Classification of Protein Sequences into Paralog and Ortholog Clusters Using Sequence Similarity Profiles of KEGG/SSDB

Yohsuke Minowa[1] minowa@kuicr.kyoto-u.ac.jp
Toshiaki Katayama[2] ktym@hgc.jp
Akihiro Nakaya[3] nakaya@k.u-tokyo.ac.jp
Susumu Goto[1] goto@kuicr.kyoto-u.ac.jp
Minoru Kanehisa[1] kanehisa@kuicr.kyoto-u.ac.jp

[1] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan.
[2] Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 152-8552, Japan
[3] Department of Computational Biology, University of Tokyo, Kashiwanoha, Kashiwa, Chiba 277-8562, Japan.

Keywords: KEGG/SSDB, multi-domain problem, paralog, ortholog

1 Introduction

We are constructing KEGG/OC (Ortholog Clusters) from KEGG/SSDB (Sequence Similarity DataBase) [2]. KEGG/SSDB contains exhaustive protein sequence similarity scores of completed and nearly completed genomes calculated by the SSEARCH program [3]. KEGG/OC is constructed automatically from the graph analysis of searching cliques with an appropriate definition for the profiles of similarity scores. But there are two major problems to construct the current version of KEGG/OC. First, the current procedure leaves numerous singletons, which should in fact be included in the clusters of related proteins. Second, there are many clusters which contain evolutionarily unrelated proteins. These problems are mainly due to the so-called multi-domain problem. Here, we tried to overcome these problems by using the correlation coefficients for the profiles of KEGG/SSDB sequence similarity scores during classification of paralog genes, and compared our results to the current versions of KEGG/OC and COG [4]. Evaluation of clustering results was made by the degree of consistent motif structuress in the clusters.

2 Methods

The similarity profile of each gene is a vector of bit scores, which are 1 if the p-value of the sequence alignment score is lower than 1.0e-8, and 0, otherwise. The bit scores are defined for each gene against all genes in KEGG/SSDB. In the current version of KEGG/OC, we define similarity scores between two protein sequences a and b by summation of sequence alignment similarity scores between a and x, and between b and x (x is protein sequences which have significant sequence alignment similarity scores against both a and b). In this definition, even if each alignment similarity score of a-x or b-x is small, but there are numerous x's shared by a and b, then the summation of the alignment similarity score would be artificially large. In this case, genes a and b that are evolutionarily unrelated are assigned into the same cluster. In contrast, if two paralog genes a and b have few genes x's which are very similar to both a and b, these may be assigned into two different paralog clusters. To overcome

this situation, we examined correlation coefficients between two profiles to define the similarities of protein sequences.

To define ortholog clusters, we first construct paralog clusters in each organism by clique extraction from the SSDB graph whose nodes are protein sequences and whose edges are correlation coefficients of above profiles. Then each paralog cluster is treated as a single node, and multiple organisms are considered at a time where nodes are connected by symmetrical best hit relations. We apply clique extraction again to define ortholog clusters.

Evolutionarily related proteins should share at least one functional domain, therefore we examined the validity of different definitions of ortholog clusters by how many motifs they share within each cluster. First, we divided ortholog (paralog) clusters into two groups, one for sharing a Pfam motif [1] by at least 80% (high quality cluster) or not (low quality cluster). We set threshold 80%, because there are some proteins which don't have Pfam motifs, or there may be proteins which have Pfam motifs, but not detected. If protein sequences are properly classified, the proportion of clusters which share domains should be high, and the proportion of clusters which don't share domains should be low. We also checked singletons which should be assigned to clusters by comparing these domain structures.

3 Results and Discussion

Here, we show the result of paralog clusters of *Saccharomyces cerevisiae* as an example. Table 1 shows the size distributions of the clusters which are calculated by the three different methods the current method of using summation of scores, our new method of using correlation coefficients and the COG method. The proportion of the genes which belong to "high quality clusters" is the best with the correlation coefficient method, followed by COG, and the worst was the current OC (Table 2). Next, we investigated the cases where singletons should be assigned to clusters (Table 3). We compared all singletons against all "high quality clusters". The number of singletons which should be merged into these clusters are the largest with the current OC, and the smallest with the correlation coefficient method (Table 3). These results indicate that by defining paralog clusters by the correlation coefficient score of sequence similarity profiles, we can better classify protein sequences with small numbers of false positives and false negatives.

From these results, we conclude that the definition of "paralog cluster" was improved by the correlation coefficient score. But there are still some proteins which seem to be assigned to a wrong cluster. These might be caused by the clustering evaluation procedure using Pfam. Of course, Pfam doesn't cover all protein domains, and to evaluate clustering using exhaustive protein domains, we first need to construct complete ortholog clusters, define protein domains from the clusters. For this purpose, we are calculating ortholog clusters using the correlation coefficient scores for all completed or nearly completed genomes.

Table 1: The size distribution of the clusters.

	The number of clusters	Singletons	The largest cluster	The mean of cluster size
[1]CC	[2]4,278 (6,343)	3,418	(92)	(1.48)
Current OC	4,298 (6,343)	3,515	(98)	(1.47)
COG	3008 (4,781)	2,176	(90)	(1.55)

Table 2: The number of clusters which have common Pfam motifs (>= 80%) within each cluster.

	The number of clusters (without singleton)	Clusters which have common motifs	Clusters which don't have common motifs	Clusters which don't have Pfam motifs
[1]CC	860 (2,925)	646 (2,407)	30 (82)	184 (436)
Current OC	783 (2,828)	602 (2,268)	23 (133)	158 (427)
COG	832 (2,605)	643 (2,063)	13 (35)	176 (507)

Table 3: Singletons which have common motifs with a single "high quality cluster".

	singletons	Singletons which have common motifs with a single "high quality cluster"	Singletons which don't have Pfam motifs
[1]CC	3,418	211	1,770
Current OC	3,515	283	1,811
COG	2,176	241	392

[1]CC (Correlation Coefficients)
[2]The number of the clusters (the number of the genes)

Acknowledgments

This work was supported by the grants from the Ministry of Education, Culture, Sports, Science and Technology of Japan, the Japan Society for the Promotion of Science, and the Japan Science and Technology Corporation. The computational resource was provided by the Bioinformatics Center, Institute for Chemical Research, Kyoto University.

References

[1] Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M., and Sonnhammer, E.L., The Pfam protein families database, *Nucleic Acids Res.*, 30(1):276–280, 2002.

[2] Kanehisa, M., Goto, S., Kawashima, S., and Nakaya A., The KEGG databases at GenomeNet, *Nucleic Acids Res.*, 30(1):42–46, 2002.

[3] Pearson, W.R., Effective protein sequence comparison, *Methods Enzymol.*, 266:227–258, 1996.

[4] Tatusov, R.L., Koonin, E.V., and Lipman, D.J., A genomic perspective on protein families, *Science*, 278:631–637, 1997.