Visualization of Logic Analysis of Phylogenetic Profiles (LAPP)

Kay A. Robbins
Department of Computer Science
Cajal Neuroscience Research Institute
University of Texas at San Antonio
krobbins@cs.utsa.edu

Li Zhao
Department of Computer Science
University of Texas at San Antonio
lzhao@cs.utsa.edu

Abstract

Recently Bowers et al. [1] analyzed triplet logic relationships among 4873 Clusters of Orthologous Groups (COGS) from 67 fully sequenced organisms by calculating how well logic relationships between proteins a and b predicted the presence or absence of protein c (the uncertainty). The log of the normalized uncertainty distribution follows an approximately linear relationship for uncertainties in the interval [0.1, 0.9]. Using fitted parameters of this relationship as a characterization, we develop four types of visual analysis for LAPP data: distributions of uncertainty over logical relation type, distributions of uncertainty over functional categories, relationships of uncertainty of the overall population to known network relationships of a particular organism, and relationships of uncertainty distributions to groups obtained by standard clustering techniques. The purpose of this study is two-fold: to better understand the implications of uncertainty predictions for automatic protein network generation and to create new visualization tools for looking at this type of data.

1. Introduction

Structurally similar genes or homologs can arise from duplication events within the same genome (paralogs) or by evolution from a common ancestral gene during speciation (orthologs) [2]. The COG (Clusters of Orthologous Groups) system [2, 3] is a database of 138,458 proteins clustered into 4873 orthogolous groups. A profile for a particular COG representing an orthologous group is the binary vector representing the presence(1)/absence(0) of a member of the COG for each genome. Phylogenetic profiling infers functional relationships between pairs of COGs by examining the similarities of their profiles [4].

Recently Bowers et al. [1] examined triplet logic relationships among 4873 COGs from 67 fully sequenced organisms by calculating how well 8 different symmetric logic relationships *f*(*a*, *b*) between groups *a* and *b* predicted the presence or absence of group *c* using uncertainty as a metric. Here uncertainty is defined by:

$$U(x \mid y) = [H(x) + H(y) - H(x, y)] / H(x)$$

where *H* is the entropy [5]. *U* has a value between 0 and 1, where values close to 1 indicate that *y* predicts *x* with a high degree of certainty. For pairwise uncertainty, *x* and *y* are COG profiles. For triplet uncertainties, *x* is a COG profile and *y* is the profile of a logic relationship such as *f*(*a*, *b*) = *a* AND *b*. The authors show examples of network predictions based on logic relationships that have high values of uncertainty. For the 4,873 profiles represented in the COG database, there are 925 billion triplet logic combinations. For this reason, most of the computations in [1] were based on samples of 750,000 triplets. The authors of [1] suggest that this type of logic analysis of phylogenetic profiles (LAPP) could be extended to higher order logic relationships and to other types of genomic data. The potential of LAPP for automating the discovery of probable protein network relationships motivated the current work.

2. Methodogy

For this work we used the COG profiles of [1], consisting of vectors representing presence/absence for 67 species over 4873 groups. We began by directly assessing the characteristics of the COG profiles and the similarity of the profiles within the database. Profiles with a large number of 1's represent families that are highly conserved. 29% of the profiles have at least twenty 1's in the profile. Among the 4873 COG profiles there were 3685 unique profiles. The duplicate profiles fell into 325 distinct classes with an average of 4.7 duplicates per class. The maximum number of duplicates per class was 268. For this short study, we restrict our analysis to the 1513 genes that fell into these classes, reasoning that it was more likely to find network relationships among COG groups that fell within these classes.

Fig. 1 compares the distribution of pair-wise distances between the 325 COG profiles that represented duplicates (black histogram) with the overall distance distribution using number of bit differences as the distance metric. We note that the 72% of the pair-wise distances among the 325 profiles representing duplicates were less than 20 as compared with 45% when all of the unique profiles were considered, suggesting a greater likelihood of meaningful similarities among profiles with duplicates.

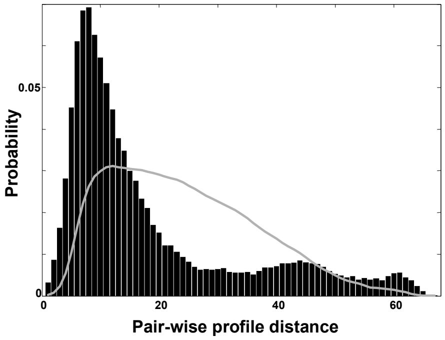

Figure 1. Distribution of pair-wise distances (bar graph) for duplicate profile classes overlaid by pair-wise distance distribution for all unique profiles (gray).

Fig. 2 compares pair-wise uncertainty of the overall distribution with that among the 325 duplicate classes. A greater percentage of the pair-wise uncertainties among the 325 duplicates are above 0.6 than for the distribution as a whole.

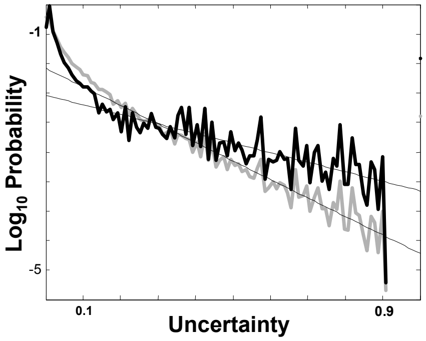

Figure 2. Distribution of pair-wise uncertainty for duplicate profiles (black) versus pair-wise uncertainty for entire data set (gray).

The log of the normalized uncertainty follows an approximately linear relationship in the interval [0.1, 0.9] as illustrated in Fig. 2. We use fitted parameters of this relationship to compare behavior of uncertainty over logical relation types and over functional categories for the population as a whole and within the duplicate classes.

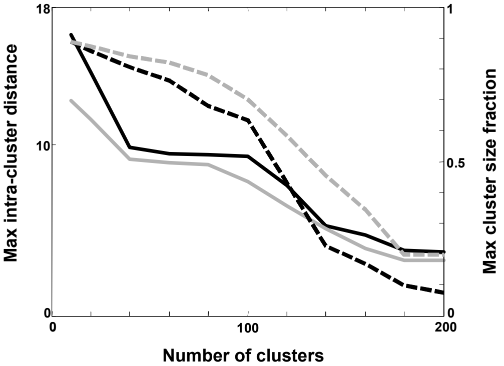

Figure 3. Maximum intra-cluster distance among unique duplicate profiles (black) and duplicate profiles (gray) as well as maximum cluster sizes (dashed).

When clustering is applied to the phylogenetic profiles, the intra-cluster distance is between 20 and 30, with most of the profiles falling into a single cluster. The intra-cluster profile distance among the unique duplicates is much smaller, although the largest cluster still contains most of the profiles.

Finally, to examine uncertainty relationships for a particular genome, we selected Saccharomyces cerevisiae or baker's yeast (organism 65 in the Bowers study) because its gene network is well-documented and because extensive expression-level from microarray experiments is available.

Acknowledgements

This work was partially supported by NIH (2-G12 RR13646) and an NSF/NIH collaborative grant in computational neuroscience NSF (EIA-0217884).

References

[1] P. M. Bowers, S. J. Cokus, D. Eisenberg and T. O. Yeates, “Use of logic relationships to decipher protein network organization,” *Science*, 2004, 306:2246-2249.

[2] R. L. Tatusov, E. V. Kookin, D. J. Lipman, “A genomic perspective on protein families,” *Science*, 1997, 279:631-637.

[3] R. L. Tatusov et al., “The COG database: an updated version includes eukaryotes,” *BMC Bioinformatics*, 2003, 4:http://www.biomedcentral.com/1471-2105/4/41.

[4] M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, T. O. Yeates, “Assigning protein functions by comparative genome alaysis: Protein phylogenetic profiles,” *PNAS*, 1999, 96:4285-4288.

[5] T. Cover and J. Thomas, *Elements of Information Theory*, Wiley-Interscience, 1991.