Letters

On the desirability of models for inferring genome phylogenies

James O. McInerney

Department of Biology, National University of Ireland, Maynooth, County Kildare, Ireland

Genomes are clearly suited for inferring common ancestry and for understanding ancestor–descendent relationships and interspecies gene transfer. Genomic evolutionary models can tell us a great deal about the processes that drive genome evolution, the mutational and selective pressures that lead to the genesis of biochemical pathways and operons, and the nature and extent of lateral gene transfer (LGT). Simultaneously, a robust phylogeny can be constructed that depicts the evolutionary relationships of the organisms in which the genomes are found.

Several approaches have been employed to infer species phylogenies at the genome level. In general terms, these can be divided into *ad hoc* summary statistics based on genome content, the use of concatenated alignments and the use of consensus methods (i.e. phylogenetic supertrees [1]).

The basic premise of methods based on summary statistics is that genomes are compared and a gene content matrix is compiled. Then, either a distance is estimated between all pairs of taxa and entered into a distance matrix that is summarized using a clustering algorithm, or a dendrogram is inferred using maximum parsimony. This is usually referred to as the species phylogeny. The principal difference between this approach and the approaches that use concatenated alignments or supertrees is that information concerning homolog interrelationships is not used. Presence or absence of homologs is the only information that is scored, and this approach can be considered *ad hoc* in the sense that the methods are applied uniformly to all datasets and, therefore, the assumptions are not informed by the data themselves.

Unsurprisingly, the results of using summary statistics have been variable. Although many methods have recovered groups that seem sensible and have support from external biochemical or morphological data, there have been cases in which the inferred trees are unusual [2].

For example, the haloarchaea are a group of halophilic Archaea, long taken to be members of the Euryarchaeota. Wolf *et al.* [2] and Korbel *et al.* [3] placed this taxon at the base of the Archaea. In the figures of Henz *et al.* [4], this taxon was placed among the Bacteria in one instance, within the Euryarchaeota in a second example and as the deepest-branching Archaeon in a third example. Dutilh *et al.* [5] point out that the correct placement of the haloarchaea is within the Euryarchaeota and that previous methods placed this taxon erroneously as a deep-branching Archaeon. This erroneous placement is likely to be due to the large number of bacterial genes present in the haloarchaea [6]. The haloarchaea are, therefore, pulled to a position that is intermediate between the two groups from which the haloarchaea genes came. The data violate the *ad hoc* assumptions of the methods. Problems of this nature argue for the development of explicit genome evolutionary models.

Evolutionary models are statements concerning how it is thought that evolution has occurred [7]. If a model were correct, the inferred distances between two genomes would be accurate and would provide consistent estimates of the topology of the resulting phylogeny. The most desirable properties of these models are explicitness when describing the evolutionary process, realism or plausibility of the assumptions contained in the models and clarity in the interpretation of the output [8]. Usually, models are derived in a maximum likelihood framework in which the model consists of the phylogenetic tree of the genomes and the process underlying their evolution [9]. However, even when alternative models are not tested or lengthy computational optimization is not performed, an explicit model of evolution can still be assumed in calculations [10].

A realistic model of genome evolution must, as a minimum, deal with gene duplication and loss, in addition to acquisition of genes by LGT. This is not to say that all parameters are necessary for all analyses. When models differ in their numbers of free parameters and are nested, a likelihood ratio test can be used to choose the most appropriate parameter.

Gu and Zhang [11] describe a model called the extended genome content distance. This model uses the number of homologs (0, 1 or >1) to derive the genome distance. The model does not take account of horizontal gene transfer and, as a result, the authors report a position for the haloarchaea that is the same as the much simpler method of Korbel *et al.* [3]. A model has also been developed that deals with LGT, albeit in a slightly different setting [12]. Nonetheless, the development of explicit model-based approaches is to be welcomed as a useful step towards the understanding of genome evolution.

When the genomic age began, it was assumed that the huge increase in the amount of available data would result in more-accurate phylogenies. Instead, the extent of apparent genome plasticity has fueled a passionate debate

Corresponding author: McInerney, J.O. (james.o.mcinerney@nuim.ie).

 0966-842X/$ - see front matter doi:10.1016/j.tim.2005.10.006

concerning prokaryotic evolution. Sensible genome models that provide information about phylogeny and the process of evolution should be a goal for genomics and systematics.

References

1 Creevey, C.J. and McInerney, J.O. (2005) Clann: investigating phylogenetic information through supertree analyses. *Bioinformatics* 21, 390–392

2 Wolf, Y.I. *et al*. (2001) Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol. Biol.* 1, 8

3 Korbel, J.O. *et al*. (2002) SHOT: a web server for the construction of genome phylogenies. *Trends Genet.* 18, 158–162

4 Henz, S.R. *et al*. (2005) Whole-genome prokaryotic phylogeny. *Bioinformatics* 21, 2329–2335

5 Dutilh, B.E. *et al*. (2004) The consistent phylogenetic signal in genome trees revealed by reducing the impact of noise. *J. Mol. Evol.* 58, 527–539

6 Kennedy, S.P. *et al*. (2001) Understanding the adaptation of *Halobacterium* species NRC-1 to its extreme environment through computational analysis of its genome sequence. *Genome Res.* 11, 1641–1650

7 Felsenstein, J. (1988) Phylogenies from molecular sequences: inference and reliability. *Annu. Rev. Genet.* 22, 521–565

8 Huelsenbeck, J.P. and Crandall, K.A. (1997) Phylogeny estimation and hypothesis testing using maximum likelihood. *Annu. Rev. Ecol. Syst.* 28, 437–466

9 Edwards, A.W.F. (1972) *Likelihood*, 2nd edn, Johns Hopkins University Press

10 Lake, J.A. and Rivera, M.C. (2004) Deriving the genomic tree of life in the presence of horizontal gene transfer: conditioned reconstruction. *Mol. Biol. Evol.* 21, 681–690

11 Gu, X. and Zhang, H. (2004) Genome phylogenetic analysis based on extended gene contents. *Mol. Biol. Evol.* 21, 1401–1408

12 Novozhilov, A.S. *et al*. (2005) Mathematical modeling of evolution of horizontally transferred genes. *Mol. Biol. Evol.* 22, 1721–1732