LETTER

Analysis of Rare Amino Acid Replacements Supports the Coelomata Clade

Igor B. Rogozin, Yuri I. Wolf, Liran Carmel, and Eugene V. Koonin

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland

The recent analysis of a novel class of rare genomic changes, RGC_CAMs (after conserved amino acids—multiple substitutions), supported the Coelomata clade of animals as opposed to the Ecdysozoa clade (Rogozin et al. 2007). A subsequent reanalysis, with the sequences from the sea anemone *Nematostella vectensis* included in the set of outgroup species, suggested that this result was an artifact caused by reverse amino replacements and claimed support for Ecdysozoa (Irimia et al. 2007). We show that the internal branch connecting the sea anemone to the bilaterian animals is extremely short, resulting in a weak statistical support for the Coelomata clade. Direct estimation of the level of homoplasy, combined with taxon sampling with different sets of outgroup species, reinforces the support for Coelomata, whereas the effect of reversals is shown to be relatively minor.

As the set of sequenced genomes from diverse taxa rapidly grows, phylogenetic analysis is entering a new era when the reconstruction of the evolutionary history of organisms on the basis of full-scale comparison of their genomes becomes the strategy of choice. In addition to more traditional, genome-wide analysis of alignments, rare genomic changes (RGCs) that are likely to comprise derived shared characters of individual clades are increasingly used in genome-wide phylogenetic studies (Rokas and Holland 2000; Nei and Kumar 2001; Rokas et al. 2003).

We have recently proposed a new type of RGCs designated RGC_CAMs (after conserved amino acids—multiple substitutions), which are inferred using a genome-scale analysis of protein and underlying nucleotide sequence alignments (Rogozin et al. 2007). The RGC_CAM approach utilizes amino acid residues that are conserved in the major lineages within an analyzed taxonomic division (e.g., eukaryotes), with the exception of a few species comprising a putative clade. In addition, to reduce the effect of homoplasy, only those amino acid replacements that require 2 or 3 nucleotide substitutions are employed for phylogenetic inference. The RGC_CAM analysis has been combined with a procedure for rigorous statistical testing of competing phylogenetic hypotheses and shown to be robust to branch-length differences and taxon sampling. When applied to animal phylogeny, the RGC_CAM approach significantly supports the coelomate clade that unites chordates with arthropods as opposed to the ecdysozoan (molting animals) clade that encompasses arthropods and nematodes (Rogozin et al. 2007). This conclusion is compatible with some previous genome-wide phylogenetic analyses (Mushegian et al. 1998; Blair et al. 2002; Stuart and Berry 2004; Wolf et al. 2004; Philip et al. 2005) but not others (Copley et al. 2004; Dopazo and Dopazo 2005; Philippe et al. 2005) and runs against the view of animal evolution that is currently prevailing in the evolutionary developmental biology (evo-devo) community (Aguinaldo et al. 1997; Adoutte et al. 2000; Telford and Copley 2005).

Irimia et al. (2007) have further explored the RGC_CAM approach, after adding proteins from 2 recently sequenced animal genomes, the cnidarian (sea anemone) *Nematostella vectensis* and the nematode *Brugia malayi*, to the original data set of Rogozin et al. (2007). The analysis of the resulting alignments has suggested that the apparent support for the coelomate clade resulted from the rapid rate of evolution in the nematodes (Irimia et al. 2007). There are 2 types of errors that have the potential to distort the results obtained with the RGC_CAM approach, namely, reversals and parallel changes (fig. 1). Irimia et al. (2007) emphasize the effect of reversals but, effectively, ignore parallel changes; furthermore, they do not report any rigorous statistical analysis of the results.

Here we report a reanalysis of animal evolution with the RGC_CAM method, with special attention to the sources of potential artifacts, using a further amended data set. The adopted animal phylogeny is shown in figure 1, and the results of the RGC_CAM analysis of the set of 15 species are shown in the table 1 (top row). Only one RGC_CAM supported the coelomate clade, and 2 RGC_CAMs supported the ecdysozoan clade (table 1). Thus, considering the lengths of the respective branches, the coelomate clade still had a weak statistical support (table 1; see Methods for the details of the statistical test) under the assumption of the basal position of *N. vectensis* (the branch separating *N. vectensis* from the rest of the Bilateria is only 3 RGC_CAMs long [fig. 1], with no reversals). We further explored the support for different topologies provided by RGC_CAMs by performing taxon sampling of the outgroup species. All combinations of 10–15 species, that is, including from 1 to 6 outgroup species (63 combinations altogether), were analyzed. Of the 63 combinations, in 29 combinations of species, the raw number of RGC_CAMs compatible with the coelomate topology was greater than the number of RGC_CAMs compatible with the ecdysozoa topology, whereas the reverse was true of 32 combinations, with the remaining 2 combinations showing the same number of RGC_CAMs for both topologies (table 1). Considering the respective branch lengths, for 57 (91%) combinations of species, there was statistical support for the coelomate clade (table 1), whereas with the rest of the combinations (9%), none of the topologies received statistical support. Thus, the results of this extensive RGC_CAM analysis indicate

Key words: phylogenetic analysis, cladistics, rare genomic changes, coelomata, ecdysozoa.

E-mail: koonin@ncbi.nlm.nih.gov.

Mol. Biol. Evol. 24(12):2594–2597. 2007
doi:10.1093/molbev/msm218
Advance Access publication October 5, 2007

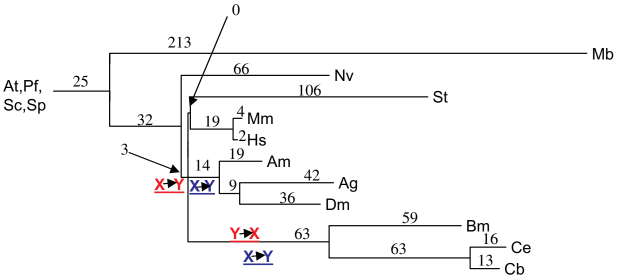

FIG. 1.—The animal phylogeny employed in this study. The node connecting Deuterostomes, nematodes, and insects is shown as a trifurcation. Branch lengths were calculated in RGC_CAM units (Rogozin et al. 2007), and the respective value is given above each branch. Reversals are shown in red, and parallel changes are shown in blue. Am, *Apis mellifera*; Ag, *Anopheles gambiae*; At, *Arabidopsis thaliana*; Bm, *Brugia malayi*; Cb, *Caenorhabditis briggsae*; Ce, *Caenorhabditis elegans*; Dm, *Drosophila melanogaster*; Hs, *Homo sapiens*; Mb, *Monosiga brevicollis*; Mm, *Mus musculus*; Nv, *Nematostella vectensis*; Pf, *Plasmodium falciparum*; Sc, *Saccharomyces cerevisiae*; Sp, *Schizosaccharomyces pombe*; St, *Strongylocentrotus purpuratus*.

support for the Coelomata topology but no significant support for the Ecdysozoa. As indicated by the results in table 1, the test loses most of its power when *N. vectensis* is included in the outgroup species set due to the very short branch connecting this species to the rest of animals. The problem could be caused by compressed cladogenesis at the base of the animal clade (Rokas et al. 2005; Rokas and Carroll 2006) although an alternative explanation, such as a whole-genome duplication with subsequent differential loss of paralogs, cannot be ruled out (Rogozin et al. 2007).

There are 2 types of evolutionary events that have the potential to produce artifacts in the RGC_CAM analysis, namely, parallel changes and reversals (fig. 1) (Irimia et al. 2007; Rogozin et al. 2007). Parallel changes are taken into account in the statistical test that was applied as part of the original RGC_CAM analysis (Rogozin et al. 2007) (table 1). However, reversals might present a substantial problem for the RGC_CAM method (Irimia et al. 2007). The RGC_CAM approach provides for the possibility to estimate the level of homoplasy directly. To obtain an estimate of the number of reversals, we employed the scheme shown in figure 2. We required the same amino acid to be shared by a pair of closely related nematodes (the 2 *Caenorhabditis* species) and outgroup species but not the rest of the animals (fig. 2). A reversal is the most parsimonious explanation for this pattern, assuming that the tree topology in the node leading to Deuterostomes, insects, and worms is a true trifurcation, and such reversals were invoked by Irimia et al. (2007) to explain the observed RGC_CAM support for the coelomate clade. If the tree topology in the node leading to Deuterostomes, insects, and worms is not a true trifurcation, 2 parallel changes, one in the internal branch leading to the coelomate clade and the other one in the *B. malayi* branch, also might explain the observed pattern. Thus, the obtained estimates give the upper bound of the number of reversals. The branches leading to the 2 *Caenorhabditis* species and to the 3 nematodes both comprise 63 RGC_CAMs (fig. 1). Thanks to this coincidence, the homoplasy level that is determined here can be directly compared with the results of the RGC_CAM analysis of the Coelomata–Ecdysozoa problem (figs. 1 and 2). For 50 of the 63 species sets obtained by sampling (see above), the number of RGC_CAMs supporting the coelomate topology is greater than the number of reversals (table 1). This excess is sufficient to reject the hypothesis that the RGC_CAMs supporting the coelomate topology are reversals with a high statistical significance (Student's *t*-test, $P = 4 \times 10^{-7}$). Thus, the support for the coelomate clade obtained using the RGC_CAM method is not explained solely by reversals.

In summary, the results of RGC_CAM analysis reported here reinforce the support for the Coelomata clade observed with this approach in the original study (Rogozin et al. 2007) and additionally emphasize the importance of the analysis of multiple outgroups for obtaining reliable results in the study of deep phylogenies. Of course, the definitive solution to the coelomate–ecdysozoa conundrum will require a much larger set of complete genome sequences representing diverse animal taxa.

Methods

Each of the 694 protein alignments constructed from selected eukaryotic orthologous groups (KOGs) (Koonin et al. 2004) analyzed here included orthologous genes from 10 eukaryotic species with completely sequenced genomes: *Homo sapiens*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Arabidopsis thaliana*, *Anopheles gambiae*, *Plasmodium falciparum*, *Caenorhabditis briggsae*, and *Mus musculus* (Rogozin et al. 2007). Amino acid sequence alignments are available at ftp://ftp.ncbi.nlm.nih.gov/pub/koonin/RGC_CAM/. To these KOGs, probable orthologs from 5 other animal genomes, namely, those of *N. vectensis*, *B. malayi*, *Apis mellifera*, *Strongylocentrotus purpuratus*, and *Monosiga brevicollis*, were added using the COGNITOR method (Tatusov et al. 1997). Briefly, all the protein sequences from the new genomes are compared with the protein sequences previously included in the KOGs; a protein is assigned to a KOG when 2 genome-specific best hits to members of the given KOG are detected. To minimize misalignment problems, only conserved, unambiguously aligned regions of the alignments constructed using the MUSCLE program (Edgar 2004) were included in the further analysis. Specifically, all positions containing a deletion or insertion in at least one sequence were removed from the protein sequence alignment together with 5 adjacent positions (Rogozin et al. 2007).

The statistical test of phylogenetic hypotheses is based on a null model under which, in a comparison of 2 alternative hypotheses, for example, ([X−Y],Z) versus ([X−Z],Y), the number of RGC_CAMs that are shared by 2 lineages due to chance (N_{XY} and N_{XZ}) is proportional to the length of the branch the position of which differs between the 2 hypotheses, that is, Y and Z, respectively. The significance of the difference between normalized numbers of RGC_CAMs was estimated using Fisher's exact test (Rogozin et al. 2007).

Table 1
RGC_CAM Analysis of the Coelomata–Ecdysozoa Problem with Sampling of the Outgroup Species

Combination of Outgroup	Hypothesis (No. of RGC_CAMs in Support)			Branch Lengths (No. of RGC_CAMs)					No. of
Species (At, Sc, Sp, Pf, Mb, Nv)	C	E	B	Deuter	Insects	Worms	Stem	*P*(C–E)	Reversals
111111	1	2	1	0	14	63	3	0.046*	1
111110	5	3	1	1	15	66	39	2×10^{-5}*	2
111101	1	2	1	0	14	78	5	0.037*	1
111011	1	4	1	1	21	105	7	0.089	1
110111	2	3	2	0	15	74	3	0.003*	1
101111	1	2	1	1	16	77	3	0.073	2
011111	1	4	1	0	16	76	3	0.062	1
111001	1	7	1	1	24	132	12	0.111	1
111100	7	3	2	1	15	81	84	$<10^{-6}$*	2
111010	5	5	1	3	22	113	62	5×10^{-5}*	4
110110	6	4	2	2	16	78	43	9×10^{-6}*	2
110101	2	3	2	0	15	94	6	0.002*	1
110011	2	5	2	2	22	126	7	0.013*	1
011110	6	5	2	2	19	80	55	2×10^{-5}*	2
001111	1	4	1	1	20	96	3	0.096	2
010111	2	5	2	0	18	89	4	0.005*	1
011011	2	7	1	1	26	140	10	0.009*	1
011101	1	4	1	0	17	97	6	0.049*	1
001011	3	8	2	2	38	185	14	0.001*	2
001101	2	6	1	1	21	126	9	0.009*	2
001110	6	5	2	5	23	101	66	6×10^{-5}*	4
010011	3	10	2	2	29	168	15	0.003*	1
010101	2	6	2	0	19	118	8	0.004*	1
010110	9	6	3	3	21	95	68	$<10^{-6}$*	2
011000	35	13	6	12	35	211	679	$<10^{-6}$*	14
011001	5	11	1	3	31	191	22	7×10^{-5}*	1
011010	11	9	2	8	30	153	114	$<10^{-6}$*	4
011100	12	5	4	3	20	102	172	$<10^{-6}$*	5
010010	18	12	3	12	34	184	212	$<10^{-6}$*	4
010001	8	17	2	5	37	244	41	10^{-6}*	1
001100	16	8	4	7	26	137	305	$<10^{-6}$*	8
010100	20	7	8	5	22	125	318	$<10^{-}6$*	10
100001	6	17	3	4	48	272	41	2×10^{-5}*	3
100010	14	12	4	10	37	214	163	$<10^{-6}$*	9
100011	3	7	2	3	35	197	17	0.002*	2
100100	19	9	4	7	24	145	318	$<10^{-6}$*	8
100101	3	5	2	1	21	134	15	5×10^{-4}*	2
100110	7	7	2	5	22	107	68	3×10^{-5}*	3
100111	2	3	2	1	21	100	8	0.005*	2
101001	3	8	1	2	30	178	19	0.001*	2
101010	8	6	2	6	26	149	83	10^{-6}*	7
101011	2	4	1	2	25	137	9	0.008*	2
101100	9	4	2	4	18	104	117	$<10^{-6}$*	4
101101	1	3	1	1	16	98	6	0.076	2
101110	5	3	1	4	17	81	45	10^{-4}*	3
110001	2	8	2	2	27	165	15	0.016*	1
110010	8	6	2	5	23	136	73	$<10^{-6}$*	4
110100	9	4	3	2	16	98	114	$<10^{-6}$*	2
111000	13	8	3	3	25	144	160	$<10^{-6}$*	5
000111	3	5	2	1	27	131	9	5×10^{-4}*	2
110000	22	9	4	5	28	180	261	$<10^{-6}$*	9
101000	26	10	5	9	33	197	269	$<10^{-6}$*	11
001010	16	11	4	11	42	205	194	$<10^{-6}$*	10
001001	8	16	2	4	45	263	45	$<10^{-6}$*	3
000110	11	9	6	7	30	140	175	$<10^{-6}$*	6
000101	6	10	2	2	30	184	34	2×10^{-6}*	2
000011	7	15	4	4	56	302	42	10^{-6}*	2
100000	71	23	12	19	53	301	1737	$<10^{-6}$*	34
010000	74	22	16	19	42	272	2727	$<10^{-6}$*	33
001000	74	22	15	21	53	298	2120	$<10^{-6}$*	42
000100	71	15	20	15	36	204	3862	$<10^{-6}$*	47
000010	55	21	13	21	64	337	1520	$<10^{-6}$*	30
000001	19	40	7	13	88	510	327	$<10^{-6}$*	8

NOTE.—The absence/presence of a species (At, Sc, Sp, Pf, Mb, and Nv) is denoted by 0/1. The following 3 phylogenetic hypotheses were analyzed: C, Coelomata, that is, (Deuterostomes, insects) nematodes; E, Ecdysozoa, that is, (insects, nematodes) Deuterostomes; B, bizarre, that is, (Deuterostomes, nematodes) insects. *P*(C–E) is the probability that the C and E hypotheses are equally likely, calculated using Fisher's exact test (Rogozin et al. 2007). Cases where the C hypothesis received a significant statistical support are indicated by asterisks. At, *Arabidopsis thaliana*; Mb, *Monosiga brevicollis*; Nv, *Nematostella vectensis*; Pf, *Plasmodium falciparum*; Sc, *Saccharomyces cerevisiae*; Sp, *Schizosaccharomyces pombe*.

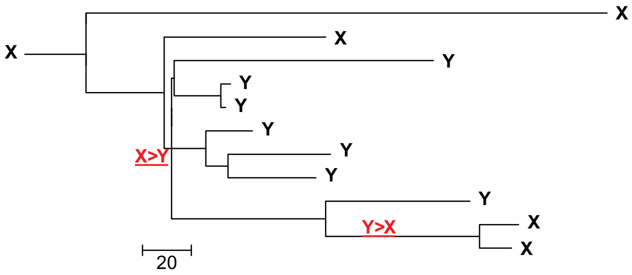

FIG. 2.—Direct determination of the number of reversals. X and Y denote 2 amino acids found in a particular position. The reversals are shown in red. The tree is the same as in figure 1, but the species names are omitted for simplicity.

Acknowledgments

We thank Scott Roy for providing his manuscript prior to publication and Miklos Csuros for helpful discussions. The sequence data for *Monosiga brevicolis* were produced by the US Department of Energy Joint Genome Institute http://www.jgi.doe.gov/. The *B. malayi* sequencing effort (http://www.tigr.org) is part of the International Brugia Genome Sequencing Project and is supported by an award from the National Institute of Allergy and Infectious Diseases, National Institutes of Health. This work was supported in part by the Intramural Research Program of the National Library of Medicine at National Institutes of Health/Department of Health and Human Services.

Literature Cited

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 USA. 97:4453–4456.

Aguinaldo AM, Turbeville JM, Linford LS, Rivera MC, Garey JR, Raff RA, Lake JA. 1997. Evidence for a clade of nematodes, arthropods and other moulting animals. Nature. 387:489–493.

Blair JE, Ikeo K, Gojobori T, Hedges SB. 2002. The evolutionary position of nematodes. BMC Evol Biol. 2:7.

Copley RR, Aloy P, Russell RB, Telford MJ. 2004. Systematic searches for molecular synapomorphies in model metazoan genomes give some support for Ecdysozoa after accounting for the idiosyncrasies of Caenorhabditis elegans. Evol Dev. 6:164–169.

Dopazo H, Dopazo J. 2005. Genome-scale evidence of the nematode-arthropod clade. Genome Biol. 6:R41.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Irimia M, Maeso I, Penny D, Garcia-Fernandez J, Roy SW. 2007. Rare coding sequence changes are consistent with Ecdysozoa, not Coelomata. Mol Biol Evol. 24:1604–1607.

Koonin EV, Fedorova ND, Jackson JD, et al. (18 co-authors). 2004. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biol. 5:R7.

Mushegian AR, Garey JR, Martin J, Liu LX. 1998. Large-scale taxonomic profiling of eukaryotic model organisms: a comparison of orthologous proteins encoded by the human, fly, nematode, and yeast genomes. Genome Res. 8:590–598.

Nei M, Kumar S. 2001. Molecular evolution and phylogenetics. Oxford: Oxford University.

Philip GK, Creevey CJ, McInerney JO. 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, Lartillot N, Brinkmann H. 2005. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa and Protostomia. Mol Biol Evol. 22:1246–1253.

Rogozin IB, Wolf YI, Carmel L, Koonin EV. 2007. Ecdysozoan clade rejected by genome-wide analysis of rare amino acid replacements. Mol Biol Evol. 24:1080–1090.

Rokas A, Carroll SB. 2006. Bushes in the tree of life. PLoS Biol. 4:e352.

Rokas A, Holland PW. 2000. Rare genomic changes as a tool for phylogenetics. Trends Ecol Evol. 15:454–459.

Rokas A, King N, Finnerty J, Carroll SB. 2003. Conflicting phylogenetic signals at the base of the metazoan tree. Evol Dev. 5:346–359.

Rokas A, Kruger D, Carroll SB. 2005. Animal evolution and the molecular signature of radiations compressed in time. Science. 310:1933–1938.

Stuart GW, Berry MW. 2004. An SVD-based comparison of nine whole eukaryotic genomes supports a coelomate rather than ecdysozoan lineage. BMC Bioinformatics. 5:204.

Tatusov RL, Koonin EV, Lipman DJ. 1997. A genomic perspective on protein families. Science. 278:631–637.

Telford MJ, Copley RR. 2005. Animal phylogeny: fatal attraction. Curr Biol. 15:R296–R299.

Wolf YI, Rogozin IB, Koonin EV. 2004. Coelomata and not Ecdysozoa: evidence from genome-wide phylogenetic analysis. Genome Res. 14:29–36.

Lauren McIntyre, Associate Editor

Accepted September 27, 2007