Evolution of viruses and cells: do we need a fourth domain of Life to explain the origin of eukaryotes?

David Moreira and Purificación López-García

Supplementary material

Materials and Methods

(a) RNAP2 sequence data set construction

We based our analyses on the RNAP2 data set made available by Williams et al. [1], who had tried to reconstruct a sequence alignment as similar as possible to that used in the original work by Boyer et al. [2]. The data set contained 80 taxa and 272 amino acid positions. We modified it in two different ways. First, we generated 10 artificial sequences consisting on random chains of amino acids with same amino acid composition as the average composition of the viral sequences in Boyer et al.'s data set [2]. We used an in-house python script to generate the random sequences, which had 272 amino acid positions. These sequences were incorporated to the multiple sequence alignment one by one or in groups of increased size (from 2 to 10). Second, we enriched the taxon sampling with new viral and eukaryotic sequences identified by BLAST [3] in the non-redundant GenBank data base and in the transcriptome data provided by the Marine Microbial Eukaryote Transcriptome Sequencing Project [4]. All sequences were aligned using MUSCLE [5]. Conserved alignment sites were identified using the very stringent method implemented in GBLOCKS [6]. Sequence composition bias was examined using the amino-acid composition test of the TREE-PUZZLE software [7].

(b) Phylogenetic reconstruction

Phylogenetic trees were reconstructed with two different methods. First, by replicating the approach used by Boyer et al. [2], who applied approximate maximum likelihood tree reconstruction with the program FASTTREE [8] and the single-matrix JTT model of

sequence evolution [9]. Second, by Bayesian inference using the program PHYLOBAYES [10] with the non-homogeneous CAT model [11]. Four independent chains were run until the maxdiff parameter was <0.1 (>100000 cycles for the smaller chain). The first 5000 trees were discarded as "burn-in" and one on two of the remaining trees from each chain were sampled to test for convergence and to compute the 50% majority rule consensus.

References

1. Williams T.A., Embley T.M., Heinz E. 2011 Informational gene phylogenies do not support a fourth domain of life for nucleocytoplasmic large DNA viruses. *PLoS One* **6**, 16.
2. Boyer M., Madoui M.A., Gimenez G., La Scola B., Raoult D. 2010 Phylogenetic and phyletic studies of informational genes in genomes highlight existence of a 4 domain of life including giant viruses. *PLoS One* **5**, e15530.
3. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J. 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402.
4. Keeling P.J., Burki F., Wilcox H.M., Allam B., Allen E.E., Amaral-Zettler L.A., Armbrust E.V., Archibald J.M., Bharti A.K., Bell C.J., et al. 2014 The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biol* **12,** 6.
5. Edgar R.C. 2004 MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792-1797.
6. Castresana J. 2000 Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17**(4), 540-552.
7. Schmidt H., Strimmer K., Vingron M., von Haeseler A. 2002 TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18**, 502-504.
8. Price M.N., Dehal P.S., Arkin A.P. 2010 FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, 0009490.
9. Jones D.T., Taylor W.R., Thornton J.M. 1992 The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* **8**, 275-282.
10. Lartillot N., Lepage T., Blanquart S. 2009 PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* **25**, 2286-2288.
11. 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.