

## LETTERS

### MtArt: A New Model of Amino Acid Replacement for Arthropoda

Federico Abascal,\*† David Posada,\* and Rafael Zardoya†

\*Departamento de Bioquímica, Genética e Inmunología, Universidad de Vigo, Vigo, Spain; and †Departamento de Biodiversidad y Biología Evolutiva, Museo Nacional de Ciencias Naturales, Madrid, Spain

A statistical approach was applied to select those models that best fit each individual mitochondrial (mt) protein at different taxonomic levels of metazoans. The existing mitochondrial replacement matrices, MtREV and MtMam, were found to be the best-fit models for the mt-proteins of vertebrates, with the exception of Nd6, at different taxonomic levels. Remarkably, existing mitochondrial matrices generally failed to best-fit invertebrate mt-proteins. In an attempt to better model the evolution of invertebrate mt-proteins, a new replacement matrix, named MtArt, was constructed based on arthropod mt-proteomes. The new model was found to best fit almost all analyzed invertebrate mt-protein data sets. The observed pattern of model fit across the different data sets indicates that no single replacement matrix is able to describe the general evolutionary properties of mt-proteins but rather that taxonomical biases and/or the existence of different mt-genetic codes have great influence on which model is selected.

The use of models of protein evolution in likelihood-based phylogenetic analyses has been hindered by the complexity imposed by the larger size of the alphabet of amino acids compared with that of nucleotides and hence by the corresponding increase in the number of possible changes. As a result, and although mechanistic models of protein evolution do exist (Yang et al. 1998), most commonly used models of amino acid replacement are empirical, that is, replacement rates are estimated from large data sets and represented in a fixed matrix. Selection between alternative protein models for phylogenetic inference has been mainly based on the nominal similarity between the data set at hand and the original data used for constructing the empirical matrices. For instance, the MtREV model has been routinely applied for phylogenetic inference based on mitochondrial (mt) protein data (Wilson et al. 2000; Hwang et al. 2001; Nardi et al. 2003). However, such assumptions are not justified (Abascal et al. 2005; Keane et al. 2006), and the choice between competing models should be made within a proper statistical framework. Throughout this manuscript, we have chosen to use the Akaike Information Criterion or AIC (Posada and Buckley 2004) for model selection.

In this work, we analyzed which existing empirical amino acid matrices best fit different mt-protein alignments of metazoans, comprising whole mt-proteome as well as mt-gene data sets. At different vertebrate taxonomic levels (fig. 1), MtREV and MtMam are the best-fit (AIC) models for most analyzed sequence data sets (model selection uncertainty was minimal; not shown). Hence, our analyses, which are based on different vertebrate and mammal species from those originally utilized to estimate MtREV and MtMam (Adachi and Hasegawa 1996; Yang et al. 1998), confirmed that these 2 matrices are the best-fit models for vertebrate and mammal mt-proteins, respectively. The main exception is Nd6, mostly best fit by JTT (Jones et al. 1992). The amniote and mammal Atp8 data sets were also best fit by JTT. The general failure of MtREV/MtMam

to best model the evolution of Nd6 may be explained by taking into account that the *nd6* gene is the only one that was not included in the original training data set used to estimate the MtREV/MtMam matrices and/or because it is the only protein-coding gene encoded on the L-strand of the vertebrate mt-genomes and hence may be subjected to particular compositional biases. In the case of Atp8, these results are likely related to its short length and high variability. As expected, most best-fit models incorporated a gamma distribution (Yang 1993) either alone (+G) or in combination with a invariable sites' distribution (+I+G; Reeves 1992) to properly model the heterogeneity in rates of evolution across sites that is characteristic of proteins (Oliveira et al. 2003; Lartillot and Philippe 2004; Rodi et al. 2004; Landau et al. 2005). Some proteins (e.g., Nd4 and Nd5) were best fit by models incorporating the observed frequencies (+F; Cao et al. 1994) rather than the frequencies associated to the MtREV and MtMam matrices.

We further investigated which replacement matrix is optimal for each of the 13 mt-proteins in different groups of metazoans. We specifically tested whether MtREV/MtMam are the best choices for mt-protein sequence data of invertebrates. From the 158 invertebrate mt-genomes available at GenBank, we randomly selected 4 data sets each of 50 species in order to represent this heterogeneous (paraphyletic) group. Figure 2 shows that the best-fit models for the whole mt-proteome and each protein were neither MtREV nor MtMam, with the exceptions of Nd1, Nd2, Nd3, and Cox2. Analyses were also conducted at different taxonomic levels (fig. 2). Most data sets in Echinodermata, Platyhelminthes, Nematoda, and Mollusca were best fit by other models than MtREV (e.g., Platyhelminthes are generally best fit by JTT), whereas Arthropoda showed a better fit to MtREV. Importantly, no particular model consistently best fit all the individual protein data sets (but see ND5 that is generally best fit by Whelan and Goldman matrix).

Our results suggest that none of the candidate models was particularly adequate for modeling the evolution of the mt-proteomes of the different invertebrate groups. The poor fit of MtREV/MtMam models outside vertebrates prompted us to estimate a new empirical model of amino acid replacement. In order to estimate a new replacement matrix, we focused on Arthropoda, rather than mixing highly heterogeneous invertebrate groups. Arthropoda is one of the most

Key words: mitochondrial genomes, protein evolution, model selection, amino acid replacement matrix, arthropods.

E-mail: rafaz@mncn.csic.es.

*Mol. Biol. Evol.* 24(1):1–5, 2007

doi:10.1093/molbev/msl136

Advance Access publication October 16, 2006

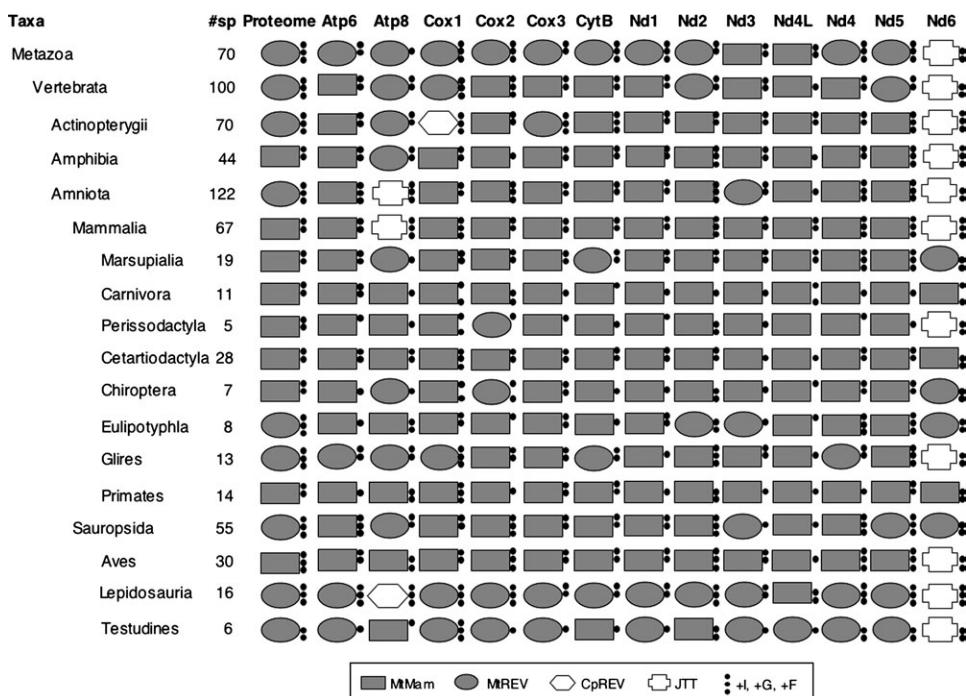


FIG. 1.—Model selection in vertebrates and mammals. The best-fit model for the proteome and each protein and for the different taxonomic groups is shown. Small black circles indicate whether the different corrections were applied for the best-fit model. The first circle from above indicates that +I was required, whereas the second and third circles refer to the +G and +F corrections, respectively. “#sp” indicates how many species were analyzed for each data set.

successful and diverse phyla and has the largest taxonomic sampling of invertebrate proteomes in the databases. The new matrix (see table S1, Supplementary Material online) was based on the whole mt-proteome of 36 arthropod species. This new model, which was named MtArt, fits the 36-arthropod data set much better than any previously existing empirical matrix: the log likelihood of MtArt+I+G (191 parameters plus 69 branch length estimates) is  $-92,522.12$ , whereas the log likelihood of MtREV+I+

G+F (21 parameters plus 69 branch length estimates) is  $-94,063.34$ , standing for a huge AIC difference of 2222.44.

As an alternative test of adequacy, we decided to investigate whether MtArt is also the best-fit model for other arthropods, as well as for other groups of invertebrates. Figure 3 shows that MtArt is not only the best-fit matrix for arthropods as a whole but also for the subphyla Crustacea, Arachnida, and Hexapoda, with the exception of the variable

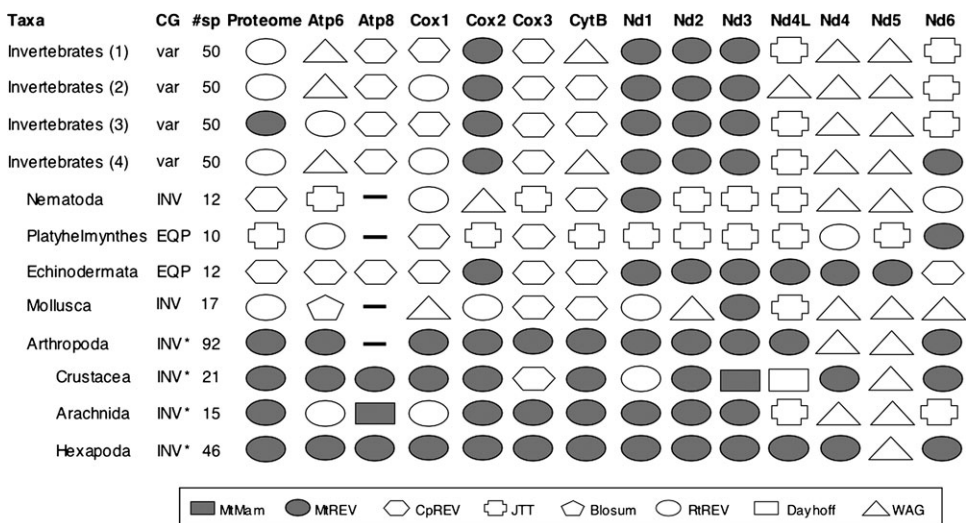


FIG. 2.—Model selection in invertebrates. The representation is similar to the one in figure 1.

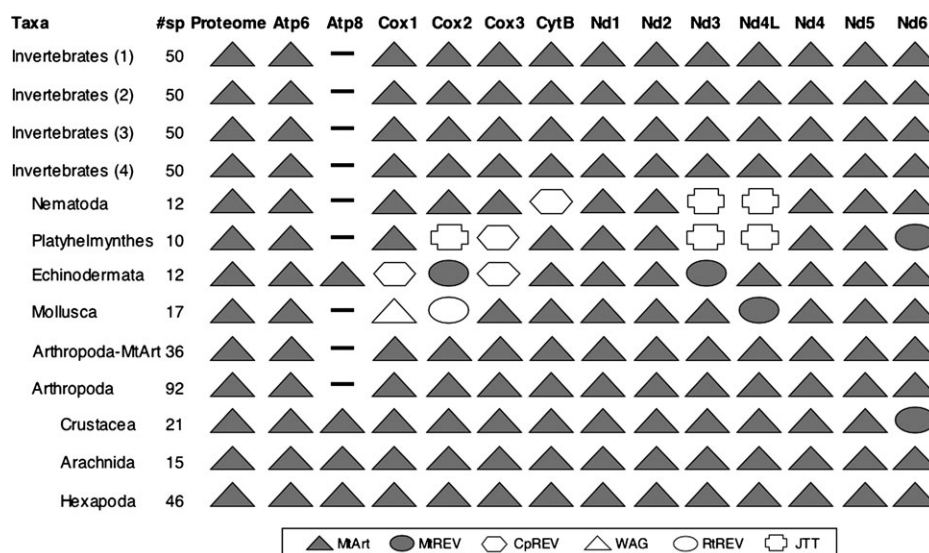


FIG. 3.—Model selection including the new MtArt model. The representation is similar to the one in figure 1.

and short (after Gblocks processing) crustacean Nd6 alignment. MtArt is also the best-fit model for the whole mt-proteome of other invertebrates, such as Mollusca, Platyhelminthes, Nematoda, and Echinodermata but not for all of their individual proteins. Moreover, the whole proteome and the individual proteins of the 4 random invertebrate data sets were all also best fit by MtArt. For vertebrate data sets, MtREV/MtMam are still the best choice (data not shown). Despite the apparent success of MtArt in modeling the different analyzed invertebrate data sets, better models may be devised indeed in the future, for example, by taking into account other features of protein evolution, like secondary and tertiary structure information (e.g., Goldman et al. 1998) or using mixture models (Lartillot and Philippe 2004).

In order to determine which are the differential features that account for the best fit of MtArt to invertebrate mt-proteins, we compared the replacement rates of the new matrix with those of MtREV, MtMam, and JTT (fig. 4). Several amino acid frequencies are particularly different between MtArt and the other matrices. For instance, serine is more frequent in MtArt than in MtREV, MtMam, and JTT (0.090 vs. 0.072, 0.072, and 0.068, respectively). This observation is congruent with the different genetic codes associated to these matrices: MtArt is based on arthropods, which use AGA codons, and in some species AGG as well, to code for serine (Abascal et al. 2006). MtREV is based on vertebrates, which use both codons as stop signals, and JTT is based on the universal translation of AGR as arginine. Phenylalanine is also found to be more frequent in MtArt, whereas threonine and alanine are less frequent. Replacements involving cysteine (e.g., cysteine ↔ valine and cysteine ↔ methionine) seem to be enriched in MtArt compared with the other matrices, whereas replacements involving histidine (e.g., histidine ↔ asparagine and histidine ↔ tyrosine) are less frequent. The replacement of the basic amino acids arginine and lysine is less frequent in MtArt than in JTT. MtMam and MtREV also show this replacement in lower frequency, an observation that has been related to the fact that this change requires

at least 2 mutations under the vertebrate mt-genetic code (because AGR do not code for arginine but instead are used as STOP codons) (Adachi and Hasegawa 1996). This argument may also apply to the lower frequency of this replacement observed in MtArt.

In the past, MtREV has been routinely used to model the evolution of mt-protein data (Wilson et al. 2000; Hwang et al. 2001; Nardi et al. 2003). It has been argued for this choice that the information embedded in MtREV is mainly related to the particular properties of mt-proteins (e.g., they are mostly transmembrane hydrophobic proteins), as well as to the differences between the standard and the vertebrate mt-genetic codes. However, there has been little investigation about these intuitive expectations. The general failure of MtREV/MtMam to best fit invertebrate mt-proteins and of MtArt to best fit vertebrate mt-proteins indicate that none of these matrices are appropriate descriptors of the general evolution of mt-proteins. These results, hence, suggest that the structural-functional properties of mt-proteins were not captured in MtREV/MtMam/MtArt and/or that the few differences between the vertebrate and invertebrate mt-genetic codes have a rather large impact on model selection (e.g., the amino acid Ser is more frequent in MtArt than in the other matrices). In addition, the pattern of model selection observed across metazoans may also indicate that the MtREV/MtMam/MtArt replacement rates could be specific of the taxonomic levels for which the matrices were trained. For example, the analysis of MtArt revealed that the replacements involving cysteine/histidine and other amino acids are, respectively, augmented/diminished in arthropods based on the comparison with MtREV, MtMam, and JTT.

## Materials and Methods

To retrieve genomes from GenBank (<http://www.ncbi.nlm.nih.gov>) at a given taxonomic level and to subsequently separate each of the protein-coding genes into separate files, we developed a program, MitoBank, based on the BioPerl library (Thompson et al. 1994; Stajich

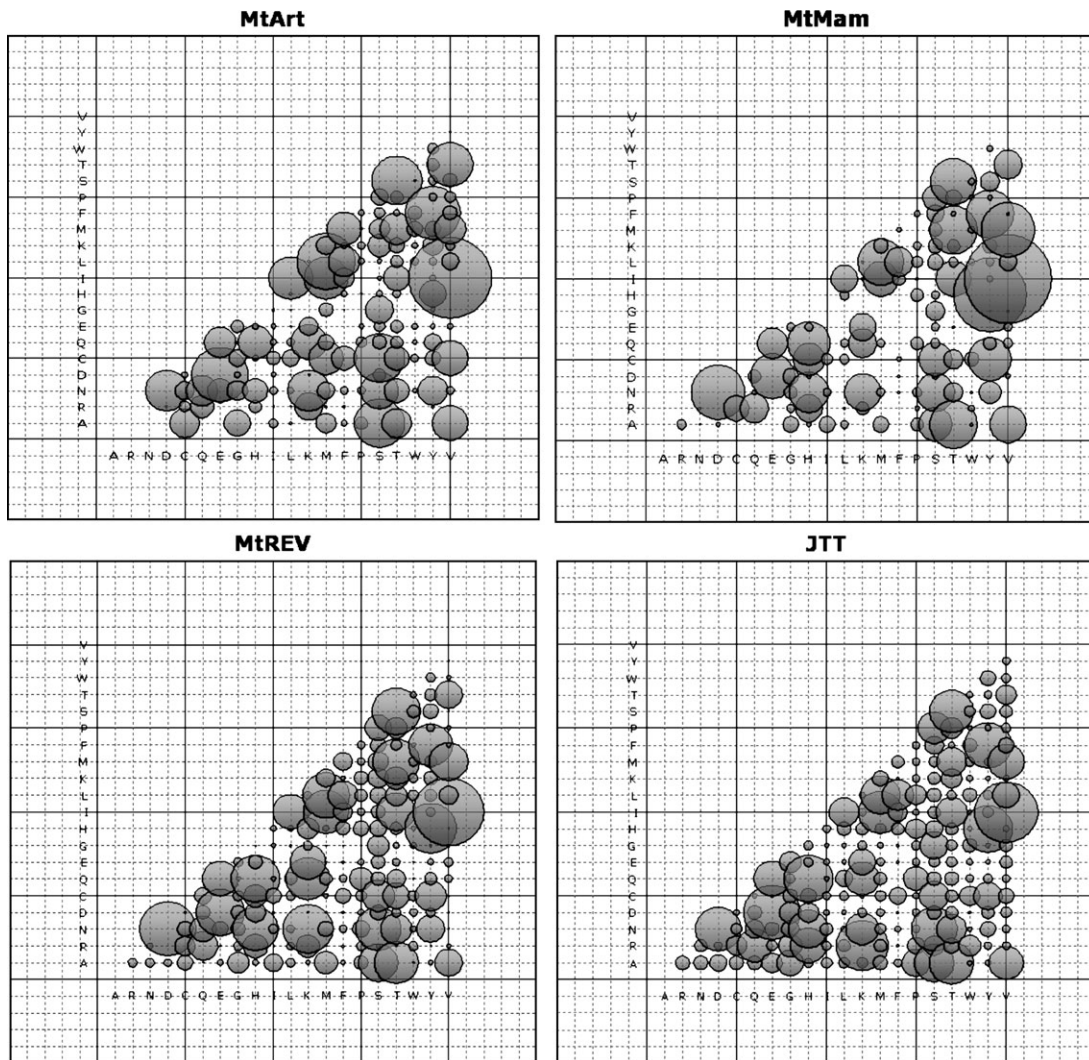


FIG. 4.—Comparison of MtArt, MtMam, MtREV, and JTT matrices. The area of the bubbles is proportional to the replacement rates.

et al. 2002) and available for download at <http://darwin.uvigo.es/>. A summary of the different taxonomic groups analyzed in this study and their relative phylogenetic position can be found in the supplementary material (fig. S1, Supplementary Material online). ClustalW (Thompson et al. 1994) was used to construct the multiple protein alignments. For those highly variable data sets that resulted in ambiguously aligned regions, the program Gblocks (Castresana 2000) was applied to objectively remove these regions from the alignment (Gblocks parameters:  $-t = p$   $-b1 = N/2 + 1$   $-b2 = N/2 + 1$   $-b3 = 8$   $-b4 = 10$   $-b5 = h$ , where  $N$  corresponds to the number of taxa). Readseq (Gilbert 2001) was intensively used to format alignments.

We used the program ProtTest version 1.2.6 (Drummond and Strimmer 2001; Guindon and Gascuel 2003; Abascal et al. 2005) to select best-fit models of protein evolution based on the Akaike criterion (Akaike 1973).

We used the PAML software package (Yang 1997) to estimate a REV model[s2] of protein evolution for arthropod mitochondria. From the 92 arthropods whose mt-genome has been sequenced to date, a balanced selection

of 36 species was performed in an attempt to cover the main groups without overrepresenting any of them. The parameters of the REV model were estimated based on a composite phylogeny (fig. S2, Supplementary Material online) that was assembled from different sources (Regier and Shultz 1997; Giribet and Ribera 2000; Giribet et al. 2001; Hwang et al. 2001; Wheeler et al. 2001; Delsuc et al. 2003; Nardi et al. 2003; Regier et al. 2005) to best reflect current knowledge of arthropod relationships. Polytomies were introduced where uncertainty existed. The replacement matrix was calculated from alignments after ambiguously aligned regions were excluded using the program Gblocks. The resulting replacement matrix, MtArt, has been included in the set of models in ProtTest version 1.2.16.

### Supplementary Material

Supplementary materials, including the MtArt replacement matrix (table S1 and MtArt.txt file), and figures S1 and S2 are available at *Molecular Biology and Evolution* online (<http://www.mbe.oxfordjournals.org>).

## Acknowledgments

We thank D. San Mauro and P. Fitze for their help with statistics. This work was supported by a research grant from the Fundación Banco Bilbao Vizcaya Argentaria [s3] (Spain). D.P. is also supported by the Ramón y Cajal programme of the Spanish Government.

## Literature Cited

- Abascal F, Posada D, Knight RD, Zardoya R. 2006. Parallel evolution of the genetic code in arthropod mitochondrial genomes. *PLoS Biol.* 4:e127.
- Abascal F, Zardoya R, Posada D. 2005. ProfTest: selection of best-fit models of protein evolution. *Bioinformatics.* 21:2104–2105.
- Adachi J, Hasegawa M. 1996. Model of amino acid substitution in proteins encoded by mitochondrial DNA. *J Mol Evol.* 42:459–468.
- Akaike H. 1973. Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csaki F, editors. *Proceedings of the 2nd International Symposium on Information Theory.* Budapest: Akademiai Kiado. p. 267–281.
- Cao Y, Adachi J, Janke A, Paabo S, Hasegawa M. 1994. Phylogenetic relationships among eutherian orders estimated from inferred sequences of mitochondrial proteins: instability of a tree based on a single gene. *J Mol Evol.* 39:519–527.
- Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol.* 17:540–552.
- Delsuc F, Phillips MJ, Penny D. 2003. Comment on “Hexapod origins: monophyletic or paraphyletic?” [author reply] *Science.* 301:1482.
- Drummond A, Strimmer K. 2001. PAL: an object-oriented programming library for molecular evolution and phylogenetics. *Bioinformatics.* 17:662–663.
- Gilbert D. 2001. ReadSeq: read & reformat biosequences [Internet]. Available from: <http://iubio.bio.indiana.edu/>.
- Giribet G, Edgecombe GD, Wheeler WC. 2001. Arthropod phylogeny based on eight molecular loci and morphology. *Nature.* 413:157–161.
- Giribet G, Ribera C. 2000. A review of arthropod phylogeny: new data based on ribosomal DNA sequences and direct character optimization. *Cladistics.* 16:204–231.
- Goldman N, Thorne JL, Jones DT. 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics.* 149:445–458.
- Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol.* 52:696–704.
- Hwang UW, Friedrich M, Tautz D, Park CJ, Kim W. 2001. Mitochondrial protein phylogeny joins myriapods with chelicerates. *Nature.* 413:154–157.
- Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci.* 8:275–282.
- Keane TM, Creevey CJ, Pentony MM, Naughton TJ, McLnerney JO. 2006. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evol Biol.* 6:29.
- Landau M, Mayrose I, Rosenberg Y, Glaser F, Martz E, Pupko T, Ben-Tal N. 2005. ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res.* 33:W299–W302.
- 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.
- Nardi F, Spinsanti G, Boore JL, Carapelli A, Dallai R, Frati F. 2003. Hexapod origins: monophyletic or paraphyletic? *Science.* 299:1887–1889.
- Oliveira L, Paiva PB, Paiva AC, Vriend G. 2003. Identification of functionally conserved residues with the use of entropy-variability plots. *Proteins.* 52:544–552.
- Posada D, Buckley TR. 2004. Model selection and model averaging in phylogenetics: advantages of AIC and bayesian approaches over likelihood ratio tests. *Syst Biol.* 53:793–808.
- Reeves JH. 1992. Heterogeneity in the substitution process of amino acid sites of proteins coded for by mitochondrial DNA. *J Mol Evol.* 35:17–31.
- Regier JC, Shultz JW. 1997. Molecular phylogeny of the major arthropod groups indicates polyphyly of crustaceans and a new hypothesis for the origin of hexapods. *Mol Biol Evol.* 14:902–913.
- Regier JC, Shultz JW, Kambic RE. 2005. Pancrustacean phylogeny: hexapods are terrestrial crustaceans and maxillopods are not monophyletic. *Proc Biol Sci.* 272:395–401.
- Rodi DJ, Mandava S, Makowski L. 2004. DIVAA: analysis of amino acid diversity in multiple aligned protein sequences. *Bioinformatics.* 20:3481–3489.
- Stajich JE, Block D, Boulez K, et al. (21 co-authors). 2002. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.* 12:1611–1618.
- Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22:4673–4680.
- Wheeler WC, Whiting M, Wheeler QD, Carpenter JM. 2001. The phylogeny of the extant hexapod orders. *Cladistics.* 17:113–169.
- Wilson K, Cahill V, Ballment E, Benzie J. 2000. The complete sequence of the mitochondrial genome of the crustacean *Penaeus monodon*: are malacostracan crustaceans more closely related to insects than to branchiopods? *Mol Biol Evol.* 17:863–874.
- Yang Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol Biol Evol.* 10:1396–1401.
- Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci.* 13:555–556.
- Yang Z, Nielsen R, Hasegawa M. 1998. Models of amino acid substitution and applications to mitochondrial protein evolution. *Mol Biol Evol.* 15:1600–1611.

Sudhir Kumar, Associate Editor

Accepted September 25, 2006