

PATRIC Standard Operating Procedures for Creation of Orthologous Gene Sets

Author: Eric E. Snyder, Maulik Shukla, Eric K. Nordberg
Version: 1.2, December 6, 2007 11:12
Supersedes Version: 1.1

Introduction

The PATRIC project places high value on the manual curation of genes and proteins. However, when faced with large numbers of closely related genomes, one must be selective in the genes targeted for this labor intensive process. To minimize redundant effort, the project has developed a strategy in which (for bacteria) a single, well-characterized genome is selected as the “reference genome”. The majority of the manual curation will be done on this genome. The remaining genomes in a genus are referred to as “associated genomes”. Only novel genes, not found in the reference genome, will be curated in the associated genomes. This begs the question: how does one determine which genes are novel and which are already represented in the reference genome? The answer lies in the identification of orthologous gene groups. Orthology implies that genes derived from a common ancestor prior to speciation of the respective organisms should have the same functional roles in those organisms, at least to a first approximation. Therefore, curation of a representative from that group should be sufficient to characterize the other members as well, given sufficient similarity in sequence composition and length. Thus, curation of all the genes/proteins within a genus can be considered complete when one member from each ortholog group has been curated. Consequently, the generation of ortholog groups is a high priority in the PATRIC project.

Purpose

When studying closely related genomes, it is important to identify which genes in one genome correspond to those in another based on common ancestry and functional and structural similarity. Specifically, we wish to identify orthologous gene pairs or, more generally, orthologous gene groups, when considering multiple genomes. Several lines of evidence can be used to infer orthology: sequence similarity, conservation of local gene order, *etc.* None of these methods are perfect, particularly in the face of gene duplications, gene rearrangements and elevated levels of repetitive elements. However, when combined with other constraints, sequence similarity is a widely used metric and well suited to the prediction of ortholog groups in sets of closely related genomes. This document describes the current procedures for the identification of orthologous gene groups (OGs) from bacterial genomes.

Application

Within a given genus, it is observed that 60 to 80% of proteins predicted for a reference genome can be paired unambiguously with their orthologous proteins from a related genome. Such a pairing is unambiguous if the two proteins align with each other over 90% of their length with a substitution rate of <10%. When these conditions are met, it is almost certain that the two proteins are functionally equivalent unless one of the proteins is inactivated by a point mutation, a condition that will be difficult to detect without significant manual follow-up. Add to this operational definition the requirement that each protein must be more similar to the indicated protein than any other in the genome and that the reverse must also be true, one can feel comfortable that any annotation assigned to one of the pair will also be appropriate for the other.

Software

Our initial in-house development effort used bidirectional best BLAST alignments as the operational definition of orthology between proteins in genome pairs. This BLAST data can be interpreted as a graph with vertices corresponding to protein sequences. An edge exists between two sequences if they are bi-directional best hits. Initial estimates of ortholog groups are produced by identifying cliques, or fully-connected subgraphs, within this larger graph. A custom set of graph analysis tools developed by Eric Nordberg as part of his Ph.D. dissertation research is used to identify cliques. Finding a maximum clique in a graph is an NP-complete problem, so heuristic approaches are used to find large cliques. This

methodology holds promise for the future. However, for production work, we have adopted the well-known program OrthoMCL¹ as our primary application for building groups of orthologous proteins. In a comparison of the Nordberg method² and OrthoMCL using rickettsial genomes, OrthoMCL yielded higher quality OGs and much faster execution times (J. Gillespie, M. Shukla, unpublished observations).

It should be noted at this point that OrthoMCL combines true orthologs and in-paralogs, paralogs arising from post-speciation gene duplication events. This is a reasonable compromise for a method based solely on sequence similarity. However, it complicates the interpretation of OGs since it is therefore possible to have some genomes represented more than once (by virtue of contributing multiple proteins). Efforts are underway by the authors to use orthogonal evidence, such as “gene neighborhood” information, to help resolve paralogous families when sequence alone is insufficient. For now, we are working on the assumption that the consequences for accurate annotation by functional inference are small.

Participants

This SOP is typically executed by a bioinformatician or software developer with input from a curator familiar with the phylogeny of the organisms under analysis.

Inputs

To generate predicted ortholog groups given a group of organisms (which typically corresponds to a PATRIC organism category), a file containing the complete list of encoded proteins (in FASTA format) is required for each genome. It is recommended that the file names be as brief as possible. For example, the seven *Brucella* genomes available as of November 26, 2007, could be abbreviated as follows:

Full Name	Abbreviation	Abbrev
<i>Brucella suis</i> 1330	Bsu1330	Bsu1
<i>Brucella abortus</i> biovar 1 str. 9-941	Bab9941	Bab9
<i>Brucella canis</i> ATCC23365	Bcan23365	Bcan
<i>Brucella melitensis</i> 16M	Bmel16M	Bmel
<i>Brucella melitensis</i> biovar Abortus 2308	Bab2308	Bab2
<i>Brucella ovis</i> ATCC 25840	Bov25840	Bov2
<i>Brucella suis</i> ATCC23445	Bsu23445	Bsu2

The desirability of short file names will become apparent in step 2.

Steps

1. Initiation

The process of ortholog group (OG) creation is initiated when one or more new genomes are added to a bacterial PATRIC organism category (*Brucella*, *Rickettsia*, *Coxiella*).

2. Execution of OrthoMCL

The Perl script **orthomcl.pl** executes both the initial all-versus-all BLAST search and follow-up clustering of proteins into orthologous groups (OGs) from a single command line. Using FASTA file names based on the seven *Brucella* genomes listed above, the command-line would be:

```
orthomcl.pl --mode 1 --fa_files Bsu1.faa,Bab9.faa,Bcan.faa,Bmel.faa,Bab2.faa,Bov2.faa,Bsu2.faa
```

It is also recommended that `stderr` and `stdout` be redirected to a file (e.g. `runtime_info`) for future reference and that the process be backgrounded on initiation. Both actions can be accomplished by appending the following to the above command-line (assuming the user's shell is `csh` or `tcsh`):

```
>& runtime_info &
```

¹ Li L, Stoeckert CJ Jr, Roos DS. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 2003 Sep;13(9):2178-89. PMID: [12952885](https://pubmed.ncbi.nlm.nih.gov/12952885/).

² As of Spring, 2007.

Note that program parameters, including the directory into which output files will be written, are hard-coded in the `orthomcl_module.pm` file. Thus far, we have obtained satisfactory results using default parameters.

3. Processing OrthoMCL Output

OrthoMCL produces numerous, voluminous output files, some of which will be described in the following section. The most important of these files is `all_orthomcl.out`, which contains information describing the ortholog groups built by the program. This file is best described as “machine readable”. It can be converted into the more user-friendly “.og” format, which has become our *de facto* standard for reporting OG data on the PATRIC website by using the script **groups2og.pl**:

```
groups2og.pl all_orthomcl.out all.faa OG_PREFIX > output.og
```

Here, `all.faa` is simply a FASTA file containing all of the input proteins in one³. In other words:

```
cat Bsu1.faa Bab9.faa Bcan.faa Bmel.faa Bab2.faa Bov2.faa Bsu2.faa > all.faa
```

OG_PREFIX is typically a mnemonic string representative of the organism category (or genus); it is used to generate the ortholog group name. For example, “Bru”, appropriate in this example, would result in OG names such as `Bru_1234`.

Outputs

1. OrthoMCL Files

`all_orthomcl.out`

Summary of ortholog groups: Each line starts with OG ID in the form: ORTHOMCL#, where # is a serial integer starting at zero. Immediately following, in parenthesis is the number of genes and taxa involved in the group, followed by a colon. Then follows a space delimited list containing the ID of each protein in the group with the name of the FASTA file from whence it came in parenthesis, immediately following.

`all_blast.bbh`

Summary of bidirectional best hits: Each line contains four fields, two protein IDs from different genomes followed by their respective alignment E-values.

`orthomcl.log`

Run-time information from `orthomcl.pl`, similar to `stdout`.

`parameter.log`

Summary of execution parameters.

2. Groups2og.pl Output

`*.og`

Human readable summary of ortholog groups: A tab-delimited file containing the following fields.

4. GROUP_NAME
5. GENE_SYMBOL
6. PRODUCT
7. NUMBER_OF_MEMBERS
8. NUMBER_OF_TAXA
9. MEMBER_LOCUS_TAG
10. GENOME_NAME
11. MEMBER_GENE_SYMBOL
12. MEMBER_PRODUCT
13. MEMBER_PROTEIN_LENGTH

The ability of the program to properly populate these fields is somewhat dependent on the formatting of the FASTA definition line from the initial `.faa` files. It is likely that `.faa` files dumped from the PATRIC

³ Now if *I* had written this program, I would have allowed the user to use the same files used to invoke `orthomcl.pl`.;-)

database will be properly formatted. If your data comes from a different source, your mileage may vary. Consult the source code for details.

Exit Criteria

We are currently in the process of developing formal acceptance criteria for updating OGs with new genome data. Although it is highly dependent on the evolutionary distance between the new genome and the existing data, we expect a general trend towards a decreasing number of new OGs with each new genome. Deviation from this trend may throw suspicion on the quality of the CDS annotation of the new genome. However, real increases in novel gene content could result from horizontal gene transfer, particularly for genomic islands, or lineage-specific transposon activity, or if the available genomes are biased with respect to the true diversity of the taxon. That said, while it is important to have clear expectations for the results of each new OG build, one must not blindly reject results that do not meet expectations. More than any other single method, the knowledge of an organism group represented by its gene families is a sensitive metric that captures our understanding of it in precise and quantifiable ways. It is far more important to understand the cases where our expectations are challenged, for that is the path to new and fundamental knowledge.