Implementation of VirtualPlant's Comparative Genomics Analysis

Arthur Goldberg, Dennis Shasha, Manpreet Katari

Summary

The lab's NSF Grant *Conceptual Data Integration for the Virtual Plant, Request for 2 Year Supplement*, subsection *AIM 2. COMPARATIVE GENOMICS ANALYSIS: Adaptation of VirtualPlant to encompass other fully sequenced species* identifies 3 tools that “we will develop for comparative analysis”:

- A visualization tool that will allow us to study and compare gene networks between species.
- A data analysis tool that looks at the behavior of orthologous genes in equivalent experiments in two species.
- A visualization tool showing conserved cis-binding sites.

This implementation plan presents a software design for the first two tools. We present the design from bottom to top, starting with generic functionality for managing orthologs, next presenting general purpose tools for analyzing sets of orthologs and experiments that span multiple species, and then concluding with some possible applications for biologists to use.

Orthologous Gene Mapping

Both the Comparative Gene Networks and Comparative Equivalent Experiments tools require that orthologous genes be mapped between species of interest. Fortunately, our users only study a small number of species – Arabidopsis thaliana, Rice, Grape, Maize and perhaps Soybean – so we can cache the mappings in a database.

An underlying layer uses the database to map between orthologs. Given the locus tag of a gene Gs in species S and a target species T the mapping can return the gene Gt in T that is orthologous to Gs, if an ortholog exists. More generally, given a gene Gs in species S the mapping can return the all genes orthologous to Gs, which will be a set of the form Gt, Gv, Gw, … for each species (t, v and w in this example) that has an ortholog to Gs.

We assume that orthology is a symmetric and transitive relation. Thus, if Gs is orthologous to Gt, then Gt must be orthologous to Gs. Also, in the previous example Gs, Gt, Gv, and Gw are all orthologous to each other. (We also initially assume that orthology between a pair of species is one-to-one, and not one-to-many. That is, any Gs in species S has at most one ortholog in species T. A one-to-many mapping would greatly complicate some analyses and data presentations.)

This orthology mapping will utilize popular existing ortholog determination methods, such as BLAST, OrthologID, INPARANOID, and/or COG. See below for details on how we will retrieve data from these sources.

The following database schema design will efficiently represent an orthologous gene mapping that involves multiple species:

1. The orthology mappings will be stored in a new VirtualPlant database called XSPECIES_DB, stored on our MySQL server. XSPECIES_DB will use a schema similar to those in our existing

databases, containing the tables OBJECT, OBJECT_CONNECTION and OBJECT_ATTRIBUTE.

2. Records will have the following content and meanings:

   OBJECT(oid, class, value) = (OID, 'Ortholog', '') will indicate the existence of an ortholog with identifier OID,

   some records in OBJECT_CONNECTION(ocid, oid1, relationship, oid2, db2) of the form (OCID, oid, 'Ortholog2Member', geneid, species), will indicate that the gene identified by 'geneid' (an OID) in the species identified by 'species' (which will identify the database storing information about the species) is a member of the ortholog identified by oid,

   some records in OBJECT_ATTRIBUTE(oaid, oid, type, value) of the form (OAID, oid, <attribute>, <value>), will mean that the ortholog set identified by oid has an attribute with the given value, and

   some CONNECTION_ATTRIBUTE records will describe attributes of the OBJECT_CONNECTION records.

3. The attributes for an Ortholog OBJECT_CONNECTION should include

   Source: the name of the internet source for the orthology mapping

   Date: the date that the mapping was downloaded from the internet

   Obsolete: a boolean that indicates that the entry in the OBJECT_CONNECTION has been updated by a latter mapping from the internet source

4. To speed up the mapping, the database indexes the OID, OCID, OID2 and TYPE columns.

Canned, parameterized queries will retrieve orthology mappings from the database.

A configuration module identifies internet data sources that store orthology mappings among species of interest. It will search for modifications to the sources, probably via periodical probes. When it detects a modification it will notify an administrator to download the new orthology mappings to the local database. The download will mark existing records as Obsolete.

Data Sources

We will cache the orthology information of one or more existing orthology determination mechanisms. We discuss the primary mechanisms here.

BLAST

We will start by using orthologs defined by reverse top BLAST hits.

OrthologID

On the Web OrthologID (http://nypg.bio.nyu.edu/orthologid/search.html) only provides a single gene locus lookup, which returns a phylogenetic tree that contains genes from Arabidopsis thaliana, Oryza sativa, Populus trichocarpa, Chlamydomonas reinhardtii (an outgroup) and possibly some other species. However, it doesn't

1. provide any interface other than a single gene query
2. report on the strength of an ortholog match
3. provide any meta-information about a search, other than what is provided in the article

However, Ernie says "... if all you need is the set of all At-Os orthologs, it should be easy for me to extract them from the existing gene family trees."

Clusters of Orthologous Groups of proteins (COGs)

COG is on the Web at http://www.ncbi.nlm.nih.gov/COG/. It provides Eukaryotic clusters among Arabidopsis thaliana and half a dozen other eukaryotes, but none of them are plants. It lists Rice as an 'Upcoming eukaryotic genome'. So COGs doesn't appear useful for our purposes.

Other sources

We will also investigate two other potential sources for orthologs:

GreenPhylDB: A phylogenomic database for plant comparative genomics

Inparanoid: a comprehensive database of eukaryotic orthologs

Issue

One tricky complication is that orthology determination mechanisms employ some thresholds that determine orthology (e.g., a match may only have x misalignments). We might wish to adjust the threshold for some of our analyses. However, that won't be possible with orthologs cached in a DBMS, unless the DBMS stores multiple orthologies, each described by the threshold employed to calculate it. Another, more dynamic, approach to adjusting the orthology thresholds would involve replacing a DBMS cache of orthologs with direct access to an orthology determination mechanism. We will not implement this functionality until later versions of our comparative genomics analysis.

Underlying Analyses

Underlying analyses will use the orthology mappings to 1) compare gene networks and 2) compare equivalent experiments. We want to define analyses that will be generally useful to biologists. Although the DBMS stores information about multiple species, initially we assume that an analysis compares only two species.

Gene Sets and Networks

Suppose we have gene network 1 (or gene set 1) for species 1 and gene network 2 (or gene set 2) for species 2.

Analysis steps could include:

1. Orthology mappings

   A. Ortholog ID: Given a gene network or set, map all of the genes that are members of an ortholog group to their orthlogs IDs. This can be done to a gene set or the genes in a network. (A gene network is just a set of triples {gene1, gene2, type of edge}.) As mentioned above, the underlying mapping is one-to-one which ensures that each gene maps to at most one ortholog.

B. Target species: Given some genes and a species S, map the genes to their orthologs in S, when they exist. This will be used, for example, to prepare for an analysis of genes in species S, such as displaying some gene sets, the GO terms that annotate them, and the GO hierarchy in Sungear. We call S the *reference species*.

2. Set operations: these operate on a pair of gene sets (GS1 and GS2) or a pair of gene networks (GN1 and GN2) from two species (1 and 2).

   A. Intersection: Find the set of genes that are orthologs between the two species. For example, if GS1 = { A1, A2, A3 } and GS2 = { R1, R2, R5 } and the pairs of orthologs are (A1, R1) and (A2, R2) then $GS1 \cap GS2 = (A1, R1), (A2, R2)$. When operating on networks, a further refinement would filter the intersection to contain only certain types of edges – for example, the intersection might contain only edges that had the same edge type in each input gene network. Such filters can be arbitrary functions on a pair of edges that output a boolean.

   B. Difference: Compute the gene set (GS1 - GS2) that contains the genes in GS1 that do not have orthologs in GS2, or vice versa. Similarly, compute the gene network (GN1 - GN2) that contains triples for which at least one gene does not have an ortholog in GN2, or vice versa. The latter could also be refined by an edge matching rule.

   C. Union: Calculate the union of two gene sets or networks, mapping orthologous genes into their orthologs. For consistency, the union of a pair of gene sets or networks must be equivalent to the union of their intersection and two differences. That is, $GS1 \cup GS2 = (GS1 \cap GS2) \cup (GS1 - GS2) \cup (GS2 - GS1)$.

3. Statistics: these operations provide statistics on the gene sets or networks.

   A. Existence: Given a pair of gene sets (GS1 and GS2) from two species, what's the probability that the observed number of orthologs between the sets (the cardinality of $GS1 \cap GS2$) could occur by chance? The probability can be calculated by a *monte carlo* experiment that repeatedly picks a random set of genes from each species – gene sets GStest1 and GStest2 – and calculates the size of their ortholog intersection. The random sets should be constrained by any prior constraint that applied to the experiments that generated GS1 and GS2.

B. Connectivity: Given a pair of gene networks, what's the probability that the observed number of edges between orthologs could happen by random chance? (The experiment would be to calculate the p value as follows: take the n edges that have orthologs on both ends and permute one column of those edges; then compute the intersection of the resulting edges with the edges from net2. That would give a p value of the number of edges in common between net1 and net2 compared to chance. <Dennis, I left this as you wrote it, as I don't understand it.>) As above, a refinement can filter edges by their characteristics.

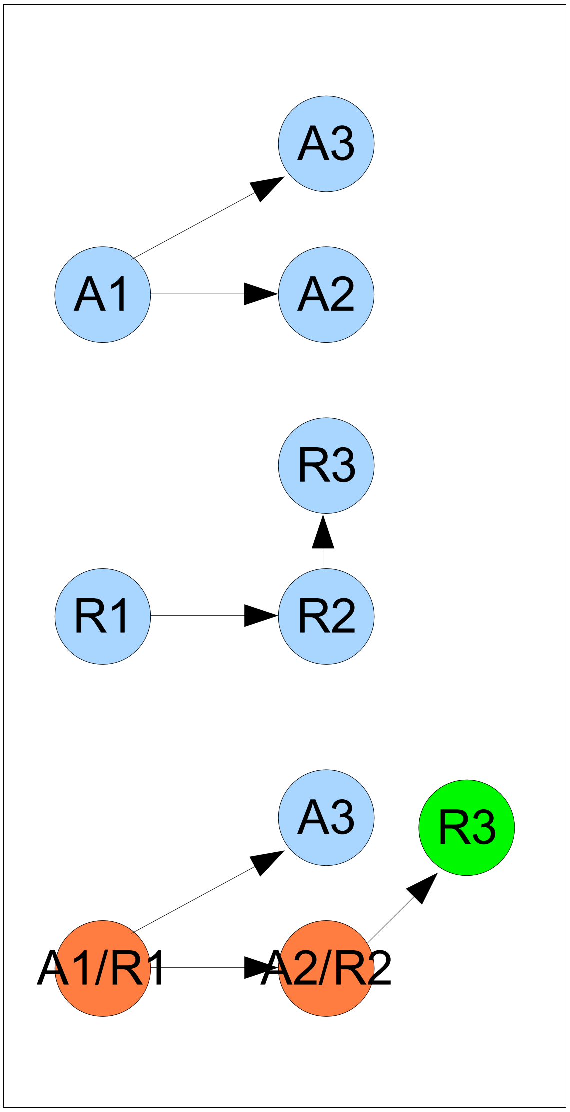
Perl Implementation

Following VirtualPlant's existing design, orthologs will be represented by a Perl object called Ortholog, and the operations above will input GeneSets and output OrthologSets or GeneSets.

User Applications

This section proposes user applications to be built upon the underlying analyses described above.

Comparative Gene Networks



A comparative gene network would provide an interactive graph displaying a pair of gene networks (GN1 and GN2) from two species. Assume that the sets $GS1 \cap GS2$, $GS2 - GS1$ and $GS1 - GS2$ are all non-null. Then nodes in each of these calculated sets would display differently, perhaps like the figure at right, which assumes the gene set examples above.

We've combined orthologous nodes (in orange) and edges of the same type.

Otherwise, edges are retained. That is, all edges in the two individual gene networks are included in the comparative gene network, with edges that end at nodes representing orthologous genes connected to those nodes.

Comparative Equivalent Experiments

We want to compare the gene expression levels of equivalent experiments. For example, consider the following set of 10 experiments:

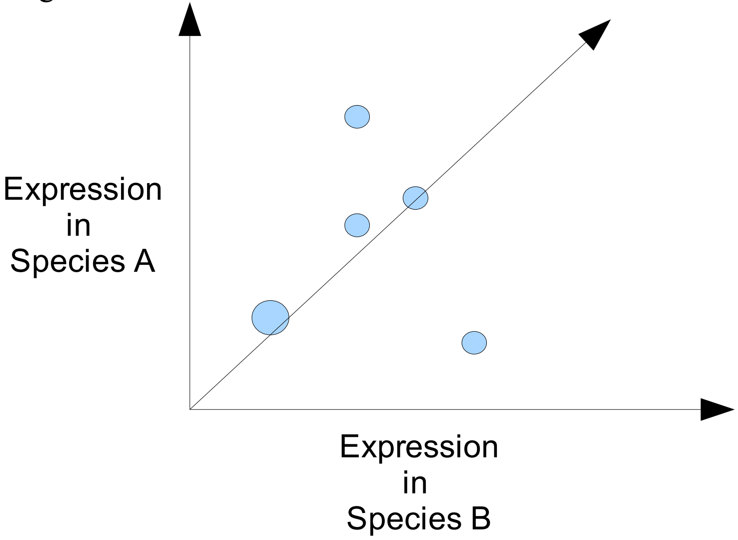
Species	**Condition**	**Experiments (replicas)**
A	C1	A1R1, A1R2
A	C2	A2R1, A2R2, A2R3
B	C1	B1R1, B1R2, B1R3
B	C2	B2R1, B2R2

Suppose that each experiment measures the gene expression levels. We would like to perform statistical analyses on these experiments, comparing the expression levels of orthologous genes.

In particular, one can ask the following questions:

- For experiments conducted under identical conditions, how do expression levels compare among orthologous genes in different species? Statistically, one can ask “Is the set of orthologous genes that are expressed significant?” Graphically, the comparison can be illustrated by a graph that plots expression levels in species A versus expression levels in species B, as below.

  It can be quantified by calculating the correlation between these expression levels.
- Another question that can be asked is “Given genes that are orthologous and have the same biological function (as determined by GO terms) how do their expression levels compare?”
- These analyses will provide insight into the similarities and differences between the functionality of orthologous genes in related plant species.


Expression
in
Species A
Expression
in
Species B