Comparative analysis of metabolism in parasitic worms

CURRENT STATUS: POSTED

Makedonka Mitreva
MGI, Washington University in St. Louis
mmitreva@wustl.edu*Corresponding Author*

Rahul Tyagi
Mitreva Lab (MGI, Washington University in St. Louis)

Swapna Seshadri
Hospital for Sick Children, Toronto, Ontario, Canada

John Parkinson
Hospital for Sick Children, Toronto, Ontario, Canada

DOI:
10.1038/protex.2018.048

SUBJECT AREAS
Computational biology and bioinformatics *Biotechnology*

KEYWORDS
metabolism, enzymes, helminths, nematoda, platyhelminthes, metabolic pathways, auxotrophy, metabolic pathway coverage, metabolic pathway diversity

Abstract

Ascertaining metabolic potential of parasites is an important step in understanding their biology, and getting insights into host-parasite interactions. Here we describe a computational protocol to use a set of high confidence predictions for enzymes encoded in worm genomes for comparative analyses at different levels of pathway resolution. This also includes analysis of chokepoint enzymes in the organism’s metabolic pathway.

Introduction

Comparative analyses of metabolism of various species leads to better understanding of similarities and differences in biology among them. This could lead to insights into their specific niche-related adaptations. Moreover, when contrasted with host organism’s metabolism, such comparisons are likely to yield important insights that may be leveraged to specifically target parasites.

See figure in Figures section.

Figure 1. Overview of the protocol. The dashed lines represent an optional step to ascertain lenient completion of modules for applications sensitive to false negatives.

As shown in **Figure 1**, many different approaches can be taken for metabolic comparisons. In this protocol we describe multiple different analysis, especially in context of the work done for the 50 helminth initiative. Since the primary objective was to see a large scale overview of metabolic potential for different groups of helminths, many of the approaches included a phylogenetic analysis at the end.

Pathway Tools package[1] was used to reconstruct individual metabolic network of each species based on reference pathways in the Biocyc database[2]. These can then be analyzed to discover vitamin and amino acid auxotrophies.

KEGG database[3] also has reference pathways. Many of these are relevant to helminth biology. These pathways can be reconstructed using the input ECs (Enzyme Commission numbers) for each species. These species-specific metabolic networks can then be compared either based on overall coverage (i.e. % of enzymes present), or diversity of extent of coverage in a species group. The networks can also be analyzed to identify species-specific chokepoint enzymes. A chokepoint in a directed network

is a node with either a single outgoing or a single incoming edge. This means that the chokepoint enzyme either uniquely consumes or uniquely produces a substrate. This makes the enzyme an especially interesting drug target. The identified chokepoints can then be used to generate a phylogenetic tree to see whether closely related species share chokepoints in general.

KEGG also defines smaller networks, which may be part of multiple larger pathways. These are called “metabolic modules”. Since these are relatively smaller, they are more amenable to topological analysis. We use every species’ enzyme annotations to find which KEGG metabolic modules are “complete” in the species (i.e. every enzyme needed to produce the final substrate, if given the initial substrate(s), is annotated). This analysis follows Tyagi et al.[4]. The identified complete modules can then be used to generate a phylogenetic tree which can be used to identify any unexpected evolutionary patterns in the metabolic potential of helminths. Since every single indispensable enzymatic step in a module needs to be present for the module to be deemed complete, this analysis is especially sensitive to false negatives. To guard against this, a “lenient completion” is also analyzed, which allows absence of up to 1 enzyme potentially due to misannotation or missed genecalls4. Any phylogenetic peculiarities identified using module completion that are also supported by lenient completion analysis are likely to be true.

Using just the EC numbers associated with each species, one could directly generate phylogenetic trees which offer a simple overview of metabolic potential evolution. In general, however, the data at this level is more noisy and more insights are obtained by looking at the data at the module or pathway level.

Reagents

KEGG database v70.

Biocyc Database.

modDFS tool.

Equipment

Computer cluster.

Procedure

Step A: Pathway Tools based reconstruction to find auxotrophies

The input to Step A is a high-confidence EC set for the parasitic worms.

1. Run Pathway Tools pipeline for each species using Biocyc database. It uses a set of rules to assign evidence scores for pathway predictions based on: presence of most of the ECs for a pathway, presence of unique ECs, presence of the first two steps (for a degradation pathway), presence of the last two steps (for a biosynthetic pathway), presence of >50% enzymes (for energy metabolism pathways). It also uses taxonomic pruning, wherever information is available, to reduce false-positives
2. Analyze biosynthesis pathways for vitamins and amino acids. These can then be compared to identify species or species-group specific auxotrophies. The output from Step A is a table showing each species' biosynthetic capability of vitamins and amino acids.

Step B: Reconstructed KEGG pathways and chokepoints comparison

The input to Step B is a high-confidence EC set for the parasitic worms.

1. From the reference pathways in KEGG database, remove those that aren't relevant to helminths. This is done by including only the KEGG pathways that have at least one reference pathway for a nematode/platyhelminth species in the KEGG database. This meant excluding pathways such as 'Carbon fixation in photosynthetic organisms', even if some of the enzymes implicated in these pathways are found in helminths. In addition, some manual curation may be needed. E.g. excluding caffeine metabolism, which does have a reference pathway for some nematodes (*C. elegans* and *C. briggsae* KEGG v70) but is deemed unlikely to be of relevance to most helminths studied by us. For KEGG v70, this leaves 65 KEGG pathways deemed to be 'helminth-relevant'.
2. Coverage is defined as the fraction of all reference pathway ECs that are annotated in a given species. Calculate coverage for every helminth-relevant KEGG metabolic pathway.

3. Coverage is compared separately among different groups of worms. These groups are either phylum level (platyhelminths and nematodes) or subsets thereof (cestodes and trematodes for platyhelminths and parasites from different clades of nematodes). Some comparisons are done among only the parasites. This means defining groups like ‘Clade IVa-’ (all Clade IVa worms except the free living e.g. ^{Rhabditophanes}) and ‘Clade V-’ (all Clade V worms except the free living e.g. $^{C.\ elegans}$ and $^{P.\ pacificus}$). The comparisons are performed using Wilcoxon tests, and FDR corrected P-values (corrected using the Benjamini-Hochberg procedure) are used to assign significance ($P<0.05$).
4. Coverage diversity is compared between different worm groups. The coefficient of variation of pathway coverage is used to measure the variation in coverage of these pathways among these groups. Comparisons are also performed across all worms between different ‘superpathways’ (e.g. combining all ‘amino acid metabolism’ pathways together). Wilcoxon test over the distribution of coefficient of variation is performed for these comparisons.
5. The chokepoint enzymes are identified according to Taylor et al[5], with the following modification: the metabolic networks analyzed are not the entire reference reaction sets in KEGG, but only the subnetworks formed by the reactions annotated in the species of interest (including ECs from hole-filling), resulting in more organism-specific metabolic networks. Chokepoints are reported in context of these species-specific networks.
6. Clustering of species based on detected chokepoint enzymes is performed. This is just a presence-absence based clustering using the Jaccard similarity index and Ward-linkage method.

7. Generate a phylogenetic tree to see whether closely related species share chokepoints in general. The output from Step B is statistical metrics (P-values) for comparison between worm groups for pathway coverage extent and coverage diversity. It also yields a set of species-specific chokepoints and a phylogenetic tree based on presence-absence of those chokepoints.

Step C: KEGG metabolic module completion analysis.

The input to Step C is a high-confidence EC set for the parasitic worms.

1. Presence/absence of KEGG metabolic 'modules' are ascertained using the modDFS algorithm[4]. The algorithm starts from the final product of the module and systematically traverses all those nodes which can produce this product by a chain of substrate-product relations.
2. Species clustering based on presence/absence of modules is performed using Ward-linkage based on the Jaccard similarity index. Generate a phylogenetic tree to see whether closely related species share metabolic modules in general.
3. For applications that are sensitive to false negatives, a "lenient completion" is also ascertained[4].
4. Species clustering based on lenient presence/absence of modules is performed using Ward-linkage based on the Jaccard similarity index. Generate a phylogenetic tree to see whether closely related species share metabolic modules in general. The output from Step C is a table showing which KEGG metabolic modules are either strictly or leniently complete in which worm species. When interpreting these results it must be remembered to only use the lenient completion results to remove potential false negatives.

Step D: Species and group comparison based on EC content.

The input to Step D is a high-confidence EC set for the parasitic worms.

1. Species clustering based on presence/absence of ECs is performed using Ward-linkage based on the Jaccard similarity index. Generate a phylogenetic tree to see whether closely related species share metabolic potential defined simply by presence of individual ECs. The output from Step D is a phylogenetic tree which can be used to find potential peculiarities in metabolic potential evolution.

Anticipated Results

The output from the protocol is a set of multiple comparisons between worm groups based on high confidence EC annotations.

References

1. Karp, P.D. et al. Pathway Tools version 19.0 update: software for pathway/genome informatics and systems biology. Brief Bioinform 17, 877-90 (2016).
2. Caspi, R. et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res 40, D742-53 (2012).
3. Kanehisa, M. et al. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res 42, D199-205 (2014).
4. Tyagi, R., Rosa, B.A., Lewis, W.G. & Mitreva, M. Pan-phylum Comparison of Nematode Metabolic Potential. PLoS Negl Trop Dis 9, e0003788 (2015).
5. Taylor, C.M. et al. Discovery of anthelmintic drug targets and drugs using chokepoints in nematode metabolic pathways. PLoS Pathog 9, e1003505 (2013).

Figures

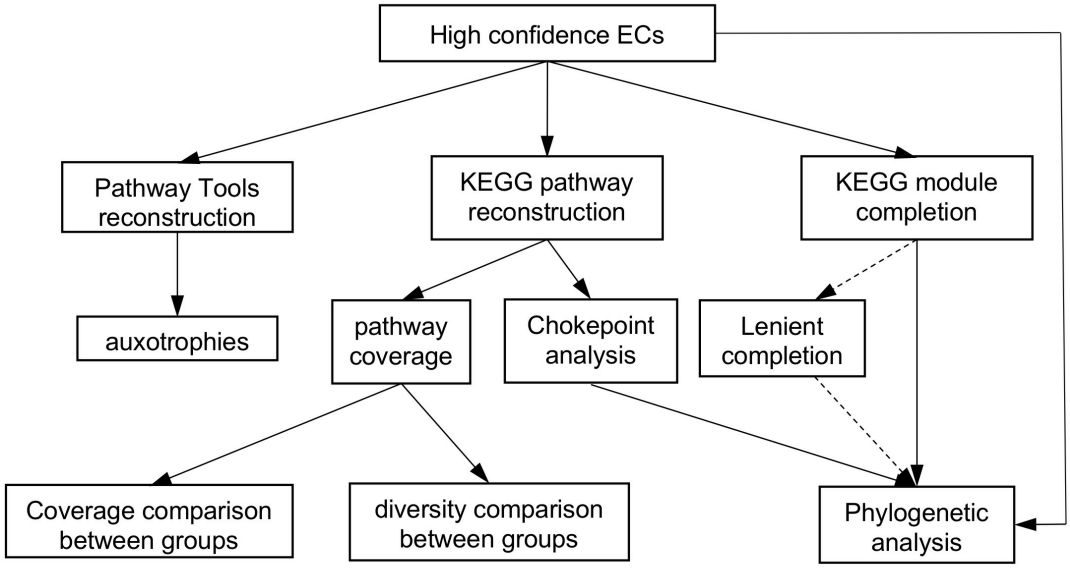

Figure 1

Overview of the protocol The dashed lines represent an optional step to ascertain lenient

completion of modules for applications sensitive to false negatives.