Automatic Protein Function Annotation through Candidate Ortholog Clusters from Incomplete Genomes

Akshay Vashist[1] , Casimir Kulikowski[1], Ilya Muchnik[1,2]
[1]Department of Computer Science, [2]DIMACS
Rutgers University, Piscataway, NJ - 08854 U.S.A
{vashisht, kulikows}@cs.rutgers.edu, muchnik@dimacs.rutgers.edu

Abstract

Annotation of protein function often arises in the context of partially complete genomes but is not adequately addressed. We present an annotation method by extracting ortholog clusters from incomplete genomes that are evolutionary closely related to the genome of interest. To construct clusters, our method focuses on sequence similarities across genomes rather than similarities between sequences within a genome. We use the quasi-concave set function optimization for extracting the ortholog clusters as extreme groups of sequences such that similarity of the least similar sequence in this group is maximum. A protein sequence is annotated with the ortholog cluster whose average similarity is highest. We have applied this method for annotating the Rice proteome based on clusters constructed on four partially complete cereal proteomes and the complete proteome from Arabidopsis.

1. Introduction

A widely recognized method for protein function annotation is to find a group of (orthologous) proteins, from closely related species, that have evolved from a common ancestral protein. So, computational methods to functionally annotate newly sequenced proteins rely on finding orthologous proteins from well-studied related species. The function annotation problem often arises in the context of partially complete genomes where the limitations imposed by the quantity of data are challenging but the data quality is promising due to high quality protein function annotation experimentally determined by biologists. However, current popular ortholog detection approaches require complete genomes as they identify reciprocal best hits from a pair of genomes as seed ortholog clusters [3]. Additionally, these approaches are not robust for detecting orthologs for protein families that have undergone recent duplications.

To overcome the limitations from incomplete genomes in ortholog detection, we recently proposed a combinatorial optimization method that uses a measure of similarity between a protein and a subset of proteins from other species [4]. The results of applying this method to 43 prokaryote genomes compare well with those of manually curated ortholog clusters in the Clusters of orthologous genes, COG [3]. Furthermore, this parameter-free method is computationally efficient and automatically determines the number of ortholog clusters in the input data.

Protein annotation using ortholog clusters requires a measure of similarity between a query protein and the ortholog clusters. Although it is customary to annotate a query protein with annotation of the cluster that contains the best Blast [1] hit for the query protein, this method is known to be error-prone as best Blast hit may not be the nearest neighbor (closest homolog) of the query protein [2]. To address this we have developed a stringent criterion for annotating a protein with an ortholog cluster. The method is applied for annotating the Rice proteome using ortholog clusters constructed from the partially complete cereal genomes and the model plant, *Arabidopsis*.

2. Ortholog Detection

The clustering method finds a group of similar proteins from different species using combinatorial optimization on a graph representing similarity relationships between proteins from different species. Thus, the underlying graph is a multipartite graph where each vertex class represents a species and protein sequences correspond to vertices in a vertex class. By suppressing the similarities between sequences within a species, it avoids detection of anciently duplicated paralogs, or similar sequences with different function in a species. The essence of the method is a function to score any arbitrary subset of sequences, then the subset with the highest score value is the ortholog cluster. So, ours is an ensemble approach for ortholog detection in contrast to the popular reciprocal blast hit approaches.

Let $V = \cup_{k=1}^{n} V_k$ be the set of all proteins from n species where V_k is the set of proteins from the species k. An arbitrary subset H of V can be decomposed as $H = \cup_{k=1}^{l} H_k$

where H_k is nonempty subset of V_k; $l \geq 2$. Then, we introduce a coefficient $\pi(i, H)$ to estimate the degree of orthologous membership of the protein i to the proteins in H:

$$\pi(i, H) = \sum_{t=1:t\neq s(i)}^{l} p(s(i), t) * \sum_{j \in H_t} m_{ij} \quad (1)$$

where m_{ij} is the Blast similarity between proteins i (from species $s(i)$) and j and $p(s(i), t)$ is a distance between the species $s(i)$ and t defined on the phylogenetic tree. Using the coefficient of similarity in (1) between a protein and a subset of proteins from other species, any arbitrary subset H of proteins from multiple species is associated with a score, $F(H)$, that quantifies the strength of the orthologous relationship among proteins in H.

$$F(H) = \min_{i \in H} \pi(i, H) \quad (2)$$

Then, a candidate ortholog cluster H^* is defined as the subset that has the maximum score over all possible subsets of proteins from the set of all proteins, V.

$$H^* = \arg \max_{H \subseteq V} F(H) \quad (3)$$

Although this optimization problem is hard in general, for the formulation given here, there is an efficient algorithm that runs in time $O(|E| + |V| \log |V|)$ where E is the set of edges in the multipartite graph [4]. This procedure outputs one ortholog cluster which is removed from the set of sequences and new ortholog cluster is found in the remaining set allowing an iterative procedure to find all clusters [4].

Criterion for Function Annotation : Annotating a query sequence with an ortholog cluster requires finding an ortholog cluster whose proteins are orthologous to the query protein. In an extreme case, it may be desirable to reconstruct the ortholog clusters considering the query sequences as part of the input data. On the other hand, if the annotation criterion is strong enough, one would expect that annotated sequences to be extensions of ortholog clusters.

To annotate a target sequence, we measure its aggregate similarity to all the sequences in a cluster. A protein sequence is annotated with the ortholog cluster whose average Blast similarity is highest. Furthermore, if this average e-value is worse than $1e^{-20}$, no annotation is assigned.

3. Results

The set of proteins in the partially completed genomes from *Zea mays* (3,138 protein sequences), *Sorghum bicolor* (468), *Triticum aestivum* (1,693), and *Hordeum vulgare* (1,112) were downloaded from PlantGDB (http://www.plantgdb.org/). The *Arabidopsis thaliana* (26,639) and *Oryza sativa* (61,250) proteomes were downloaded from MIPS (http://mips.gsf.de/proj/thal/db/) and TIGR (ftp://ftp.tigr.org/pub/data/) respectively.

As the observed similarity between orthologs from recently diverged species is larger relative to that between orthologs from anciently diverged species, we corrected the observed similarities using a phylogenetic distance as in (1). This distance between a pair of species is defined as the height of the subtree containing those species. We also used the Pfam (http://pfam.wustl.org) database as a source of annotations for the protein sequences to evaluate the quality of our clustering and function annotations.

Applying the clustering method to the 33,227 sequences from the five plant genomes, we found 1,440 ortholog clusters containing 10,785 sequences. There are 49 candidate clusters that contain sequences from all 5 species. Most (33 clusters contain more than 10 sequences) of these clusters are large and contain more than one protein sequence (paralogs) from each species. A manual inspection of existing annotations for sequences (available as part of the input data) within a candidate ortholog cluster reveals their consistency and that they are involved in some of the life critical processes such as *transcription* and *RNA synthesis*, and to plants specific pathways such as *chlorophyll, phytochromes, gibberlin, starch branching enzymes* etc. This confirms the obvious: genomic regions known to contain sequences of interest are likely to have been the first sequenced, and so these well-studied proteins from partially complete genomes are critical to function annotation.

We were able to annotate 15,523 proteins out of the 61,250 proteins in rice using 1,164 of 1,440 ortholog clusters. Clearly, more than one rice protein was annotated with most clusters which is consistent with the widespread duplications in the cereal genomes. An assessment of the quality of annotation using Pfam annotations shows that 92% of annotated rice sequences are consistent in Pfam annotation with their respective clusters and more than 5% sequences and their corresponding clusters did not have any Pfam annotation associated with them. Almost a half (7,662) of the sequences are annotated by 22 clusters related to large families of genetic elements, such as transposable elements (8 clusters) and retrotransposons (10 clusters including a polyprotein related cluster that annotates 1425 sequences) known to be widely present in plants.

References

[1] S. Altschul *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–3402, 1997.

[2] L. Koski and G. Golding. The closest blast hit is often not the nearest neighbor. *J Mol Biol*, 52:540–542, 2001.

[3] R. Tatusov, E. Koonin, and D. Lipmann. A genomic perspective on protein families. *Science*, 278:631–637, 1997.

[4] A. Vashist, C. Kulikowski, and I. Muchnik. Automatic screening for groups of orthologous genes in comparative genomics using multiple-component clustering. Technical Report 2004-33, DIMACS, 2004.