Bioinformatics Advance Access published January 21, 2009

Phylogenetics

Gclust: *trans*-kingdom classification of proteins using automatic individual threshold setting

Naoki Sato[1,*]

[1] Department of Life Sciences, Graduate School of Arts and Sciences, University of Tokyo, Komaba, Meguro-ku, Tokyo, 153-8902, Japan

Associate Editor: Prof. Martin Bishop

ABSTRACT

Motivation: *Trans*-kingdom protein clustering remained difficult because of large sequence divergence between eukaryotes and prokaryotes and the presence of a transit sequence in organellar proteins. A large-scale protein clustering including such divergent organisms needs a heuristic to efficiently select similar proteins by setting a proper threshold for homologs of each protein. Here a method is described using two similarity measures and organism count.

Results: The Gclust software constructs minimal homolog groups using all-against-all BLASTP results by single-linkage clustering. Major points include (1) estimation of domain structure of proteins, (2) exclusion of multi-domain proteins, (3) explicit consideration of transit peptides, and (4) heuristic estimation of a similarity threshold for homologs of each protein by entropy-optimized organism count method. The resultant clusters were evaluated in the light of power law. The software was used to construct protein clusters for up to 95 organisms.

Availability: Software and data are available at http://gclust.c.u-tokyo.ac.jp/Gclust_Download.html.

Contact: naokisat@bio.c.u-tokyo.ac.jp

1 INTRODUCTION

Protein clustering based on sequence similarity is a method of identifying orthologous proteins over various organisms. The COG (clusters of orthologous proteins) (Tatusov et al. 2003) is a classical example of such clusters. The clustering uses sequence similarity such as those represented by the E-value of BLASTP (Altschul et al. 1997). Clustering uses either similarity score or some secondary information such as bidirectional best hit. The latter is, however, difficult to define when there are many highly similar paralogs. Clustering of similarity data requires a threshold for cutoff. If a single threshold value is applied for all proteins, large clusters consisting of many proteins that share partial similarities are formed. To circumvent this, various criteria are applied to subdivide the clusters, such as quasi-clique (Fujibushi et al. 2000, Vashist et al. 2007) or phylogenetic analysis (Engelhardt et al. 2006, Chiu et al. 2006, Poptsova and Gogarten 2007). Complete phylogenetic analysis of large clusters is time- and resource-consuming, and is not practical for large-scale clustering. Instead, a heuristic such as InParanoid (Remm et al. 2001, Kersey et al. 2005) that distinguishes efficiently in-paralogs from out-paralogs was introduced. Nevertheless, clustering becomes more and more difficult as the number of genomes increases. Even though various elaborate methods are to be used in the main clustering, we always need to select, by some threshold or heuristic, initial groups of proteins that are to be analyzed further. Setting the initial threshold is not a simple matter, because some protein groups are highly conserved while others are not. For example, the photosynthetic reaction center proteins PsaA (and PsaB) in various organisms have a very low E-value such as 10^{-150}, whereas the similarity of the Mn-binding protein involved in oxygen evolution, PsbO, is about 10^{-45}. The similarity of the small protein in the photosynthetic reaction center, PsbL (39 aa), is about 10^{-10}. In this regard, another criterion such as overall length of regions having similarity is useful. Another strategy is to search orthologs with gradually lowering the threshold until the proteins of all organisms are included. If these criteria are properly used to select initial groups of proteins, further detailed clustering will be effective. This is the basic motivation of the present study.

Protein clustering has been performed separately in bacteria and eukaryotes, but *trans*-kingdom clustering remained difficult because of large sequence divergence between eukaryotes and prokaryotes and the presence of transit (or signal) sequence in organellar proteins. This hampered identification of proteins of endosymbiont origin. Mitochondria and chloroplasts (plastids) are believed to be descendents of ancestral endosymbionts related to α-proteobacteria (Lang et al. 1999, Fitzpatrick et al. 2005) and cyanobacteria (Martin et al. 2002, Sato 2006), respectively. Many of organellar proteins, either of endosymbiont origin or of eukaryotic origin, are encoded by the nuclear genome and remain unidentified (Chinnery 2003, Sato 2006). Targeting prediction is an approach of finding candidates for organellar proteins. An alternative way of identifying nuclear-encoded organellar proteins that originated from endosymbionts is a comparative genomic approach that identifies conserved hypothetical proteins in eukaryotes and α-proteobacteria, or those in plants, algae and cyanobacteria, which requires *trans*-kingdom protein clustering.

In this regard, existing comparative genomic databases are divided in prokaryotic and eukaryotic databases, such as the COG (clusters of orthologous groups of proteins) (Tatusov et al. 2003),

*To whom correspondence should be addressed.

MBGD (for prokaryotes) (Uchiyama 2003), and Inparanoid (for eukaryotes) (Kersey et al. 2005). The current COG database is composed of the traditional COG of 66 unicellular organisms and the KOG of seven eukaryotes, and is intended to infer proteins originated from the common ancestor. Unfortunately, the COG database includes almost no proteins involved in photosynthesis, because each COG cluster is a homologous protein cluster that is conserved in more than three different categories of organisms.

In the present study, I describe a method for *trans*-kingdom protein clustering that uses the following criteria. First, domain composition is analyzed to exclude multidomain proteins. Second, initial protein groups were selected by entropy-optimized organism count (EOOC) method, which is a heuristic involving E-value, overall similarity and organism count. Third, transit peptides (N-terminal extension of nuclear-encoded organellar proteins) were explicitly considered in comparing prokaryotic and eukaryotic proteins. The algorithm is implemented as Gclust software, which is available as a source code from the Gclust database web site. A preliminary version of the software was described in a previous paper (Sato et al. 2005).

2 METHODS

2.1 Preparation of sequence data

A set of genomes is defined as a dataset. The GenBank format files in the GenBank or RefSeq databases were downloaded from the GenomeNet mirror site. The *Cyanidioschyzon* (a unicellular red alga, see Matsuzaki et al. 2004) database was downloaded from the Cyanidioschyzon Genome Project server (http://merolae.s.u-tokyo.ac.jp/). The protein models of various genomes in the Joint Genome Institute (JGI) were downloaded from the JGI web site (http://www.jgi.doe.gov/). Then, Perl (version 5.8) scripts and SISEQ software (Sato 2000; now available from http://nsato4.c.u-tokyo.ac.jp/old/Siseq.html) were used to prepare a complete protein sequence file in the FASTA format and an annotation table for the entire dataset. For all-against-all BLASTP (Altschul et al. 1997: version 2.2.15 or 18) analysis, the protein sequence file was divided into 20 - 250 parts by the SISEQ software, and these were searched against the entire sequence file in parallel. The results with E-values less than 10^{-3} were obtained as a table (-m8 option) for further analysis.

2.2 Pre-processing

In the first step, all segment-to-segment homology links were assembled for each protein (called 'subject'). This includes all homology links of the subject plus all queries that present homology to the subject. Here, similarities having an E value lower than 1×10^{-6} were used for proteins with ≥100 aa. For smaller proteins with less than 100, 70 and 40 aa, the cut-off was set as 1×10^{-5}, 1×10^{-4} and 1×10^{-3}, respectively, if 'tapering mode' is used. Domain structure of each protein was then estimated by delimiting the boundary using the homologous segments as detected by the BLASTP analysis. Boundaries were successively added using a rule, in which a new boundary is allowed if it is spaced at least 50 aa apart from the existing boundaries. This simple rule gives satisfactory results for most data. Multidomain proteins were recognized as such if two domains of a single protein had links to different groups of proteins that did not share direct similarity links. Eukaryotic proteins that have an additional N-terminal domain with respect to prokaryotic homologs were considered to have a putative transit peptide. Large proteins (typically, >2,200 aa) or proteins having too many homology links (typically, >35,000) were marked as such and were destined to be excluded from the following analysis. The results were stored as an intermediate file (data.out). The flow of data processing is detailed in a supplementary material (Supplementary Table S3).

2.3 Estimation of a threshold for individual protein groups

In the second step, the data.out file and the list of organisms (org_list) were used to estimate initial valid homologs for each protein by the 'entropy-optimized organism count' (EOOC) method. Here, three measures were used, namely, E-value, overlap score, and organism count.

1. Two similarity measures, 'best score' and 'overlap score', between all protein pairs $P_i - P_j$ were calculated from the segment-to-segment homology data. The former is the best of all the segment-to-segment similarity scores (E-values) defined between a protein pair. The overlap score OS_{ij} of proteins P_i and P_j is defined using $R_i(P_j)$, the total length of segments in P_i that show similarity to P_j, and $R_j(P_i)$ as follows:

$$OS_{ij} = [\, R_i(P_j) + R_j(P_i) \,] \,/\, (\, L_i + L_j \,)$$

where L_i and L_j are the lengths of proteins P_i and P_j, respectively. E-values are not considered for calculating OS_{ij}.

2. For each protein, a two-dimensional histogram of homologs is constructed (Fig. 1A), using best score (y) and overlap score (x) as the two axes.

3. Optimal peak of the distribution was selected within the 2D histogram using information content and organism count. In an ideal case, the first peak from the top right corner was to be selected. But if the peak consists of proteins of a single organism, for example, another peak was taken as a candidate to include as many species as possible. There are many different cases of distribution, but the problem can be solved by choosing a peak that is located above the line: $y - delta\_min = x - 10$, in which *delta_min* was determined by the EOOC method based on information content and organism count (Fig. 1B) as described in Appendix (Supplementary document). Then, a boundary was defined around the peak. The homologs mapped within the boundary were taken as initial valid homologs.

2.4 Final processing

In the third step, initial valid homologs were assembled by single-linkage clustering. Duplication was resolved either by merging smaller clusters to a larger cluster or decomposing the smaller clusters depending on how many overlaps and cross-links are present between the larger and the smaller clusters. Decomposed homologs and singletons were added to appropriate clusters later. Merging and sorting were repeated to form quasi-cliques. The clusters were sorted according to their size.

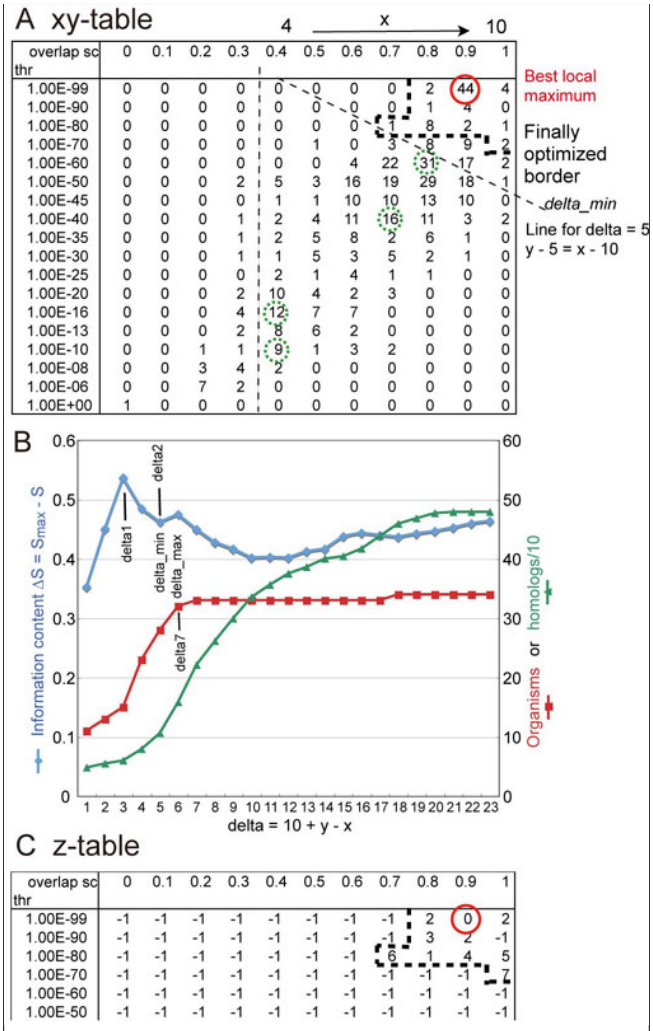

Fig. 1. Method of setting of a threshold for homologs of each protein.

A. Two dimensional histogram of homologs for initial selection (xy-table). B. Identification of the best local maximum by information content and organism count. C. A table used to delineate edges of a peak (z-table). See

In the final step, homology matrices were produced. The similarity data that were not used for cluster construction were retained as a list of relatives for each cluster. Relatives were further merged depending on organism count in re-grouping.

2.5 Implementation

The single-linkage protein clustering with individual threshold setting by the EOOC method was implemented in the Gclust software version 3.5.5z3, which was written in C and compiled by GCC.

2.6 Test data

The original protein data for constructing the COG were obtained from the NCBI web site (as of the year 2006). They were clustered by the Gclust software and the resultant homolog groups were compared with the COG. The dataset CZ36 (see **Results**) was used for the evaluation of various parameters of Gclust. This dataset can be obtained from the web site with the software.

3 RESULTS

3.1 Clustering with a single threshold

First, a simple clustering using a single threshold E-value was evaluated as a reference using the test dataset CZ36, which consists of 221,764 proteins in 14 cyanobacteria, five photosynthetic bacteria, seven plastid-containing organisms (plants and algae, plus malaria parasite), four non-photosynthetic eukaryotes, four non-photosynthetic bacteria, and two Archaea. The proteins were clustered by single-linkage clustering with different single E-value thresholds, using the Gclust software with the single threshold option implemented since its old version (Sato 2002). The results are shown in the upper part of Tables 1 and 2 (Line IDs 1-4).

Computation time increased sharply with increase in threshold E-value. E = 10^{-20} was a practical limit of processing. As the E-value threshold increased, the size of the largest cluster (Cluster 1) increased dramatically, due to inclusion of unrelated proteins that showed marginal similarity. At the same time, the number of total clusters decreased.

To evaluate the quality of clustering, some representative proteins that are conserved in plants/algae and cyanobacteria or bacteria were examined. Here, it should be noted that plant and algal chloroplast proteins are encoded by either nuclear or chloroplast genome. These include small proteins with low conservation, such as PsbU, PsbV, RpsU, and Ssl3451 homologs. MinE and RbcS are conserved proteins that are present in both prokaryotes and plants/algae. These six proteins are small proteins, which has a relatively long transit sequence, and suitable for checking *trans*-kingdom clustering (see top rows in Table 2). Ycf82 (DgdA or Slr1508 homologs ; Sakurai et al. 2007) and MgdA (Sll1377 homologs; Awai et al. 2006) are recently identified glycosyltransferases, which are difficult to classify, because there are many similar homologs with different substrate specificities. LipB was also used as a widely conserved protein in prokaryotes and chloroplasts having wide-range similarities.

It is clear that the simple single-linkage clustering resulted in incorrect groups for all these proteins. MinE, Ssl3451, PsbU and RpsU were not clustered. Other proteins formed large clusters (with asterisks). These results indicate that a single threshold at any level does not effectively classify proteins.

3.2 Entropy-optimized organism count method

Construction of *trans*-kingdom clusters using tens of organisms over different kingdoms requires, therefore, a more efficient and accurate way of finding authentic homologs. An important point to consider is the presence of transit peptide that is located at the N-terminus of nuclear-encoded organellar proteins. The sizes of transit peptides are usually less than 100 aa, but are significantly large for small proteins such as MinE (compare the sizes of homologs in *Synechocystis* and *Arabidopsis* in Table 2, top rows). For the recognition of transit peptides, targeting prediction was not

used, but a protruding N-terminal domain was assumed to be the transit peptide.

Another, more essential, point is the initial recognition of homologs. In the present study, I constructed homolog groups (House and Fitz-Gibbon 2002) rather than ortholog groups, but the number of paralogs per genome within a homolog group was kept to a minimum. Such homolog groups correspond roughly to protein families with or without known functions. To construct such minimal homolog groups, a proper threshold should be used for each group. Some protein groups are highly conserved, with a threshold E-value of less than 10^{-150}, whereas the similarity of some others is about 10^{-10}. To find highly related homologs, namely, proteins having similarity over their entire length, I introduced an additional measure, 'overlap score' as defined in Methods. Many authentic homologs (orthologs and highly related paralogs) have a high overlap score (>0.9), independent of the E-value. The identification of transit peptides as proposed in the preceding paragraph and the calculation of overlap score described here were both deduced from the homologous regions in homolog pairs (see Methods, procedure 1 in step 2).

As phylogenetic relationship of organisms to be analyzed is not apparent *a priori*, the only feasible search for authentic homologs is to count organisms in the list of homologs arranged in the order of similarity (Fig. 1), and the selection is stopped if homologs of nearly all organisms are included. The criterion for 'nearly all' is dependent on various cases in the real situations. In an extreme case, in which some proteins are highly similar to one another while many others are divergent, then we must put the first ones into a single group excluding the others, even though the number of organism count is not enough. In another extreme case, in which many homologs in various different organisms are highly similar, it is necessary to select many paralogs per genome. The number of homologs and the organism count should be compromised depending on cases. A heuristic, called 'entropy-optimized organism count' or EOOC method, to solve this problem is detailed in **Appendix** (Supplementary document). The use of organism count in selecting initial homologs could reduce the computational load of subsequent clustering, irrespective of which algorithm is to be used in the main clustering. In the present study, single-linkage clustering (with post-processing) rather than more sophisticated clustering was used to evaluate the effectiveness of EOOC method.

After the single linkage clustering with quasi-clique modification, some clusters remain split. Regrouping of such split clusters was attempted by using the initial homology information before the homolog selection. If two groups had significant number of similarity links to each other but not to others, they were candidates for joining. The joining can be achieved by various different methods such as phylogenetic analysis, but, for simplicity, it was done by the Jensen-Shannon divergence measure with respect to organism count in the present study. Namely, if the two groups having enough similarity links contained different sets of organisms, they were joined. This led to maximize number of organisms in each minimal homolog group.

3.3 Evaluation of EOOC method

The dataset CZ36 was again used to evaluate the EOOC method using the Gclust software in the clique mode (lower part of Tables 1 and 2). The use of EOOC method reduced both total number of clusters and the size of the largest cluster. The former was due to reduction in singletons. The number of clusters except singletons was rather stable at about 25,000-28,000 with various parameter settings, and was significantly larger than that in single threshold clustering. The clustering was dependent on *num_thr*, the total number of divisions for E-value in assessing the 2D histograms. The size of the largest cluster was reduced to 556 with 14 divisions without tapering option, and to 814 with 14 divisions with tapering option. Total clusters were around 83,458 - 88,086 without regrouping. After regrouping, some small clusters became merged, and the number of total clusters was reduced to 77,837 – 82,124.

The quality of clustering was evaluated using various known protein families as described above. The EOOC method gave reasonable clusters in most cases. Regrouping resulted in a more complete homolog groups. The effect of tapering option was difficult to assess. It resulted in larger clusters of DgdA and MgdA, whereas smaller clusters were obtained for PsbU, RpsU and LipB. The number of divisions in the 2D table (*num_thr*) had some effects. Divisions of 10 gave inefficient clustering of DgdA and PsbV. Supplementary Table S5 describes taxonomic distributions of protein families that are most difficult to classify, which were clustered with different parameters. Among all the data, the results obtained by the clique mode with 14 divisions with (ID = 12, 13) or without (ID = 6) tapering option were reasonable. The options must be adjusted for different datasets.

3.4 Characteristics of Gclust clusters with respect to other clusters

The Gclust clustering was also tested using the original sequence data for COG. The dataset includes 192,988 proteins of 52 eubacteria, eleven Archaea and three unicellular eukaryotes. The COG database includes 4,873 clusters with 144,320 proteins. The remaining 48,668 proteins have not been used for constructing the COG database, because they are not conserved in more than or equal to three organism groups. Gclust clusters were constructed from the identical 192,988 proteins, and obtained 56,836 clusters. 36,665 of them were singletons, and 8,212 were doublets. The 144,320 proteins of the COG database were included in 24,236 Gclust clusters. Among them, 12,069 were singletons (51.1% by clusters, 9.1% by sequences). The Gclust clusters were generally smaller than the COG clusters (Fig. 2 and Supplementary Table S4). Namely, in 11,689 Gclust clusters (49.5% by clusters, 75.8% by sequences: the number without singletons), all members (see the sector marked '100%') are related to a single COG cluster (Fig. 2A). In contrast, 2,127 COG clusters (43.6% by clusters, 23.0% by sequences) have a single Gclust counterpart (Fig. 2B). In COG0500 (SAM-dependent methylase), for example, the 786 members are dispersed in 252 different Gclust clusters (Supplementary Table S4, sheet C). These are highly divergent family of proteins with different substrate specificities. In the COG database, all proteins having a methylase domain are considered as members of COG0500, whereas the Gclust software distinguishes differences in the overall structure. That is why some large COG clusters correspond to many Gclust clusters. These results suggest that the Gclust clusters are useful in finding conserved hypothetical proteins (Galperin and Koonin 2004), because they correspond roughly to single functional groups.

Table 2. Effects of various modes and parameters on cluster size for representative genes introduced into plants and algae by endosymbiosis.

	Gene		*minE*	*ssl3451*	*rbcS*	*dgdA*	*mgdA*	*psbU*	*psbV*	*rpsU*	*lipB*
	Protein size	*Syn*	97	81	113	389	479	131	160	60	227
		Cya	none	none	cp: 138	cp: 371	none	154	cp: 150	106	265, cp:195
ID		*Ara*	229	108	174-181	none	none	none	none	183, 185	235, 276
1			1	1	6	15	15	3	8	1	15
2			1	1	8	15	15	3	13	1	27
3			6	1	8	15	16	6	14	1	**678**
4			14	6	24	**23115**	**115**	11	14	9	**23115**
5			22	15	30	14	16	14	18	28	43
6			22	15	30	14	18	16	18	32	43
7			22	15	30	14	18	16	18	32	43
8			21	15	30	1	15	13	6	26	41
9			21	15	30	1	15	13	16	28	41
10			21	15	30	1	**40**	13	16	28	41
11			21	15	30	20	18	13	18	26	41
12			21	15	30	20	20	13	18	28	41
13			21	15	30	20	20	13	18	28	41
14			21	15	30	16	18	13	18	26	41
15			21	15	30	16	20	13	18	28	41
16			21	15	30	16	20	13	18	28	41

dgdA and *mgdA* in *Synechocystis* are *slr1508* and *sll1377*, respectively. These names were not included in the GenBank database at the time of construction of CZ36.

Bold face indicates unusually large clusters formed by merging different protein groups.

Syn, Synechocystis; Cya, Cyanidioschyzon; Ara, Arabidopsis.

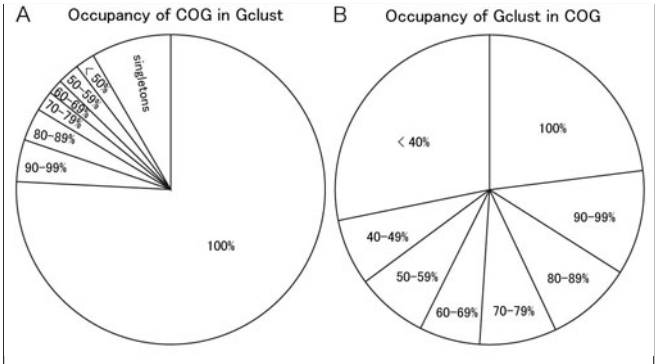

Fig. 2. Comparison of Gclust with COG.

The sequences used in constructing COG was clustered with Gclust. The clustering was performed with the tapering option and with *num_thr*=14. A, Statistic of occupancy of the most abundant COG cluster within each Gclust cluster; B, Statistic of occupancy of the most abundant Gclust cluster within each COG cluster. The number is expressed in percentage of total number of sequences in each cluster. In A, clusters having a single se-

3.5 Statistical evaluation of the clusters

The results of various clustering methods were analyzed by double logarithmic plot. Figure 3A-C shows the result with the CZ36 dataset. Clustering with a single threshold yielded clusters that follow a single power function (Fig. 3A), except the topmost and lowermost values. The exponent was -2.06 ± 0.05 with various different threshold values. The clusters obtained by the EOOC method with regrouping, however, consisted of two groups that were characterized by two different exponents (Fig. 3C, see also Supplementary Fig. S4), -1.97 ± 0.02 and -3.26 ± 0.25, respectively. The cluster size corresponding to the transition point was 35.5 ± 2.9, which roughly matches the number of organisms (36 in this case) in the dataset. The biphasic character was less evident with data without regrouping, giving an exponent of -2.12 ± 0.03 above the transition. These results indicate that the protein clusters can be regarded essentially scale-free (Albert 2005), but the property changes at the homolog per organism ratio of 1.0.

The clustering of COG was essentially similar (Fig. 3D and E). The result obtained by EOOC method was approximated by two power functions, having exponents -1.716 and -3.606, respectively. The transition was observed at a cluster size 56.3, again roughly similar to the number of organisms, 66. In the original COG clusters, the exponents were -1.136 and -2.503, respectively. The frequency of small clusters was limited, probably because COG cluster is shared by at least three groups of organisms (Tatusov et al. 2003). Despite various differences, the transition point was at 52.3, which was similar to the corresponding value obtained by Gclust.

4 DISCUSSION

4.1 *Trans*-kingdom homolog groups

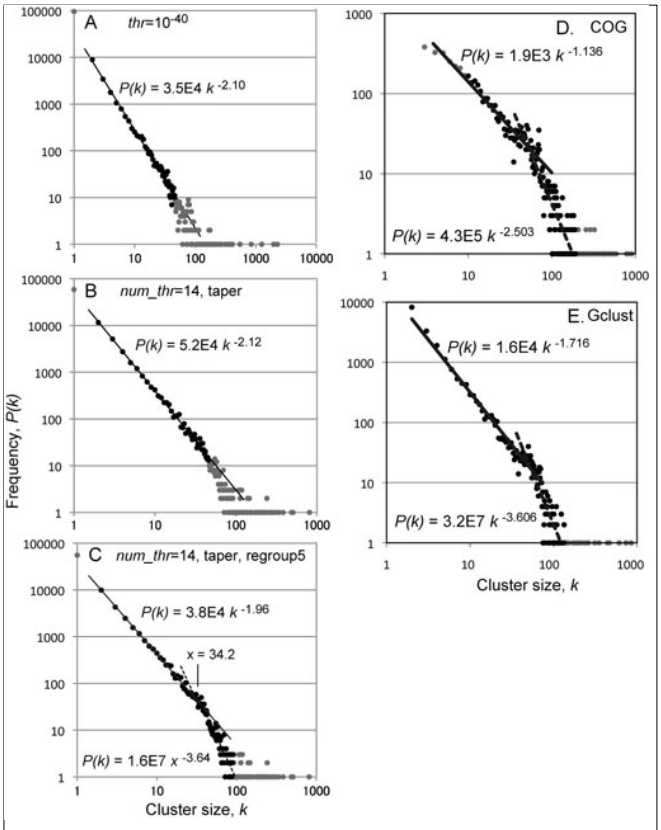

Fig. 3. Statistics of cluster size distribution.

The frequency distribution of cluster size is presented. A-C, CZ36 data. A, single threshold clustering with E = 10^{-40}; B, clustering by the EOOC method without regrouping; C, clustering by the EOOC method with regrouping at level 5. D and E, COG data. D, original COG; E, clusters con-

Trans-kingdom (or *trans*-domain) homolog comparison is important for phylogenomics (Sato et al. 2005, Galperin and Koonin 2004, De Crécy-Legard and Hanson 2007) involving endosymbiotic evolution. In order to find hypothetical proteins that are conserved in organisms performing oxygenic photosynthesis, we need to compare cyanobacteria, photosynthetic eukaryotes, and non-photosynthetic reference organisms. In this respect, we need a new database of homologs.

Such *trans*-kingdom or *trans*-domain comparison has been difficult, for the following reasons: (1) Conservation of protein sequences is higher within each kingdom, domain or phylum than among different kingdoms, domains or phyla. (2) There are families of proteins with high conservation and families of proteins with higher rates of evolution. (3) Eukaryotic organellar proteins have a transit sequence, which is not present in the bacterial homologs. This is especially the problem in clustering small proteins, in which the transit sequence is comparable in length with or longer than the mature sequence. The problem of transit sequence can be solved by explicitly considering the N-terminal extra sequence. However, this has not been done in previous protein clustering. EOOC method was developed to solve the two other problems (2) and (3). The results of clustering CZ36 dataset (Table 2) indicated that the clustering by the EOOC method gives

a satisfactory result in most protein clusters. The software has been tested to construct protein clusters for up to 95 organisms.

4.2 Scale-free property of homolog clusters

The results of statistical analysis in Fig. 3 indicate that the homolog clusters are scale-free. Scale-free nature of network has been known in metabolic network and protein-protein interaction network (see Albert 2005 for a review). A scale-free network is characterized by an exponent γ in the distribution function $P(k) = A\ k^{-\gamma}$. Although a set of protein clusters are different from a network, clustering with a single threshold resulted in a scale-free clusters with γ = 2.06. In the clustering by the EOOC method with regrouping, γ =1.97 if $k \leq N_0$, where k was the cluster size and N_0 was the number of organisms. In the region where $k > N_0$, γ = 3.26. This transition can be regarded as a result of natural property of protein clusters, because the plot of COG data (Fig. 3D) also showed such a transition. Since the quality of the clusters was excellent in the data for Fig. 3C as judged from the results in Tables 1 and 2, we can regard the plot in Fig. 3C as representing the natural distribution of homolog clusters. The value of γ = 2 is characteristic of hierarchical network of embedded modularity (Ravasz et al. 2002), whereas γ = 3 is characteristic of growing network with preferential attachment (Barabási and Albert 1999). However, it is difficult to give a meaning to the γ values observed in protein clustering.

4.3 Use of homolog groups in phylogenomic discovery

The protein clusters can be used for phylogenomic analysis of conserved proteins (Pellegrini et al. 1999, De Crécy-Legard 2007). A preliminary version of such analysis for the discovery of chloroplast proteins of endosymbiont origin was published (Sato et al. 2005). The use of the current clustering method will provide a tool for large-scale discovery of conserved hypothetical proteins (Galperin and Koonin 2004, Sjölander 2004).

ACKNOWLEDGEMENTS

The author thanks Mr. N. Sasaki in the laboratory for his helpful discussions.

Funding: This work was supported in part by Grants-in-Aid for Scientific Research (MEXT, Japan) (nos. 17018010, 18017005, 20017006 and 16GS0304).

REFERENCES

Albert,R. (2005) Scale-free networks in cell biology. *J. Cell Sci.*, **118**, 4947-4957.

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

Awai.K. *et al.* (2006) Comparative genomic analysis revealed a gene for monoglucosyldiacylglycerol synthase, an enzyme for photosynthetic membrane lipid synthesis in cyanobacteria. *Plant Physiol.,* **141**, 1120-1127.

Barabási,A.-L. and Albert,R. (1999) Emergence of scaling in random networks. *Science,* **286**, 509-512.

Chinnery,P.F. (2003) Searching for nuclear-mitochondrial genes. *Trends Genet*., **19**, 60-62.

Chiu,J.C. *et al.* (2006) OrthologID: automation of genome-scale ortholog identification within a parsimony framework. *Bioinformatics,* **22**, 699-707.

De Crécy-Legard,V. and Hanson,A.D. (2007) Finding novel metabolic genes through plant-prokaryote phylogenomics. *Trends Microbiol.*, **15**, 563-570.

Engelhardt,B.E. *et al.* (2006) Protein molecular function prediction by Bayesian phylogenomics. *PLoS Comput. Biol,*. **1**, e45, 0432-0445

Fitzpatrick,D.A. *et al.* (2005) Genome phylogenies indicate a meaningful α-proteobacterial phylogeny and support a grouping of the mitochondria with the Rickettsiales. *Mol. Biol. Evol.*, **23**, 74-85.

Fujibushi,W. *et al*. (2000) Automatic detection of conserved gene clusters in multiple genomes by graph comparison and p-quasi grouping. *Nucleic Acids Res.,* **28**, 4029-4036.

Galperin,M.Y. and Koonin,E.V. (2004) 'Conserved hypothetical' proteins: prioritization of targets for experimental study. *Nucleic Acids Res.*, **32**, 5452-5463.

House,C.H. and Fitz-Gibbon,S.T. (2002) Using homolog groups to create a whole-genomic tree of free-living organisms: an update. *J. Mol. Evol.,* **54**, 539-547.

Kersey,P. *et al*. (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.,* **33**, D476-D480.

Lang,B.F. *et al.* (1999) Mitochondrial genome evolution and the origin of eukaryotes. *Annu. Rev. Genet.*, **33**, 351-397.

Martin,W. *et al*. (2002) Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc. Nat. Acad. Sci. USA,* **99**, 12246-12251.

Matsuzaki,M. *et al*. (2004) Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D. *Nature,* **428**, 653-657.

Pellegrini,M. *et al.* (1999) Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA,* **96**, 4285-4288.

Poptsova,M.S. and Gogarten,J.P. (2007) BranchClust: A phylogenetic algorithm for selecting gene families. *BMC Bioinformatics*, **8**, 120

Ravasz,E. *et al*. (2002) Hierarchical organization of modularity in metabolic networks. *Science*, **297**, 1551-1555.

Remm,M. *et al.* (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.,* **314**, 1041-1052.

Sakurai,I. *et al*. (2007) Digalactosyldiacylglycerol is required for stabilization of the oxygen-evolving complex in photosystem II. *Plant Physiol.,* **145**, 1361-1379.

Sato,N. (2000) SISEQ: Manipulation of multiple sequence and large database files for common platforms. *Bioinformatics* **16**, 180-181.

Sato,N. (2002) Comparative analysis of the genomes of cyanobacteria and plants. *Genome Informatics,* **13**, 173-182.

Sato,N. *et al*. (2005) Mass identification of chloroplast proteins of endosymbiont origin by phylogenetic profiling based on organism-optimized homologous protein groups. *Genome Informatics,* **16**, 56-68.

Sato,N. (2006) Origin and Evolution of Plastids: Genomic View on the Unification and Diversity of Plastids. In Wise,R.R. and Hoober,J.K. (eds), *The Structure and Function of Plastids*. Springer, Dordrecht, pp. 75-102.

Sjölander,K. (2004) Phylogenomic inference of protein molecular function: advances and challenges. *Bioinformatics*, **20**, 170-179.

Tatusov,R.L. *et al*. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics,* **4**, 41.

Vashist,A. *et al.* (2007) Ortholog clustering on a multipartite graph. *IEEE/ACM Trans. Comput. Biol. Bioinform.,* **4**, 17-27.

Uchiyama,I. (2003) MBGD: microbial genome database for comparative analysis. *Nucleic Acids Res.,* **31**, 58-62.