Annotating metabolic enzymes in parasitic worm proteomes

John Parkinson (jparkin@sickkids.ca)
Hospital for Sick Children, Toronto, Canada
Swapna Seshadri
Hospital for Sick Children, Toronto, Canada
Rahul Tyagi
Mitreva Lab (MGI, Washington University in St. Louis)
Makedonka Mitreva
MGI, Washington University in St. Louis

Method Article

Keywords: enzyme annotation, genome annotation, metabolism, enzymes, helminths, nematoda, platyhelminthes

Posted Date: April 17th, 2018

DOI: https://doi.org/10.1038/protex.2018.047

Abstract

Ascertaining metabolic potential of parasites is an important step in understanding their biology, and getting insights into host-parasite interactions. This process starts with identifying enzymes using protein sequences encoded in their genome. Here we describe a computational protocol to obtain a set of high confidence predictions for enzymes encoded in worm genomes using multiple enzyme annotation methods. This also includes an optional gene-family based inference method to expand the results by adding a set of relatively lower confidence predictions, if needed. This work was done jointly between the "Mitreva Lab":https://www.nature.com/protocolexchange/labgroups/1124733, and the Parkinson Lab \(Hospital for Sick Children, Toronto) Website: "http://www.compsysbio.org":http://www.compsysbio.org.

Introduction

Enzyme annotation using amino acid sequence isn't a simple process and multiple methods have been published that accomplish this using different approaches. This protocol uses some of these complementary methods to obtain a high-confidence set of enzymes encoded in the genome. See figure in Figures section. **Figure 1: Overview of the protocol.** EC sets in green box are high confidence ECs. EC set C \(in red box) is a lower confidence set with less false negatives. As shown in **Figure 1**, the first step is to annotate enzymes using the protein sequences and multiple annotation methods. Most of these methods use different characteristics of the protein sequence and assign an enzyme activity, formally represented as an Enzyme Commission number \(EC) to it if it passes certain qualification thresholds. KAAS[1], which uses the KEGG database[2] assigns KEGG Orthology \(KO) IDs, which can then be translated to corresponding EC IDs using information from the KEGG database. PRIAM[3] uses position-specific scoring matrices derived for ECs in the ENZYME database[4]. Since, we consider KAAS and PRIAM to have a potential for greater false positive rates, we only use an intersection of their EC annotations. On the other hand, predictions from the BRENDA database[5], comprising literature-curated predictions, and the DETECT pipeline[6], a prediction approach accounting for sequence diversity across enzyme families, are considered robust. Thus KAAS/PRIAM predictions are combined with annotations from DETECTv1 and BRENDA to yield a high-confidence set of ECs \(EC set A in **Figure** **1**).This set is unlikely to suffer from high false-positive rates, because DETECT considers probabilities of all sequences and BRENDA is a literature-curated set of enzymes. However, due to stringency of annotation parameter, it is possible that some ECs will be missed in some draft genomes \(i.e. false negatives). For applications that are sensitive to false negatives \(e.g. determining amino acid auxotrophies), a second step is introduced to add additional ECs to EC set A. This is done using the pathway hole-filling algorithm of the Pathway Tools package[7]. Pathway Tools reconstructs metabolic networks using reference pathways, together with an input EC set \(EC set A in this case). Based on the coverage of reference pathways and input genome sequences, the hole-filling algorithm identifies genes that are likely to encode for candidate false negative ECs. These ECs are further pruned by considering only those supported by KEGG, PRIAM, and EFICaZ[8] and are subsequently added to EC set A. Together the set of high-confidence EC annotations \(EC set A) and those identified through the hole filling procedure constitute the final set of high-quality

EC predictions \(EC set B). For large scale comparative studies, we also use gene family information to infer additional ECs that were not identified by the annotation pipeline. Since sequence orthology does not always correlate with shared enzyme activity, we only infer ECs when there are no conflicts in the orthogroup's EC annotation, and the annotation is supported by some high quality genomes as part of the gene family. Note, these low-confidence ECs \(EC set C) are only used to validate results obtained from EC set B, rather than for de-novo analysis.

Reagents

KEGG database v70. Biocyc Database. Compara Database generated using species of interest. Local installation or access to any relevant webservers for Pathway Tools, DETECT v1.0, KAAS v2, PRIAM, BRENDA, EFICAz v2.5.

Equipment

Computer cluster.

Procedure

Step A: Annotate ECs using 4 different methods The input to Step A is the protein sequences for the parasitic worms. 1. Use BRENDA for predicting ECs. Sequences from SWISSPROT[9] annotated with EC numbers in BRENDA for each organism are mapped to sequences in the genome via a two-way reciprocal BLAST-based search strategy. 2. Use DETECT v1.0 to predict ECs. Here we used a high confidence cutoff ILS of 0.9 \(inclusive), requiring at least 5 positive matches. Further processing is used to include hits with ILS>0.2: sequences annotated with the same EC in DETECT and BLAST against SWISSPROT database are retained, whereas discordant ECs with a higher E-value in BLAST \(>1e-30) were added if: 1) it was a different EC; or 2) replaced if the ECs differed only in the fourth digit of the classification hierarchy. 3. Use PRIAM to predict ECs. The parameter set to be used : minimum probability >0.5, profile coverage >70%, check catalytic - TRUE) 4. Use KAAS \(KEGG Automatic Annotation Server) to predict ECs. For this, a locally installed version 2 of KAAS can be used with default settings \(i.e. bi-directional best hit with bit-score threshold of 35). The KOs \(KEGG Orthologs) annotated by KAAS are then associated with corresponding ECs using the KO and EC definitions in KEGGv70. 5. For each species, find the ECs common between predictions of PRIAM and KAAS from steps 3 and 4 above. Call this set of ECs "EC_set_1". 6. For each species, combine the ECs obtained using BRENDA and DETECT in steps 1 and 2 above with EC_set_1 \(i.e. take union of these ECs). This is "EC set A". The output from Step A is a set of high confidence ECs for each species whose proteome was used for predictions. **Step B: Pathway hole filling using Pathway Tools** The input to Step B is "EC set A" for all the species of interest. Also required is an installation of Pathway Tools pipeline \(v18.5) along with definitions of metabolic pathways from Biocyc database[3]. 1. Run Pathway Tools pipeline for each species. 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. 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'. 3. For these helminth-relevant pathways, use the pathway hole-filler component \(default settings) of Pathway Tools to assign genes to pathway holes. 4. Only include Predictions for those genes from step 3 above that either \(a) had no assigned ECs by Step A above; or \(b) were assigned a different EC by that method but also had support for the alternative EC \(i.e. the one supported by pathway hole-filler) based on at least one of KAAS, PRIAM, or EFICAz v2.5 predictions \(default settings). 5. Combine any ECs obtained by step 4 above with "EC set A" \(i.e. take union of the sets) to yield "EC set B". The output from Step B is an expanded set of high quality ECs for each species whose proteome was used for predictions. **Step C: Using Compara database to expand the set of ECs**. Step C can be optionally used when a Compara database[10] is available that has been generated for the set of helminth species used in Steps A and B. The resulting ECs could have higher false positive rates, and should be used for applications sensitive to false negatives. The input to Step C is "EC set B" obtained after step B above along with the corresponding EC-gene mappings for each species. 1. From the helminth species analyzed, identify a set of "Tier 1" species whose genomes and gene models are deemed to have high quality. 2. From the Compara gene families, remove those that have different ECs assigned after Step B to at least 2 genes from Tier 1 species. 3. From the remaining Compara gene families, remove those that have less than 2 genes from Tier 1 species annotated with an EC after Step B. 4. For each of the remaining Compara gene family, there should be a unique EC annotated to Tier 1 genes. Assign this EC to any genes in the family without EC annotation after Step B. This new set of ECs for each of the species is the "EC set C". The output from Step C is an expanded set of relatively lower confidence ECs for each species whose proteome was used for predictions.

Anticipated Results

The output from the protocol is a set of high confidence ECs annotated for each helminth species whose proteome was input. A lower confidence set of ECs is also obtained after optional Step C, which can be used for validation of results obtained using the high confidence ECs that may be sensitive to false negatives.

References

1. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A.C. & Kanehisa, M. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res 35, W182-5 \(2007). 2. Kanehisa, M. et al. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res 42, D199-205 \(2014). 3. Claudel-Renard, C., Chevalet, C., Faraut, T. & Kahn, D. Enzyme-specific profiles for genome annotation: PRIAM. Nucleic Acids Res 31, 6633-9 \(2003). 4. Bairoch, A. The ENZYME database in 2000. Nucleic Acids Res 28, 304-5 \(2000). 5. Schomburg, I. et al. BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. Nucleic Acids Res 41, D764-72 \(2013). 6. Hung, S.S., Wasmuth, J., Sanford, C. & Parkinson, J. DETECT–a density estimation tool for enzyme classification and its application to Plasmodium falciparum. Bioinformatics 26, 1690-8 \(2010). 7. 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). 8. Kumar, N. & Skolnick, J. EFICAz2.5: application of a high-precision enzyme function predictor to 396 proteomes. Bioinformatics 28, 2687-8 \(2012). 9. Boutet, E. et al. UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View. Methods Mol Biol 1374, 23-54 \(2016). 10. Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A.C. & Kanehisa, M. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res 35, W182-5 \(2007).

Figures

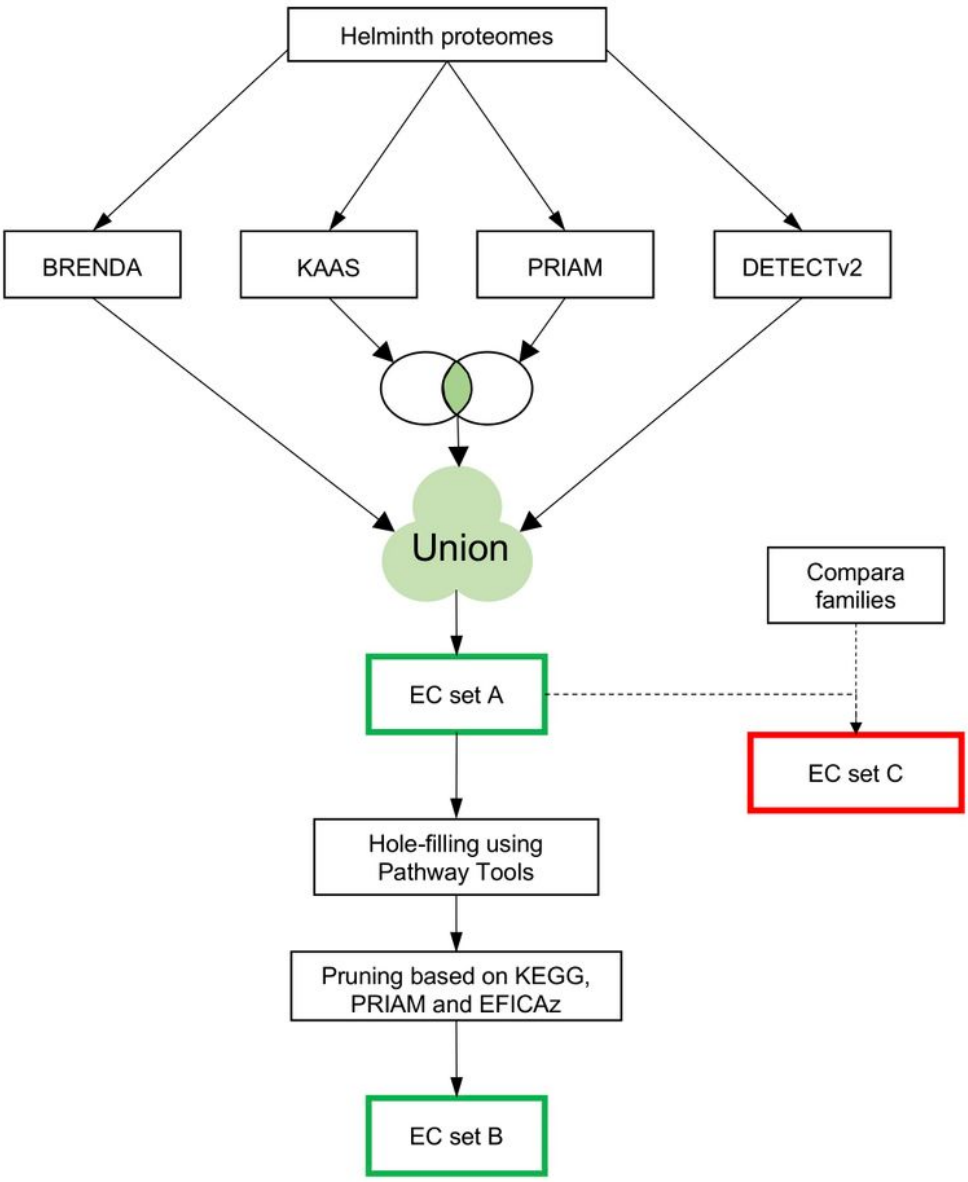

Figure 1

Overview of the protocol EC sets in green box are high confidence ECs. EC set C (in red box) is a lower confidence set with less false negatives.