Orthology Analysis

part of “Graphen und Netzwerke in der Biologie”

Sonja Prohaska

Computational EvoDevo
University Leipzig

Leipzig, SS 2011

Deriving orthology information from multiple genomes

- **given**: gene inventory of multiple genomes
- **to be found**: sets of orthologous genes
- **approach**: pairwise reciprocal best alignment heuristic
- **idea**: compute alignments between genes of different genomes, construct a graph $\vec{\Upsilon}$ with genes as nodes, alignments as edges and edge weights $\omega_{x \to y}$ holding the bit score (similarity measure).

Translating the task into a graph theoretic problem

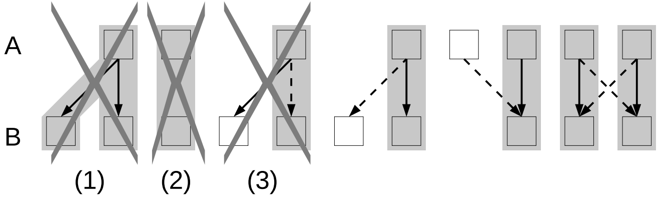
Orthologous sets detection $=$ finding *nearly* disjoint maximal *nearly*-complete multipartite subgraphs in the edge-weigthed directed graph $\vec{\Upsilon}$.

multipartite graph ... a graph whose vertices can be divided into n disjoint sets such that every edge connects a vertex from A to a vertex from B for all pairs of sets, while A and B is such a pair of disjoint sets and $B \neq A$ (n is the number of genomes, each U is the set of genes of one species).

complete directed graph ... directed graph in which every pair of distinct vertices is connected by two edge, one in either direction.

idealized dataset

- each protein x from species A has at most one ortholog in any other species $B \neq A$
- if $y \in B$ is an ortholog of $x \in A$,
  - then a search of x against B yields at least one alignment
  - and the unique best alignment of query x against B is the true ortholog y of x



grey shadows: true orthology relations
solid arrows: refer to the best alignment
dotted arrows: refer to alignments other than the best one
cases (1)-(3) cannot occure by definition of an idealized dataset

reciprocal pairwise best alignment heuristic

construct a subgraph $\vec{\Upsilon}_{RBAH}$ of $\vec{\Upsilon}$ such that for each protein x in species A and a given species $B \neq A$ only the edge with maximal weight is retained:

$$(x \to y) \in \vec{\Upsilon}_{RBAH} \text{ iff } \omega_{x \to y} = \max_{y' \in B} \omega_{x \to y'} \tag{1}$$

The symmetric subgraph of $\vec{\Upsilon}_{RBAH}$, containing only reciprocal best alignments, can be regarded as an undirected graph Υ_{RBAH}.

A directed graph D is symmetric

iff, for every directed edge $(x \to y)$ in D the corresponding inverted edge $(y \to x)$ also belongs to the graph D.

Properties of Υ_{RBAH} and $\vec{\Upsilon}_{RBAH}$

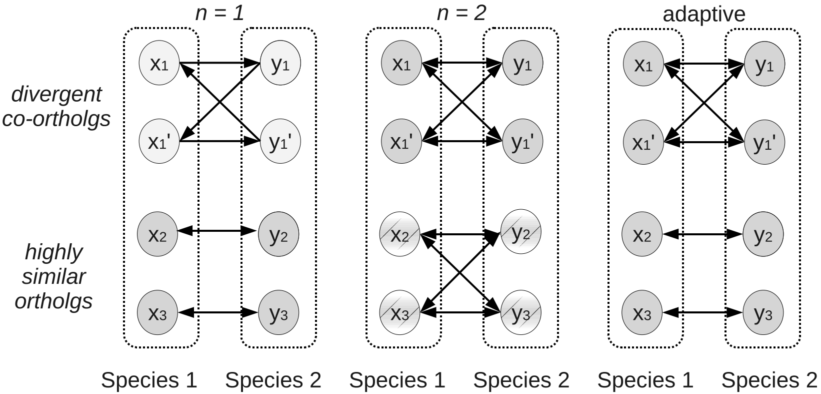
What did we achieve so far?

- any two vertices are connected by edges in Υ_{RBAH} if and only if they are orthologs
- a set of orthologs is a complete multipartite subgraph of $\vec{\Upsilon}_{RBAH}$ in which every species is represented at most once.
- sets of orthologs correspond to the connected components of $\vec{\Upsilon}_{RBAH}$

Holds for ideal data sets only!

Handling of co-orthologs

- one-to-many and many-to-many orthology relations could be handled iff they all scored maximal
- they will show slightly different scores in real data



n ... number of best alignments to include per protein (number of expected co-orthologs)

Adaptive selection of the number of best alignments to include

introduce a cut-off value that depends on the quality of the matches, namely the maximal value multiplied with a factor $f < 1$ (generally close to 1).

$$(x \to y) \in \vec{\Upsilon}^{*} \text{ iff } \omega_{x\to y} \geq f \max_{y' \in B} \omega_{x \to y'} \tag{2}$$

This increases the number of edges among co-orthologs/in-paralogs and reduces spurious edges. (Some false positive and false negative edges remain.)