

Inferring deep duplication nodes in vertebrate protein families

Karin S. Kassahn¹
k.kassahn@imb.uq.edu.au

Mark A. Ragan¹
m.ragan@imb.uq.edu.au

¹ The University of Queensland, Institute for Molecular Bioscience and ARC Centre of Excellence in Bioinformatics, Brisbane, QLD 4072, Australia

Abstract

Amino acid site saturation due to long divergence times can complicate the inference of vertebrate protein family relationships. These families often contain anciently duplicated loci for which relationships are incorrectly recovered, producing so-called “outgroup topologies”. We investigated the effects of excluding potentially non-homologous and/or fast-evolving sites from the alignment, and inferred phylogenetic trees using Bayesian methods with or without topological constraints. We find that the most successful strategy to recover ancient duplication nodes is to contrast the results from constrained and unconstrained analyses using Bayes factors.

Key words: phylogenetics, amino acid site saturation, gene duplication, Bayes factor

1 Introduction

Teleost fishes experienced a whole-genome duplication shortly after their divergence from the tetrapod lineage some 450 MYA [1]. Identifying the correct duplication nodes for these anciently duplicated genes is critical to the analysis of how these genomes have evolved after duplication. A number of publicly available phylogenomic datasets provide protein trees for these families. However, phylogenetic inference of vertebrate protein families commonly suffers from artifacts associated with amino acid site saturation [2]. We previously performed genome-scale synteny and phylogenetic analyses to identify anciently duplicated fish genes and, in this process, identified protein families for which the Ensembl Compara v45 protein family trees show outgroup topologies (Fig. 1). Here, we selected five of these families and applied several approaches to test how phylogenetic tree inference for these families could be improved.

2 Methods and Results

We trialled the Gblocks software with default settings to eliminate poorly aligned sequence positions that may be saturated [3] and found that Gblocks excluded most sites from the alignments (Table 1). Visual inspection of the alignments revealed no obvious problems with alignment quality. We tested for compositional heterogeneity between sequences using Bowker’s and Stuart’s tests [4]. For most protein families, there was no evidence of compositional heterogeneity (Table 1). We then performed Bayesian phylogenetic analysis on these families using MrBayes [5], running one million generations and four chains for each analysis, setting a burn-in of 2500 generations, and sampling the tree space every 100 generations. We used model jumping between fixed-rate amino acid models to determine the most suitable substitution model. For each protein family we performed two independent analyses either enforcing topological constraints or not, and calculated the resulting Bayes factors comparing the relative posterior probabilities of the topologies. Topological constraints were used to impose the outgroup topology inferred in the Ensembl Compara analyses. In three of five cases, Bayesian phylogenetic inference showed higher posterior probabilities and significant Bayes factors for tree topologies describing two sister clades in fish, than for the outgroup topologies (Fig. 2 and Table 1). Excluding sites with rates in the top 10% generally resulted in a loss of information to resolve protein relationships and multifurcations in the protein family tree.

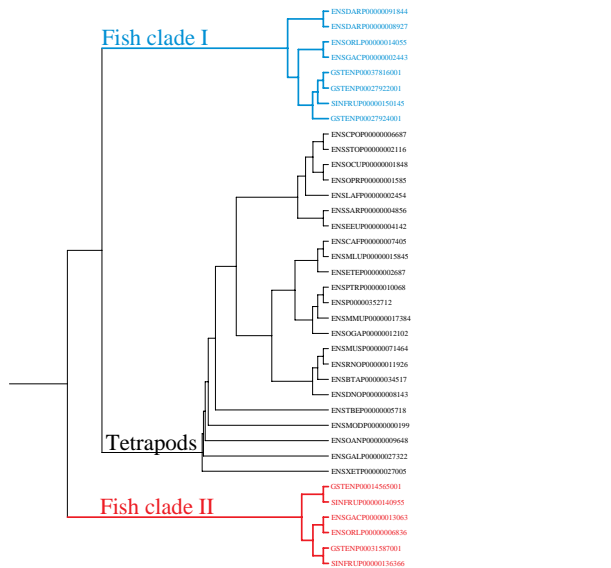


Figure 1: Outgroup topology. The blue and red fish clades are erroneously split into different clades.

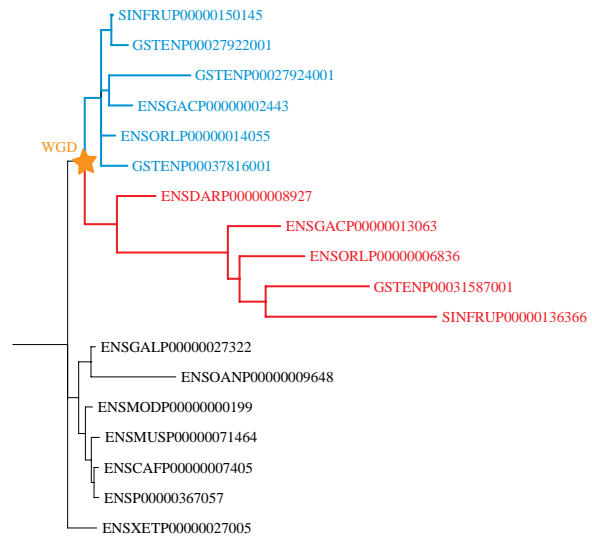


Figure 2: The correct sister clades in fish and the ancient duplication node (WGD) are recovered.

Table 1: Protein families for which phylogenetic inference could be improved. Marginal likelihoods are shown for Bayesian analyses. ‘ns’ not significant.

Family ID	Number of sequences	Number of alignment sites	Bowker's test	Gblocks (sites remaining)	Bayesian analysis		Bayes Factor
					Unconstrained	Constrained	
7681.1	43	493	ns	9	-7903.07	-7909.03	15.62
10760.2	37	1421	ns	0	-18323.9	-18339.8	31.82
9035.4	49	870	ns	0	-14291.6	-14308.4	43.18

3 Discussion

Our results show that it is possible to improve tree inference for protein families that contain anciently duplicated genes and suffer from outgroup topology artifacts. We find that Bayesian inference comparing unconstrained and constrained topologies is the most useful strategy to test the likelihood of different evolutionary histories. Exclusion of fast-evolving, and thus likely saturated, sites did not help resolve the protein family histories in the examined families. Alignment masking using the Gblocks software with default settings was too restrictive, eliminating most sites from the alignments. The development of less-restrictive masking algorithms that flag non-reliable alignment sites could help eliminate confounding noise in phylogenetic inference.

References

- [1.] Jaillon, O., Aury, J.M., Brunet, F., Petit, J.L., *et al.*, Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature*, 431(7011):946-957, 2004.
- [2.] Van de Peer, Y., Frickey, T., Taylor, J.S., Meyer, A., Dealing with saturation at the amino acid level: a case study based on anciently duplicated zebrafish genes. *Gene*, 295(2):205-211, 2002.
- [3.] Castresana, J., Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.*, 17(4):540-552, 2000.
- [4.] Ababneh, F., Jermini, L.S., Ma C., Robinson, J., Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. *Bioinformatics*, 22(10):1225-1231, 2006.
- [5.] Ronquist, F. and Huelsenbeck, J.P., MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12):1572-1574, 2003.