THE CONSISTENT SIGNAL IN GENOME TREES REVEALED BY REDUCING THE IMPACT OF NOISE

Bas E. Dutilh [(1)], Martijn A. Huynen [(1)], William J. Bruno [(2)], and Berend Snel [(1).]

[(1)] Center for Molecular and Biomolecular Informatics / Nijmegen Center for Molecular Life Sciences, University of Nijmegen, Toernooiveld 1, 6525 ED, Nijmegen, The Netherlands.

[(2)] Theoretical Biology and Biophysics, Los Alamos National Laboratory, NM, USA.

www.cmbi.kun.nl/~dutilh dutilh@cmbi.kun.nl

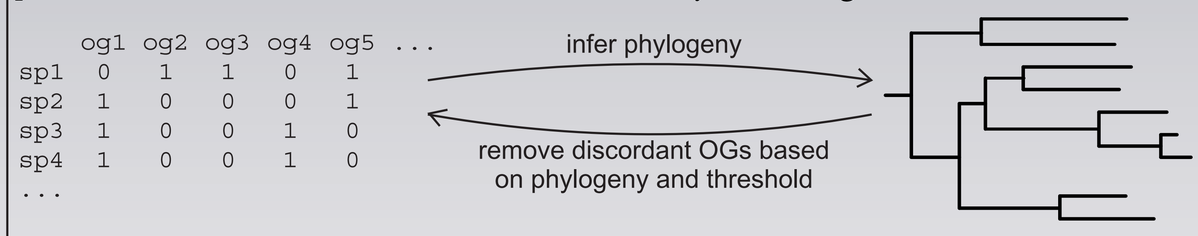
Introduction

It has been argued that phylogenetic trees based on shared gene content are unreliable because of convergence due to horizontal gene transfer and parallel gene loss.

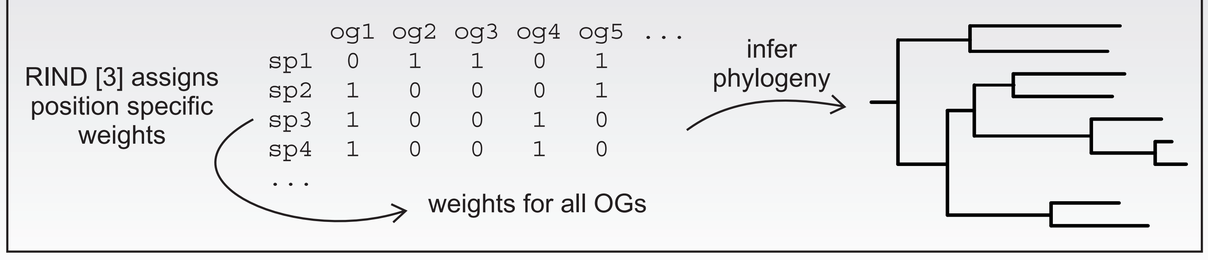
Here, we reduce the impact of noise in gene content phylogenies by two schemes that treat the presence/absence profiles of orthologous groups (OGs) in complete genomes as sequence alignments [1].

Data. A binary profile indicated the presence or absence of 19433 OGs [2] in 89 complete genomes (16 archaea, 65 bacteria and 8 eukaryotes).

Method I: iterative removal of discordant OGs. Iteratively, phylogenies were inferred and discordant OGs identified by comparing their distribution with random profiles. More and more noise could be removed by increasing the threshold.

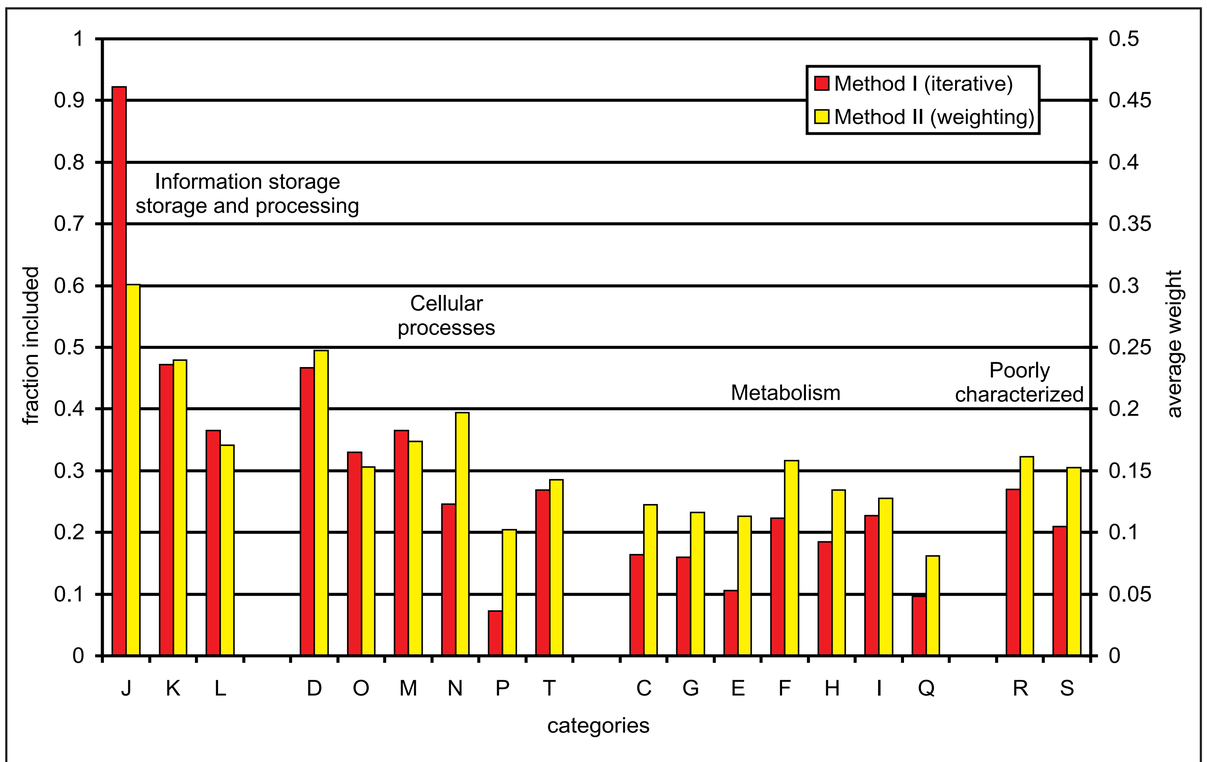


Method II: weighting OGs. RIND was originally developed for assessing amino acid sequences [3]. After assigning high weights to the taxon specific genes, and low weights to genes that evolve rapidly, a filtered gene content phylogeny was inferred.

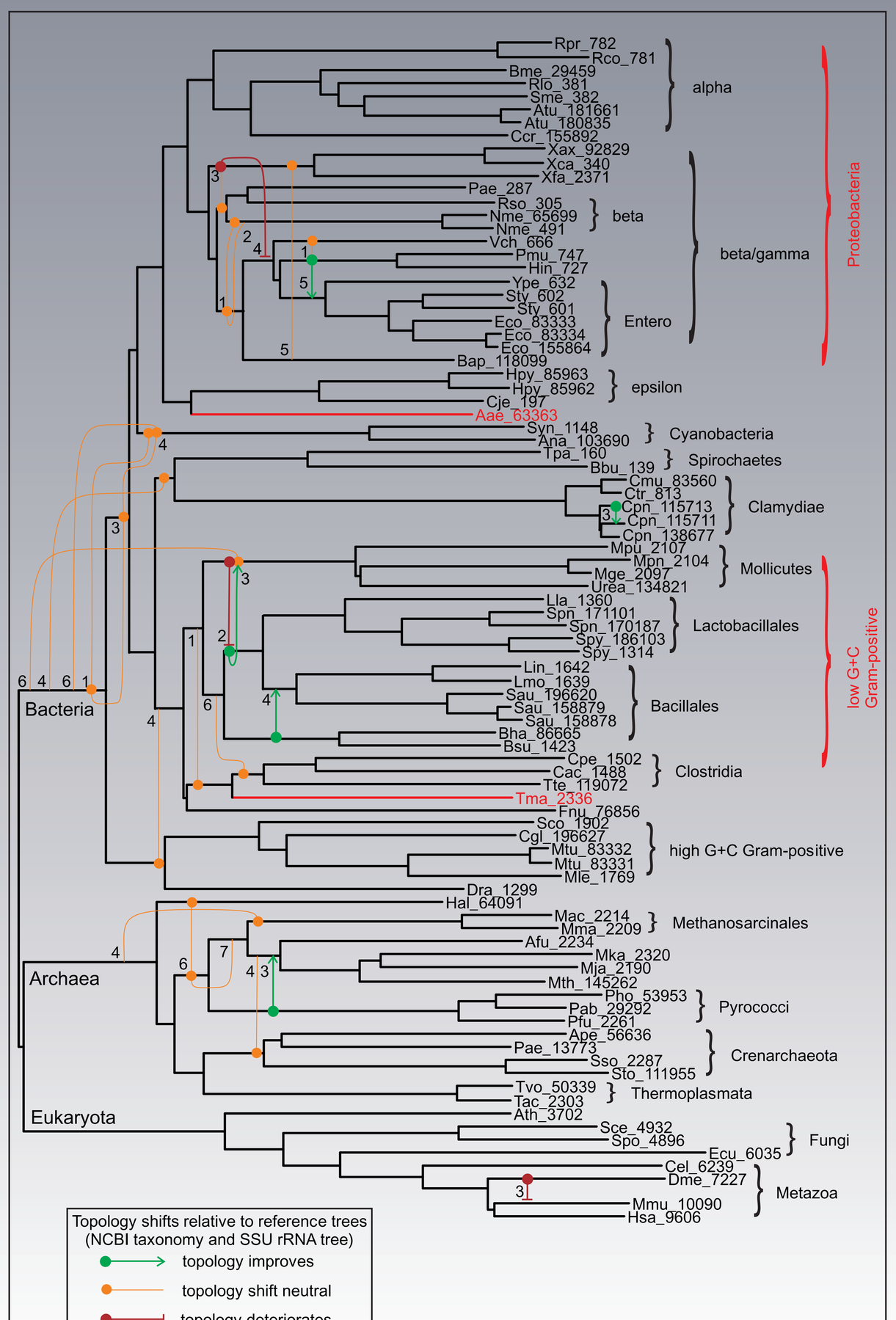


Results

- Method I shows that 69% of the OGs can be removed before the phylogenetic signal breaks down. 27% of the OGs are perfect according to the initial phyhlogeny.
- Iterations initialized with random phylogenies converge rapidly. This shows that the phylogenetic signal is the only signal present in the gene distribution across genomes.



Functional categories of the slow-evolving COGs [4]. 'Metabolism' COGs are faster evolving than 'Information storage and processing' COGs. This supports the complexity hypothesis, i.e. operational genes are transferred more readily than informational genes.



Gene content phylogeny *). The shifts that occur during the iterations of Method I are indicated. The improvements are small but consistent with Method II. The hyperthermophiles *Aquifex aeolicus* (related to the Proteobacteria) and *Thermotoga maritima* (Gram-positive bacteria) remain in their respective positions after removal of horizontal transfer candidates, supporting the hypothesis of the independent origins of eubacterial (hyper)thermophily.

*) The species abbreviations are the first letter of the family name and the first two letters of the species name, followed by the taxonomic identifier.

Conclusions

- **Noise in gene content phylogenies, which results from processes like horizontal transfer or parallel gene loss, does not cause a systematic bias in the tree topology.**
- **Filtering out this noise improves the topology of the tree.**
- **The only consistent signal in the orthologous groups is phylogenetic.**

Acknowledgements

This work was supported in part by contract QLTR-2000-01676 of the European Union, and by a grant from the Netherlands Organization for Scientific Research (NWO).

References

[1] Clarke GD, Beiko RG, Ragan MA and Charlebois RL (2002) Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and a distance matrix based on mean normalized BLASTP scores. J Bacteriol 184:2072-2080.

[2] von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P and Snel B (2003) STRING: a database of predicted functional associations between proteins. Nucleic Acids Res 31:258-261.

[3] Bruno W. J. (1996) Modeling residue usage in aligned protein sequences via maximum likelihood. Mol Biol Evol 13:1368-1374.

[4] Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND and Koonin EV (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res 29:22-28.