (M.J. Telford).

1055-7903/$ - see front matter
doi:10.1016/j.ympev.2009.01.011

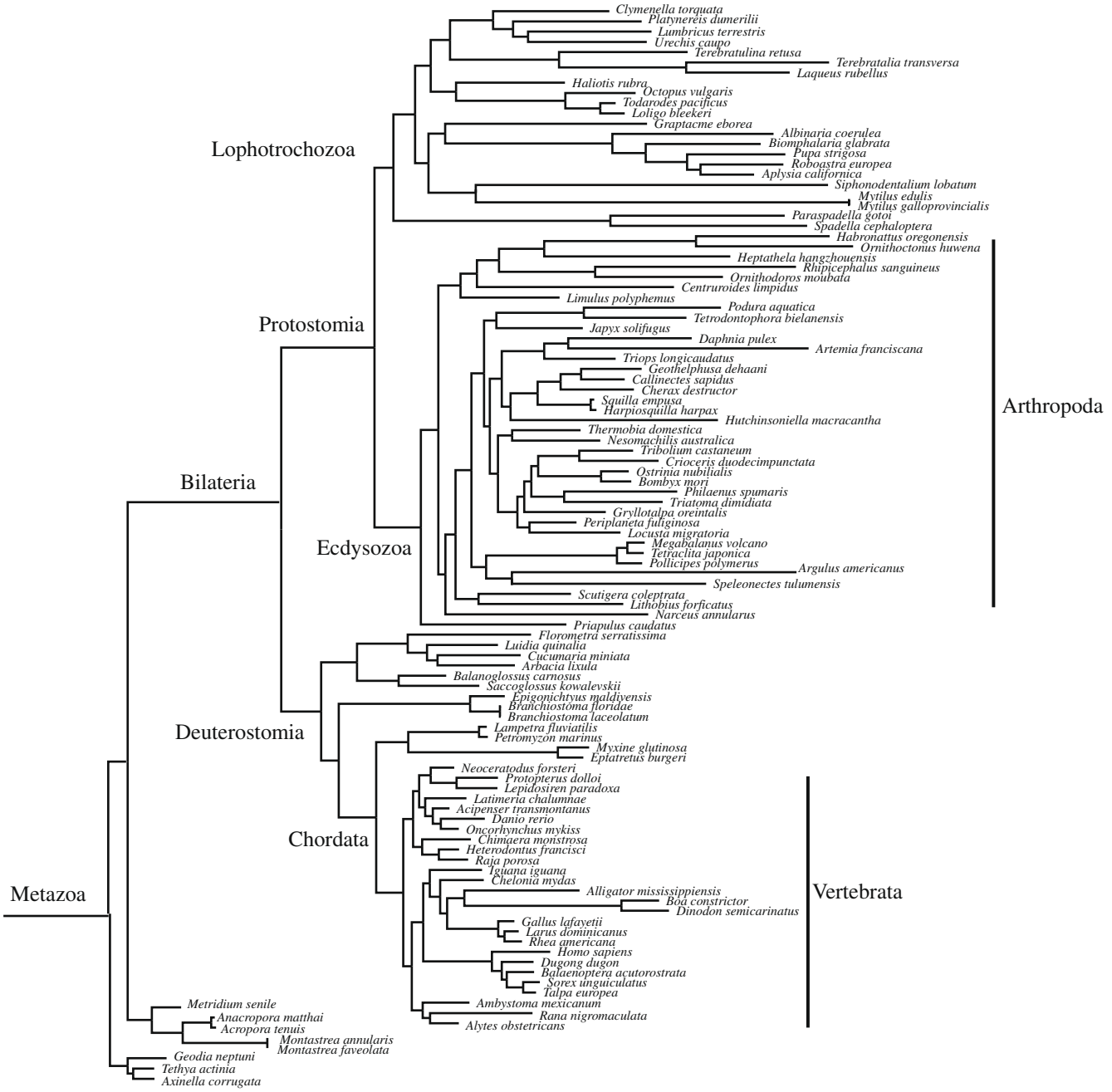

Fig. 1. Phylogenetic tree of the 108 metazoan species used to infer the MtZoa model. Note that commonly used empirical models such as MtREV (which is derived from vertebrates, indicated by vertical bar) and MtArt (derived from arthropods, indicated by vertical bar) are based on the comparison of restricted datasets. MtZoa is based on a larger and broader dataset, including lophotrochozoans, non-chordate deuterostomes and diploblastic metazoans. The topology was inferred using MrBayes under the MtREV model and some nodes have been constrained to reflect current knowledge of metazoan relationships; branch lengths have been estimated by PAML, during the inference of the MtZoa model.

1.1. Synopsis

In order to generate an empirical model that is more representative of the whole animal kingdom we estimated an empirical transition probability matrix, called MtZoa (Fig. 2A), based on the general reversible model and an alignment of 13 concatenated mitochondrial proteins from more than 100 phylogenetically diverse metazoan species. We tested how MtZoa and other models fit different metazoan datasets and show that our model is particularly indicated for the analysis of diverse metazoan, lophotrochozoan and deuterostome datasets.

2. Materials and methods

We assembled an alignment of the 13 mitochondrial proteins from 108 metazoan species, consisting of 39 deuterostomes, 22 lophotrochozoans 39 ecdysozoans and eight non-bilaterians. We constructed the corresponding tree (in Fig. 1) using MrBayes and the MtREV model and constraining some major nodes in order to reflect current knowledge of metazoan relationships and the so called "new animal phylogeny" (Webster et al., 2006; Telford et al., 2008; Dunn et al., 2008). We excluded lineages characterized by extremely accelerated substitution rate, such as urochordates

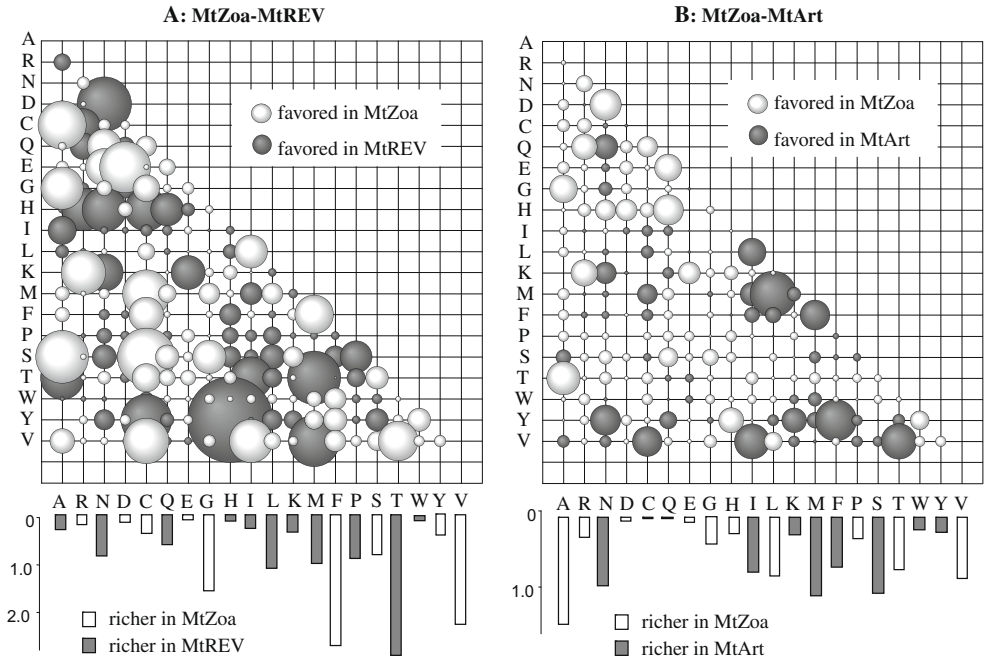

Fig. 2. Differences in replacement rates (bubbles in matrices) and stationary frequencies (bars) between (A) MtZoa and MtREV and (B) MtZoa and MtArt. Areas of bubbles are proportional to the absolute differences between replacement rates. The size of the bubbles in the legend correspond to a difference of 50. Length of bars corresponds to the absolute differences between stationary frequencies expressed as a percentage. White indicates a higher replacement rate or higher amino acid frequency in MtZoa and grey shows the reverse. Note that in (B), amino acids whose codons are rich in A and T (NIKMFY) are enriched and more replaceable in MtArt than in MtZoa.

and nematodes in order to minimize the degree of saturation of substitutions in the alignment and avoid the generation of a highly saturated substitution matrix. We excluded poorly aligned sites by manual refinement resulting in an alignment of 2589 amino acid positions. We used the maximum likelihood approach implemented in PAML (Yang, 2007) to estimate a general reversible amino acid replacement model, assuming reversibility, so that the rate matrix $Q = \{q_{ij}\}$ satisfies the condition $\pi_i q_{ij} = \pi_j q_{ji}$ for all the amino acid pairs, where π_j is the stationary frequency of amino acid j and r_{ij} is the replacement rate between amino acids i and j.

In order to highlight the differences in replacement rates and amino acid frequencies between our matrix and previous matrices, we generated a subtraction matrix, whose values correspond to the differences in replacement rate (r_{ij}) between MtZoa and MtREV (Fig. 2A upper) and between MtZoa and MtArt (Fig. 2B upper). We also calculated differences in the stationary frequencies (π_j) between the two pairs (lower parts of Fig. 2).

We recompiled MrBayes3.1 (Huelsenbeck and Ronquist, 2001) substituting our new replacement matrices (MtZoa) and also MtArt (Abascal et al., 2007) for existing ones and ran tree searches on six different metazoan datasets of concatenated mitochondrial proteins under MtZoa, MtArt, MtREV and the mechanistic GTR model. We have used two published datasets of 23 arthropods (Rota-Stabelli and Telford, 2008) and 41 mammals (Horner et al., 2007) and constructed four additional mitochondrial protein datasets: one containing 44 species from diverse metazoan groups, one with 24 lophotrochozoans, one with 30 ecdysozoans and one with 30 deuterostomes. We modeled among site rate heterogeneity with an invariable plus gamma distribution with four rate categories and ran two separate Bayesian tree searches until long after the likelihood of the sampled trees had plateaued. While the likelihood associated with empirical models converged between the two runs after few hundred generations (we have run them for a minimum of 300,000), the mechanistic GTR model required up to two million generations, depending on the dataset.

We evaluated model fit to the data using the Akaike Information Criterion (AIC) (Akaike, 1974) and Bayesian Information Criterion (BIC) (Schwartz, 1978) defined as follow: AIC = −2 log-likelihood +2 K; BIC = −2 log-likelihood +2 K log N, where K is the number of free parameters in the model and N is the number of sites in the alignment (Huelsenbeck et al., 2004). We estimated the harmonic mean with Tracer (http://beast.bio.ed.ac.uk) using the log-likelihood of the trees sampled after burn-in. In a few cases the mean log-likelihood of the two Bayesian runs were slightly different and we kept the highest in order to be more conservative for the test of model fit. The numbers of free parameters used in the AIC and BIC were determined as the number of branch lengths to be estimated plus the number of free parameters in the model (two for the empirical models and 209 for the GTR model).

3. Results and discussion

3.1. The MtZoa model

The MtZoa model is characterized by replacement rates that differ considerably from those of MtREV (Fig. 2A) and of MtArt (Fig. 2B). Replacements involving cysteine, valine and serine are more common in MtZoa than in MtREV (white bars in Fig. 2A), while those involving histidine, asparagine and tyrosine are less frequent (grey bubbles). Stationary frequencies also differ: phenylalanine and valine are more frequent in MtZoa (white bars in Fig. 1A), while threonine is distinctly less frequent than in MtREV (grey bars).

Compared to MtArt, MtZoa is impoverished in serine (grey bar in Fig. 1B), reflecting the differences between the invertebrate and the vertebrate mitochondrial genetic code (MtArt is based only on species with an invertebrate genetic code). Compared to MtArt, MtZoa is also clearly enriched in alanine (whose corresponding codon GCN is GC rich) and impoverished in methionine and asparagine (corresponding codons, ATR and AAY are AT rich; bars in Fig. 1B). Additionally, glycine, proline and arginine, whose codons

Table 1
Fit of different models to six metazoan mitochondrial datasets.

Model	Statistic	Dataset					
		Metazoa	Lophotrochozoa	Ecdysozoa	Deuterostomia	Arthropoda	Mammalia
MtArt	Δ lnl	−1217	−706	−266	−542	−46	−3094
	AIC	2018	996	116	960	BEST	5572
	BIC	1118	827	BEST	961	BEST	4933
MtZoa	Δ lnl	−658	−293	−641	−62	−277	−2364
	AIC	900	170	866	BEST	462	4312
	BIC	BEST	BEST	751	BEST	463	3473
MtREV	Δ lnl	−5607	−3072	−3571	−1294	−2055	−628
	AIC	10798	5728	6618	2464	4018	840
	BIC	9898	5559	6503	2465	4019	BEST
Mechanistic GTR	Δ lnl	Highest	Highest	Highest	Highest	Highest	Highest
	AIC	BEST	BEST	BEST	292	324	BEST
	BIC	1977	2707	2761	3125	3208	2142

Note: for each of the dataset and model we show three statistics: the differences in the harmonic mean of the log-likelihoods (Δ lnl), AIC and BIC. The highest value of the log-likelihood is set to HIGHEST and the highest value of AIC and BIC is set as the BEST. Other values are reported as the difference compared to these values.

are all enriched in G and C nucleotides, are slightly more frequent in MtZoa, while glutamate, isoleucine, tyrosine and phenylalanine (AT rich) are less frequent. Similarly and more importantly, most of the replacements involving AT rich amino acids (NKMIYF) are favoured in MtArt, while those involving GC rich amino acids (GARP) are favoured in MtZoa. This is a key difference, which seems to reflect the compositional properties of the arthropod mtDNA that is typically biased toward a high content of A and T nucleotides and suggests that MtZoa may be a more appropriate estimator than MtArt for the study of differently biased datasets such as lophotrochozoans and deuterostomes, which are less AT rich (Rota-Stabelli and Telford, 2008).

3.2. The fit of models to datasets

We used the AIC and BIC criteria to assess how the various models fit diverse metazoan mitochondrial datasets. AIC, and especially BIC sensibly penalize the model in a way that is proportional to the number of parameters and have been proved to be an appropriate tool for non-nested model selection (Posada and Buckley, 2004). For the calculation of AIC and BIC we used the harmonic mean of the log-likelihood of the trees sampled from the Bayesian analyses of 6 different mitochondrial dataset using MtREV, MtArt, MtZoa and GTR models. Results are summarized in Table 1, which show for each dataset and model the mean log-likelihood, the AIC and the BIC values. According to this table, MtZoa is the preferred empirical model when diverse metazoan, lophotrochozoan and deuterostome species are analyzed. For these datasets, the differences in AIC or BIC values between MtZoa and MtArt or MtREV are high, in the range of, respectively, 100s and 1000s. Conversely, MtArt and MtREV clearly better fit the ecdysozoan and the mammalian datasets, respectively, reinforcing the view that the taxonomic level from which the matrices are estimated and different genetic codes (Abascal et al., 2006) may play a decisive role in the assessment of the model that best fits a certain dataset.

The log-likelihoods associated with the mechanistic GTR model (whose parameters have been deduced directly from the datasets) are clearly the highest for all the datasets. This is easily explained by the 209 free parameters of the GTR model (empirical models have only two), which are responsible for a natural increase in the log-likelihood. Interestingly, at least one of the empirical models (MtZoa, MtArt or MtREV) shows a significantly better fit to the data for some (according to AIC) or all datasets (according to BIC). This result suggests that, in the cases of small datasets, the considerable computational time required for the estimation of all the parameters of mechanistic GTR model is unlikely to be justified by a relatively moderate increase in the corresponding log-likelihood. It is remarkable that, according to our tree searches, in some cases GTR required more than 100 times the computational time required by any of the empirical models.

4. Conclusions

We suggest that MtZOA should be used for the mitogenomic analysis of deuterostome and lophotrochozoan datasets and for datasets containing diverse or basal metazoan groups. Conversely, MtArt and MtREV should be used, respectively, for ecdysozoan and mammalian datasets. As a general rule, we advocate that the taxonomic set from which models are estimated plays a decisive role in the assessment of the best fit to datasets and that, in the case of poor phylogenetic signal or problematic nodes, the use of a more appropriate model which reflects the evolutionary pattern of the given taxonomic sample, results in a much higher likelihood, a better fit to the dataset and may consequently help lessen possible systematic biases. We also suggest that, according to AIC and BIC criteria, empirical models may be preferable to the mechanistic GTR one, as a moderate increase in the log-likelihood of GTR trees, may not justify the much larger amount of time needed for computation. This is particularly true for taxonomically small datasets (such as the ones we used for the test of model fit) which may not contain sufficient substitutional information for a correct estimation of the replacement rates of the GTR mechanistic model.

In the Supplementary file MtZoa.txt we provide the replacement rates and the stationary frequencies of MtZoa, which can be used as a .dat file in CODEML (PAML). It is possible to use MtZoa in MrBayes specifying "prset aarevmatpr = fixed (between brackets the 190 values of the replacement matrix separated by commas)" and "prset statefreqpr = fixed (the 20 stationary frequencies comma-separated)". It is also possible to use MtZoa in the forthcoming version of PhyloBayes (http://www.lirmm.fr/mab/article.php3?id_article 329) using the command – mtzoa.

Acknowledgment

Omar Rota-Stabelli was supported by the European Community's Marie Curie Research Training Network ZOONET under contract MRTN-CT-2004-005624.

Appendix A. Supplementary data

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.ympev.2009.01.011.

References

Abascal, F., Posada, D., Knight, R.D., Zardoya, R., 2006. Parallel evolution of the genetic code in arthropod mitochondrial genomes. PLoS Biol. 4, e127.

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

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

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Automatic Control 19 (6), 716–723.

Carapelli, A., Lio, P., Nardi, F., van der Wath, E., Frati, F., 2007. Phylogenetic analysis of mitochondrial protein coding genes confirms the reciprocal paraphyly of Hexapoda and Crustacea. BMC Evol. Biol. 7 (Suppl 2), S8.

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.

Foster, P.G., Jermiin, L.S., Hickey, D.A., 1997. Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. J. Mol. Evol. 44, 282–289.

Horner, D.S., Lefkimmiatis, K., Reyes, A., Gissi, C., Saccone, C., Pesole, G., 2007. Phylogenetic analyses of complete mitochondrial genome sequences suggest a basal divergence of the enigmatic rodent *Anomalurus*. BMC Evol. Biol. 7, 16.

Huelsenbeck, J.P., Larget, B., Alfaro, M.E., 2004. Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. Mol. Biol. Evol. 21, 1123–1133.

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

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

Rota-Stabelli, O., Telford, M.J., 2008. A multi criterion approach for the selection of optimal outgroups in phylogeny: recovering some support for Mandibulata over Myriochelata using mitogenomics. Mol. Phylogenet. Evol. 48, 103–111.

Telford, M.J., Bourlat, S.J., Economou, A., Papillon, D., Rota-Stabelli, O., 2008. The evolution of the Ecdysozoa. Philos. Trans. R. Soc. Lond. B Biol. Sci. 363, 1529–1537.

Schwartz, G., 1978. Estimating the dimension of a model. Ann. Syst. 6 (2), 461–464.

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

Webster, B.L., Copley, R.R., Jenner, R.A., Mackenzie-Dodds, J.A., Bourlat, S.J., Rota-Stabelli, O., Littlewood, D.T., Telford, M.J., 2006. Mitogenomics and phylogenomics reveal priapulid worms as extant models of the ancestral Ecdysozoan. Evol. Dev. 8, 502–510.

Yang, Z., 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24, 1586–1591.

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