Characterization of colored Best Match Graphs

Manuela Geiß

Bioinformatics Group
University of Leipzig

TBI Winterseminar
Bled, 15th February 2018
Orthology analysis is an important part of data analysis in many areas such as comparative genomics and molecular phylogenetics.

Two fundamentally different ways of orthology estimation:
1. **Indirect** approach: Infer orthology relation from a gene-tree/species-tree pair
2. **Direct** approach: Estimate orthology relation directly from data
   → Best Match Heuristics
Assumption:
”The most closely related relative of a gene is the one that is most similar” (in terms of sequence distances)
→ Molecular clock hypothesis (Zuckerkandl and Pauling)
→ Often violated, still best match heuristics perform quite well on real data
Assumption:
"The most closely related relative of a gene is the one that is most similar" (in terms of sequence distances)
→ Molecular clock hypothesis (Zuckerkandl and Pauling)
→ Often violated, still best match heuristics perform quite well on real data

Software tools like ProteinOrtho give an approximate orthology graph

**Workflow:** Sequence data → ProteinOrtho → Cograph-editing → Orthology relation and representing tree
Assumption: 
"The most closely related relative of a gene is the one that is most similar" (in terms of sequence distances)
→ Molecular clock hypothesis (Zuckerkandl and Pauling)
→ Often violated, still best match heuristics perform quite well on real data

Software tools like Protein0rtho give an approximate orthology graph

**Workflow:** Sequence data → Proteinorth0 → Cograph-editing
→ Orthology relation and representing tree

Idea: Deeper understanding of Best Match Graphs to make the process more efficient
Evolutionary relatedness as phylogenetic property:

**Definition**

The leaf $y$ is a best match of the leaf $x$ in $T$ if $\text{lca}(x, y) \preceq \text{lca}(x, y')$ for all leaves $y'$ from species $\sigma(y') = \sigma(y)$. We write $x \rightarrow y$.

$\sigma = \text{colors (\text{\, = species})}$

$lca = \text{last common ancestor}$
Definition

Given a tree $T$ and a leaf-coloring $\sigma$, the colored best match graph $G(T, \sigma)$ has vertex set $L$ and arcs $xy \in E(G)$ if $x \neq y$ and $x \to y$. Each vertex $x \in L$ obtains the color $\sigma(x)$. The rooted tree $T$ explains the vertex-colored graph $(G, \sigma)$ if $(G, \sigma)$ is the cBMG obtained from $T$.

$\sigma =$ colors (= species)
Definition
Given a tree $T$ and a leaf-coloring $\sigma$, the colored best match graph $G(T, \sigma)$ has vertex set $L$ and arcs $xy \in E(G)$ if $x \neq y$ and $x \rightarrow y$. Each vertex $x \in L$ obtains the color $\sigma(x)$. The rooted tree $T$ explains the vertex-colored graph $(G, \sigma)$ if $(G, \sigma)$ is the cBMG obtained from $T$.

$\sigma = \text{colors (} = \text{species)}$

→ Which directed graphs are Best Match Graphs?
In a colored di-graph, we define:

OUT-Neighborhood ("out-going edges"): \( N(x) = \{ z \mid xz \in E(G) \} \)

IN-Neighborhood ("in-coming edges"): \( N^-(x) = \{ z \mid zx \in E(G) \} \)

Example:

\[
N(a) = N(b) = \{ y \} \\
N^-(a) = N^-(b) = \{ x, y \} \\
N(c) = \{ x, y \} \\
N^-(c) = \emptyset
\]
**Definition**

Two vertices $x, y \in L$ are in relation $\sim$ if $N(x) = N(y)$ and $N^-(x) = N^-(y)$.

\[ \alpha = \{a, b\}, \beta = \{c\}, \gamma = \{x\}, \delta = \{y\} \]

Observation: all vertices in a class are of the same color

Monotonicity: $N(\alpha) \subseteq N(\beta) \Rightarrow N(N(\alpha)) \subseteq N(N(\beta))$
**Assumption:** There is a tree that explains \((G, \sigma)\).
The case of two colors

**Assumption:** There is a tree that explains \((G, \sigma)\).

Some nice properties:

\[(N0) \quad \beta \subseteq N(\alpha) \text{ or } \beta \cap N(\alpha) = \emptyset\]

Idea of hierarchy: for any class, one collects everything that is "below" this class and this gives the tree \(H\).

Intuition: The reachable set of \(\alpha\) is \(R(\alpha) = \alpha \cup N(\alpha) \cup N(N(\alpha))\).

→ But when does such a tree exist for a 2-colored digraph?
**The case of two colors**

**Assumption:** There is a tree that explains \((G, \sigma)\).

Some nice properties:

1. \((N0)\) \(\beta \subseteq N(\alpha)\) or \(\beta \cap N(\alpha) = \emptyset\)
2. \((N2)\) \(N(N(N(\alpha)))) \subseteq N(\alpha)\)
The case of two colors

**Assumption:** There is a tree that explains \((G, \sigma)\).

Some nice properties:

\((N0)\) \(\beta \subseteq N(\alpha)\) or \(\beta \cap N(\alpha) = \emptyset\)

\((N2)\) \(N(N(N(\alpha)))) \subseteq N(\alpha)\)

**Idea of hierarchy:** for any class, one collects everything that is ”below” this class and this gives the tree \((\rightarrow \text{Hierarchy } \mathcal{H})\)

Intuition: The *reachable set* of \(\alpha\) is

\[ R(\alpha) = \alpha \cup N(\alpha) \cup N(N(\alpha)) \]
**The case of two colors**

**Assumption:** There is a tree that explains \((G, \sigma)\).

Some nice properties:

\((N0)\) \(\beta \subseteq N(\alpha) \) or \(\beta \cap N(\alpha) = \emptyset\)

\((N2)\) \(N(N(N(\alpha)))) \subseteq N(\alpha)\)

**Idea of hierarchy:** for any class, one collects everything that is ”below” this class and this gives the tree (→ Hierarchy \(\mathcal{H}\))

Intuition: The *reachable set* of \(\alpha\) is

\[ R(\alpha) = \alpha \cup N(\alpha) \cup N(N(\alpha)) \]

→ But when does such a tree exist for a 2-colored digraph?

Manuela Geiß

Characterization of colored Best Match Graphs
Augenkrätze-Theorem

Let \((G, \sigma)\) be a 2-colored digraph. Then there exists a tree \(T\) explaining \(G\) if and only if \(G\) satisfies properties (N0), (N1), (N2), and (N3).

\(\beta \subseteq N(\alpha) \) or \(\beta \cap N(\alpha) = \emptyset\)

\(\alpha \cap N(\beta) = \beta \cap N(\alpha) = \emptyset\) implies \(N(\alpha) \cap N(N(\beta)) = N(\beta) \cap N(N(\alpha)) = \emptyset\).

\(N(N(N(\alpha))) \subseteq N(\alpha)\)

If \(\alpha \neq \beta\) with \(\alpha \cap N(N(\beta)) = \beta \cap N(N(\alpha)) = \emptyset\), then \(N(\alpha) \cap N(\beta) \neq \emptyset\) if and only if \(N(\alpha) \subseteq N(\beta)\) or \(N(\beta) \subseteq N(\alpha)\), and \(N(\alpha) - \alpha = N(\beta) - \beta\).

Before we extend these results to \(n\) colors, we need a little recap:
Augenkrätze-Theorem

Let \((G, \sigma)\) be a 2-colored digraph. Then there exists a tree \(T\) explaining \(G\) if and only if \(G\) satisfies properties (N0), (N1), (N2), and (N3).

(N0) \(\beta \subseteq N(\alpha)\) or \(\beta \cap N(\alpha) = \emptyset\)

(N1) \(\alpha \cap N(\beta) = \beta \cap N(\alpha) = \emptyset\) implies \(N(\alpha) \cap N(N(\beta)) = N(\beta) \cap N(N(\alpha)) = \emptyset\).

(N2) \(N(N(N(\alpha)))) \subseteq N(\alpha)\)

(N3) If \(\alpha \neq \beta\) with \(\alpha \cap N(N(\beta)) = \beta \cap N(N(\alpha)) = \emptyset\), then \(N(\alpha) \cap N(\beta) \neq \emptyset\) if and only if \(N(\alpha) \subseteq N(\beta)\) or \(N(\beta) \subseteq N(\alpha)\), and \(N^-(\alpha) = N^-(\beta)\).
Let \((G, \sigma)\) be a 2-colored digraph. Then there exists a tree \(T\) explaining \(G\) if and only if \(G\) satisfies properties (N0), (N1), (N2), and (N3).

(N0) \(\beta \subseteq N(\alpha)\) or \(\beta \cap N(\alpha) = \emptyset\)

(N1) \(\alpha \cap N(\beta) = \beta \cap N(\alpha) = \emptyset\) implies 
\[N(\alpha) \cap N(N(\beta)) = N(\beta) \cap N(N(\alpha)) = \emptyset.\]

(N2) \(N(N(N(\alpha)))) \subseteq N(\alpha)\)

(N3) If \(\alpha \neq \beta\) with \(\alpha \cap N(N(\beta)) = \beta \cap N(N(\alpha)) = \emptyset\), then 
\[N(\alpha) \cap N(\beta) \neq \emptyset\] if and only if \(N(\alpha) \subseteq N(\beta)\) or 
\(N(\beta) \subseteq N(\alpha)\), and \(N^{-}(\alpha) = N^{-}(\beta)\).

→ Before we extend these results to \(n\) colors, we need a little recap:
Some basics: Rooted Trees and Triples

Rooted Tree $T$:

- An acyclic, connected graph

Triples:
- $T$ displays a triple $ab|c$ if the path from $c$ to the root is not intersected by the path from $a$ to $b$.
- $\mathcal{R}(T) = \{ab|c, ab|d, ab|e\}$
Some basics: Rooted Trees and Triples

Rooted Tree $T$:

- $T$ displays a triple $ab|c$ if the path from $c$ to the root is not intersected by the path from $a$ to $b$.
- $\mathcal{R}(T) = \{ab|c, ab|d, ab|e\}$
- A set of triples $R$ is said to be consistent if there is a tree $T$ with $R \subseteq \mathcal{R}(T)$.
- Consistency-check via BUILD-algorithm in polynomial time. In case of consistency, it returns a tree $T$ with $R \subseteq \mathcal{R}(T)$. 

Triples:
Generalization to $n$ colors

All information that is needed, is contained in the 2-cBMG’s:

**Theorem**

*A colored digraph $(G, \sigma)$ is a $n$-cBMG if and only if all induced subgraphs on two colors are 2-cBMG’s and the union of the triples obtained from their least resolved trees forms a consistent set.*
Generalization to $n$ colors

All information that is needed, is contained in the 2-cBMG’s:

**Theorem**

A colored digraph $(G, \sigma)$ is a $n$-cBMG if and only if all induced subgraphs on two colors are 2-cBMG’s and the union of the triples obtained from their least resolved trees forms a consistent set.

- a) Evolutionary scenario
- b) Induced subgraphs on two colors and least resolved trees.
- c) Least resolved tree for $G$

Manuela Geiß
Characterization of colored Best Match Graphs
Algorithm for the tree-reconstruction of a \( n \)-cBMG

For every induced subgraph on two colors: check (N0)-(N3)  
→ if positive:
  - build the least-resolved tree using the hierarchy \( \mathcal{H} \)
  - collect all triples from this tree

→ Use the set of all triples as input for BUILD: consistency check and tree construction

→ The resulting tree is the least-resolved tree that explains the given graph
Summary & Outlook

What we did so far:

- Characterization of two-colored Best Match Graphs by properties (N0)-(N3) and extension to $n$ colors
- Algorithm for the tree reconstruction of colored BMGs

Next steps:

- What about reciprocal $n$-cBMG’s?
- What can we say about Cographs?
- Optimization of data analysis in the context of Proteinortho
Special Thanks to:
Peter F. Stadler
Marc Hellmuth
Edgar Chávez
Marcos González
Maribel Hernández Rosales
Alitzel López
Dulce Valdivia

Thank you for your attention!
\[ R(\alpha) = N(\alpha) \cup N(N(\alpha)) \]
\[ Q(\alpha) = \{ \beta \mid N^-(\beta) = N^-(\alpha) \text{ and } N(\beta) \subseteq N(\alpha) \} \]
\[ R'(\alpha) = R(\alpha) \cup Q(\alpha) \]
\[ \mathcal{H} := \{ R'(\alpha) \mid \alpha \in \mathcal{N} \} \]
Appendix

a) b) c)

Characterization of colored Best Match Graphs