**Aim 3: Network-based Phylogeny**

Traditionally, phylogenomic analysis has been anchored in sequence data, usually sequences of nucleotides. In many natural phenomena however, interactions are as important as the hardware. For example, birds and fish flock in similar patterns even though their genetic makeup is quite different. This has led to the notion of identifying species based on their networks and building trees based on network metrics instead of sequence metrics. We call the result a PhyloNetomic tree. This aim proposes to build a visualization and analytical tool for the construction of PhyloNetomic trees that have measured or inferred network data. This will enable users toIDENTIFY AND LOOK FOR NETWORK MOTIFS THAT ARE ANCESTRAL, VS DERIVED, COMPARE THE NETWORK VIEW TO THE TOPOLOGY OF A SEQUENCE ONLY TREE. WE MAKE THE TREE A RESOURCE AND UPDATE IT AS NEW FULLY SEQUENCED SPECIES COME ON LINE. WE ALSO ENABLE RESEARCHERS TO SELECT A SUBSET OR TO MAKE THEIR OWN TREE.

The basic tool for the construction of a tree will be a global network alignment program. The current best candidates are MI-Graal [ref 3 below], IsoRank [Rohit Singh, Jinbo Xu, and Bonnie Berger. (2008) Global alignment of multiple protein interaction networks with application to functional orthology detection, *Proc. Natl. Acad. Sci. USA*, 105:12763-12768] and IsoRankN [Ref: Chung-Shou Liao, Kanghao Lu, Michael Baym, Rohit Singh, and Bonnie Berger. (2009) IsoRankN: Spectral methods for global alignment of multiple protein networks, *Bioinformatics*, 25:i253-i258.]. These programs construct a tree based on multiway network alignment. We will provide a visual query platform on that tree.

Reference 1 http://rsif.royalsocietypublishing.org/content/7/50/1341.full compares different species based on the network topology of protein-protein graphs, but doesn't require nodes that map to one another to be sequence orthologous. We don’t think this is so relevant, because orthology is closely related to function.

Reference 2 http://bioinformatics.oxfordjournals.org/content/23/13/1631.full uses both sequence and topological similarity in an integer programming framework to try to find the best match. This could be used for pairwise alignment, but we are interested in global alignment.

Reference 3 http://bioinformatics.oxfordjournals.org/content/27/10/1390.full MI-Graal achieves a global network alignment enabling it to construct a phylogenetic tree from protein-protein interactions.

IsoRank and IsoRankN also do a multiple network alignment, but based on spectral methods.

Our tool will permit users to generate new trees based on user-specified experimental conditions, sets of genes, orthology, and p-value thresholds. It will also allow users to include inferred as well as experiment-driven edges. *Gloria: it makes more sense for users to be able to specify this. Aim 1 will apply once they have chosen the thresholds, but the thresholds may depend on the user’s view of the coverage/quality tradeoff.*

Dennis- I DON’T LIKE THE IDEA OF LETTING THE USER SET THE PARAMETERS. I THINK WE SHOULD BE PROPOSING EXPERIMENTS TO DETERMINE WHAT ARE THE BEST SETTINGS OF PARAMETERS BASED ON SOME VALIDATION TESTING. CANT WE SAY THAT WE WILL USE THE CRITERIA WE DETERMINE IN AIM 1 (E.G. ORTHOLOGY METHOD AND CORRELATION ANALYSIS) TO GIVE THE BEST VALIDATION AS DETERMINED IN AIM 1

The PhyloNetomic tree will then be constructed based on these user decisions [will need an example figure]. The tree itself will be queryable as follows:

1. At every node in the PhyloNetomic tree, it will be possible to find the network elements (for any subset of network data types) that are common to all (or a certain user-specified fraction) of the species in the clade governed by that node

2. The user may click on two nodes in the tree to find common network edges as well as the symmetric difference (edges that one has but the other doesn't).

3. A set of edges will be convertible to a set of genes for purpose of GO analysis or other such purpose.

4. A set of edges will be displayable AS A NETWORK MODULE using Cytoscape.

Such a tool will permit a network-based analysis of a set of species. For example, it will be possible to answer questions over all (currently 21) sequenced plant species such as: WE SHOULD PROPOSE TO ANSWER THESE QUESTIONS USING THE 21 SPECIES TREE.

1. How does the PhyloNetomic tree differ from the Phylogenomic one for one or more clades?

2. In which GO terms do two species (or two clades) differ most strongly in their expression correlation? In which GO terms are they most similar?

3. Which network motifs are ANCESTRAL, WHICH ONES ARE found in some clade?