Gene Phylogenies and Protein-Protein Interactions: Possible Artifacts Resulting From Shared Protein Interaction Partners

Paulo R. A. Campos [a], Viviane M. de Olivera [a], Günter P. Wagner [b], and Peter F. Stadler [c,d]

[a] *Instituto de Física "Gleb Wataghin", Universidade Estadual de Campinas Cidade Universitária Zeferino Vaz, Barão Geraldo 13083-070, Campinas SP, Brazil.* prac@ifi.unicamp.br, vivimol@ifi.unicamp.br

[b] *Department of Ecology and Evolutionary Biology Yale University, New Haven, CT 06405-8106, USA* gunter.wagner@yale.edu

[c] *Lehrstuhl für Bioinformatik am Institut für Informatik und Interdisziplinäres Zentrum für Bioinformatik, Universität Leipzig, Kreuzstraße 7b, D-04103 Leipzig, Germany.* studla@bioinf.uni-leipzig.de

[d] *Institut für Theoretische Chemie und Molekulare Strukturbiologie, Universität Wien, Währingerstraße 17, A-1090 Wien, Austria*

Abstract

The study of gene families critically depends on the correct reconstruction of gene genealogies, as for instance in the case of transcription factor genes like *Hox* genes and *Dlx* gene families. Proteins belonging to the same family are likely to share some of the same protein interaction partners and may thus face a similar selective environment. This common selective environment can induce co-evolutionary pressures and thus can give rise to correlated rates and patterns of evolution among members of a gene family. In this study we simulate the evolution of a family of sequences which share a set of interaction partners. Depending on the amount of sequence dedicated to protein-protein interaction and the relative rate parameters of sequence evolution three outcomes are possible: if the fraction of the sequence dedicated to interaction with common co-factors is low and the time since divergence is small, the trees based on sequence information tend to be correct. If the time since gene duplication is long two possible outcomes are observed in our simulations. If the rate of evolution of the interaction partner is small compared to the rate of evolution of the focal protein family, the reconstructed trees tend towards star phylogenies. As the rate of evolution of the interaction partner approaches that of the focal protein family the reconstructed phylogenies tend to be incorrectly resolved. We conclude that the genealogies of gene families can be hard to estimate, in particular if the proteins interact with a conserved set of binding partners, as is likely the case for transcription factors.

Key words: Gene phylogeny, tree reconstruction, correlated substitutions

Manuscript *28 March 2004*

1 Introduction

Correlations in the accepted mutations at different positions of the same gene form the basis e.g. of the phylogenetic approach to RNA structure prediction [8, 7]. Similar approaches have been attempted on protein structures [20], albeit with less general applicability due to correlation chaining effects [10, 6]. Analogously, correlated mutations in different proteins contain information about protein-protein interaction [16, 15]. Correlated rates of evolution can therefore be employed to discover protein-protein interactions [21].

On the other hand, certain global correlations in mutation patterns are known to interfer with the performance of most phylogeny reconstruction methods. Convergence in nucleotide composition, for instance, may cause problems because unrelated lineages show similarities due to similar nucleotide compositions, not due to shared histories, see e.g. [1] for a detailed analysis. Codon usage biases are likely to have similar effects.

In this short contribution we consider a more intricate mechanism of introducing correlation into protein sequences of paralog genes that arises through interactions of paralog proteins with common binding partners. For example, the *Hox* genes from paralog groups PG1 through PG10 gain DNA binding affinity through cooperative binding with the protein *Pbx1* [11], while the so-called posterior *Hox* genes *HoxA9*, *HoxA10*, *HoxA11*, *HoxA13*, and *HoxD12* all interact with *Meis1* [19]. The work presented here was motivated by the analysis of the *Hox* genes of the Amphioxus (*Branchiostoma floridae*) [4]. *Hox* genes code for homeodomain containing transcription factors which are homologous to the genes in the Drosophila homeotic gene clusters [12]. Vertebrates, in contrast to all invertebrates examined, have multiple Hox gene clusters that have arisen from a single ancestral cluster in the most recent common ancestor of chordates, i.e. Amphioxus and vertebrates [5, 9]. In a gene tree of e.g. the human and the Amphioxus *Hox* genes we therefore expect that (1) the up to four human *Hox* genes of each of the 14 paralog groups cluster together, (2) these subtrees cluster together with the corresponding Amphioxus *Hox* genes, and (3) the "top" of the tree above these subtrees reflects the history by which the ancestral chordate *Hox* cluster came about through a history of duplications from a single "Ur-Hox" gene. While this pattern is presented by the anterior and (mostly) the middle group *Hox* genes, the posterior genes *Hox-9* through *Hox13* surprisingly show a different pattern, as described in [4], see Fig. 1. (The *Hox-14* paralog group which has recently been discovered in shark and latimeria [17] has been lost in mammals).

There are several possible explanations for that result: It is possible that the posterior *Hox* genes of Amphioxus are not ortholog with the 5' paralog groups in gnathostomes, although at least PG-9 and PG-10 have orthologs in echino-

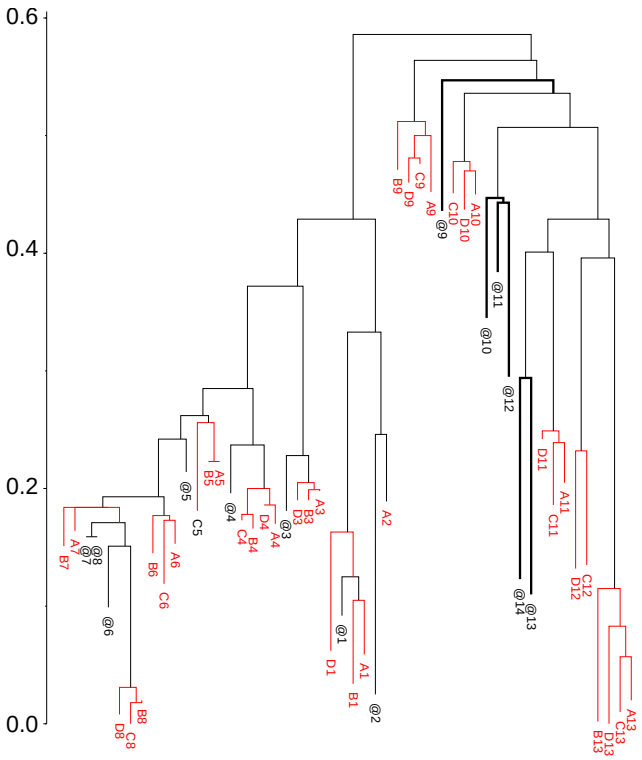

Fig. 1. Neighborjoining tree of the homeobox sequences of the *Hox* class transcriptions factors from Amphioxus (*Branchiostoma floridae*, marked by @) and for *Homo sapiens*. Gnathostomes including humans have four copies of the Hox cluster, labeled A through D. Each cluster contains a subset of the 14 distinct paralog genes that were presumably present in the primoridal chordate *Hox* cluster. On clearly sees the "anterior" *Hox* genes 10, 11, 12 as well 13 and 14 of the Amphioxus cluster together rather than with the gnathostome paralog groups as one would have expected.

derms [13]. Another possibility is that the rate of evolution of posterior genes is higher than that of anterior genes and thus obscures the phylogenetic signal, a hypothesis called *posterior flexibility* [4]. An alternative explanation is that the reconstructed gene tree in Fig. 1 does not represent the true evolutionary history of these genes. Strong biases in sequence composition or strong difference in codon usage between anterior and posterior genes do not seem to account for the distortion of the tree. We therefore have to search for other possible mechanisms that could generate artifacts in reconstructed phylogenetic trees.

We discuss here a simple mechanistic model that explains *correlated substitutions* in groups of paralog genes which, as we shall see, can lead to erroneously grouping together genes from within the same organism.

2 The Model

We consider a set of N paralog genes x^k, which for simplicity we model as 0/1 strings of length L, and a collection of N_c binding motifs of length L_c. Each gene x^k interacts with a randomly chosen fraction χ of n possible interaction partners. Each binding motif ξ is located at a fixed (randomly assigned) sequence position in sequence x^k provided that x^k interacts with the corre-

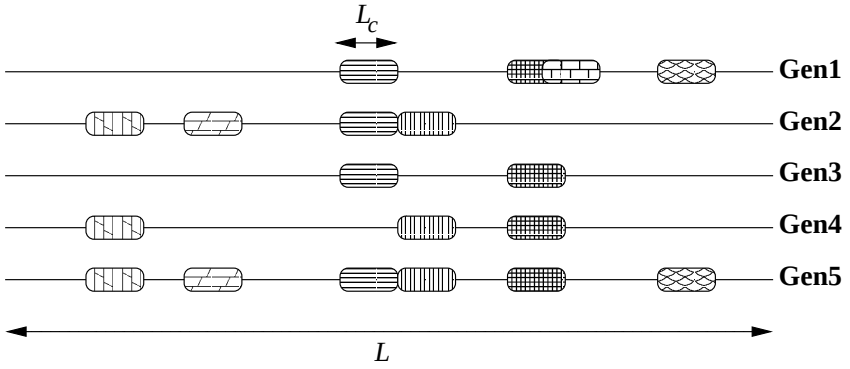

Fig. 2. Description of the model. We assume N paralog genes of length L along which N_c potential binding sites of length L_c are randomly placed. Each binding site is functional in a particular gene with a probability χ. In this case the sequence of the binding site must stay within a distance D of the target sequence necessary to bind the interaction partner. This binding sequence is subject to mutation with a rate p^* per site and generation, the paralog genes themselves evolve at a rate p.

sponding interaction partner. Let us denote the sequence of the binding site on gene x^k by $x^k[\xi]$. The "ideal" sequence of the binding site in ξ^*, Fig. 2. We assume that the Hamming distance $d(x^k[\xi], \xi^*)$ must not exceed a threshold D for any of the binding sites on gene x^k in a viable organism. This assumption in essence places a selection pressure on the gene x^k since $d(x^k[\xi], \xi^*) > D$ is lethal.

The evolution of the N paralog genes is modeled as a simple random walk whereby substitutions are independently at all sites with a probability p in each generation. At the same time the interaction partners, i.e., the ideal binding site sequences evolve with a different rate p^*.

A single speciation event is modeled by copying both the genes and interaction sites at time t_{div} and then using independent random numbers for simulating the mutations in both copies from this point on. A pair of genes that arose from the same ancestral gene in this duplication event will be referred to as *first-order paralogs* below.

We measure the pairwise Hamming distances of all genes in both organisms in regular intervals and construct a pair-distance matrix. The Neighbor-Joining algorithm [18] (implemented in the `phylip` package [3]) is used to reconstruct a common gene tree of both copies of the system. By construction we know which genes are orthologs. Immediately after the duplication the two "species" are in fact identical. For small divergence times we therefore obtain trees in which the ortholog pairs always group together. The parameter P_{cb} measures the fraction of "correct branches", i.e., the fraction of pairs of first-order paralog genes that correctly appear as neighbors in the reconstructed tree.

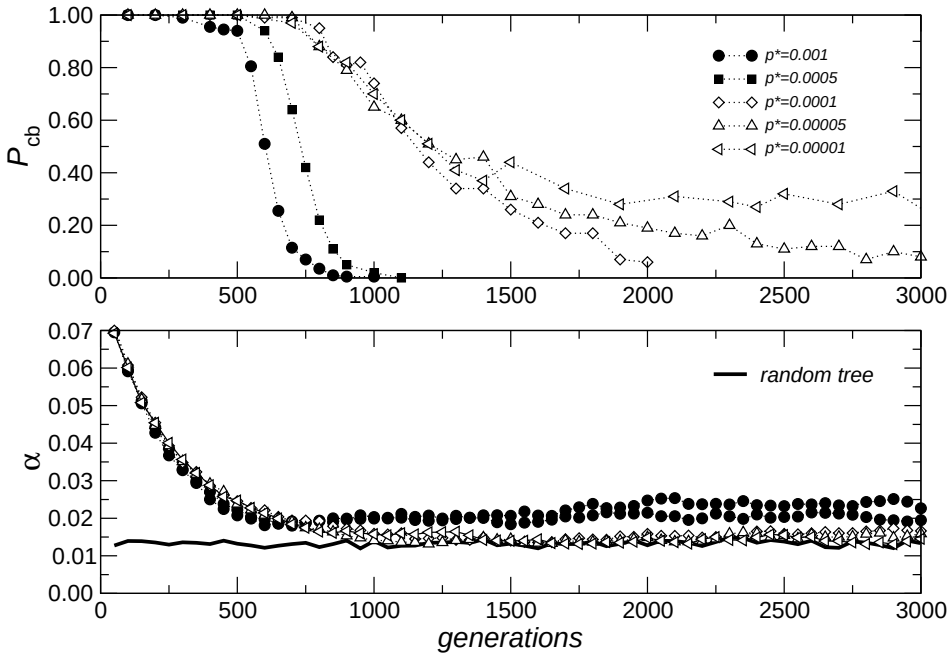

Fig. 3. Transition from trees that correctly identify first order paralogs to star-like trees (for small p^*, open symbols $\circ$) and to incorrect trees that deviate significantly from stars $*$ (for large values of p^*, filled symbols, $\bullet$). The fraction of $P_{\rm cb}$ of correctly branching trees averaged over 20 independent simulations with the same parameters is shown in the upper panel, the tree-likeness parameter α is displayed in the lower panel. Simulation parameters: $L = 2000$, $N = 20$, $L_c = 50$, $\chi = 0.5$ $D = 10$, $p = 0.001$, p^* varies.

Star-likeness of a tree is measured in terms of the parameters of statistical geometry [14]. We display α, the relative length of the longer of the two non-trivial splits averaged over all quadruples of sequences in the tree. This quantity decreases to a fixed value (depending on the sequence length) close to 0, see Fig. 3.

3 Simulation Results

Fig. 3 shows that correlated mutations may have a strong influence on the shape of the reconstructed tree. If p^* is small, the star-likeness parameter monotonically approaches the limit given by a tree reconstructed from random sequences while the fraction of correctly placed first-order paralogs slowly decreases.

For larger values of p^* there is a sharp transition from correct to incorrect branching while the trees stay significantly different from stars and random-trees as signified by an α-parameter that is about 50% larger than in the first

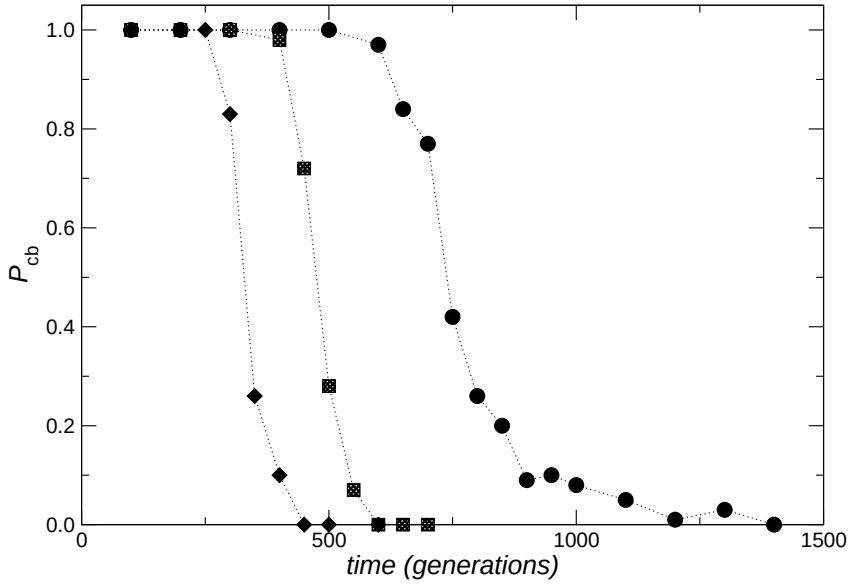

Fig. 4. Transition from correct to incorrect trees for three different numbers of interaction sites in each gene $\chi = 0.1$ (•), 0.5 (■) and 0.9 (⧫). The other simulation parameters are $L = 2000$, $N = 40$, $L_c = 50$, $p = p^* = 0.001$, $D = 10$.

case.

For values of $1 \lessapprox L_c N\chi$ almost all sequence positions are located within binding sites. As a consequence we observe only convergent evolution and all genes eventually have pairwise Hamming distances less than D. The time at which the transition to an incorrect tree occurs thus approaches 0 in the limit $p \to 1$.

4 Concluding Remarks

In this contribution we have introduced a simple model of correlations in sequence evolution that arise from the interactions of protein sequences with common binding partners. Our simulations show that reconstructed phylogenies can be affected significantly by this effect at least in cases of proteins that are part of a complicated regulation network involving multiple protein-protein interactions [2], such as transcription factors.

We emphasize that we *do not* claim here that the correlated evolution of amino acids explains the observed trees of *Hox* genes in Fig. 1. Rather, we have present here a feasible mechanism that *can* lead to artifacts in tree reconstructions that *could* be responsible for the observed trees. More detailed models, that most likely will require at least crude knowlege of the three-dimensional structure of the *Hox* proteins and their interaction partners as well as the locations of the binding sites on these structures, will be necessary

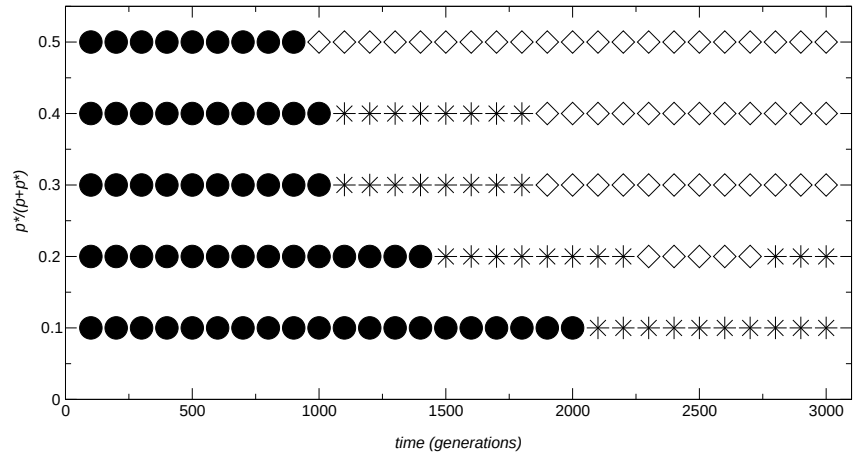

Fig. 5. Typical tree obtained at a given time for given strenght of the correlations measures as $p^*/(p+p^*)$ in 5 runs with $p = 0.001$. Symbols: • correct tree, ♢ incorrect tree that is not a start with at least some first-order paralogs not grouped together, $*$ denotes star-trees. The other simulation parameters are $L = 2000$, $N = 20$, $L_c = 50$, $D = 10$, $\chi = 0.5$.

to confirm or reject the hypothesis that the tree in Fig. 1 was indeed shaped by correlated substitutions and that these hypothetical correlations are indeed caused by a selection pressure to maintain common binding site patterns among co-regulated paralog genes. Another way to test whether a correlated pattern of substitutions may be responsible for unexpected gene trees is to test for signs of co-evolutionary *dynamics* among the suspect sequences [21]: If the lengths of branches on a constrained tree (which is based on other data and reflects the most likely species tree), are highly correlated among the members of a gene family, a co-evolutionary dynamics is likely, which may influence the gene tree reconstruction.

Acknowledgments

This work is supported by the Fundação de Amparo a Pesquisa do Estado de São Paulo (FAPESP) under project no. 03.00182-0 (PRAC, VMO), the National Science Foundation (NSF), grant INB-0321470 (GPW), and the Bioinformatics Initiative of the Deutsche Forschungsgemeinschaft (DFG), grant no. BIZ-6/1-2 (PFS).

References

[1] G. C. Conant and P. O. Lewis. Effects of nucleotide composition bias on the success of the parsimony criterion in phylogenetic inference. *Mol. Biol. Evol.*, 18:1024–1033, 2001.

[2] E. Davidson. *Genomic Regulatory Systems.* Academic Press, San Diego, 2001.

[3] J. Felsenstein. Phylip – phylogeny inference package (version 3.2). *Cladistics*, 5:164–166, 1989.

[4] D. E. K. Ferrier, C. Minguillón, P. W. H. Holland, and J. Garcia-Fernández. The amphioxus Hox cluster: deuterostome posterior flexibility and *Hox14*. *Evol. Dev.*, 2:284–293, 2000.

[5] J. Garcia-Fernández and P. W. Holland. Archetypal organization of the amphioxus hox gene cluster. *Nature*, 370:563–566, 1994.

[6] B. G. Giraud, A. S. Lapedes, and L. C. Liu. Analysis of correlations between sites in models of protein sequences. *Phys. Rev. E*, 58:6312–6322, 1998.

[7] R. R. Gutell, A. Power, G. Z. Hertz, E. J. Putz, and G. D. Stormo. Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *Nucl. Acids Res.*, 20:5785–5795, 1992.

[8] R. R. Gutell and C. R. Woese. Higher order structural elements in ribosomal RNAs: Pseudo-knots and the use of noncanonical pairs. *Proc. Natl. Acad. Sci. USA*, 87:663–667, 1990.

[9] C. Kappen, K. Schughart, and F. H. Ruddle. Two steps in the evolution of antennapedia-class vertebrate homeobox genes. *Proc. Natl. Acad. Sci. USA*, 86:5459–5463, 1989.

[10] A. S. Lapedes, B. G. Giraud, L. C. Liu, and G. D. Stormo. Correlated mutations in protein sequences: phylogenetic and structural effects. In *Proceedings of the AMS/SIAM Conference on statistic in Molecular Biology, Seattle, WA.*, 1997.

[11] R. S. Mann and S.-K. Chan. Extra specificity from *extradenticle*: the partnership between HOX and PBX/EXD homeodomain proteins. *Trends Genet.*, 12:258–262, 1996.

[12] W. McGinnis and R. Krumlauf. Homeobox genes and axial patterning. *Cell*, 68:283–302, 1992.

[13] V. B. Morris, J. Brammall, M. Byrne, and M. Frommer. cDNA *hox* sequences 3' of the homeobox isolated from the sea urchin *Holopneustes purpurescens* are definitive for sea urchin hox orthologues. *DNA Seq.*, 2002:185–193, 13.

[14] K. Nieselt-Struwe. Graphs in sequence spaces: a review of statistical geometry. *Biophysical Chemistry*, 66:111–131, 1997.

[15] Y. Ofran and B. Rost. Predicted protein-protein interaction sites from local sequence information. *FEBS Let.*, 544:236–239, 2003.

[16] F. Pazos, M. Helmer-Citterich, G. Ausiello, and A. Valencia. Correlated mutations contain information about protein-protein interaction. *J. Mol. Biol.*, 271:511–523, 1997.

[17] T. P. Powers and C. T. Amemiya. Evidence for a hox14 paralog group in vertebrates. *Current Biol.*, 14:R183–R184, 2004.

[18] N. Saitou and M. Nei. The neighbor-joining method: a new method for

reconstructing phylogenetic trees. *Mol Biol. Evol.*, 4:406–425, 1987.

[19] W. F. Shen, J. C. Montgomery, S. Rozenfeld, J. J. Moskow, H. J. Lawrence, A. M. Buchberg, and C. Largman. *AbdB*-like hox proteins stabilize DNA binding by the *Meis1* homeodomain proteins. *Mol. Cell. Biol.*, 17:6448–6458, 1997.

[20] I. N. Shindyalov, N. A. Kolchanov, and C. Sander. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Eng.*, 7:349–358, 1994.

[21] A. Valencia and F. Pazos. Computational methods for the prediction of protein interactions. *Curr. Op. Struct. Biol.*, 12:368–373, 2002.