Functional annotation of parasitic worm genomes, by assigning protein names and GO terms

Matthew Berriman (mb4@sanger.ac.uk)
Berriman Lab Group, Sanger Institute
Avril Coghlan
Berriman Lab Group, Sanger Institute

Method Article

Keywords: genomes, annotation, gene functions, protein names, GO terms, Gene Ontology, parasites, parasitic worms, helminths, nematodes, flatworms, platyhelminths, Nematoda, Platyhelminthes

Posted Date: May 15th, 2018

DOI: https://doi.org/10.1038/protex.2018.055

Abstract

Given a set of predicted protein-coding genes for a newly sequenced genome, functional annotation involves assigning putative functions to the predicted genes. Two ways in which this can be done are assigning protein names and Gene Ontology \(GO) terms to the predicted proteins. Here we describe a computational pipeline for assigning protein names and GO terms to predicted proteins in parasitic worm \(nematode and platyhelminth) genomes, which transfers names and GO terms from orthologues in other species.

Introduction

Given a set of predicted protein-coding genes for a newly sequenced genome, functional annotation involves assigning putative functions to the predicted genes. Two ways in which this can be done are assigning protein names and Gene Ontology \(GO; Gene Ontology Consortium, 2010) terms to the predicted proteins. Here we describe a computational pipeline for assigning protein names and GO terms to predicted proteins in parasitic worm \(nematode and platyhelminth) genomes, which transfers names and GO terms from orthologues in other species. When assigning protein names, UniProt protein naming rules \(www.uniprot.org/docs/nameprot) are followed where possible. This recommends that a good and stable name for a protein is "as neutral as possible"; that a protein name "should be, as far as possible, unique and attributed to all orthologs"; and a protein name "should not contain a specific characteristic of the protein, and in particular it should not reflect the function or role of the protein, nor its subcellular location, its domain structure, its tissue specificity, its molecular weight or its species of origin”. In our protocol, a protein name is assigned to each predicted protein based on curated names in UniProt \(Bairoch & Apweiler, 2000) for human, zebrafish, _Drosophila melanogaster_, _Caenorhabditis elegans_, and _Schistosoma mansoni_ orthologues identified from a database of gene families \(e.g. built using Ensembl Compara; Vilella _et al._ 2009), or \(if no information is found from orthologues) based on InterPro \(Hunter _et al._ 2012) domains. Figure 1 shows an example of using our protein naming pipeline for four _Strongyloides ratti_ genes that belong to the tubulin polyglutamylase family \(underlined in pink), where four different protein names were assigned to them \(in pink), based on names of their _C. elegans_ or human orthologues. Since each of the _S. ratti_ genes belonged to a different subfamily of the tubulin polyglutamylase family, they were assigned different names. Advantages of our approach are that it avoids taking the protein name from the top BLAST \(Altschul _et al._ 1997) hit \(which may not have a meaningful name, and/or may not be an orthologue if the gene of interest belongs to a large gene family); it transfers protein names from curated UniProt entries, so these names should be well constructed; it transfers protein names from orthologs, as recommended by UniProt; and although the pipeline was designed with parasitic worms in mind, it can be easily adapted for other taxonomic groups \(e.g. protozoans). Previous approaches for assigning names to predicted proteins include assigning a name based on top BLAST hits. For example, for the _Echinococcus multilocularis_ gene set, Tsai _et al._ \(2013) found the top ten BLASTP hits of a predicted protein in GenBank \(Benson 2018) and found a consensus between their protein names, downweighting

uninformative names and giving higher weight to the parts of names that agree between hits. A similar approach to ours is used by Ensembl \(Zerbino _et al._ 2018) to transfer gene names \(as opposed to protein names) from curated databases \(HGNC \(Yates _et al._ 2017), MGI \(Eppig _et al._ 2017), ZFIN \(Bradford _et al._ 2015)) to orthologues in other species. In our protocol, GO terms are assigned by transferring GO terms from human, zebrafish, _C. elegans_, and _Drosophila melanogaster_ orthologues \(again, identified from a database of gene families), and using InterProScan \(Jones _et al._ 2014), i.e. InterPro2GO. To maximise the amount of GO annotation, terms are transferred from all orthologues, not just one-to-one orthologues \(and therefore different usage to annotation of vertebrate orthologues by the Ensembl Compara GO-transfer approach; Vilella _et al._ 2009). The Compara GO-transfer pipeline is designed to transfer GO terms between relatively closely related vertebrate species. In contrast, our pipeline is designed to transfer GO terms across animal phyla \(e.g. from _D. melanogaster_ or human to a parasitic worm). Instead of transferring GO terms directly between orthologues, the last common ancestor terms of orthologues from two different species \(e.g. a _C. elegans_ orthologue and a _D. rerio_ orthologue) are transferred. These GO terms are more likely to be conserved across the more distantly related species in this data set, and thus more likely to be accurate predictions for the query protein \(e.g. from a parasitic worm such as _Brugia timori_). Like our approach for protein names, our pipeline for assigning GO terms was designed with parasitic worms in mind, but could easily be adapted for other taxonomic groups. Vilella _et al._ 2009 showed that the similar approach developed by Ensembl, for transferring GO terms between one-to-one orthologues among vertebrates, gave more detailed GO terms than InterPro2GO, when they transferred GO terms from human and mouse genes to vertebrate orthologues. Thus, we believe that our approach for transferring GO terms between animal orthologues, supplemented by GO terms from InterPro2GO, will give an accurate and complete set of GO annotations for a parasitic worm genome.

Reagents

Gene Ontology website UniProt database website InterPro database website Ensembl Compara pipeline GeneDB database website

Equipment

Computer cluster.

Procedure

This protocol requires a database of gene families that includes all predicted proteins of the species of interest \(e.g. _Brugia timori_), as well as _C. elegans_, _S. mansoni_, human, and zebrafish. This could be built for example using the Ensembl Compara pipeline \(Vilella _et al._ 2009), which identifies gene families and orthologues in those families. **Step 1: Protein Names** 1. For each predicted gene in the genome of interest \(e.g. _B. timori_), we first identify its one-to-one or many-to-one orthologues in human, zebrafish, _Drosophila melanogaster_, _Caenorhabditis elegans_, and _Schistosoma mansoni_ \(e.g.

many- _B. timori_ -to-one- C. _elegans_ orthologues). The orthologues are identified based on the in-house database of gene families. 2. Check whether the predicted gene of interest \(e.g. in _B. timori_) has a human, zebrafish, _Drosophila melanogaster_, _Caenorhabditis elegans_, or _Schistosoma mansoni_ orthologue with a manually curated protein name in UniProt \(Bairoch & Apweiler 2000), or a _S. mansoni_ orthologue with a manually curated protein name in GeneDB \(Logan-Klumpler _et al._ 2012). For this, if Ensembl accessions were used in the in-house database of gene families, the correspondence between UniProt accessions and Ensembl accessions can be downloaded from the UniProt website. 3. Taking the orthologues with curated names found in \(2) above, in order of preference, select an orthologue from: _C. elegans_, _S. mansoni_, human, _D. melanogaster_ and then zebrafish. If an orthologue with a manually curated protein name from the most preferred species \(_C. elegans_) is not found, check the second-most preferred species \(_S. mansoni_), and so on. 4. Taking the selected orthologue found in \(3) above, if it is from UniProt the UniProt 'recommended name' \(RN) of the orthologue is used as the protein name, if it is from GeneDB the GeneDB 'product description' is used as the protein name. 5. If no orthologue with a manually curated protein name is identified in step \(2) above, then check if the predicted gene of interest has an orthologue with a non-curated protein name \(i.e. from a TrEMBL entry; Bairoch & Apweiler, 2000). 6. Transfer the selected protein name from \(4) or \(5) above \(e.g. 'caveolin') to the predicted gene of interest \(e.g. in _B. timori_) and record the UniProt/GeneDB accession of the source protein, along with the evidence code ECO:0000265 \('sequence orthology evidence used in automatic assertion'), from the Evidence Code Ontology \(ECO; www.evidenceontology.org). If several genes in the species of interest \(e.g. _B. timori_) are assigned the same protein name \(for example, because of many-to-one orthology to the same _C. elegans_ gene), number them \(e.g. 'caveolin-1', 'caveolin-2', etc.) to ensure they are given unique names. 7. To identify InterPro \(Hunter _et al._ 2012) domains in the proteins of the query species \(e.g. _B. timori_), run InterProScan \(Quevillon _et al._ 2005) version 5.0.7 on the protein sequences of all predicted genes for that species. 8. If a particular query protein \(e.g. from _B. timori_) is not assigned any protein name based on its orthologues, then assign a protein name based on InterPro domains in the protein \(e.g. 'ankyrin repeat and SAM-domain-containing protein'). Note the InterPro accession\(s) of the source domains, and record the evidence code for the protein name as ECO:0000259 \('match to InterPro signature evidence used in automatic assertion'). 9. If a query protein \(e.g. from _B. timori_) is not assigned a protein name based on either orthologues or InterPro domains, name it 'Hypothetical protein'. 10. Add the protein names to the protein fasta file headers for the species of interest \(e.g. _B. timori_). **Step 2: GO Terms** 1. For each predicted gene in the genome of interest \(e.g. _B. timori_), we first identify its one-to-one, one-to-many, and many-to-many orthologues in human, zebrafish, _Drosophila melanogaster_, and _Caenorhabditis elegans_. The orthologues are identified based on the in-house database of gene families. 2. Download manually curated GO annotations for human, zebrafish, _C. elegans_, and _D. melanogaster_ from the Gene Ontology website \(Gene Ontology Consortium, 2010), and filter them to exclude annotations not based on experimental evidence \(i.e. only those with evidence codes IDA/IEP/IGI/IMP/IPI should be retained), annotations with a 'NOT' qualifier, and annotations to the GO:0005515 \('protein binding') term, following the criteria used by the Compara project for projecting GO terms to vertebrate orthologues \(Vilella _et al._ 2009). 3. Download the GO hierarchy from the Gene

Ontology website \(Gene Ontology Consortium, 2010). 4. Check whether the predicted gene of interest \(e.g. in _B. timori_) has human, zebrafish, _Drosophila melanogaster_, or _Caenorhabditis elegans_ orthologues with manually curated GO terms. 5. Taking the orthologues and their manually curated GO terms from \(4) above, take each pair of orthologues \(A, B) from two different species \(e.g. a _C. elegans_ orthologue and a _D. rerio_ orthologue, but not two _C. elegans_ orthologues), and use a breadth-first search algorithm to find the last common ancestors of their GO terms in the GO hierarchy. For example, if A has GO terms \{A1, A2, A3} and B has GO terms \{B1, B2}, we find the last common ancestors of A1+B1, A1+B2, A2+B1, A2+B2, A3+B1, and A3+B2. The GO terms assigned to the \(e.g. _B. timori_) query gene are the union of the last common ancestors of GO terms for all pairs of orthologues from two different species. Figure 2 shows an example of such a breadth-first search, where the nodes N1-N14 represent nodes in the GO hierarchy \(e.g. N14=GO:0030336, etc.). If two orthologues in different species, gene 1 and gene 2, have GO terms \{N1, N10} and \{N2, N11} respectively, then the ancestors of N1 and N2 are \{N5, N7}, the ancestors of N1 and N11 are \{N14}, the ancestors of N10 and N11 are \{N13}, and the ancestors of N10 and N2 are \{N14}. Therefore the union of all ancestors is \{N5, N7, N13, N14}. 6. Taking the set of GO terms assigned in \(5) above, remove any GO term from this set that is an ancestor \(in the GO hierarchy) of another term in the set. For example, in the example in Figure 2, we remove N14 as it is an ancestor of N13, leaving the final set of last common ancestors of genes 1 and 2 to be \{N5, N7, N13}. 7. For each GO term in \(6), note the UniProt accession of the source \(orthologue) protein, and record the evidence code for the GO term as IEA \('inferred from electronic annotation'). 8. Assign GO terms of the three possible types \(molecular function, cellular component and biological process) to each query protein \(e.g. from _B. timori_) in this way \(using \(1)-\(7) above). 9. For each query protein \(e.g. from _B. timori_), also identify GO terms using InterproScan \(Jones _et al._ 2014), which identifies InterPro \(Hunter _et al._ 2012) domains in the protein and maps GO terms to the domains \(using InterPro2GO). Note the InterPro accession\(s) of the source domains, and record the evidence code for the GO terms as IEA.

Troubleshooting

A possible disadvantage of our protocol is that it transfers a protein name from the orthologue with the curated name \(e.g. a human orthologue), which might not be phylogenetically closest orthologue \(e.g. _C. elegans_ orthologue) if the phylogenetic closest orthologue lacks a curated protein name in UniProt. We made this design choice because we felt that curated protein names were most likely to be reliable. A second possible disadvantage is that our approach relies on orthologues from a certain set of species \(zebrafish, human, _D. melanogaster_, _C. elegans_, _S. mansoni_, chosen because they are five relatively well studied species), so will not have used curated protein names for orthologues from less studied species, even if they exist.

Anticipated Results

The output from the protocol is a protein name \(and its source protein and evidence code) for each predicted protein in our genome assembly of interest \(e.g. _B. timori_), as well as \(zero, one or multiple) GO terms \(and their source genes and evidence codes) assigned to each predicted gene in our genome of interest.

References

Altschul, S. F. _et al._ Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. _Nucleic acids res._ **25**, 3389-402 \(1997). Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. _Nucleic acids res._ **28**, 45-48 \(2000). Benson, D.A. _et al._ GenBank. _Nucleic acids res._ **46**, D41-D47 \(2018). Eppig, J. T. _et al._ Mouse Genome Informatics \(MGI): Resources for Mining Mouse Genetic, Genomic, and Biological Data in Support of Primary and Translational Research. _Methods mol biol._ **1488**, 47-73 \(2017). Gene Ontology Consortium. The Gene Ontology in 2010: extensions and refinements. _Nucleic acids res._ **38**, D331-335 \(2010). Hunter, S. _et al._ InterPro in 2011: new developments in the family and domain prediction database. _Nucleic acids res._ **40**, D306-312 \(2012). Jones, P. _et al._ InterProScan 5: genome-scale protein function classification. _Bioinformatics_ **30**, 1236-1240 \(2014). Logan-Klumpler, F. J. _et al._ GeneDB–an annotation database for pathogens. _Nucleic acids res._ **40**, D98-108 \(2012). Quevillon, E. _et al._ InterProScan: protein domains identifier. _Nucleic acids res._ **33**, W116-120 \(2005). Ruzicka, L. _et al._ ZFIN, The zebrafish model organism database: Updates and new directions. _Genesis_ **53**, 498-509 \(2015). Tsai, I. J. _et al._ The genomes of four tapeworm species reveal adaptations to parasitism. _Nature_ **496**, 57-63 \(2013). Vilella, A. J. _et al._ EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. _Genome Res._ **19**, 327-335 \(2009). Yates, B. _et al._ Genenames.org: the HGNC and VGNC resources in 2017. _Nucleic acids res._ **45**, D619-D625 \(2017). Zerbino, D. R. _et al._ Ensembl 2018. _Nucleic acids res._ **46**, D754-D761 \(2018)

Acknowledgements

Thank you to Eleanor Stanley for discussion, in particular for bringing the UniProt protein naming rules \(www.uniprot.org/docs/nameprot) to our attention.

Figures

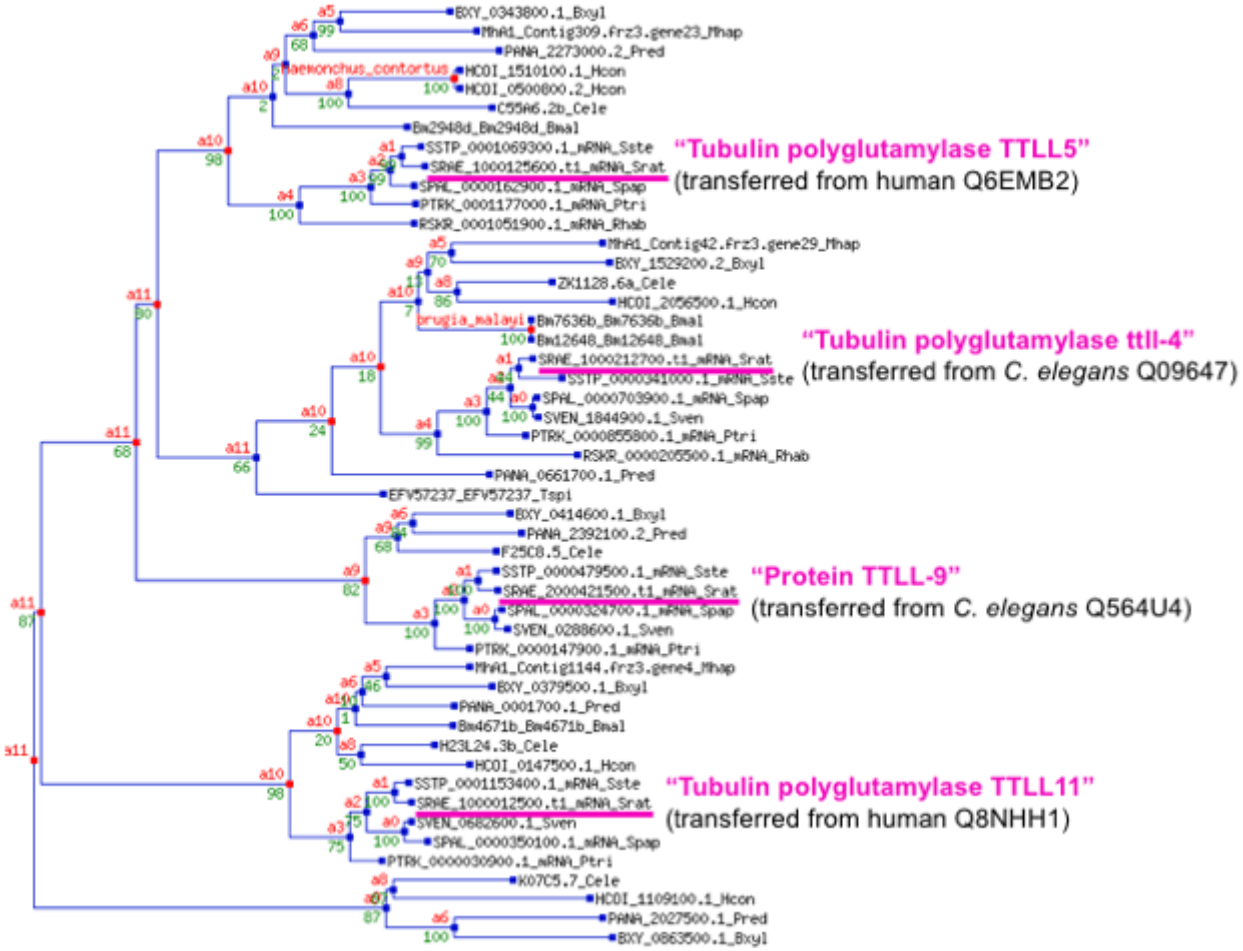

Figure 1

Example of protein naming

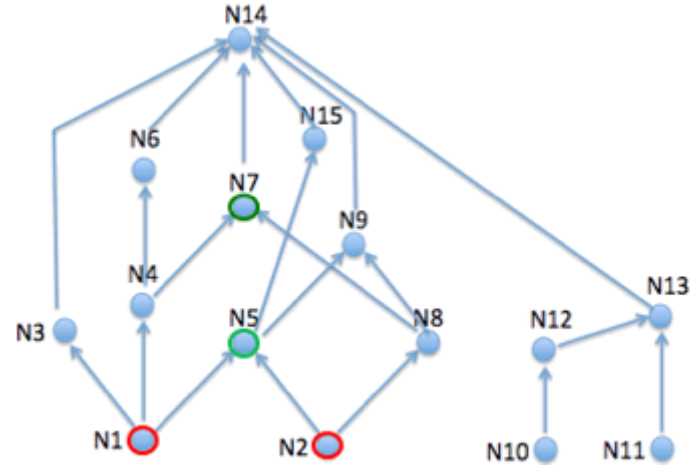

Figure 2

Example of breadth-first search.