Best Match Graphs:
Reconstruction of gene family phylogenies

Manuela Geiß[3], Edgar Chávez[1], Marcos González[1], Alitzel López[1], Bärbel M. R. Stadler[5], Dulce I. Valdivia[2], Marc Hellmuth[4], Maribel Hernández Rosales[1], and Peter F. Stadler[6]

[1] CONACYT-Instituto de Matemáticas, UNAM Juriquilla, Blvd. Juriquilla 3001, 76230 Juriquilla, Querétaro, QRO, México
maribel@im.unam.mx

[2] Universidad Autónoma de Aguascalientes, Centro de Ciencias Básicas, Av. Universidad 940, 20131 Aguascalientes, AGS, México; Instituto de Matemáticas, UNAM Juriquilla, Blvd. Juriquilla 3001, 76230 Juriquilla, Querétaro, QRO, México.

[3] Bioinformatics Group, Department of Computer Science; and Interdisciplinary Center of Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany
manuela@bioinf.uni-leipzig.de

[4] Institute of Mathematics and Computer Science, University of Greifswald, Walther-Rathenau-Straße 47, D-17487 Greifswald, Germany; Center for Bioinformatics, Saarland University, Building E 2.1, P.O. Box 151150, D-66041 Saarbrücken, Germany
mhellmuth@mailbox.org

[5] Max-Planck-Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig
baer@bioinf.uni-leipzig.de

[6] Bioinformatics Group, Department of Computer Science; Leipzig University, Härtelstraße 16-18, D-04107 Leipzig; and studla@bioinf.uni-leipzig.de

Abstract. A phylogenetic tree is a graph without cycles which does not contain internal vertices of degree two, and whose the leaves represent biological entities, such as genes. From the relationships between the genes in this tree, we build a digraph where nodes represent genes and we draw an arc from gene x to gene y if the species where x and y reside are different, and, moreover, y is a gene that is one of the closest evolutionary relatives of x, when compared to all others genes residing in the same species as y. We call this digraph a best match graph and we show here the properties of this type of graphs in order to be derived from a phylogenetic tree.

Keywords: Phylogenetic Combinatorics · Colored digraph · Reachable sets · Hierarchy · Hasse diagram · Rooted triples · Supertrees

1 Introduction

One of the most used starting points for inferring phylogenetic relationships among genes are Symmetric Best Matches[1], also known as Best Bidirectional Hits[2]. Their connection to the evolutionary history of a gene family relies in the concept of closest evolutionary relatedness. This concept is defined relative to the phylogenetic tree T of the genes under consideration.

1.1 Best Match Relations Properties

Let T be a phylogenetic gene tree with leaf set L. For each gene $x \in L$ we denote the species to which it belongs by $\sigma(x) \in S$, where S is a set of species. We can now define a best match relation as:

Definition 1. *The leaf y is a **best match** of the leaf x in T: $x \to y$ if and only if $lca(x,y) \preceq lca(x,y')$ for all leaves y' from the same species as y, i.e. $\sigma(y') = \sigma(y)$.*

Note that this definition can also be expressed in terms of divergence times as: y is a best match for x if and only if $t(x,y) \leq t(x,y')$ for all y' with $\sigma(y') = \sigma(y)$. For this reason, the *best match relation* serves as a generalization of *evolutionary closeness.*

Definition 2. *Given a tree and a colour map $\sigma : L \to S$, the **coloured Best Match Graph (cBMG)** $G(T,\sigma)$ is a digraph which has vertex set L and arcs $xy \in E(G)$ if $x \neq y$ and $x \to y$. Each vertex $x \in L$ has the colour $\sigma(x)$.*

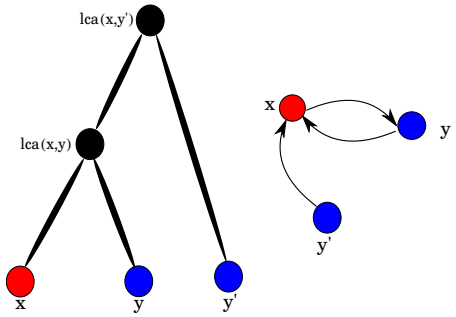

Fig. 1. A tree T with a colour map σ and its associated Best Match Graph $G(T,\sigma)$.

Definition 3. *Two vertices $x,y \in L$ are in relation $\overset{\bullet}{\sim}$ if $N(x) = N(y)$ and $N^-(x) = N^-(y)$. Where $N(x)$ is defined as the set of out-neighbours of x and $N^-(x)$ the set of the in-neighbours of x.*

The $\overset{\bullet}{\sim}$ relation gives us a partition on the set of vertices of the Best Match Graph and induces as well as hierarchy that helps us to infer a phylogenetic tree.

For the case of a Best Match Graph in two colors, we have the followin definition:

Definition 4. *For any connected 2-cBMG (G,σ) there exists a **Unique Least Resolved Tree** (T',σ) that explains (G,σ). (T',σ) is obtained by contraction of all redundant edges in an arbitrary tree (T,σ) that explains (G,σ) .*

The least resolved tree can be constructed in cubic time as shown in [3].

References

1. Tatusov, R.L. R.L., Koonin, E.V., Lipman, D.J.: A genomic perspective on protein families. Science 278, 631637 (1997).
2. Overbeek, R., Fonstein, M., DSouza, M., Pusch, G.D., Maltsev, N.: The use of gene clusters to infer functional coupling. Proc Natl Acad Sci USA 96, 28962901 (1999).https://doi.org/doi:10.1073/pnas.96.6.2896
3. Gei, M., Chávez, E., González, M., López, A., Stadler, B.M.R., Valdivia, D., Hellmuth, M., Hernández Rosales, M., Stadler, P.F.: Best match graphs. J. Math. Biol. (2018).