博士論文

# Computational analysis of orthologous genes: refined identification, database construction, and functional analysis

（オーソログ遺伝子のコンピュータ解析：精密な同定、データベース構築、および機能解析）

千葉啓和

# Abstract

Owing to rapid progress in sequencing technologies, an increasing number of genomes have been sequenced. To discover biological knowledge from such growing genomic data, genome comparison based on the gene orthology relation is a promising approach. In this thesis, I describe the following three computational methods for analyzing orthologous genes: refinement of ortholog clustering at the domain level; construction of an ortholog database as a platform of integrative analysis; and comparison of protein-coding and promoter regions of orthologs.

Although several computational methods have been developed to create ortholog groups, most of those methods do not evaluate orthology at the sub-gene level. In a previous method for domain-level ortholog clustering, DomClust, proteins are split into domains on the basis of alignment boundaries identified by all-against-all pairwise comparison. However, this method often fails to determine appropriate boundaries. Together with a collaborator, I have developed a method to improve domain-level ortholog classification using multiple alignment information. This method is based on a scoring scheme, the domain-specific sum-of-pairs (DSP) score, which evaluates ortholog clustering results at the domain level as the sum total of domain-level alignment scores. We developed a refinement pipeline, DomRefine, to improve domain-level clustering by optimizing the DSP score. We applied DomRefine to domain-level ortholog groups created by DomClust using a dataset obtained from the Microbial Genome Database for Comparative Analysis (MBGD). We then evaluated the results using COG clusters and TIGRFAMs models as the reference data. We observed that the agreement between the resulting classification and the classifications in the reference databases was improved at almost every step in the refinement pipeline. Moreover, the refined classification showed better agreement than the classifications in the eggNOG databases when the TIGRFAMs models were used as the reference data. Thus, DomRefine is a useful tool for improving the quality of domain-level ortholog classification among microbial genomes. Combining with a rapid domain-level ortholog clustering method, such as DomClust, it can be used to create a high-quality ortholog database that can serve as a solid basis for various comparative genome analyses.

To discover biological knowledge by utilizing growing heterogeneous data, including genomic sequences, a flexible framework for data integration is necessary. The Semantic Web provides a key technology for the flexible integration of heterogeneous data using ortholog information as a central resource for interlinking corresponding genes among different organisms. Together with my collaborators, I have constructed an ortholog database using the Semantic Web technology, aiming at the integration of numerous genomic data and various types of biological information. To formalize the structure of the

ortholog information in the Semantic Web, we have constructed the Ortholog Ontology (OrthO). While the OrthO is a compact ontology for general use, it is designed to be extended to the description of database-specific concepts. On the basis of the OrthO, we described the ortholog information from the MBGD in the form of Resource Description Framework (RDF) and made it available through the SPARQL endpoint, which accepts arbitrary queries specified by users. In this framework based on the OrthO, the biological data of different organisms can be integrated using the ortholog information as a hub. Furthermore, the ortholog information from different data sources can be compared with each other using the OrthO as a shared ontology. We showed some examples demonstrating that the ortholog information described in RDF can be used to link various biological data, such as taxonomy information and Gene Ontology. Thus, the ortholog database using the Semantic Web technology can contribute to biological knowledge discovery through integrative data analysis.

A number of studies have compared protein sequences or promoter sequences between mammalian species, which provided many insights into genomics. However, the correlation between protein conservation and promoter conservation remains controversial. Along with my collaborators, I examined both protein conservation and promoter conservation for human and mouse orthologous genes, and observed a very weak correlation between them. We further investigated their relationship by decomposing it based on functional categories, and then identified categories with significant tendencies. Remarkably, the "ribosome" category showed significantly low promoter conservation despite its high protein conservation, and the "extracellular matrix" category showed significantly high promoter conservation despite its low protein conservation. These results show the relation of gene function to protein conservation and promoter conservation, revealing that there seem to be nonparallel components between protein and promoter sequence evolution.

In summary, I developed a method for detecting ortholog groups at the domain level with higher accuracy than previous methods, and then constructed an ortholog database that can work as a platform for integrative data analysis. These works will provide a basis for a wide range of comparative analysis based on the refined orthology information, thereby enhancing biological knowledge discovery from genomic sequences. In addition, I conducted an analysis of sequence conservation in protein-coding and promoter regions, which presents a novel viewpoint of comparative analysis based on gene orthology information.

# Contents

# Chapter 1 Introduction

Identification of orthologs constitutes the basis for comparative analysis of multiple genomes. It provides not only a foundation for inferring the evolutionary history of genes and genomes but also an important clue for inferring protein functions [1]. Originally, orthologs were defined as a pair of genes diverged from the same ancestral gene by speciation, whereas paralogs are a pair of genes diverged by gene duplication [2]. Because the functions of orthologs are typically more conserved than those of paralogs, orthology relationships are often used to transfer functional annotations between organisms [3,4]. The concept of orthology has been extended from pairs of organisms to multiple organisms by clustering orthologs into ortholog groups [5]. Ortholog groups are a vital resource for comparative analysis of multiple genomes and provide a basis for the analysis of phylogenetic profiles (i.e., the presence and absence patterns of genes in genomes) [6].

Owing to rapid progress in sequencing technologies, an increasing number of genomes have been sequenced. In particular, the accumulation of microbial genome data is remarkable [7]; several thousand genomes across diverse taxa have already been sequenced, and even more data have been generated as metagenomes from various environmental samples. A reliable method for identifying ortholog groups among multiple genomes is needed for comparative analysis of this huge amount of microbial data. Although several computational methods have been developed to create ortholog groups, most of those methods do not evaluate orthology at the sub-gene level. In a previous method for domain-level ortholog clustering, DomClust [8], proteins are split into domains on the basis of alignment boundaries identified by all-against-all pairwise comparison. However, this method often fails to determine appropriate boundaries. In Chapter 2, I present a method for improving ortholog classification at the domain level using multiple alignment information [9]. Together with a collaborator, I designed a scoring scheme to evaluate the inferred domain organization on the basis of multiple alignments and developed procedures to improve the inference by optimizing the score.

In addition to genomic sequence, various types of biological data have been rapidly accumulating because of the rapid progress of biotechnology; therefore, the effective computational management of such data appears to be a challenging issue in biological data analysis. In particular, the heterogeneity of biological data makes the integration required for data analysis a significant challenge. To achieve the integration of such growing heterogeneous data, there is an urgent need for consolidating key information that links biologically related resources to each other. Among the various biological resources, ortholog information can play a central role in integrating the biological data of multiple species. Biological functions of orthologs are usually conserved [4]; thus, ortholog information is a useful resource to link the corresponding genes of different species and transfer the biological knowledge of model organisms to organisms with newly sequenced genomes. In this era where numerous novel genome sequences are being determined, the concept of such computational knowledge transfer is becoming increasingly valuable. Interlinking biological resources using ortholog information as a hub structure is a powerful approach for genomic data integration and biological knowledge discovery.

For the integration of biological data derived from different data sources, the use of the Semantic Web technology [10] is a promising approach [11,12]. In the past few years, there has been a continuous effort to apply the Semantic Web to biological databases in order to enhance their interoperability [11,13]. Restructuring the ortholog database as a hub of the biological database network based on the Semantic Web will have a significant impact for biological database integration. In Chapter 3, I present the construction of an ortholog database using the Semantic Web technology [14]. In this work, my collaborators and I proposed a general model for describing ortholog information on the basis of our novel ontology. Using this model, we expressed the ortholog data of the Microbial Genome Database for Comparative Analysis (MBGD) [15] and made them available through the SPARQL endpoint. I show several examples of SPARQL queries to demonstrate that our ortholog database could work as a hub for integrating several genomic data resources and support knowledge discovery through its search functionalities.

Comparative analysis is a powerful approach to extract functional or evolutionary information from biological sequences (reviewed in [16-18]). There were many pioneering works on the molecular evolution of mammalian protein sequences [19], which were followed by large-scale comparative analyses between species [20-22]. These studies revealed that the evolutionary rates of protein sequences depend on the protein functions. Furthermore, the complete sequences of mammalian genomes [23-25] facilitated large-scale comparisons of non-coding sequences, which provided insights about regulatory sequences [26-28].

While many efforts have been made to examine protein sequence conservation or regulatory sequence conservation, the relationships between them are poorly understood. In Chapter 4, I present comparative analysis of protein and promoter sequences for human and mouse orthologous genes aiming to elucidate what kinds of relationships exist between promoter conservation and protein conservation in mammals [29]. In this work, together with my collaborators I investigated the relationship by decomposing it based on the functional categories of genes. The results revealed that there seem to be nonparallel components between protein and promoter sequence evolution.

On the basis of the studies presented in Chapter 2–4, I will conclude the thesis and provide future directions in Chapter 5.

# Chapter 2 Refinement of ortholog clustering at the domain level

## 2.1 Background

Several previous studies have developed orthology inference algorithms and ortholog databases [30,31]. One of the most basic algorithms to identify orthologs is the bidirectional best hit (BBH) approach for a pair of species [16]. The BBH approach was extended to deal with multiple species by applying clustering methods to the graph of BBH relationships; this approach for creating ortholog groups is known as a graph-based method [5,32-35]. The Clusters of Orthologous Groups (COGs) database is a pioneering study of graph-based methods and is still one of the most popular ortholog databases, although it is no longer updated [5,32]. The eggNOG database was later constructed by extending COGs incrementally using a computational method [33]. Another approach for creating ortholog groups is based on the phylogenetic tree of genes and is called a tree-based method. Such a method produces more reliable results than graph-based methods but at the expense of higher computational costs [34,36-38]. The DomClust algorithm [8], which is used to create ortholog groups in the MBGD database, adopts an intermediate approach, where ortholog groups are identified on the basis of hierarchical clustering trees created from a graph of all-against-all pairwise similarity relationships. In prokaryotes, the prevalence of horizontal gene transfers (HGTs) makes accurate ortholog inference infeasible [39]. Therefore, a relaxed condition, i.e., closest homologs in different species regardless of HGTs, is usually used as an alternative definition of orthology for prokaryotic genome comparison [4].

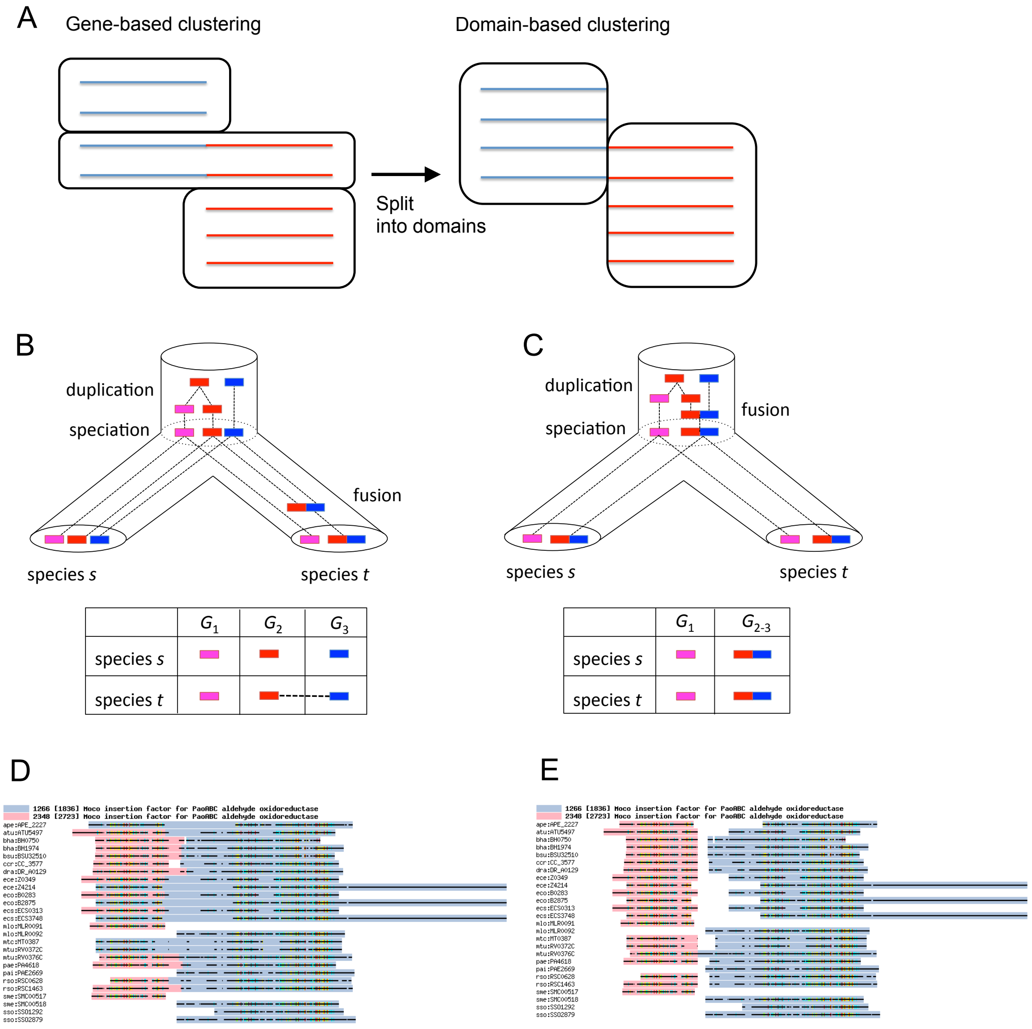
Among numerous methods proposed to create ortholog groups, only a few methods consider orthology relationships at the sub-gene level. Figure 1A is a schematic illustration of ortholog clustering at the domain level, where fusion proteins comprising originally distinct proteins are included. With a simple clustering method that does not consider sub-gene level classification, a fused protein will be assigned to exclusively one of the clusters (Figure 1A, left). However, considering that each domain in the fused protein can have a distinct function that is shared among the corresponding orthologs, a natural method of grouping them is to split the fused proteins into domains and treat them separately (Figure 1A, right). Such a clustering procedure, called domain-level ortholog clustering, is a challenging problem because not only the cluster members but also the set of fusion proteins and domain boundaries within them must be identified. Some methods such as HOPS [40] use information of known domains such as those included in the Pfam database to identify domains and then identify orthologs within each domain. However, such approaches are unsuitable for comprehensive ortholog classification of the entire set of proteins because of their dependency on the existing domain database.

The orthologous domains considered here are orthologous gene subsequences that have been stable (unsplit) during evolution after speciation from a common ancestor. To clarify the difference between orthologous domains and conventional homologous domains, let us consider the following evolutionary scenarios (Figure 1B, C). In Figure 1B, a gene fusion event occurred after speciation. In this case, the fused gene is split into two subsequences in the orthologous domain classification. In Figure 1C, a gene fusion event occurred before speciation. In this case, full-length fused genes are classified in the orthologous domain group because the fused form is stable after speciation. In either scenario, there are two homologous domain groups: one is the blue domain and the other includes both the red and pink domains that are paralogous to each other. These examples illustrate that orthologous domains can be longer than homologous domains if domain reorganization occurs before speciation.

Note that the full length of a gene can be an orthologous domain. If the domain-reorganization event after speciation is either gene fusion or gene fission, the orthologous domain should correspond to the full length of a gene in at least one of the species (Figure 1B). Thus, the orthologous domain defined here is a suitable unit for functional annotation in comparative genomics, with gene fusion/fission events taken into consideration and seems well consistent with manually curated ortholog databases such as COGs, although there are no clear-cut criteria for splitting genes into subsequences in the COG construction procedure [41]. DomClust automatically detects a domain-reorganization event and splits a cluster into orthologous domains during the process of hierarchical clustering [8].

In practical applications, the determination of orthologous domains becomes more complicated because of several factors, including insertions/deletions of promiscuous domains and random disruption of coding sequences due to loss of function. These factors fragment orthologous domains into smaller pieces than expected as a unit of functional annotation. To avoid this over-splitting problem, the DomClust algorithm tries to split genes into the minimum number of domains required for ortholog clustering, i.e., a gene is split only when a different set of genes is putatively orthologous to each split segment with sufficiently large scores [8]. Moreover, DomClust merges two adjacent domains in its final step when genes in the fission form are much fewer than those in the fusion form [8]. However, such approaches do not always work well. Figure 1D illustrates a simple but typical example, where domain boundaries determined by DomClust are inconsistent in a multiple sequence alignment. Such inconsistent alignment boundaries are problematic because they not only cause incorrect sequence grouping but also lead to failure of the above mechanisms of DomClust to avoid over-splitting. This problem arises presumably because DomClust determines the boundaries using pairwise, rather than multiple, sequence alignments. Thus, utilizing multiple alignment information supposedly improves the accuracy of domain-level ortholog clustering (Figure 1E).

In this chapter, I present a method for improving domain-level ortholog classification using multiple alignment information [9]. Together with a collaborator, I designed a scoring scheme to evaluate the inferred domain organization on the basis of multiple alignments and developed procedures to improve the inference by optimizing the score. The improvement procedures included the merge of adjacent domains to fix the over-splitting problem and determination of optimal domain boundaries. In addition, a phylogenetic tree was created for each cluster to check the cluster members in terms of orthology relation. To evaluate the improvements, we compared the obtained ortholog groups with the original ones by examining the agreement with COG and TIGRFAMs [42], which are the manually curated reference databases.



|  | $G_1$ | $G_2$ | $G_3$ |
|---|---|---|---|
| species *s* | 🟪 | 🟥 | 🟦 |
| species *t* | 🟪 | 🟥- - - - | - - - -🟦 |

|  | $G_1$ | $G_{2\text{-}3}$ |
|---|---|---|
| species *s* | 🟪 | 🟥🟦 |
| species *t* | 🟪 | 🟥🟦 |

**Figure 1. The concept and examples of domain-level ortholog clustering.**

(A) Schematic illustration of ortholog clusters containing fusion proteins. The lines represent protein sequences, and red and blue colors represent two distinct domains of the proteins. (B, C) Groups of orthologous domains in two evolutionary scenarios: the case of gene fusion after speciation (B) and gene fusion before speciation (C). (D, E) Appropriate re-splitting of proteins refines the domain-level ortholog clustering. Examples of inconsistent domain boundaries (D) and a refined version of the boundaries (E) are shown in multiple alignments, where two adjacent domains are colored in light blue and pink, respectively (see 2.2 Methods for details of the alignment visualization tool).

## 2.2 Methods

### 2.2.1 Overview of the refinement pipeline

In the study presented in this chapter, we assumed DomClust results as the input to our method, although any other domain-level clustering could have been applied. As illustrated in Figure 1A, a split of a protein sequence during domain-level ortholog clustering leads to the creation of adjacent domains that belong to different clusters (adjacent clusters). Pairs of adjacent clusters were the targets of our refinement procedure. For each pair of adjacent clusters in the input, a multiple alignment of protein sequences contained in either cluster was created and used in our refinement procedure. A domain-specific sum-of-pairs (DSP) score was introduced to evaluate the domain organization. The DSP score is based on the sum-of-pairs (SP) score [43]. However, it is calculated for each domain and inconsistencies in domain boundaries are evaluated as gaps so that the sum of the DSP scores in the alignments of adjacent clusters reflects the quality of domain classification. We defined five basic operations to modify and improve the domain organization by maximizing the DSP score and compiled them as a pipeline named DomRefine (Figure 2). The first two procedures in the pipeline (*merge* and *merge_divide_tree*) were designed to solve the over-splitting problem; *merge* determined whether two adjacent clusters should be merged, whereas *merge_divide_tree* temporarily merged the adjacent clusters and then divided them into two groups (rather than split into two domains). The next two procedures (*move_boundary* and *create_boundary*) determined the optimized boundaries between the domains; the *move_boundary* procedure moved existing domain boundaries, whereas the *create_boundary* procedure introduced new boundaries. All the four procedures improved the domain organization on the basis of the maximization of the DSP score. In contrast, the last procedure (*divide_tree*) is a type of conventional tree-based approach for ortholog classification; it divided a cluster into subgroups along with the phylogenetic tree if the subgroups shared intraspecies paralogs.

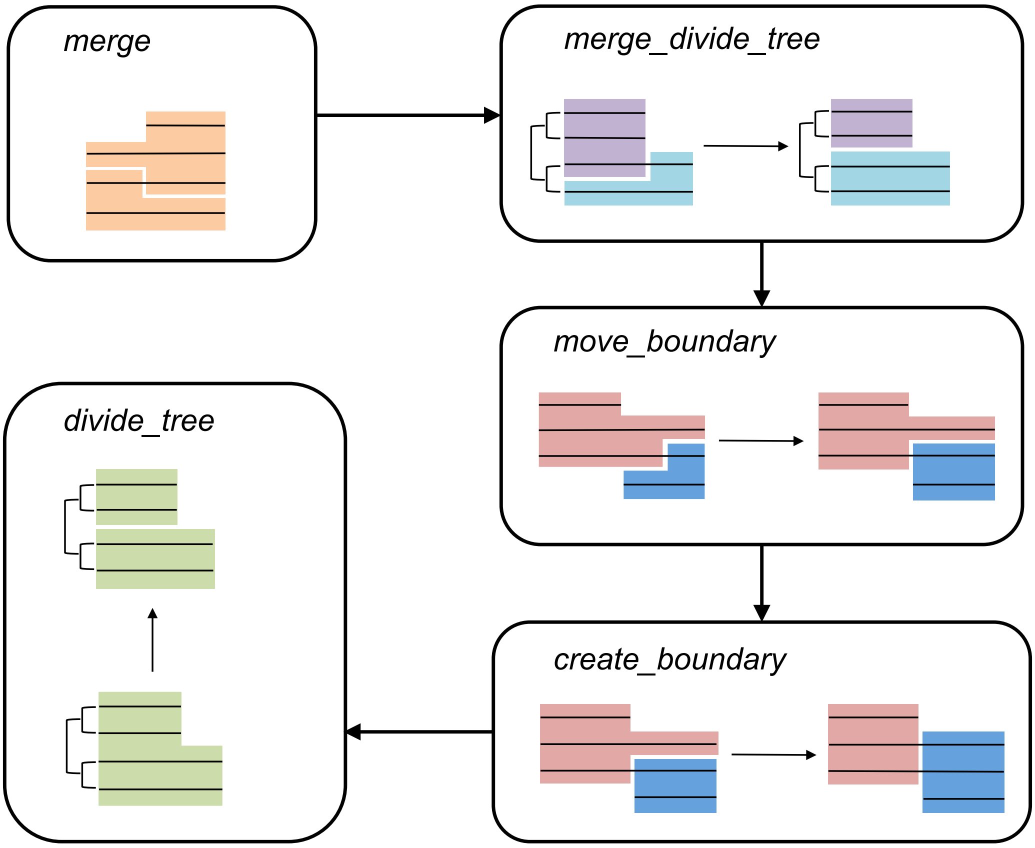


**Figure 2. The DomRefine pipeline.**

The pipeline is given a domain-level ortholog clustering result and modifies domain organizations using five procedures. Domain organizations are illustrated using the different colors. Multiple alignments of amino acid sequences are represented by sets of aligned horizontal lines. Adjacent clusters are merged if the score increases by merging the clusters (*merge*). Given a pair of adjacent clusters, adjacent domains are temporarily merged and then divided into clusters considering score changes on the phylogenetic tree (*merge_divide_tree*). Existing boundaries are moved (*move_boundary*), and new boundaries are created (*create_boundary*). When species overlap between sub-clusters on the phylogenetic tree is detected, the cluster is divided into subgroups (*divide_tree*).

### 2.2.2 Definition of the score

The DSP score is calculated on the basis of multiple alignments. The score evaluates the consistency of domain-level ortholog clusters and multiple alignments. The basic idea is the sum-of-pairs score of a multiple alignment, which is a standard measure of evaluation of protein sequence alignment [44]. The unique idea of our score is that the calculation of the sum of pairs is restricted to specific domains, and that inconsistencies in the domain boundary positions are treated as gaps. Consider the alignment in the form of matrix $A = (a_{ij})$, $i = 1, .., N_{seq}$, $j = 1, \ldots, N_{pos}$, where $a_{ij}$ represents an amino acid or a gap, $N_{seq}$ is the number of sequences, and $N_{pos}$ is the number of positions in the alignment. The positions of a domain on the amino acid sequences are also defined in the form of matrix $D = (d_{ij})$ of the same size of $A$, where $d_{ij} = 1$ if $a_{ij}$ is within the domain or otherwise $d_{ij} = 0$. The DSP score of domain $D$ in multiple alignment $A$ is given by

$$S_A(D) = \sum_{i<i'}^{N_{seq}} \left[ \sum_{j=1}^{N_{pos}} \{ s_{dom}(a_{ij}, a_{i'j}, d_{ij}, d_{i'j}) \} - n_{G_{open}}(a_{i\cdot}, a_{i'\cdot}, d_{i\cdot}, d_{i'\cdot}) G_{open} \right],$$

where $G_{open}$ is the gap-opening penalty. $n_{G_{open}}(a_{i\cdot}, a_{i'\cdot}, d_{i\cdot}, d_{i'\cdot})$ is the number of open gaps between the $i$-th sequence and $i'$-th sequence, where the open gaps are counted in the regions of $d_{ij} = 1$ and $d_{i'j} = 1$, and the mismatches of the domain terminal positions are also counted as open gaps. $s_{dom}$ is a function similar to a commonly used score matrix, but it returns a value depending on the domain as follows:

$$s_{dom}(a, a', d, d') = \begin{cases} s_{mat}(a, a'), & \text{if} \quad b(a)d = 1 \text{ and } b(a')d' = 1 \\ G_{ext}, & \text{else if} \quad b(a)d = 1 \text{ or } b(a')d' = 1, \\ 0, & \text{else if} \quad b(a)d = 0 \text{ and } b(a')d' = 0 \end{cases}$$

where $s_{mat}$ is a commonly used score matrix such as the BLOSUM score matrices, $G_{ext}$ is the gap extension penalty, and $b(a) = 1$ if $a$ represents an amino acid or otherwise $b(a) = 0$. Therefore, a higher DSP score is obtained when the domain organization is such that sequence regions similar to each other (i.e., aligned with a positive score) belong to the same domain and sequence regions dissimilar to each other (i.e., aligned with a negative score) belong to different domains, because the DSP score counts similarity scores only between sequences belonging to the same domain. If the domain boundaries are not consistent with each other in the multiple alignment, they are penalized as external gaps, decreasing the score. Thus, an increase in the DSP score denotes that the domain boundaries are more consistent with each other in multiple alignment and/or the sequences belonging to the same domain produce a higher sum-of-pairs score. To normalize the DSP score with respect to the number of sequences and sequence lengths, we divide the DSP scores or the differences in the DSP scores by $N_{seq}$ and $N_{aa}$, where $N_{aa}$ is the total number of amino acids included in the alignment.

### 2.2.3 The *merge* procedure

In the *merge* procedure, all the split proteins in the dataset are re-examined in multiple alignments. Consider a pair of clusters that share at least one common protein whose sub-sequences are members of each cluster. We define two clusters as adjacent if they have a shared protein whose sub-sequences in each cluster are adjacent to each other in the shared protein sequence. To determine whether a pair of adjacent clusters should be merged, the DSP scores are evaluated before and after the merge. First, the score is calculated for each of the clusters before the merge. Then, the clusters are merged by canceling the split between the clusters. The clusters are to be merged under the condition of the normalized score change $(S' - S)/(N_{seq}\, N_{aa}) > S_\delta$, where $S$ and $S'$ are the scores before and after the merge, respectively, and $S_\delta$ is a threshold for the merge. Following the examination of adjacent cluster pairs, all the pairs to be merged are merged at once.

### 2.2.4 The *merge_divide_tree* procedure

The *merge_divide_tree* procedure temporarily merges a pair of adjacent clusters and then divides them into two groups as a split of a phylogenetic tree. Because this procedure is preceded by the *merge* procedure, we assume that clusters that should be merged are already merged.

A motivating example of this procedure is as follows: suppose there are two domains A and B. Some proteins have both domains (domain organization A + B) and the others have only domain A (domain organization A). In this case, we may want to classify these proteins into two groups corresponding to the two domain organizations, A + B and A, instead of the original domain-level classification, A and B. The *merge_divide_tree* procedure adopts the modified classification only when the resulting subgroups are consistent with the gene phylogeny, i.e., when they correspond to a split of the gene tree, as well as when the resulting DSP score becomes higher than before.

More precisely, this procedure re-defines the two groups on a phylogenetic tree as follows. If a root of the tree is determined, two subgroups are produced. The initial domain patterns are compared between the newly defined subgroups, and the difference is quantified as follows:

$$t_{diff}(G_1, G_2, t_1, t_2) = \big||g_1 \cap t_1| + |g_2 \cap t_2| - |g_1 \cap t_2| - |g_2 \cap t_1|\big| + \big||g_{12} \cap t_1| - |g_{12} \cap t_2|\big|,$$

where $G_1$ and $G_2$ represent initial clusters, $t_1$ and $t_2$ represent newly defined subgroups, $g_{12}$ is the set of genes in both $G_1$ and $G_2$, $g_1$ is the set of genes in $G_1$ but not in $G_2$, and $g_2$ is the set of genes in $G_2$ but not in $G_1$. We calculated $t_{diff}$ for all candidate roots and selected the root showing the largest $t_{diff}$. If several candidate roots show the same value of $t_{diff}$, the root with the longest edge among them is selected. Finally, the DSP score change was calculated comparing the original and resulting states, and the modification was executed only when the score increases.

### 2.2.5 The *move_boundary* and *create_boundary* procedures

The *move_boundary* procedure moves the set of domain boundaries between two adjacent clusters at the same time, keeping them in the same column on the multiple alignment. By moving the position from the N terminus to the C terminus on the multiple alignment, the position showing the highest score is selected. If the best score is higher than the score of the initial state, the move of the boundaries is retained.

The *create_boundary* procedure creates a new boundary on candidate sequences, which are not split into domains in the initial state. Following the examination of all the protein sequences without splits, if the set of newly introduced splits increases the DSP score, boundary creation is applied.

### 2.2.6 The *divide_tree* procedure

The *divide_tree* procedure checks whether the resulting clusters contain paralogous genes using a species overlap criterion that is used in DomClust as well as several tree-based ortholog classification methods. For this purpose, using FastTree, we created phylogenetic trees on the basis of multiple alignments produced by Clustal Omega. Although the obtained tree is unrooted, the root is placed on one of the edges so that the height of the resulting rooted tree is minimized. Division of a cluster into subgroups is determined by a species overlap rule as follows: $|S_{sp}(t_1) \cap S_{sp}(t_2)|/|S_{sp}(t_1) \cup S_{sp}(t_2)| \geq R_{sp}$, where $t_1$ and $t_2$ represent candidate subgroups of the phylogenetic tree, $R_{sp}$ is a threshold, and $S_{sp}(t_i)$ represents the set of species included in $t_i$.

### 2.2.7 Dataset

The 2002 version of the COG database (COG02) contains genes from 43 species in 3307 clusters. We excluded ortholog groups comprising genes of fewer than three phylogenetically distinct organisms, retaining 3192 clusters, as described previously [8]. The 2003 version of the COG database (COG03) contains genes from 66 species in 4873 clusters [32]. Using the same filter applied to COG02, the number of clusters was reduced to 4814. DomClust was executed using the following parameters: *ao* (member overlap for merging adjacent clusters) of 0.8, *ai* (member overlap for absorbing adjacent small clusters) of 0.95, *V* (alignment coverage for domain split) of 0.6, and *C* (cutoff score for domain split) of 80. For the execution of the DomRefine pipeline, the following parameters were set: $G_{open}$ of 10, $G_{extension}$ of 0.5, $S_d$ of −0.05, and $R_{\mathrm{sp}}$ of 0.5, and BLOSUM45 was used as the score matrix $s_{mat}$. In the tests to recover COG classification by DomClust, an additional parameter was used to specify a condition that at least three phylogenetically distinct organisms must be included in each cluster, as described previously [8].

The FAMILY dataset was created using the MBGD database [15]. Using NCBI taxonomy information, one representative genome was selected from each family. The resulting number of genomes was 309. COG and NOG clusters included in the eggNOG database v3.0 [45] were concatenated and designated as eggNOG in this study. To compare eggNOG classification with our classification based on the FAMILY dataset, we compared the list of genes between the FAMILY dataset and eggNOG v3.0 using NCBI taxonomy ID for organisms and locus ID for genes and extracted the intersection of these gene sets, obtaining a total of 587,463 genes from 210 organisms. Note that the eggNOG cluster sizes in the resulting FAMILY210 dataset were reduced from the original one because the species subset was extracted.

### 2.2.8 Evaluation criteria

If overlapping fragments are observed between a COG cluster $C_i$ and a DomClust cluster $D_j$, whereas no overlapping fragments are observed between $C_i$ and $D_{j'}$ and between $C_{i'}$ and $D_j$ for any $j' \neq j$ and $i' \neq i$, then the relation of $C_i$ and $D_j$ is called a one-to-one relationship. When we have two clustering results, we can evaluate the consistency between them using the number of one-to-one relationships between them. To evaluate clustering results showing moderate agreement with the reference classification more appropriately than counting the number of one-to-one relationships, the agreement of clustering results was quantified as follows. The overlap ratio of fragment $c \in C_i$ and fragment $d \in D_j$ is calculated as $r_{over} = |c \cap d|/\max(|c|, |d|)$. The mean overlap ratio $\bar{r}_{over}$ is obtained by averaging $r_{over}$ for the overlapping fragments.

### 2.2.9 Software used

The core part of the pipeline that calculates the DSP score was implemented in the C language. Other parts of the pipeline are implemented in the Perl language. The programs were executed on Linux. DomClust [8] was used to obtain the initial clustering results. The pipeline accepts the DomClust default format, which includes the cluster members and the regions of the member domains. The DomRefine output is obtained in the same format as the input. Clustal Omega [46] was used to create multiple alignment with *auto* option. FastTree [47] was used to create a phylogenetic tree based on the multiple alignment produced by Clustal Omega. For visualizing domain-level clustering results on multiple alignments, I developed a visualization tool using Perl and the GD library (http://search.cpan.org/dist/GD/). The tool colors the amino acid residues according to the conservation rate $p_{cons}$ in the multiple alignment: red for $p_{cons} \geq 70\%$, yellow for $70\% > p_{cons} \geq 50\%$, and cyan for $50\% > p_{cons} \geq 30\%$. The scatter plot was created using R (http://www.r-project.org/). A significance test of the results obtained by the binomial test was performed using the *binom.test* function of R considering gains and losses as successes and failures in trials, respectively. TIGRFAMs release 13.0 [42] was used as protein models. For searching the protein sequences using the protein models, HMMER3 [48] was used with the "trusted cutoff " of each model. DomRefine including the visualization tool can be downloaded from the following link: http://mbgd.genome.ad.jp/domrefine/

## 2.3 Results

### 2.3.1 Illustrative examples: refinements at the domain level

Figure 3 illustrates the examples of improved domain organization obtained by DomRefine. In the original classification by DomClust (Figure 3A), several proteins are split into domains, but the splitting pattern is inconsistent in the multiple alignment. In this case, canceling those splits to merge two clusters seemed to produce better classification. Indeed, the *merge* procedure merged these clusters because of the increase in the DSP score after merge, which resulted from the gain of the SP score between the newly aligned residues in the merged alignment and the disappearance of gaps owing to inconsistent domain boundaries. Figure 3B illustrates another example where the inconsistent domain boundaries were modified to lie at more appropriate positions. As a reference, the regions determined by the TIGRFAMs models are also illustrated. In the original classification, some proteins are split into domains, but the resulting domain boundaries did not coincide with the region detected by TIGRFAMs models. In addition, two proteins that also matched the same TIGRFAMs model are not split in the original classification. The *move_boundary* procedure moved all the existing boundaries at the same time in the multiple alignment to the best position on the basis of the DSP score. The subsequent *create_boundary* procedure created new boundaries, and the creation of these boundaries increased the DSP score. As a result of these procedures, we obtained domain boundaries that perfectly matched the region detected by TIGRFAMs models (Figure 3B).

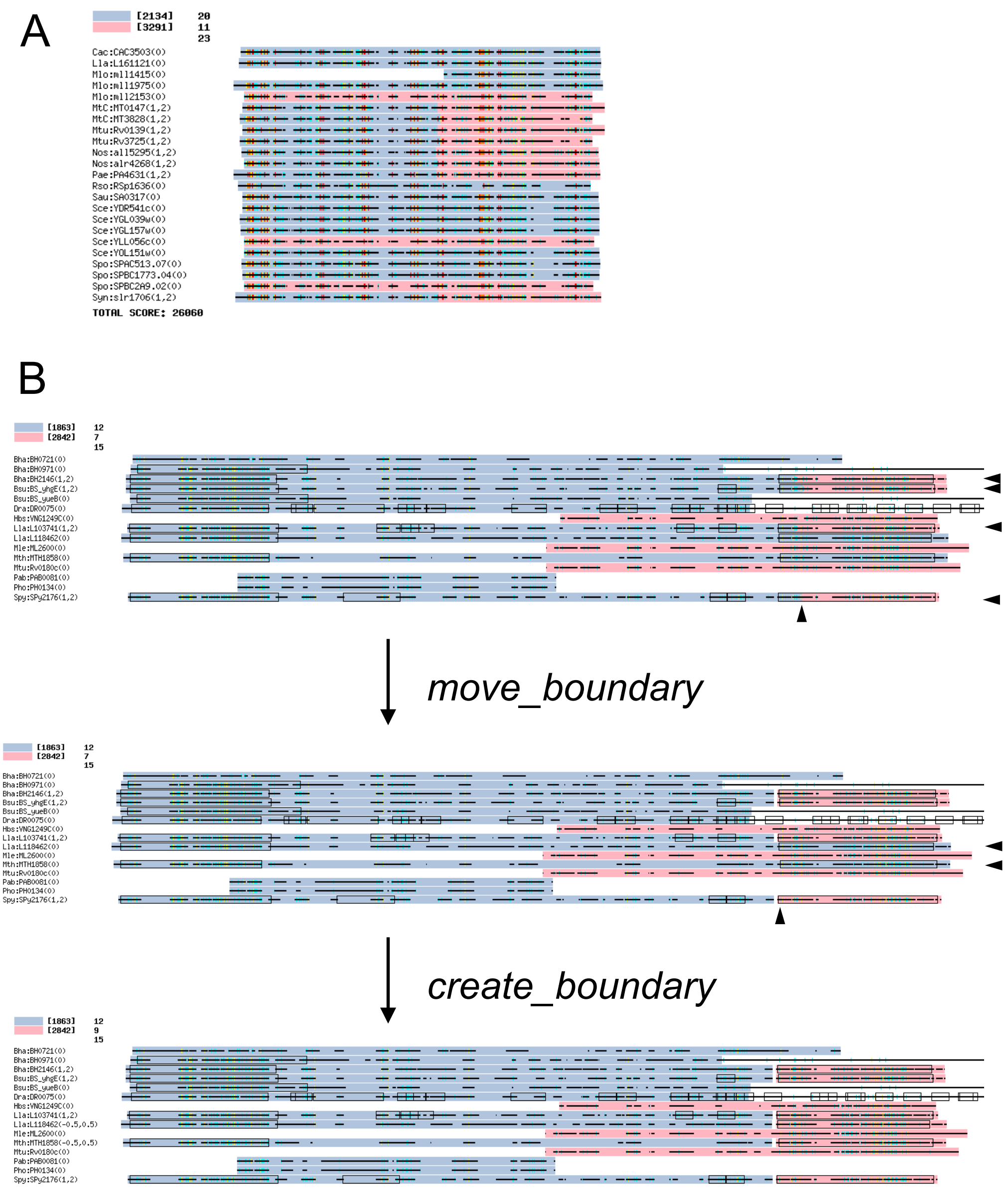


**Figure 3. Examples of improvement in domain-level ortholog clusters.**

Examples of improvement by merge (A) and move_boundary and create_boundary (B) procedures are shown with multiple alignments, where two adjacent domains are colored in light blue and pink, respectively. The arrowheads indicate the domain boundaries to be modified. The black rectangles represent the matches of the TIGRFAMs models.

### 2.3.2 Statistics of domain-level ortholog clustering results

The method was tested on proteome sets retrieved from the COG and MBGD databases. The protein sequences from the COG03 dataset (including 66 organisms) were clustered into ortholog groups by our method, and the results were compared with the manually curated COG clusters for evaluation. To test the utility of our method in a more practical situation, we also constructed a larger dataset (the FAMILY dataset including 309 organisms) by selecting a representative organism from each taxonomic family of the MBGD database. For each of the COG03 and FAMILY datasets, we first applied DomClust to classify genes into ortholog groups and then applied the DomRefine pipeline to improve the classification. For the FAMILY dataset, we compared our results with eggNOG, which was constructed by computationally extending COG. In the comparison with eggNOG, we extracted the common proteome between FAMILY and eggNOG (FAMILY210 dataset including 210 organisms).

Table 1 summarizes the statistics of the ortholog clustering results. Although DomRefine had limited effects on the total number of clusters [from 7503 to 7307 (97.4%) for COG03; from 60775 to 57644 (94.8%) for FAMILY210], it caused significant changes in the number of split clusters. For the COG03 dataset, the number of split clusters produced by DomClust alone was higher than that in the original COG, reflecting the over-splitting problem of DomClust. After DomRefine was applied, however, the number of split clusters decreased drastically [from 2439 to 1562 (64.0%)] to approximately the same number as COG. This result was in line with expectations, given that DomRefine was designed to fix over-splitting problems. Similarly, in the FAMILY210 dataset, the number of split clusters was decreased from 15879 to 10942 (68.9%). In contrast, the number of split clusters in eggNOG was remarkably small (2333, which is only 3.6% of the total number of clusters) compared with the number in COG, DomClust, and DomRefine (range, 19%–33%). In particular, the number of split clusters in eggNOG is considerably lower than that in COG, on which it is based, presumably because of the lack of a procedure for splitting clusters into domains when creating new clusters not included in COG, i.e., non-supervised orthologous groups (NOGs) during the construction of eggNOG.

For more detail, we also examined the distribution of the cluster size (the number of proteins in each cluster) (Figure 4). In general, the distributions of the cluster size show a near-linear relationship on a log–log plot, indicating that cluster sizes approximately follow a power-law distribution. For the COG03 dataset, the distributions of COG and DomClust show similar trends: the distributions deviate downward from the linear relationship at cluster sizes lower than 10 (Figure 4A) as observed previously [49]. This is because they retain only ortholog groups that have more than three members from (not closely related) different species (for results with smaller groups, see Figure 5). However, this trend is considerably prominent in COG than in DomClust, probably reflecting the feature of the COG classification that ortholog groups often contain small inparalog groups that should be separated according to a rigorous definition of orthology.

For the FAMILY dataset, the DomClust distribution follows a linear relationship in the log–log plot ($\log_{10} y = -1.499 \log_{10} x + 4.206$, $R^2 = 0.90$, Figure 4B), whereas the eggNOG distribution deviates from a linear relationship (for the fitted line, see Figure 5B). When the eggNOG clusters are separated into COG-derived clusters and NOG, their distributions are substantially different (Figure 4B, upper right). The COG-derived cluster exhibits a curved distribution, deviating downward from the linear relationship at cluster sizes lower than 100. The NOG distribution has a steeper negative slope than DomClust (for the fitted line, see Figure 5B) and deviates downward at cluster sizes greater than 10. In summary, DomClust, a fully automated clustering method, exhibited a power-law distribution in cluster size, whereas eggNOG, a combined approach of manual and automated methods, produced two different types of clusters and thus exhibited a relatively skewed size distribution.

**Table 1. Statistics of domain-level ortholog clustering results.**

| COG03 dataset | | | FAMILY210 dataset | | |
|---|---|---|---|---|---|
| **Method** | **No. of clusters** | | **Method** | **No. of clusters** | |
| | $N_{clust}$ | $N_{clust}^{split}$ | | $N_{clust}$ | $N_{clust}^{split}$ |
| **COG** | 4814 | 1389 (29%) | **eggNOG** | 64983 | 2333 (3.6%) |
| **DomClust** | 7503 | 2439 (33%) | **DomClust** | 60775 | 15879 (26%) |
| **DomRefine** | 7308 | 1562 (21%) | **DomRefine** | 57644 | 10942 (19%) |

$N_{clust}$ denotes the total number of clusters. $N_{clust}^{split}$ denotes the number of clusters that include proteins split into domains. The ratio of split clusters to the total number of clusters is shown in parenthesis.

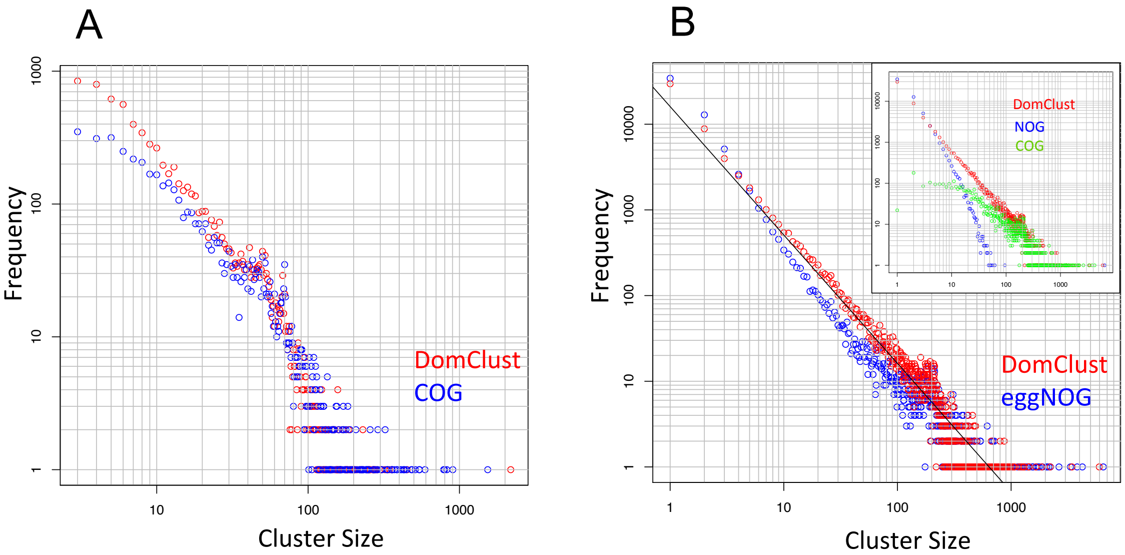


**Figure 4. Cluster size distributions of domain-level ortholog clusters.**

Clustering results for the COG03 dataset (A) and the FAMILY210 dataset (B). The red circles represent DomClust results. The blue circles represent COG data in (A) and eggNOG in the main plot area of (B). In the upper right window of (B), the eggNOG distribution is divided into COG-derived clusters (green circles) and NOG clusters (blue circles). The line represents log10 y = −1.499 log10 x + 4.206, obtained by the linear regression of the DomClust distribution on the log–log plot (B).

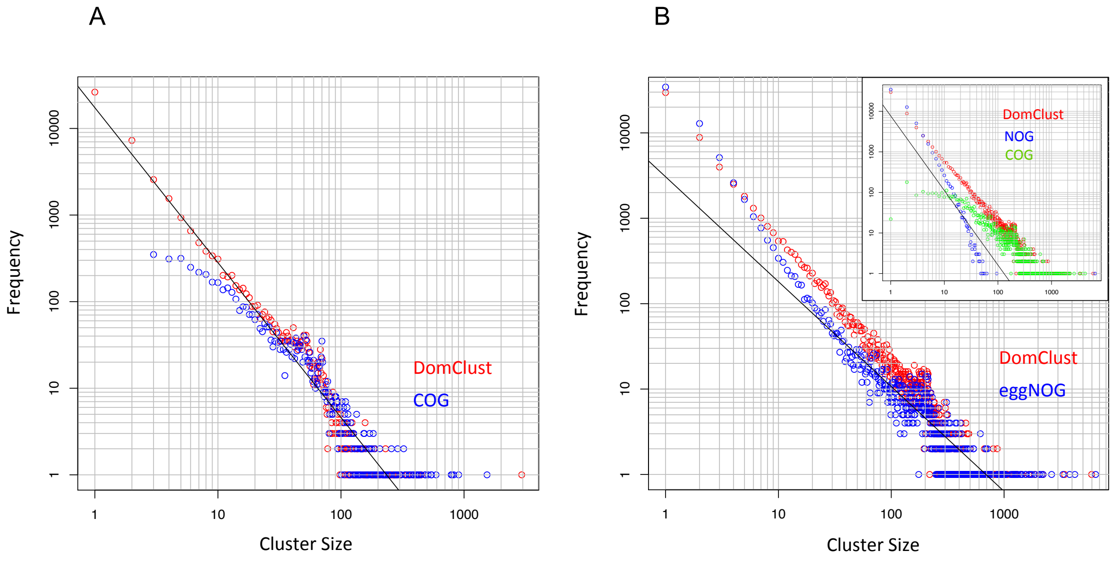


**Figure 5. Supplement to the cluster size distributions of domain-level ortholog clusters.** Clustering results for COG03 dataset (A) and FAMILY210 dataset (B). (A) This analysis is same as that for Figure 4A, except that DomClust was here executed with options which allow generation of ortholog groups with less than three members (domclust -n1 -ne1). A line is fitted to the DomClust distribution by linear regression ($\log_{10} y = -1.789 \log_{10} x + 4.240$, $R^2$=0.90). (B) This analysis is same as that for Figure 4B, but the line was fitted to the eggNOG distribution ($\log_{10} y = -1.228 \log_{10} x + 3.486$, $R^2$=0.85) and NOG distribution ($\log_{10} y = -1.838 \log_{10} x + 3.885$, $R^2$=0.71).

### 2.3.3 Assessment of the pipeline

To assess our refinement method, we examined whether our fully automated procedures could recover the manually curated COG database (COG02 including 43 organisms and COG03 including 66 organisms). To quantify the agreement of the clustering results between two methods (ours and COG) at the domain level, we first identified corresponding clusters as cluster pairs sharing at least one overlapping domain of the same protein and then extracted only those cluster pairs that had one-to-one correspondence (see 2.2 Methods for details). The number of one-to-one corresponding cluster pairs against COG ($N_{COG}^{1to1}$) was then used as an indication for the agreement between two clustering results. Figure 6 presents the changes in $N_{COG}^{1to1}$ during the DomRefine procedures. An increase in the agreement with COG was observed during the *merge* and *merge_divide_tree* procedures (Figure 6A, B). These procedures exhibited greater changes than the subsequent procedures to modify boundaries (*move_boundary* and *create_boundary*). This is probably because increasing one-to-one relationships by moving a boundary requires exact matches of boundary positions; thus, $N_{COG}^{1to1}$ is not a sensitive measure for capturing a moderate improvement in boundary positions. On the other hand, the consistency with COGs was decreased in the last procedure, *divide_tree*, which divides a cluster into subgroups to separate paralogs rather than modifying the domain organization. However, this result does not necessarily mean that *divide_tree* failed to improve ortholog classification, considering that a COG cluster often includes obvious outparalogs as members, resulting in a larger cluster than that produced by more rigorous ortholog grouping (see 2.4 Discussion).

Next, we examined the contribution of the DSP score to the refinement in the *merge* procedure. To quantify moderate agreement between two clustering results, we calculated the mean overlap ratio of corresponding domains ($\bar{r}_{over}$). For each pair of adjacent clusters, we calculated the changes in the DSP score and the changes in $\bar{r}_{over}$ after the merge for 2029 pairs of adjacent clusters and examined the correlation between them (Figure 7). A positive correlation was observed between them (Pearson's correlation coefficient $r$ = 0.51, $P$-value of <1E-15). This observation supports an assumption that the DSP score is able to quantify the quality of domain-level ortholog classification in terms of consistency using the COG database as a reference. We drew a LOWESS curve to reveal the details of the relationship between the score changes and $\bar{r}_{over}$ changes. When the score changes were positive, $\bar{r}_{over}$ changes were mostly positive (128 pairs in positive and 22 in negative). Thus, we could safely merge clusters if the resulting score change was positive. In contrast, when the score changes were negative, $\bar{r}_{over}$ changes varied, spanning positive (639 pairs) and negative (1185 pairs), meaning that some cluster pairs that should be merged may show negative score changes after the merge. In fact, the LOWESS curve demonstrated that when the score changes were small negative values, $\bar{r}_{over}$ changes were slightly positive on average (for score changes between −0.05 and 0; the mean $\bar{r}_{over}$ change was 0.06), suggesting that the threshold of the DSP score change for merging adjacent clusters should be a negative value rather than zero. This was desirable for avoiding the over-splitting problem because in this case, a domain split was introduced only when the splitting caused a sufficient score gain. On the basis of Figure 7, we used −0.05 as the threshold for the DSP score change to decide merges.

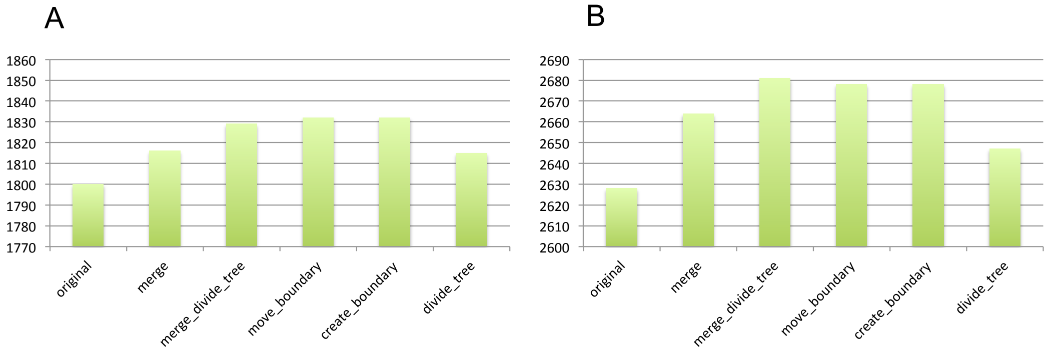


**Figure 6. Consistencies of the resulting clusters with COG clusters.**

(A) COG02 dataset and (B) COG03 dataset. Vertical axes represent the numbers of one-to-one relationships.

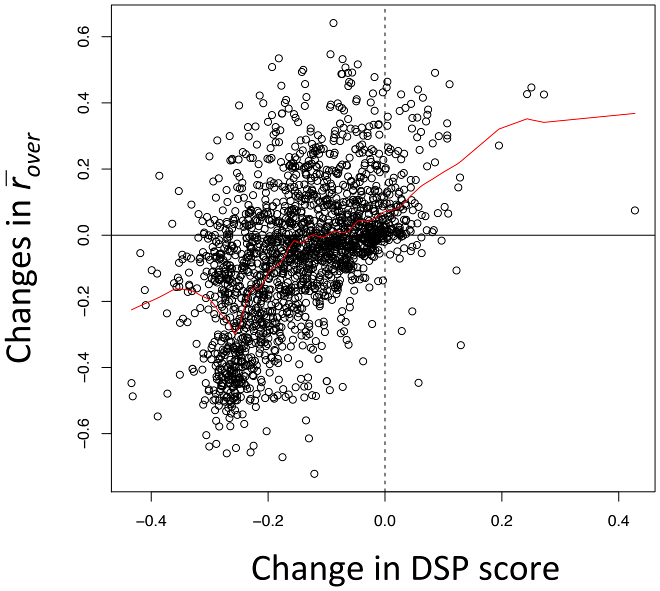


**Figure 7. Correlation of the DSP score and consistency with COG.**

Each circle represents a pair of ortholog clusters that was one of the targets of the merge procedure. The red line was drawn by the *lowess* function of R with parameter *f* = 0.1.

### 2.3.4 Practical application

To demonstrate the utility of our method in a more practical situation, we applied the method to the FAMILY dataset that covers the diversity of currently sequenced microbial genomes, in addition to the COG03 dataset. We here used the TIGRFAMs database instead of the COG database to evaluate the clustering result. TIGRFAMs is a database containing the profile hidden Markov models (HMMs) constructed from manually curated multiple alignments of functionally equivalent protein families (equivalogs) [42] with "trusted cutoff" information for searching sequences with HMM using the HMMER program [48]. Thus, TIGRFAMs can be used to classify any set of protein sequences using the HMMER program. In addition, equivalogs defined in TIGRFAMs are a suitable reference classification for evaluating our ortholog classification, in that the main aim of the ortholog classification is to infer gene functions.

We applied our method (DomClust and DomRefine) to the COG03 and FAMILY datasets to classify genes and evaluated the resulting clusters using the TIGRFAMs database as a reference. As in the previous section, we considered the number of one-to-one corresponding cluster pairs against TIGRFAMs ($N_{TIGR}^{1to1}$) as a measure of consistency between two classifications. We examined the changes in $N_{TIGR}^{1to1}$ during the DomRefine procedure (Figure 8A, B) and again observed gradual increases during the DomRefine procedures in both the COG03 and the FAMILY210 datasets. In total, $N_{TIGR}^{1to1}$ was increased from 1235 to 1272 for the COG03 dataset and from 1375 to 1448 for the FAMILY210 dataset (Table 2).

However, some differences were observed between the results of this test (Figure 8A) and that of the previous test (Figure 6B), where the same COG03 dataset was used as a classification target, but COG instead of TIGRFAMs was used as the reference database. In particular, $N_{TIGR}^{1to1}$ was increased by the *divide_tree* procedure (Figure 8A), whereas $N_{COG}^{1to1}$ was decreased in the previous test (Figure 6B). In addition, $N_{TIGR}^{1to1}$ was less increased in the *merge* and *merge_divide_tree* steps, but more increased in the *move_boundary* step. Changes in the number of one-to-one ortholog relationships, illustrated in Figure 8, were analyzed in more detail by decomposing the change into gains and losses of one-to-one relationships (Figure 9). Although occasionally a one-to-one relationship can be lost during the procedure, the gain of new relations significantly ($P < 0.05$ by binomial test) exceeds the losses in total and in most steps that have sufficient numbers of modifications (Figure 9).

To compare the classification performance, we also evaluated the COG and eggNOG classifications in terms of the agreement with the TIGRFAMs models ($N_{TIGR}^{1to1}$). For the COG03 dataset, $N_{TIGR}^{1to1}$ of the original COG classification was 1107, whereas for the FAMILY210 dataset, $N_{TIGR}^{1to1}$ of the eggNOG classification was 1149 (Table 2). Both these values were even lower than those of the

DomClust classification before refinement (1235 and 1375, respectively; Table 2). Thus, DomClust/DomRefine classifications showed better agreement than the COG/eggNOG classifications when evaluated on the basis of the agreement with the TIGRFAMs classification.

To examine the inclusion relationships between corresponding ortholog groups in different ortholog classification systems, including DomClust/DomRefine, COG/eggNOG, and TIGRFAMs groups, we considered three additional concepts, equivalent, supergroup and subgroup that were introduced in the previous work [50] (Table 3). The inclusion relationships among them tend to be COG > DomClust/DomRefine > TIGRFAMs > NOG, where A > B indicates that clusters in A tend to be supergroups of clusters in B. Note that a TIGRFAMs group can be a subgroup of a real orthologous group because of a strict trusted cutoff value, but the evaluation measure $N_{TIGR}^{1to1}$ is effective even in such a case, provided that there is a one-to-one relationship between the TIGRFAMs group and the corresponding target group.

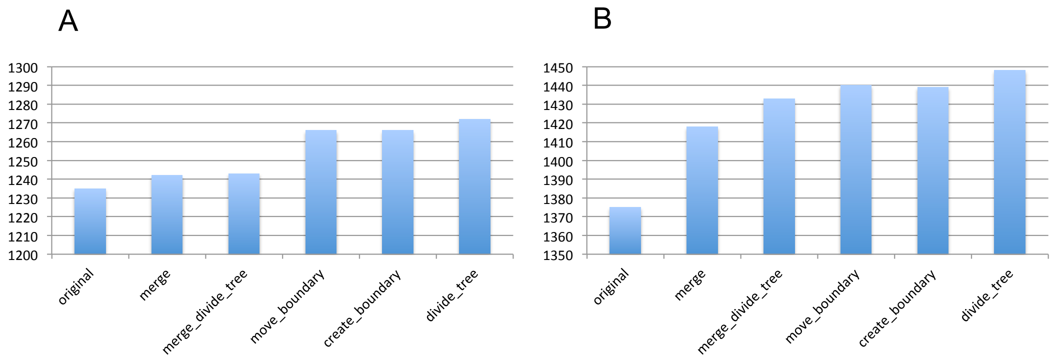
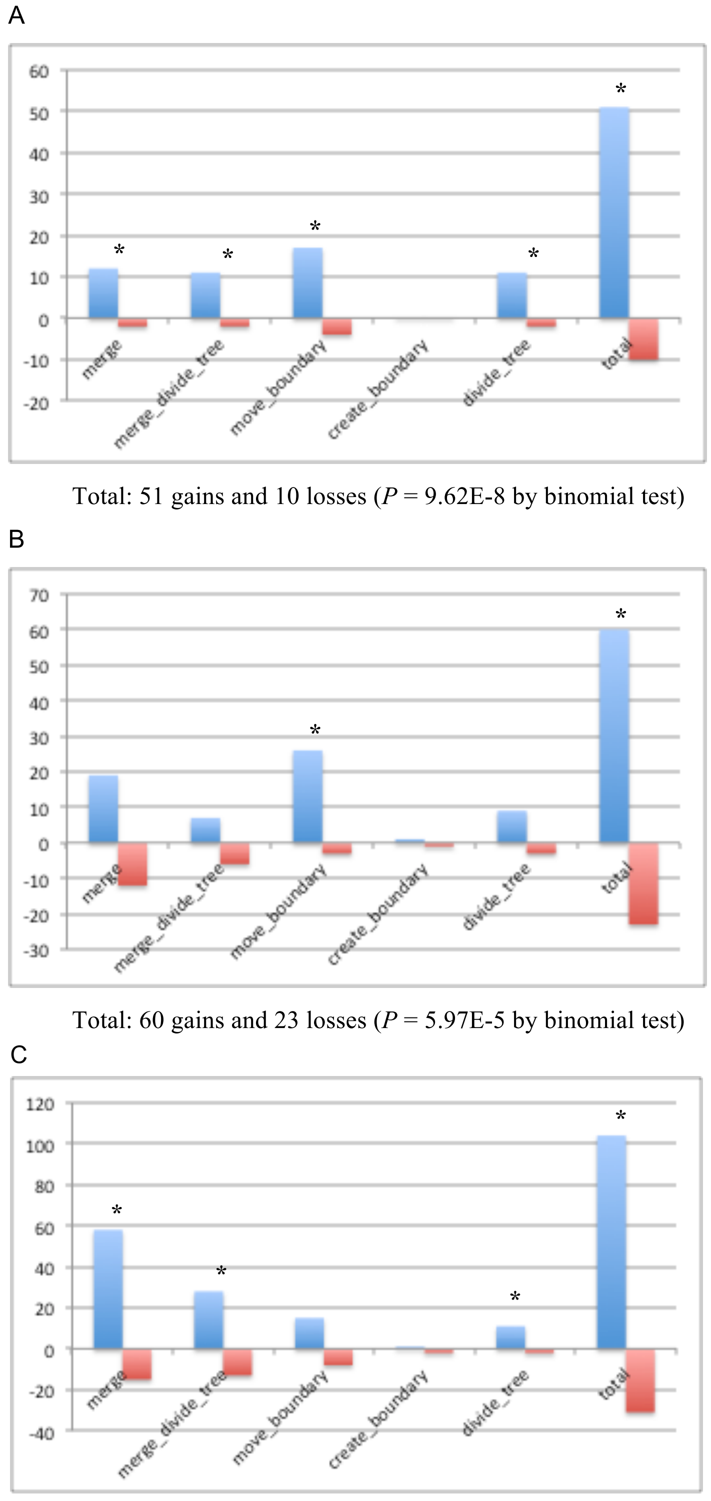


**Figure 8. Consistencies of the resulting clusters with TIGRFAMs models.**

(A) COG03 dataset and (B) FAMILY210 dataset. Vertical axes represent the numbers of one-to-one ortholog relationships.



Total: 51 gains and 10 losses (*P* = 9.62E-8 by binomial test)

Total: 60 gains and 23 losses (*P* = 5.97E-5 by binomial test)

Total: 104 gains and 31 losses (*P* = 2.06E-10 by binomial test)

**Figure 9. Gains and losses of correspondences between the resulting clusters and TIGRFAMs models.**

(A) COG02 dataset, (B) COG03 dataset and (C) FAMILY210 dataset. The blue bars represent gains of one-to-one relationships and the red represents losses. Significant differences of the gains and losses with *P* < 0.05 by binomial test are indicated by *.

**Table 2. Number of consistent clusters with TIGRFAMs models.**

| COG03 dataset | | FAMILY210 dataset | |
|---|---|---|---|
| **Method** | $N_{TIGR}^{1to1}$ | **Method** | $N_{TIGR}^{1to1}$ |
| **COG** | 1107 | **eggNOG** | 1149 |
| **DomClust** | 1235 (1.12) | **DomClust** | 1375 (1.20) |
| **DomRefine** | 1272 (1.15) | **DomRefine** | 1448 (1.26) |
| **TIGRFAMs*** | 3576 | **TIGRFAMs*** | 3924 |

The ratio of $N_{TIGR}^{1to1}$ to COG or eggNOG is shown in parenthesis. *Number of TIGRFAMs models with hits in the corresponding dataset, which is the possible maximum number of $N_{TIGR}^{1to1}$.

**Table 3. Number of reference clusters corresponding to the obtained clusters.**

| | | **COG03 dataset** | | | |
|---|---|---|---|---|---|
| **Clusters** | **Reference** | $N_{ref}^{equiv}$ | $N_{ref}^{sub}$ | $N_{ref}^{super}$ | $N_{clust}$ |
| **COG** | **TIGRFAMs** | 1271 | 1678 | 55 | 4814 |
| **DomClust** | **TIGRFAMs** | 1364 | 1342 | 102 | 7503 |
| **DomRefine** | **TIGRFAMs** | 1386 | 1389 | 106 | 7308 |
| **DomRefine** | **COG** | 3618 | 359 | 779 | 7308 |

| | | **FAMILY210 dataset** | | | |
|---|---|---|---|---|---|
| **Clusters** | **Reference** | $N_{ref}^{equiv}$ | $N_{ref}^{sub}$ | $N_{ref}^{super}$ | $N_{clust}$ |
| **eggNOG** | **TIGRFAMs** | 1448 | 1828 | 587 | 64983 |
| COG* | **TIGRFAMs** | 1004 | 1721 | 84 | 4873 |
| NOG* | **TIGRFAMs** | 444 | 107 | 564 | 60110 |
| **DomClust** | **TIGRFAMs** | 1652 | 1524 | 306 | 60775 |
| **DomRefine** | **TIGRFAMs** | 1674 | 1674 | 308 | 57644 |
| **DomRefine** | **eggNOG** | 35542 | 26691 | 4806 | 57644 |
| **DomRefine** | COG* | 3763 | 735 | 1998 | 57644 |
| **DomRefine** | NOG* | 31779 | 25956 | 2808 | 57644 |

$N_{ref}^{equiv}$, $N_{ref}^{sub}$ and $N_{ref}^{super}$ represent the number of reference clusters that are equivalent, subgroup and supergroup of the cluster, respectively. $N_{clust}$ is the total number of clusters obtained by each method. Let $C \wedge R$ denote a set of corresponding segment pairs between a cluster $C$ and a reference cluster $R$. Here, we considered that a segment $s_c \in C$ corresponds to a reference segment $s_r \in R$ if $|s_c \cap s_r|/|s_r| \geq 0.9$. Let $p_c = |C \wedge R|/|C|$, $p_r = |C \wedge R|/|R|$ and $F = 2p_c p_r/(p_c+p_r)$. We defined $R$ as being equivalent to $C$ if $F \geq 0.7$ ; otherwise, $R$ is a subgroup of $C$ if $p_r \geq 0.7$ or a supergroup of $C$ if $p_c \geq 0.7$.

Each raw represents the result of comparison between obtained clusters and reference clusters. If $N_{ref}^{sub} > N_{ref}^{super}$, then the obtained clusters tend to be larger than the reference clusters. If $N_{ref}^{sub} < N_{ref}^{super}$, then the obtained clusters tend to be smaller than the reference clusters.

*eggNOG clusters were divided into COG-derived clusters and NOG clusters.

### 2.3.5 Examples of obtained ortholog groups

On the basis of the resulting number of clusters for FAMILY210 (Table 1), the DomRefine result included a larger number of split clusters than eggNOG (10942 against 2333). We here focused on the genes split in the DomRefine result but not in eggNOG. Figure 10A presents an example of the clusters containing such genes, where two adjacent clusters corresponded to TIGRFAMs domains TIGR03546 and TIGR03545, respectively, both of which were functionally uncharacterized protein families. Although DomClust split a fused gene, nam:NAMH_0533 (*Nautilia profundicola*), into two domains, it failed to split another plausible fused gene, ftu:FTT_0505 (*Francisella tularensis*). However, DomRefine corrected the classification (Figure 10A). When the members of the clusters were compared to eggNOG, they overlapped three NOG clusters: NOG12793 ($N$ = 6473), NOG44136 ($N$ = 7), and NOG145366 ($N$ = 2), where $N$ indicates the cluster sizes in the FAMILY210 dataset. eggNOG did not split the two plausible fused genes, ftu:FTT_0505 and nam:NAM_0533; it assigned ftu:FTT_0505 to NOG145366 and nam:NAMH_0533 to NOG12793. As a result, proteins with the same TIGRFAMs hits were separated into different clusters. In contrast, NOG12793 was the largest eggNOG cluster containing proteins with many different TIGRFAMs hits (97 families), indicating that it is too large in terms of grouping corresponding genes among organisms.

Figure 10B presents another example, where the proteins had hits to TIGR00324 (*endA*: tRNA intron endonuclease). Here genes of FAMILY210 were extracted to demonstrate the subset of the alignment. Of 35 proteins, 12 had two domains both of which correspond to TIGR00324, whereas in several species, these domains are coded as two separate genes. Some other species, such as *Methanocaldococcus jannaschii*, contain only one gene consisting of one domain (mja:MJ_1424). It is known that the two tandemly repeated domains, N-terminal repeat (NR) and C-terminal repeat (CR), have distinct functional roles and were suggested to have arisen by gene duplication and subfunctionalization [51]. Thus, it is reasonable to cluster these homologous domains into two distinct ortholog groups. When we created a phylogenetic tree using both the domains, we discovered distinct clusters corresponding to NR and CR. DomClust successfully clustered these domains except for two genes (Figure 11), but DomRefine failed to refine these, in that the boundary modification reduced the agreement with TIGRFAMs hits (Figure 10B). One reason for this failure could be that the presence of tandemly repeated domains confounded the alignment, and DomRefine based on an incorrect alignment may fail to refine the domain boundary. In fact, in this case, single-domain proteins of *Nitrosopumilus maritimus*, nmr:NMAR_0450 and nmr:NMAR_1039, which were assigned to the NR and CR clusters, respectively, were both located in the C-terminal half in the alignment. Another problem affecting the alignment was the presence of unconserved sequences in the N-terminal regions of eukaryotic genes, such as cdu:CD36_42500 (*Candida dubliniensis*). In domain inferences of DomClust and DomRefine, these regions are treated as C-terminal groups (colored in light blue). Influenced by such an unconserved region, regions such as nmr:NMAR_0450 are prevented from being aligned to the N-terminal region and are consequently aligned to the C-terminal region.

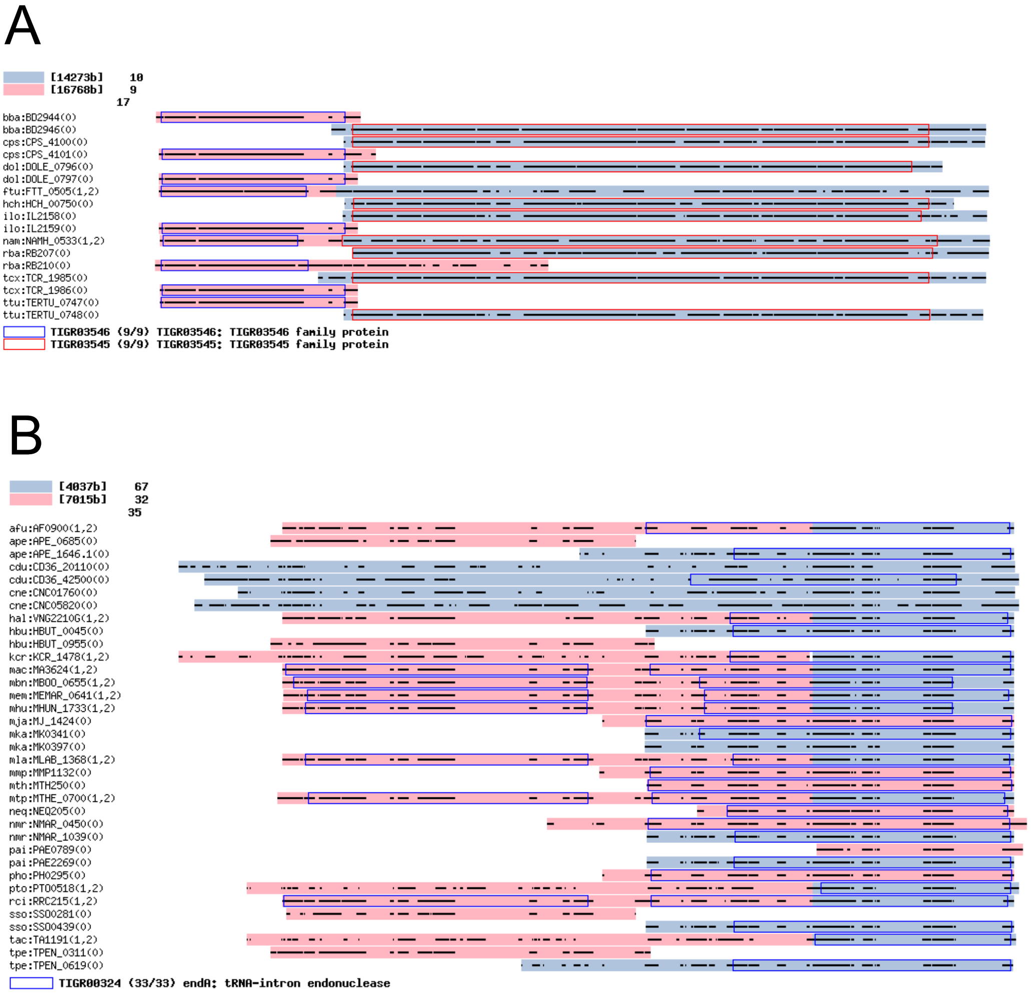


**Figure 10. Examples of the resulting ortholog clusters.**

Examples of ortholog clusters obtained by DomRefine applied to the FAMILY210 dataset. (A) Clusters including genes split in the DomRefine result but not in eggNOG. (B) Clusters including genes with tandemly repeated domains. In these figures, coloring of each residue according to the conservation rate is disabled in order to simplify the representation.

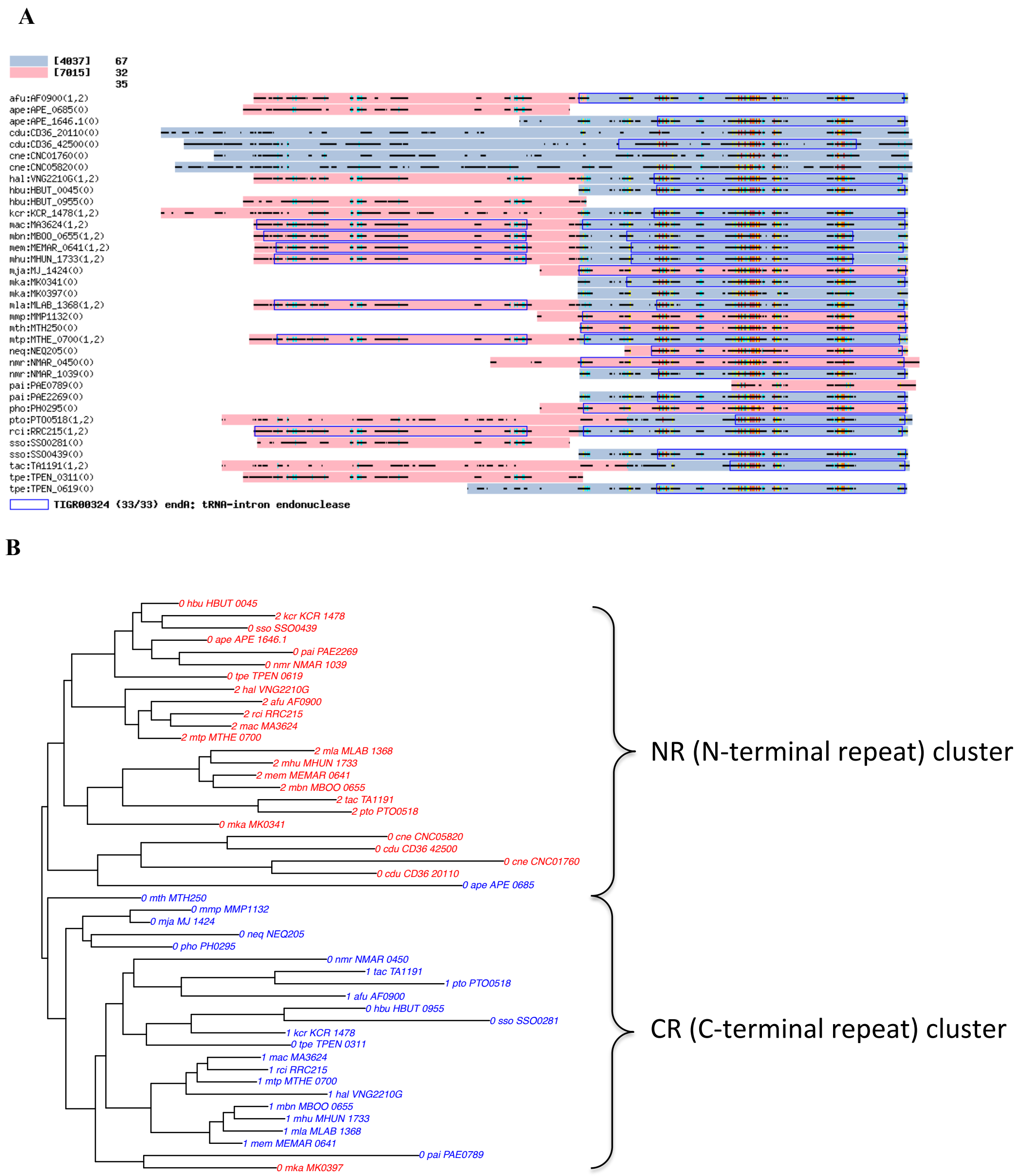


**Figure 11. Supplement to the examples of ortholog clusters.**

The proteins contained in these examples are same as those in Figure 8B. Results of DomClust before applying DomRefine are shown. (A) The protein sequences are aligned by Clustal Omega. Domains are colored in pink or light blue acocording to the DomClust results. (B) After the proteins are split into domains, those domains are aligned by Clustal Omega (domain by domain), and the phylogenetic tree of them are created by FastTree. In the tree, we found distinct clusters corresponding to N-terminal repeat (NR) and C-terminal repeat (CR). The leaves colored in red and blue correspond to the DomClust cluster colored in pink and light blue in (A), respectively. DomClust successfully clustered the domains except two genes.

## 2.4 Discussion

In the study presented in this chapter, we developed a method, DomRefine, to improve domain-level ortholog classification and applied the method to refine the ortholog classification created by DomClust, using the proteome sets extracted from the COG and MBGD databases. We demonstrated that our method was able to achieve improvements when the results were evaluated on the basis of COG and TIGRFAMs, which are the manually curated reference databases. Although COG and TIGRFAMs clusters have different characteristics (as discussed below), DomClust clusters became more consistent with both COG and TIGRFAMs after the *merge* procedure of DomRefine (Figure 6, Figure 8), suggesting that the over-splitting problem in orthologous domains mentioned in 2.1 Background were alleviated.

The TIGRFAMs database consists of HMMs constructed from curated multiple sequence alignments and is designed mainly for detecting functionally equivalent homologous proteins (equivalogs) among prokaryotic genomes [42]. Therefore, validating the obtained orthologous domains by TIGRFAMs models is reasonable in that the main aim of the ortholog database among prokaryotic genomes is to infer protein functions. In addition to the TIGRFAMs database, we used the COG database, a manually curated ortholog database for microbial genomes, as the reference database. However, when the same classification results of the COG03 dataset were evaluated using the different reference databases, COG and TIGRFAMs, different tendencies were observed between them (Figure 6B and Figure 8A). In particular, the agreement with COG decreased after the *divide_tree* procedure (Figure 6B), whereas that with TIGRFAMs increased (Figure 8A). This difference is probably caused by the known COG problem that a substantial fraction of COG groups contain non-orthologous (or out-paralogous) genes [52]; thus, division of groups using the *divide_tree* procedure such that paralogous genes are appropriately separated can reduce the consistency with the COG classification. Another difference is that the *move_boundary* procedure improved domain boundaries in terms of their correspondence with TIGRFAMs (Figure 8A), whereas it failed to improve them in terms of their correspondence with COG (Figure 6B). This was observed because TIGRFAMs is constructed from the HMMs of well-conserved and well-characterized protein families, whereas COG was originally constructed from a clustering result based on all-against-all similarities. Consequently, the *move_boundary* procedure modified the domain boundaries to improve the coverage of well-conserved domain boundaries defined in TIGRFAMs, but may not have improved the correspondence with COG boundaries. In either case, we consider TIGRFAMs as a better reference dataset than COG to evaluate orthologous domain classification.

The goal of this study was to construct a fully automated and reliable procedure to create ortholog database, a necessary resource in the era of huge amounts of genomic data. In this respect, the eggNOG database, which was constructed by computational extension of COG, is another ortholog database that covers the currently sequenced genomes and is periodically updated. However, eggNOG consists of two different types of ortholog groups, i.e., the extension of the original COGs and the remaining NOGs, because of the nature of its incremental updating procedure. COG-derived clusters tend to be larger, whereas the NOG clusters tend to be smaller (Table 3). As a result of the mixture of the two different distributions, the cluster size distribution of eggNOG appears to be deviated from the power-law distribution, which has been observed in various types of protein clusters [49] (Figure 4B).

To compare the classification performance, we also evaluated the COG and eggNOG clusters in terms of the agreement with the TIGRFAMs models ($N_{TIGR}^{1to1}$) and discovered that our method showed better agreement than the COG/eggNOG classifications (Table 2). The original DomClust classification already showed better agreement than the COG classification partly because of the abovementioned problem that some COG groups contain non-orthologous genes. In the eggNOG classification, additional problems caused by its incremental updating procedure can magnify the difference. In fact, the increasing rate of $N_{TIGR}^{1to1}$ from the eggNOG classification to the DomClust classification using the FAMILY210 dataset (20%) was higher than that from the COG classification to the DomClust classification (12%) (Table 2). The increasing rates were further increased when the COG/eggNOG classifications were compared to the classifications after refinement (15% and 26%, respectively; Table 2).

One of the problems with incremental updating in the eggNOG classification is that a new domain split appears to be rarely introduced during the NOG classification in contrast to the original COG classification (Table 1). The DomClust/DomRefine procedure identified a substantial number of clusters that are not defined in COG, where domain splitting was needed for valid ortholog classification, as in the examples illustrated in Figure 10A. As illustrated in Figure 1A, a clustering method without domain splitting generally tends to create clusters with smaller sizes than that with domain splitting when fused proteins are included in the dataset. This may partly explain the smaller size distribution of the NOG clusters observed in Figure 4B.

Although numerous methods have been developed for identifying orthologs, few methods have focused on classification at the sub-gene level. Our method splits proteins into domains in the course of clustering with the aim of detecting the correct grouping of proteins (Figure 1A). The resulting splits of proteins suggest domain fusion/fission events in evolutionary history, which may result in functional divergence among orthologous proteins. In this sense, domain-level ortholog classification provides a valuable source for evolutionary analysis.

In this study, we evaluated the *merge* procedure using the COG database as a reference to estimate a reasonable threshold for the DSP score change (Figure 7). The DomRefine pipeline also depends on the other settings, including the parameters of DomClust for initial clustering and the selection of the score matrix for calculating the DSP score. The optimal settings for them remain to be explored to pursue the better clustering results.

The example in Figure 10B showed the limitation of the current version of DomRefine. Theoretically, our system is applicable to eukaryotic protein classification. However, given the abundance of complex multidomain architectures among eukaryotic proteins and the frequent differences in domain composition among apparent orthologs [23,53], domain-level clustering of eukaryotic proteomes is more challenging than prokaryotic proteomes. In particular, a tandem repeat of homologous domains within a protein, which is quite common in eukaryotic proteins, may confound the multiple alignment, possibly leading to a failure of DomRefine to refine domain boundaries. As far as I tested, handling of tandemly duplicated domains seems to be more or less a common problem in existing alignment programs, although Clustal Omega used in this study demonstrated a relatively better performance with respect to this point. Thus, a special procedure may be required to handle such tandem repeats correctly as a pre- or postprocessing step of an alignment program unless improved versions of the alignment programs are available.

Although the current DomRefine pipeline requires much larger computational time than that required by DomClust, the parallelization technique enables the execution of the pipeline in a feasible time (Table 4). Of the required time, the calculation of the DSP score comprises only a small fraction, and most of the computational time is spent performing multiple alignments. Possible approaches to address this bottleneck include incremental calculation of large multiple alignments using alignments for subsets of the cluster, if available. It is notable that the obtained multiple alignment information will be a useful resource not only for the DomRefine pipeline but also for various other applications. Therefore, it is worth computing and storing the multiple alignment information for general use.

In a recent update of MBGD database [54], the DomRefine pipeline was used to improve the standard ortholog datable. Future works include applying our method to various other ortholog datasets. Our method will enhance the reliability of ortholog databases and thereby contribute to comparative analyses using them.

**Table 4. The execution time of DomClust and DomRefine for COG03 dataset.**

| **Method** | **Time** (minutes) |
|---|---|
| **DomClust**[a] | 1.90 |
| **DomRefine**[b] | 352.48 |
| Total[*] | 16507.08 ( 100%) |
| Clustal Omega[*] | 13353.60 (80.9%) |
| FastTree[*] | 744.42 ( 4.5%) |
| DSP score[*] | 212.56 ( 1.3%) |
| Others[*] | 2196.50 (13.3%) |

[a] DomClust was executed on a single core of Intel Xeron 2.7 GHz. The calculation of DomClust does not include the construction of all-against-all similarity data.

[b] DomRefine was executed on a parallel environment including a single core of Intel Xeon 2.7 GHz and 100 cores of Intel Xeron 2.8 GHz through a job management system based on Sun Grid Engine. The real time required to finish the computation on the environment was measured.

[*] In the case of DomRefine, the execution time measured on each core was totalized for all the processes (Total), or for a specific type of processes (Clustal Omega, FastTree, DSP score and Others).

## 2.5 Conclusions

We developed a method for improving domain-level ortholog classification on the basis of the optimization of a score and demonstrated the effectiveness of the method using the manually curated reference databases. For this purpose, we designed a score for evaluating ortholog clusters at the domain level on multiple alignments and demonstrated that the method contributes to the improvement of the clusters. This method will enhance the reliability of ortholog databases and thereby contribute to comparative analyses using them.

# Chapter 3 Ortholog database for integrative analysis of genomic data

## 3.1 Background

Among the various ortholog databases currently available, the MBGD provides a system for users to select specific sets of species to be compared, thus providing a flexible mechanism for finding orthologs [15]. Although MBGD and other ortholog databases provide Web browser interfaces to efficiently retrieve ortholog information and related data, such interfaces are not sufficient for users who want to retrieve various information using the orthology relation as a hub of links.

For the integration of biological data derived from different data sources, the use of the Semantic Web technology [10] is a promising approach [11,12]. In the Semantic Web, all the information is described in the Resource Description Framework (RDF) [16], in which the Uniform Resource Identifier (URI) assures the uniqueness of each resource worldwide and contributes to valid data integration of data collected from different sources. The Semantic Web technology also provides a search functionality using SPARQL [4] standardized by the World Wide Web Consortium (W3C), which includes a protocol to access the data across the Web. Thus, constructing a database using the Semantic Web that accepts SPARQL queries means that the data are not only locally available but also accessible through arbitrary queries specified by users across the Web. An additional merit of using the Semantic Web is that data modeling is based on ontologies, which define the relations between the terms and work as a translation layer to unite different terminologies used by different resource providers. In the past few years, there has been a continuous effort to apply the Semantic Web to biological databases for enhancing their interoperability [11,13]. Restructuring the ortholog database as a hub of the biological database network based on the Semantic Web will have a significant impact for biological database integration.

In this chapter, I present the construction of an ortholog database using the Semantic Web technology [14]. Here I show a general RDF model we developed for describing ortholog information using ontologies. On the basis of this model, ortholog data can be converted into RDF to construct an integrative database available through the SPARQL endpoint. I show several examples of SPARQL queries to demonstrate that our database can work as a hub for integrating several genomic data resources and support knowledge discovery through its search functionalities.

## 3.2 Methods

### 3.2.1 Construction of ontologies

The ortholog ontology (OrthO), an ontology for MBGD (MBGD-O), and an ontology for GO annotation (GOA-O) were created on the ontology editor, Protégé [55] Desktop 4.3 OS X application bundle, which was obtained from http://protege.stanford.edu. The ontology files were saved in Turtle [56] format. Afterwards, the ontology files were manually edited using text editors. The created ontologies are available at http://mbgd.genome.ad.jp/ontology/. For covering the concepts defined by OrthoXML [57] (http://orthoxml.org/), we inspected an example described in the OrthoXML documentation (http://orthoxml.org/0.3/orthoxml_doc_v0.3.html) and listed the representations used therein. The terms to be included in OrthO were then determined according to this list. Among these terms, we designed a hierarchical class structure, if necessary.

### 3.2.2 Preparation of datasets

The ortholog information and the related data about genes, genomes, and organisms included in MBGD release 2014-01 [15] were converted to RDF (downloadable at http://mbgd.genome.ad.jp/rdf/archive in Turtle format) and loaded to Virtuoso [58]. In this study, COG and NOG clusters included in the eggNOG database v3.0 [45] were concatenated and designated as eggNOG. Taxonomy information and gene ontology represented in RDF were downloaded from the UniProt [34] FTP site (ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/rdf/taxonomy.rdf.gz, ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/rdf/go.rdf.gz), and loaded to Virtuoso. GO annotation with evidence codes was downloaded from the UniProt-GOA database (ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/UNIPROT/gene_association.goa_uniprot.gz). The converters from the original data to RDF were written in Perl. Raptor RDF Syntax Library (http://librdf.org/raptor/) version 2.0.13 was used to count the numbers of triples of RDF files.

### 3.2.3 Settings of the RDF store

RDF data and ontologies were loaded to an RDF store and made accessible through the SPARQL endpoint. Virtuoso Open-Source Edition (http://www.openlinksw.com/dataspace/doc/dav/wiki/Main/) 7.1.0 was installed into Linux and was used as an RDF store and the SPARQL endpoint. *DB.DBA.TTLP_MT()* function was used to load the data in Turtle format and *DB.DBA.RDF_LOAD_RDFXML_MT()* function for RDF/XML format, through the *isql* interface of Virtuoso. To load RDF in Turtle format containing triples larger than 10 million triples, the triples were divided into smaller files containing less than 10 million triples and loaded in parallel using the *ld_dir()*

and *rdf_loader_run()* functions with a parallelization degree of four. To make inference rules defined in ontologies executable in Virtuoso, the *rdfs_rule_set()* function was used through the *isql* interface of Virtuoso. To enable inference for a SPARQL query to Virtuoso, the following line should be specified at the beginning of the query,

```
define input:inference "mbgd"
```

where "mbgd" is the rule set including OrthO, MBGD-O, and GOA-O. There is another rule set "ontologies" that includes all the ontologies stored in the MBGD SPARQL endpoint. The functionality of Virtuoso was restricted by setting several options. The maximum number of returned results was set to 10,000. The maximum system memory usage was set to 8 GB. Virtuoso is accessible as the MBGD SPARQL endpoint at http://sparql.nibb.ac.jp/sparql.

### 3.2.4 Usage of the SPARQL endpoint

When calculating the query execution time, Virtuoso was restarted before each calculation to refresh the cache. To analyze the data by accessing the SPARQL endpoint from local computers, the SPARQL package (http://cran.r-project.org/web/packages/SPARQL/) of R (http://www.r-project.org/) was used as the SPARQL client.

### 3.2.5 Browsing the RDF

Each resource URI under the namespace of http://mbgd.genome.ad.jp/rdf/resource/ was made dereferenceable by converting the HTTP access into a query to the SPARQL endpoint using a Perl CGI. For example, HTTP access to http://mbgd.genome.ad.jp/rdf/resource/gene/eco:B0002 is converted to the following SPARQL query,

```
SELECT ?subject ?predicate ?object
WHERE {
  { <http://mbgd.genome.ad.jp/rdf/resource/gene/eco:B0002> ?predicate ?object }
  UNION
  { ?subject ?predicate <http://mbgd.genome.ad.jp/rdf/resource/gene/eco:B0002> }
}
```

which enables the instant browse of the RDF graph under the MBGD RDF name space by clicking the URI on Web browsers.

## 3.3 Results

### 3.3.1 RDF model of ortholog information

Together with my collaborators, I developed a general RDF model for describing ortholog information [14]. Our model can accept different ortholog databases in the same framework for their interoperable use. As the basis of the RDF model, we defined the Ortholog Ontology (OrthO), which comprises the basic terms required for describing the ortholog information (available at http://purl.jp/bio/11/orth). The terms in OrthO are defined using the Web Ontology Language (OWL) [42]: each of the defined terms is either a class to be used for representing a specific group of resources or a property for representing a specific relationship from resources to resources or to data values. The terms in OrthO have hierarchical relationships shown in Figure 12A. On the basis of OrthO, the ortholog information can be described in RDF and thereby is connectable to other resources that are also described in the form of RDF (Figure 12B). The OrthO and other ontologies used in this study are listed in Table 5 with their namespaces and their abbreviated forms (prefix).

Among the most important classes in the OrthO is *OrthologGroup*, which is defined as a set of homologous sequences derived from a common ancestral sequence by speciation. On the other hand, the OrthO also defines *ParalogGroup* as a set of homologous sequences derived from a common ancestral sequence by duplication and *HomologGroup* as a super class of the *OrthologGroup* and *ParalogGroup*. *SequenceUnit* is a class for generally representing each member of such groups. Typically, *SequenceUnit* is *Gene* or *Protein* in most ortholog databases, but it can be any sequence element between which homology relation can be defined. *SequenceUnit* is linked to its source organism (class *Organism*), which is essential information constituting the ortholog data. To represent the entire set of ortholog groups, the OrthO defines a class *Dataset*. One of the main applications of OrthO is the description of ortholog datasets from different sources within the same framework for integrative use. In such situations, distinct ortholog datasets can be represented as different instances of *Dataset*. Similarly, ortholog classifications for different sets of target organisms or different versions of datasets can also be represented as different instances of *Dataset*.

The most important property in the OrthO is *member*, which typically links *Group* to *SequenceUnit*. In the case of the hierarchical grouping of orthologs [59], however, *member* may link *Group* to *Group* to form a tree structure in which internal nodes are *Group* and leaves are *SequenceUnit*. The property *crossReference* represents the correspondence between two instances of *SequenceUnit* derived from different databases. The property *organism* is used to reference the source organism (*taxon* is used more specifically to reference an NCBI taxonomy ID). The OrthO also includes the properties *ortholog* and *paralog* for describing pairwise relationships between two instances of *SequenceUnit*.

In comparison to a preceding study of ontology for orthology (OGO) [60], our model is designed to be more general and conforming to various ortholog databases. In particular, in our OrthO, sequence units constituting orthologs may be a gene, a protein, a transcript, or even a newly defined element on genome sequences (such as orthologous domains [8,9,59] or regulatory regions). In addition, the OrthO can be applied to a hierarchical grouping of orthologs and paralogs created by assigning a speciation or duplication event to each node [61], thus enabling a more elaborate description of evolutionary relationship between group members.

In addition to its unique characteristics, the OrthO has compatibility and interchangeability to other ontologies, which increases its usability. To achieve the compatibility for the concepts commonly existing in the OrthO and OGO, the corresponding classes or properties are associated in the OrthO, thus enabling automatic translation between terminologies when inference is enabled in RDF stores. For example, *orth:member* and *ogo:hasOrthologous* are associated by *owl:equivalentProperty*; thus, ortholog information described in the OrthO can be searched by OGO. Furthermore, the terms in OrthO are also associated with other generally used ontologies such as Sequence Ontology (SO) [62] and Semanticscience Integrated Ontology (SIO) [63].

In addition to the association with other ontologies, the OrthO also has basic compatibility with OrthoXML [57] (http://orthoxml.org/), which was proposed as a community standard for formatting ortholog information. The terms of OrthO conform to the tag names of OrthoXML, including *orthologGroup*, *paralogGroup*, and *gene*, which have attributes such as *geneId*, *proteinId*, or *transcriptId*. Because of the similarity of the concepts between the OrthoXML and OrthO, ortholog information described in the OrthoXML can also be described using the OrthO. For example, an OrthoXML example shown in the OrthoXML documentation (http://orthoxml.org/0.3/orthoxml_doc_v0.3.html) can be described in OrthO (Figure 13).

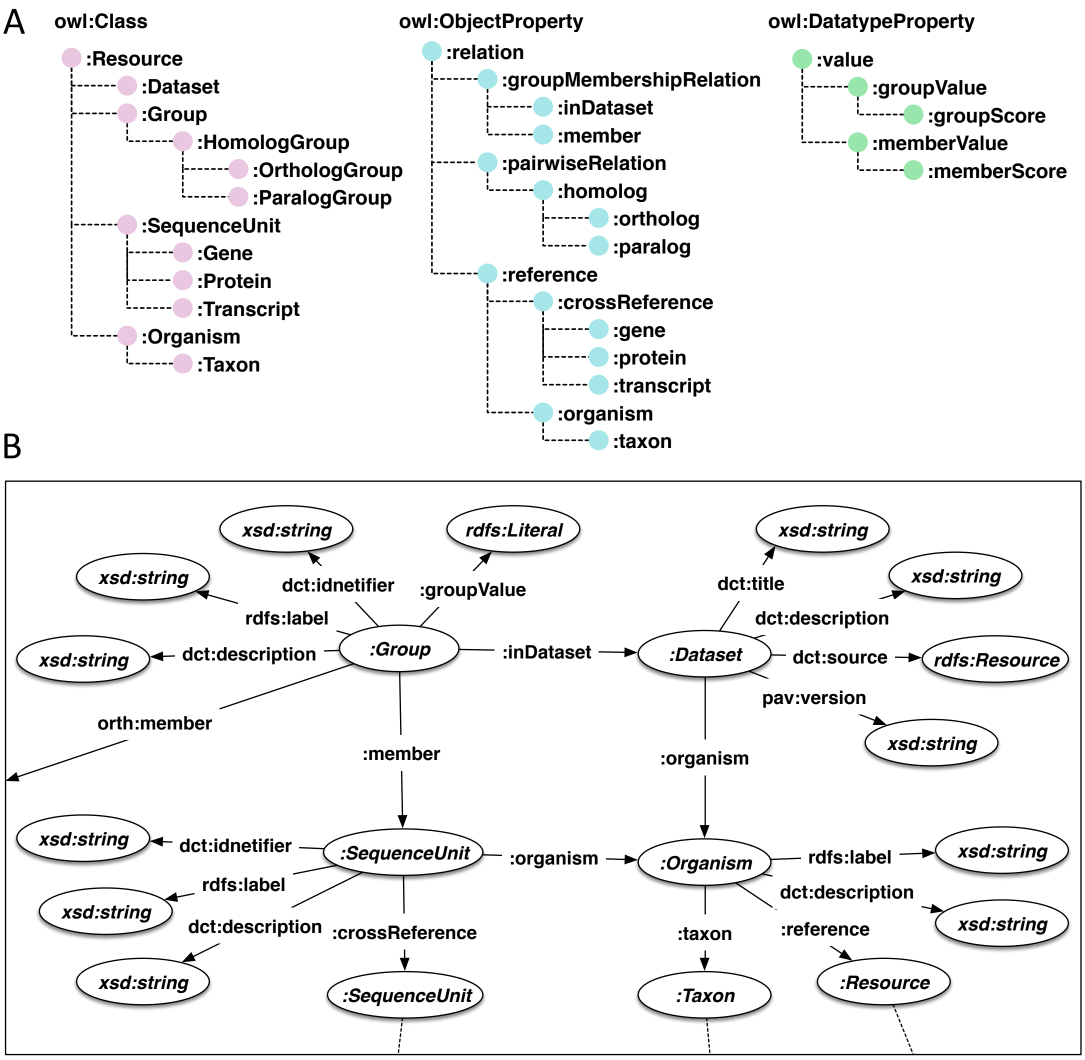


**Figure 12. RDF model of ortholog information based on OrthO.**

(A) Hierarchical structure of classes and properties in OrthO. OrthO includes 12 classes (*owl:Class*) and 20 properties (15 of *owl:ObjectProperty* and 5 of *owl:DatatypeProperty*). (B) Schematic representation of RDF graph structure of ortholog information described using OrthO. The elliptical nodes represent instances of classes. The directed edges represent properties. The dotted lines represent possible links to other resources.

**Table 5. List of ontologies available at the MBGD SPARQL endpoint.**

| Ontology title | Prefix | Namespace |
|---|---|---|
| Ortholog Ontology (OrthO) | orth: | http://purl.jp/bio/11/orth# |
| An ontology for MBGD | mbgd: | http://purl.jp/bio/11/mbgd# |
| An ontology for GO annotation | goa: | http://purl.jp/bio/11/goa# |
| The RDF Concepts Vocabulary (RDF) | rdf: | http://www.w3.org/1999/02/22-rdf-syntax-ns# |
| The RDF Schema vocabulary (RDFS) | rdfs: | http://www.w3.org/2000/01/rdf-schema# |
| The OWL 2 Schema vocabulary (OWL 2) | owl: | http://www.w3.org/2002/07/owl# |
| Dublin Core Metadata Element Set, Version 1.1 | dc: | http://purl.org/dc/elements/1.1/ |
| DCMI Metadata Terms | dct: | http://purl.org/dc/terms/ |
| Vocabulary of Interlinked Datasets (VoID) | void: | http://rdfs.org/ns/void# |
| SKOS Vocabulary | skos: | http://www.w3.org/2004/02/skos/core# |
| Provenance, Authoring and Versioning (PAV) | pav: | http://purl.org/pav/ |
| Ontological Gene Orthology (OGO) | ogo: | http://miuras.inf.um.es/ontologies/OGO.owl |
| FALDO: Feature Annotation Location Description Ontology | faldo: | http://biohackathon.org/resource/faldo# |
| UniProt core ontology | up: | http://purl.uniprot.org/core/ |
| RDF representation of taxonomy | tax: | http://purl.uniprot.org/taxonomy/ |
| RDF representation of GO | go: | http://purl.uniprot.org/go/ |

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix dct: <http://purl.org/dc/terms/> .
@prefix void: <http://rdfs.org/ns/void#> .
@prefix pav: <http://purl.org/pav/> .
@prefix orth: <http://purl.jp/bio/11/orth#> .
@prefix : <http://mbgd.genome.ad.jp/rdf/resource/orthoxml_example/> .
@prefix wgene: <http://www.wormbase.org/db/gene/gene?name=> .
@prefix wprot: <http://www.wormbase.org/db/seq/protein?name=WP:> .
@prefix hgene: <http://Dec2008.archive.ensembl.org/Homo_sapiens/geneview?gene=> .
@prefix hprot: <http://Dec2008.archive.ensembl.org/Homo_sapiens/protview?peptide=> .
@prefix tax: <http://purl.uniprot.org/taxonomy/> .

<http://mbgd.genome.ad.jp/rdf/resource/orthoxml_example>
    a orth:Dataset ;
    dct:title "Examle" ;
    dct:description "Stripped down version of a real InParanoid 7.0 file." ;
    dct:source "inparanoid" ;
    pav:version "7.0" ;
    pav:derivedFrom <http://orthoxml.org/0.3/orthoxml_doc_v0.3.html#example> ;
    void:dataDump <http://mbgd.genome.ad.jp/rdf/archive/orthoxml_example.ttl> ;
    orth:organism :worm , :human .

:worm
    a orth:Organism ;
    rdfs:label "Caenorhabditis elegans" ;
    dct:source "WormBase" ;
    pav:derivedFrom "Caenorhabditis-elegans_WormBase_WS199_protein-all.fa" ;
    orth:taxon tax:6239 .

:human
    a orth:Organism ;
    rdfs:label "Homo sapiens" ;
    dct:source "Ensembl" ;
    pav:derivedFrom "Homo_sapiens.NCBI36.52.pep.all.fa" ;
    orth:taxon tax:9606 .

:group1
    a orth:OrthologGroup ;
    dct:identifier "1";
    orth:inDataset <http://mbgd.genome.ad.jp/rdf/resource/orthoxml_example> ;
    :bit "5093"^^xsd:integer ;
    orth:member :geneRef1 ,:geneRef2 .

:group3
    a orth:OrthologGroup ;
    dct:identifier "3";
    orth:inDataset <http://mbgd.genome.ad.jp/rdf/resource/orthoxml_example> ;
    orth:member :geneRef5 , :geneRef6 , :geneRef7 .

:geneRef1
    a orth:Gene ;
    dct:identifier "1";
    :inparalog "1"^^xsd:integer ;
    :bootstrap "1.00"^^xsd:decimal ;
    orth:organism :worm ;
    orth:gene wgene:WBGene00000962 ;
    orth:protein wprot:CE23997 .

:geneRef2
    a orth:Gene ;
    dct:identifier "2";
    :inparalog "1"^^xsd:integer ;
    :bootstrap "1.00"^^xsd:decimal ;
    orth:organism :worm ;
    orth:gene hgene:ENSG00000197102 ;
    orth:protein hprot:ENSP00000348965 .

:geneRef5
    a orth:Gene ;
    dct:identifier "5";
    :inparalog "1"^^xsd:integer ;
    :bootstrap "1.00"^^xsd:decimal ;
    orth:organism :worm ;
    orth:gene wgene:WBGene00006801 ;
    orth:protein wprot:CE43332 .

:geneRef7
    a orth:Gene ;
    dct:identifier "7";
    :bootstrap "0.4781"^^xsd:decimal ;
    orth:organism :human ;
    orth:protein hprot:ENSP00000373884 .

:bit
    rdfs:subPropertyOf orth:groupScore ;
    dct:description "BLAST score in bits of seed orthologs" .

:inparalog
    rdfs:subPropertyOf orth:memberScore ;
    dct:description "Distance between edge seed ortholog" .

:bootstrap
    rdfs:subPropertyOf orth:memberScore ;
    dct:description "Reliability of seed orthologs" .
```

**Figure 13. RDF representation of the OrthoXML example.**

### 3.3.2 Ortholog database using RDF

Whereas the OrthO can describe basic information commonly contained in most ortholog datasets, each dataset often contains database-specific concepts. The OrthO can be extended to describe such database-specific concepts. For describing MBGD data, we have constructed a database-specific ontology for MBGD (MBGD-O, available at http://purl.jp/bio/11/mbgd, see Figure 14 for hierarchical structure). The ontology includes specific terms such as *mbgd:Domain* to represent a sub-sequence of a protein as a sequence unit of classification and *mbgd:Chromosome* to represent a chromosome containing each gene. Such terms are designed to conform to the existing architecture of the MBGD database, enabling easy conversion from the existing database files to their RDF representation. The graph structure of RDF described by OrthO and MBGD-O is shown in Figure 15. In the MBGD-O, terms related to orthologs are defined based on the OrthO (i.e., associated with the OrthO terms using *rdfs:subClassOf* or *rdfs:subPropertyOf*). Because of these associations between the different levels of ontologies, the database can be searched either by database-specific terms (e.g., *mbgd:uniprot* representing a cross-reference to UniProt) or more general terms (e.g., *orth:crossReference*). Because the OrthO can provide a basis shared by different ontologies, the ortholog data described in different terminologies can be compared/merged by way of the OrthO.

To further demonstrate the ability of this model to describe the ortholog information, we applied the model to other ortholog databases. Among the previously developed ortholog databases [30,31], many databases have been represented using OrthoXML. Such data can be described by our model. On the other hand, the eggNOG database [33] contains positional information of orthologous regions at sub-gene level. This does not conform to OrthoXML but is common to MBGD. In fact, we could describe the data obtained from the eggNOG database using the OrthO and MBGD-O to create the RDF version of eggNOG, which was used for comparison with our database below (see 3.3.4 Comparison of ortholog information from different data sources).

To realize the retrieval of ortholog information described in RDF, we used an RDF store, Virtuoso [58]. The list of stored graphs and number of triples contained in each of them are shown in Table 2. The total number of triples stored is 1,150,394,708. The RDF data stored in Virtuoso can be retrieved by SPARQL across the Web using HTTP (http://sparql.nibb.ac.jp/sparql).

To provide users with easy access to the RDF data, we created a portal site for searching the database (http://mbgd.genome.ad.jp/sparql), which includes a schematic illustration of the RDF structure, query examples, ontology downloads, RDF archives, and documentations (Figure 16). On the portal site, the retrieval of ortholog information can be executed by entering a SPARQL query in the text box, and typical example queries are shown alongside. Those examples are clickable and provide an easy test environment for starters. All examples in this chapter are included in this portal site. Experienced users can access the SPARQL endpoint from their own programs via HTTP access. Alternatively, users can specify the SPARQL endpoint as the target of *SERVICE* keyword in a federated SPARQL query. In addition, access to each resource URI under the namespace of MBGD is dynamically converted into a query to the SPARQL endpoint (see 3.2 Methods), enabling an instant browse of MBGD RDF. The MBGD RDF data created in this work is downloadable from http://mbgd.genome.ad.jp/rdf/archive/ and available under Creative Commons Attribution Share Alike (CC BY-SA 3.0).

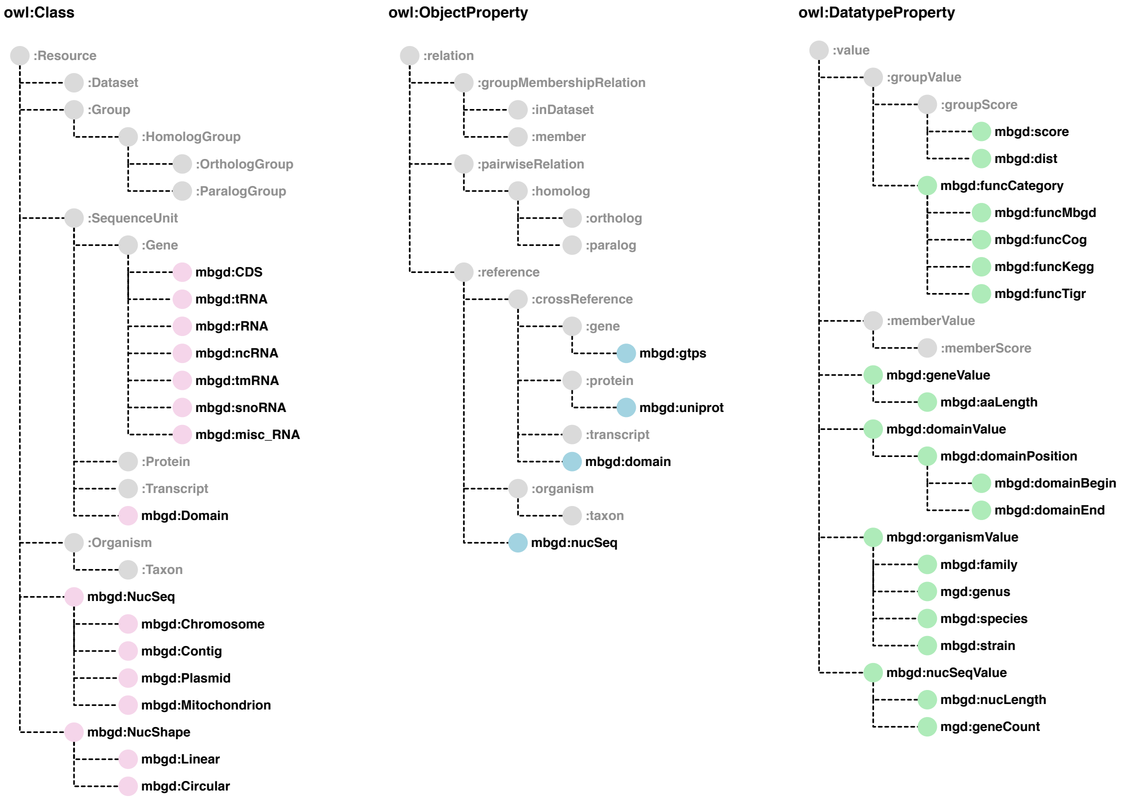


**Figure 14. Hierarchical structure of classes and properties in MBGD-O.**

MBGD-O includes 16 classes (*owl:Class*) and 25 properties (4 of *owl:ObjectProperty* and 21 of *owl:DatatypeProperty*). Terms of OrthO are shown in gray.

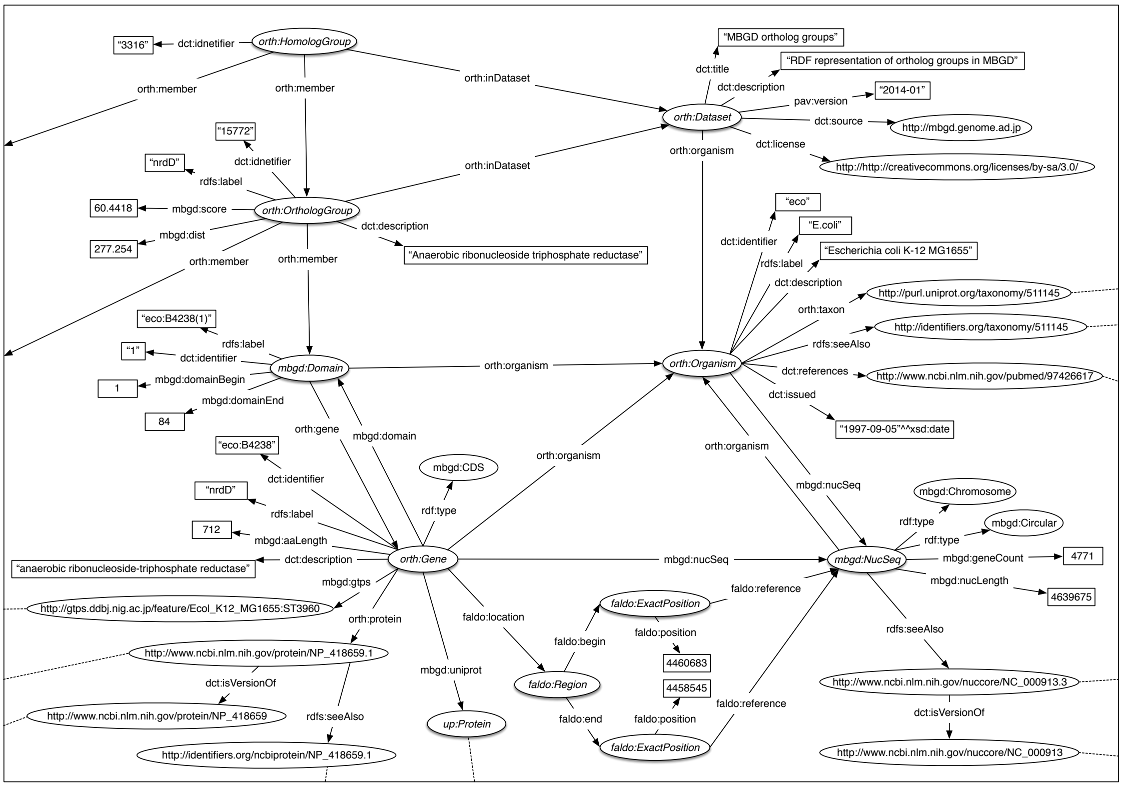
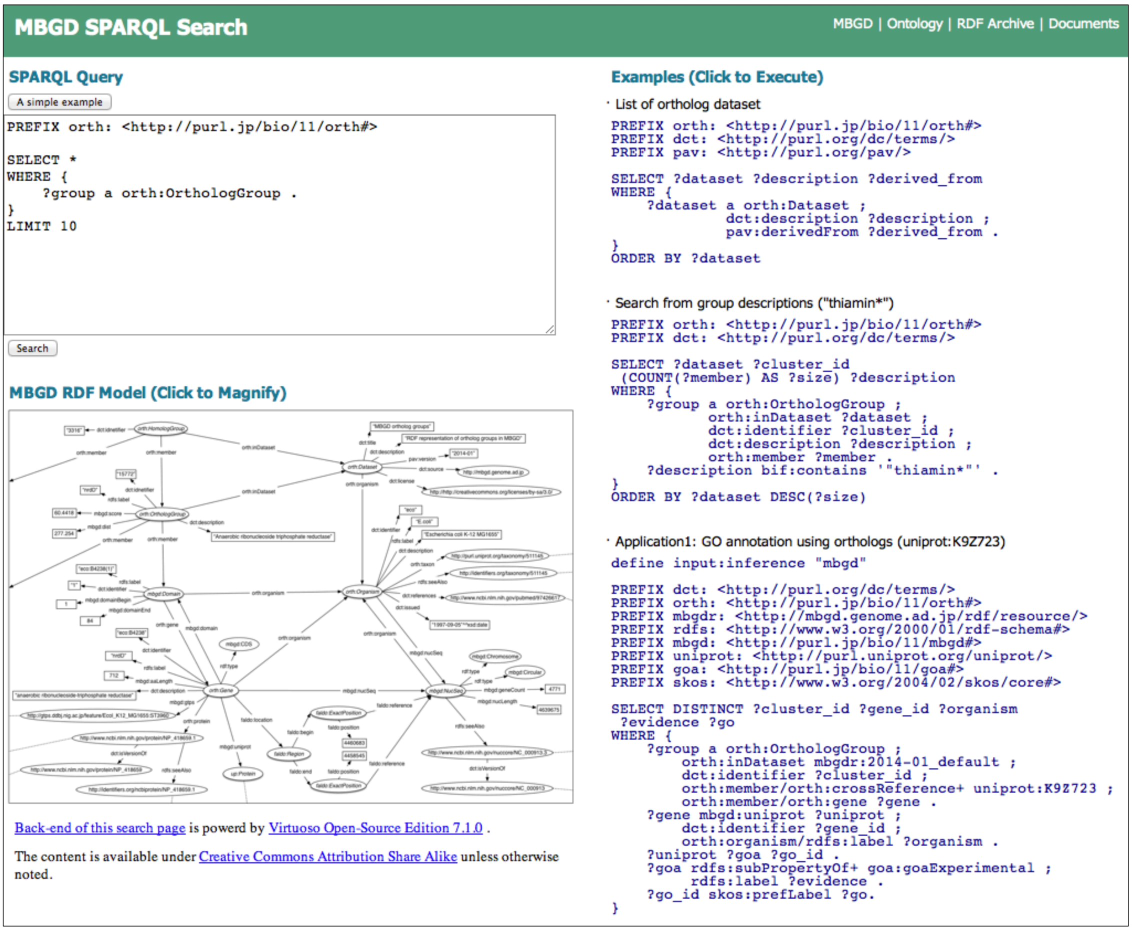


**Figure 15. Schematic diagram of the RDF representation of MBGD data.**

The elliptical nodes represent resources. Specifically, the shaded elliptical nodes where classes are shown in italics represent instances of the classes. In the unshaded elliptical nodes, the URIs of the resources are directly shown. The rectangular nodes represent literals. The directed edges represent properties. The dotted lines represent possible links to other resources.



MBGD SPARQL Search

MBGD | Ontology | RDF Archive | Documents

SPARQL Query

A simple example

```
PREFIX orth: <http://purl.jp/bio/11/orth#>

SELECT *
WHERE {
    ?group a orth:OrthologGroup .
}
LIMIT 10
```

Search

MBGD RDF Model (Click to Magnify)

Back-end of this search page is powerd by Virtuoso Open-Source Edition 7.1.0 .

The content is available under Creative Commons Attribution Share Alike unless otherwise noted.

Examples (Click to Execute)

- List of ortholog dataset

```
PREFIX orth: <http://purl.jp/bio/11/orth#>
PREFIX dct: <http://purl.org/dc/terms/>
PREFIX pav: <http://purl.org/pav/>

SELECT ?dataset ?description ?derived_from
WHERE {
    ?dataset a orth:Dataset ;
             dct:description ?description ;
             pav:derivedFrom ?derived_from .
}
ORDER BY ?dataset
```

- Search from group descriptions ("thiamin*")

```
PREFIX orth: <http://purl.jp/bio/11/orth#>
PREFIX dct: <http://purl.org/dc/terms/>

SELECT ?dataset ?cluster_id
 (COUNT(?member) AS ?size) ?description
WHERE {
    ?group a orth:OrthologGroup ;
           orth:inDataset ?dataset ;
           dct:identifier ?cluster_id ;
           dct:description ?description ;
           orth:member ?member .
    ?description bif:contains '"thiamin*"' .
}
ORDER BY ?dataset DESC(?size)
```

- Application1: GO annotation using orthologs (uniprot:K9Z723)

```
define input:inference "mbgd"

PREFIX dct: <http://purl.org/dc/terms/>
PREFIX orth: <http://purl.jp/bio/11/orth#>
PREFIX mbgdr: <http://mbgd.genome.ad.jp/rdf/resource/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX mbgd: <http://purl.jp/bio/11/mbgd#>
PREFIX uniprot: <http://purl.uniprot.org/uniprot/>
PREFIX goa: <http://purl.jp/bio/11/goa#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>

SELECT DISTINCT ?cluster_id ?gene_id ?organism
 ?evidence ?go
WHERE {
    ?group a orth:OrthologGroup ;
        orth:inDataset mbgdr:2014-01_default ;
        dct:identifier ?cluster_id ;
        orth:member/orth:crossReference+ uniprot:K9Z723 ;
        orth:member/orth:gene ?gene .
    ?gene mbgd:uniprot ?uniprot ;
        dct:identifier ?gene_id ;
        orth:organism/rdfs:label ?organism .
    ?uniprot ?goa ?go_id .
    ?goa rdfs:subPropertyOf+ goa:goaExperimental ;
         rdfs:label ?evidence .
    ?go_id skos:prefLabel ?go.
}
```

**Figure 16. The portal page of MBGD SPARQL Search.**

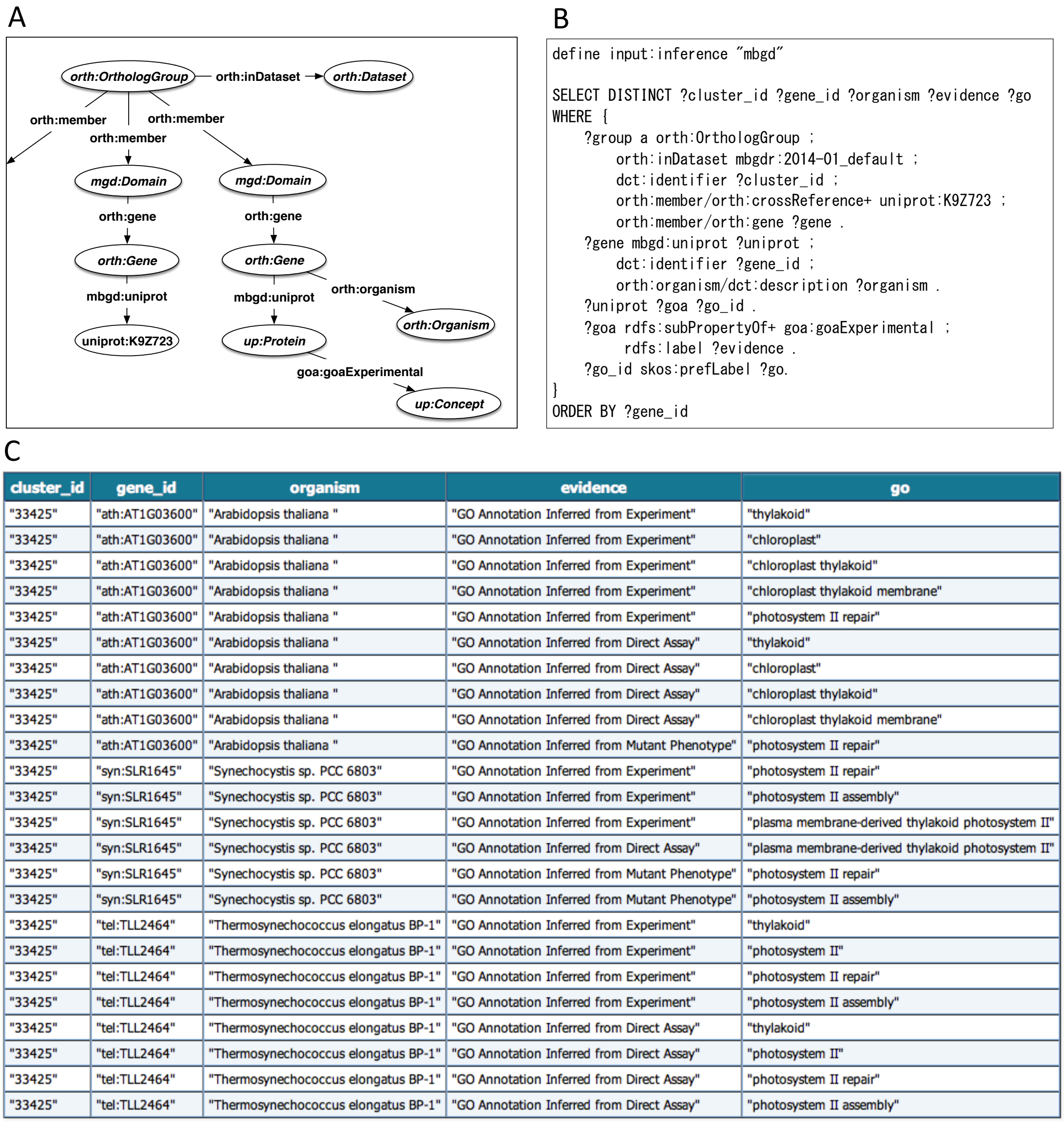
**Table 6. List of the datasets available at the MBGD SPARQL endpoint.**

| Dataset title | Graph name | Triples |
|---|---|---|
| MBGD ortholog groups | http://mbgd.genome.ad.jp/rdf/resource/2014-01_default | 76,155,196 |
| MBGD genes | http://mbgd.genome.ad.jp/rdf/resource/2014-01_gene | 686,902,009 |
| MBGD organism | http://mbgd.genome.ad.jp/rdf/resource/2014-01_organism | 31,397 |
| MBGD chromosomes and plasmids | http://mbgd.genome.ad.jp/rdf/resource/2014-01_nucseq | 6,796,757 |
| Cross-references from MBGD to UniProt | http://mbgd.genome.ad.jp/rdf/resource/2014-01_xref_uniprot | 8,012,666 |
| eggNOG COG | http://mbgd.genome.ad.jp/rdf/resource/eggnog_3.0_COG | 42,787,220 |
| eggNOG NOG | http://mbgd.genome.ad.jp/rdf/resource/eggnog_3.0_NOG | 21,469,150 |
| eggNOG proteins | http://mbgd.genome.ad.jp/rdf/resource/eggnog_3.0_protein | 24,572,358 |
| eggNOG organisms | http://mbgd.genome.ad.jp/rdf/resource/eggnog_3.0_organism | 1,144 |
| UniProt-GOA | http://mbgd.genome.ad.jp/rdf/resource/uniprot-goa | 274,338,183 |

### 3.3.3 Retrieving ortholog information of a specific protein

A typical use of an ortholog database is transferring functional annotations from known genes in model organisms to genes of unknown function in other organisms, on the basis of the conjecture that orthologs are usually functionally conserved. To demonstrate such an application in our database, we showed a query to retrieve ortholog information of a specified protein. Here, we specified a UniProt ID to obtain ortholog information. For describing functional categories of genes, we used Gene Ontology (GO) [64]. The UniProt GO Annotation (UniProt-GOA) database [65] (http://www.ebi.ac.uk/GOA) provides GO term assignment to proteins with evidence codes (http://www.geneontology.org/GO.evidence.shtml). We created an ontology for GO annotation (GOA-O, Table 5, http://purl.jp/bio/11/goa) and described UniProt-GOA data in RDF using it (Table 6). If some model organisms have experimentally verified GO annotations, we can transfer such a validated annotation to orthologs of other organisms.

Figure 17 shows an example SPARQL query to retrieve experimentally verified GO annotations assigned to some orthologs of the query protein UniProt K9Z723; Figure 17A shows the RDF data structure related to this query and Figure 17B shows the SPARQL code. While this protein of *Cyanobacterium aponinum* PCC 10605 does not have any GO annotation with experimental evidence codes, the ortholog information provides corresponding proteins of other organisms, including *Synechocystis* sp. PCC 6803, which have GO annotations, including "photosystem II repair," with experimental evidence codes such as "GO Annotation Inferred from Experiment" and "GO Annotation Inferred from Direct Assay," which are represented by subproperties of *goa:goaExperimental* (Figure 17C).



```
define input:inference "mbgd"

SELECT DISTINCT ?cluster_id ?gene_id ?organism ?evidence ?go
WHERE {
    ?group a orth:OrthologGroup ;
        orth:inDataset mbgdr:2014-01_default ;
        dct:identifier ?cluster_id ;
        orth:member/orth:crossReference+ uniprot:K9Z723 ;
        orth:member/orth:gene ?gene .
    ?gene mbgd:uniprot ?uniprot ;
        dct:identifier ?gene_id ;
        orth:organism/dct:description ?organism .
    ?uniprot ?goa ?go_id .
    ?goa rdfs:subPropertyOf+ goa:goaExperimental ;
         rdfs:label ?evidence .
    ?go_id skos:prefLabel ?go.
}
ORDER BY ?gene_id
```

| cluster_id | gene_id | organism | evidence | go |
|---|---|---|---|---|
| "33425" | "ath:AT1G03600" | "Arabidopsis thaliana " | "GO Annotation Inferred from Experiment" | "thylakoid" |
| "33425" | "ath:AT1G03600" | "Arabidopsis thaliana " | "GO Annotation Inferred from Experiment" | "chloroplast" |
| "33425" | "ath:AT1G03600" | "Arabidopsis thaliana " | "GO Annotation Inferred from Experiment" | "chloroplast thylakoid" |
| "33425" | "ath:AT1G03600" | "Arabidopsis thaliana " | "GO Annotation Inferred from Experiment" | "chloroplast thylakoid membrane" |
| "33425" | "ath:AT1G03600" | "Arabidopsis thaliana " | "GO Annotation Inferred from Experiment" | "photosystem II repair" |
| "33425" | "ath:AT1G03600" | "Arabidopsis thaliana " | "GO Annotation Inferred from Direct Assay" | "thylakoid" |
| "33425" | "ath:AT1G03600" | "Arabidopsis thaliana " | "GO Annotation Inferred from Direct Assay" | "chloroplast" |
| "33425" | "ath:AT1G03600" | "Arabidopsis thaliana " | "GO Annotation Inferred from Direct Assay" | "chloroplast thylakoid" |
| "33425" | "ath:AT1G03600" | "Arabidopsis thaliana " | "GO Annotation Inferred from Direct Assay" | "chloroplast thylakoid membrane" |
| "33425" | "ath:AT1G03600" | "Arabidopsis thaliana " | "GO Annotation Inferred from Mutant Phenotype" | "photosystem II repair" |
| "33425" | "syn:SLR1645" | "Synechocystis sp. PCC 6803" | "GO Annotation Inferred from Experiment" | "photosystem II repair" |
| "33425" | "syn:SLR1645" | "Synechocystis sp. PCC 6803" | "GO Annotation Inferred from Experiment" | "photosystem II assembly" |
| "33425" | "syn:SLR1645" | "Synechocystis sp. PCC 6803" | "GO Annotation Inferred from Experiment" | "plasma membrane-derived thylakoid photosystem II" |
| "33425" | "syn:SLR1645" | "Synechocystis sp. PCC 6803" | "GO Annotation Inferred from Direct Assay" | "plasma membrane-derived thylakoid photosystem II" |
| "33425" | "syn:SLR1645" | "Synechocystis sp. PCC 6803" | "GO Annotation Inferred from Mutant Phenotype" | "photosystem II repair" |
| "33425" | "syn:SLR1645" | "Synechocystis sp. PCC 6803" | "GO Annotation Inferred from Mutant Phenotype" | "photosystem II assembly" |
| "33425" | "tel:TLL2464" | "Thermosynechococcus elongatus BP-1" | "GO Annotation Inferred from Experiment" | "thylakoid" |
| "33425" | "tel:TLL2464" | "Thermosynechococcus elongatus BP-1" | "GO Annotation Inferred from Experiment" | "photosystem II" |
| "33425" | "tel:TLL2464" | "Thermosynechococcus elongatus BP-1" | "GO Annotation Inferred from Experiment" | "photosystem II repair" |
| "33425" | "tel:TLL2464" | "Thermosynechococcus elongatus BP-1" | "GO Annotation Inferred from Experiment" | "photosystem II assembly" |
| "33425" | "tel:TLL2464" | "Thermosynechococcus elongatus BP-1" | "GO Annotation Inferred from Direct Assay" | "thylakoid" |
| "33425" | "tel:TLL2464" | "Thermosynechococcus elongatus BP-1" | "GO Annotation Inferred from Direct Assay" | "photosystem II" |
| "33425" | "tel:TLL2464" | "Thermosynechococcus elongatus BP-1" | "GO Annotation Inferred from Direct Assay" | "photosystem II repair" |
| "33425" | "tel:TLL2464" | "Thermosynechococcus elongatus BP-1" | "GO Annotation Inferred from Direct Assay" | "photosystem II assembly" |

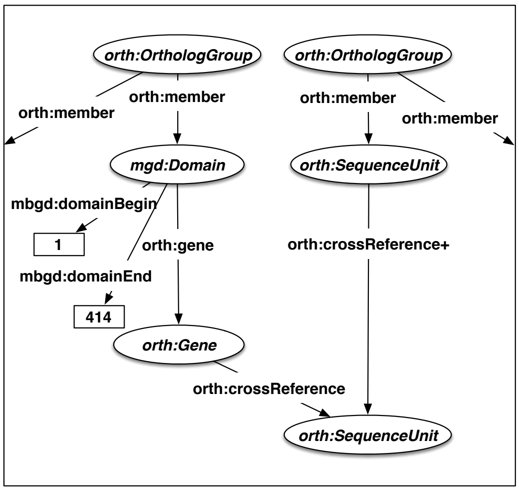
**Figure 17. Retrieval of ortholog information of a specific protein.**

(A) Schematic diagram of the RDF graph structure related to the query in B. The elliptical nodes represent resources. Specifically, the shaded elliptical nodes where classes are shown in italics represent the instances of the classes. In the unshaded elliptical node, the URI of the resource is directly shown. (B) SPARQL query to get GO annotation of an ortholog group. The prefix declarations are omitted for readability. (C) Search results of the query shown in B.

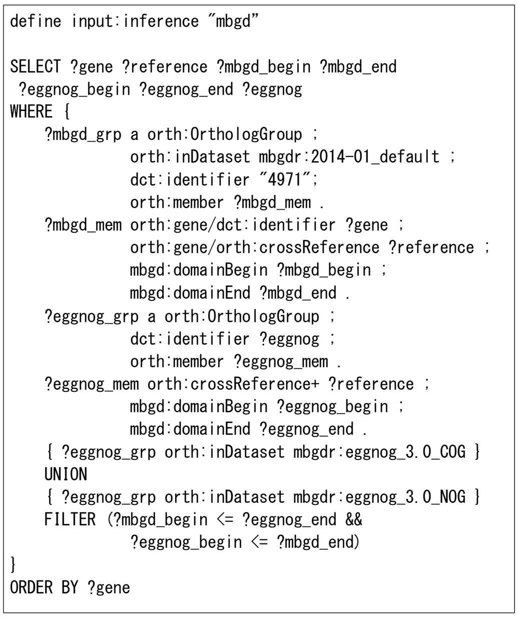
### 3.3.4 Comparison of ortholog information from different data sources

There are various ortholog databases that are constructed based on different methods and different sets of genomes. The users' concerns when using them may include comparing or merging ortholog groups derived from different data sources. However, the differences in the resource IDs (e.g., gene IDs and organism IDs) between them could hamper this task because finding identical members between corresponding groups is not straightforward. In our framework, even if independent gene IDs are used in different ortholog databases, cross-reference information assigned to each gene in each database can indirectly create a linkage between the corresponding genes through a common cross-reference. To test the comparability between data from different sources, we used the RDF version of ortholog information of eggNOG in addition to MBGD. The SPARQL query shown in Figure 18 finds eggNOG clusters corresponding to a given MBGD cluster. Although gene IDs in MBGD and those of corresponding genes in eggNOG are different, cross-references to the common database entry IDs, in this case UniProt IDs or RefSeq IDs, make it possible to interlink corresponding entries in MBGD and eggNOG. Although the query in Figure 18B does not explicitly specify any database name for cross-reference, it can find corresponding entries through either UniProt ID or RefSeq ID because the general property *orth:crossReference* is a super-property of both *mbgd:uniprot* (referring to UniProt ID) and *mbgd:protein* (referring to RefSeq ID) and *orth:crossReference+* allows arbitrary times (one or more) of any cross-references. Thus, the abstraction mechanism based on the ontology enhances the integration of different datasets by hiding implementation details in each database. The query compares the MBGD and eggNOG cluster members (both of which are domains, i.e., sub-sequences of proteins), and finds overlapping segments within the same genes (Figure 18C). It is possible to make more useful linkages between ortholog groups from different databases using a more complicated query (Figure 19); the number of common members is divided by each group size to produce overlap ratios, which are then used to define the relations between the ortholog groups, such as equivalent, subgroup, and supergroup with a similar criterion to that for the cross-reference section in the MBGD database [50].

A

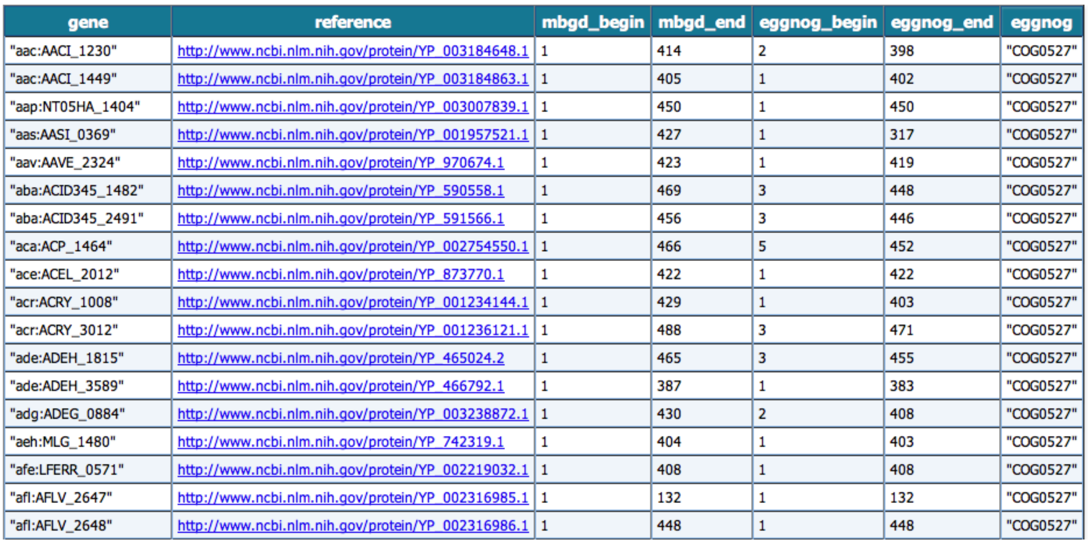


B



```
define input:inference "mbgd"

SELECT ?gene ?reference ?mbgd_begin ?mbgd_end
 ?eggnog_begin ?eggnog_end ?eggnog
WHERE {
    ?mbgd_grp a orth:OrthologGroup ;
              orth:inDataset mbgdr:2014-01_default ;
              dct:identifier "4971";
              orth:member ?mbgd_mem .
    ?mbgd_mem orth:gene/dct:identifier ?gene ;
              orth:gene/orth:crossReference ?reference ;
              mbgd:domainBegin ?mbgd_begin ;
              mbgd:domainEnd ?mbgd_end .
    ?eggnog_grp a orth:OrthologGroup ;
              dct:identifier ?eggnog ;
              orth:member ?eggnog_mem .
    ?eggnog_mem orth:crossReference+ ?reference ;
              mbgd:domainBegin ?eggnog_begin ;
              mbgd:domainEnd ?eggnog_end .
    { ?eggnog_grp orth:inDataset mbgdr:eggnog_3.0_COG }
    UNION
    { ?eggnog_grp orth:inDataset mbgdr:eggnog_3.0_NOG }
    FILTER (?mbgd_begin <= ?eggnog_end &&
              ?eggnog_begin <= ?mbgd_end)
}
ORDER BY ?gene
```

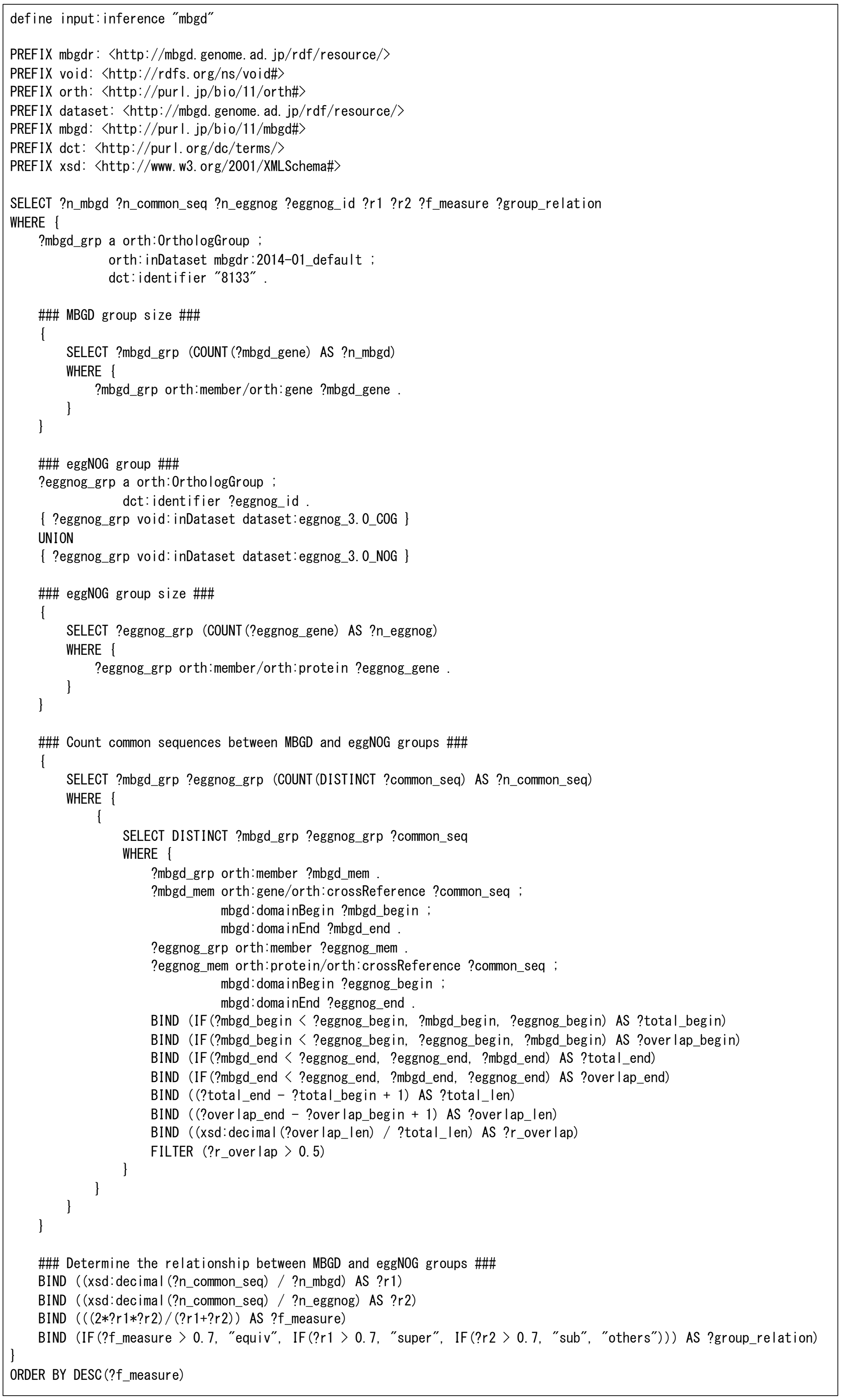
C



| gene | reference | mbgd_begin | mbgd_end | eggnog_begin | eggnog_end | eggnog |
|---|---|---|---|---|---|---|
| "aac:AACI_1230" | http://www.ncbi.nlm.nih.gov/protein/YP_003184648.1 | 1 | 414 | 2 | 398 | "COG0527" |
| "aac:AACI_1449" | http://www.ncbi.nlm.nih.gov/protein/YP_003184863.1 | 1 | 405 | 1 | 402 | "COG0527" |
| "aap:NT05HA_1404" | http://www.ncbi.nlm.nih.gov/protein/YP_003007839.1 | 1 | 450 | 1 | 450 | "COG0527" |
| "aas:AASI_0369" | http://www.ncbi.nlm.nih.gov/protein/YP_001957521.1 | 1 | 427 | 1 | 317 | "COG0527" |
| "aav:AAVE_2324" | http://www.ncbi.nlm.nih.gov/protein/YP_970674.1 | 1 | 423 | 1 | 419 | "COG0527" |
| "aba:ACID345_1482" | http://www.ncbi.nlm.nih.gov/protein/YP_590558.1 | 1 | 469 | 3 | 448 | "COG0527" |
| "aba:ACID345_2491" | http://www.ncbi.nlm.nih.gov/protein/YP_591566.1 | 1 | 456 | 3 | 446 | "COG0527" |
| "aca:ACP_1464" | http://www.ncbi.nlm.nih.gov/protein/YP_002754550.1 | 1 | 466 | 5 | 452 | "COG0527" |
| "ace:ACEL_2012" | http://www.ncbi.nlm.nih.gov/protein/YP_873770.1 | 1 | 422 | 1 | 422 | "COG0527" |
| "acr:ACRY_1008" | http://www.ncbi.nlm.nih.gov/protein/YP_001234144.1 | 1 | 429 | 1 | 403 | "COG0527" |
| "acr:ACRY_3012" | http://www.ncbi.nlm.nih.gov/protein/YP_001236121.1 | 1 | 488 | 3 | 471 | "COG0527" |
| "ade:ADEH_1815" | http://www.ncbi.nlm.nih.gov/protein/YP_465024.2 | 1 | 465 | 3 | 455 | "COG0527" |
| "ade:ADEH_3589" | http://www.ncbi.nlm.nih.gov/protein/YP_466792.1 | 1 | 387 | 1 | 383 | "COG0527" |
| "adg:ADEG_0884" | http://www.ncbi.nlm.nih.gov/protein/YP_003238872.1 | 1 | 430 | 2 | 408 | "COG0527" |
| "aeh:MLG_1480" | http://www.ncbi.nlm.nih.gov/protein/YP_742319.1 | 1 | 404 | 1 | 403 | "COG0527" |
| "afe:LFERR_0571" | http://www.ncbi.nlm.nih.gov/protein/YP_002219032.1 | 1 | 408 | 1 | 408 | "COG0527" |
| "afl:AFLV_2647" | http://www.ncbi.nlm.nih.gov/protein/YP_002316985.1 | 1 | 132 | 1 | 132 | "COG0527" |
| "afl:AFLV_2648" | http://www.ncbi.nlm.nih.gov/protein/YP_002316986.1 | 1 | 448 | 1 | 448 | "COG0527" |

**Figure 18. Comparison of ortholog information from different data sources.**

(A) Schematic diagram of the RDF graph structure related to the query in B. The elliptical nodes represent instances of classes. The rectangular nodes represent literals (integers in this example). (B) SPARQL query to compare orthologs between MBGD and eggNOG. The first line enables the inference based on sub-class and sub-property relations (see 3.2 Methods). (C) Search results of the query shown in B.



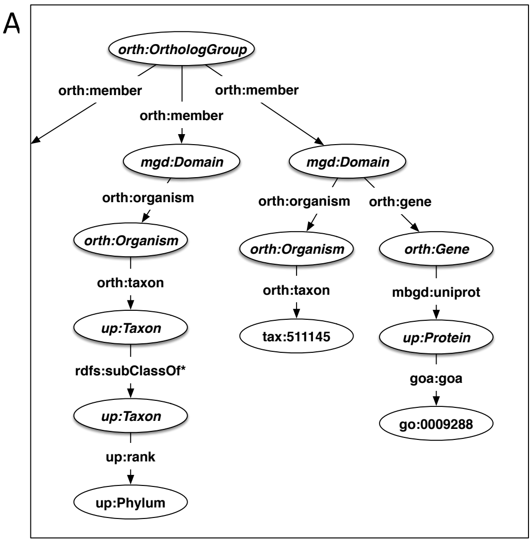
```
define input:inference "mbgd"

PREFIX mbgdr: <http://mbgd.genome.ad.jp/rdf/resource/>
PREFIX void: <http://rdfs.org/ns/void#>
PREFIX orth: <http://purl.jp/bio/11/orth#>
PREFIX dataset: <http://mbgd.genome.ad.jp/rdf/resource/>
PREFIX mbgd: <http://purl.jp/bio/11/mbgd#>
PREFIX dct: <http://purl.org/dc/terms/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

SELECT ?n_mbgd ?n_common_seq ?n_eggnog ?eggnog_id ?r1 ?r2 ?f_measure ?group_relation
WHERE {
    ?mbgd_grp a orth:OrthologGroup ;
              orth:inDataset mbgdr:2014-01_default ;
              dct:identifier "8133" .

    ### MBGD group size ###
    {
        SELECT ?mbgd_grp (COUNT(?mbgd_gene) AS ?n_mbgd)
        WHERE {
            ?mbgd_grp orth:member/orth:gene ?mbgd_gene .
        }
    }

    ### eggNOG group ###
    ?eggnog_grp a orth:OrthologGroup ;
                dct:identifier ?eggnog_id .
    { ?eggnog_grp void:inDataset dataset:eggnog_3.0_COG }
    UNION
    { ?eggnog_grp void:inDataset dataset:eggnog_3.0_NOG }

    ### eggNOG group size ###
    {
        SELECT ?eggnog_grp (COUNT(?eggnog_gene) AS ?n_eggnog)
        WHERE {
            ?eggnog_grp orth:member/orth:protein ?eggnog_gene .
        }
    }

    ### Count common sequences between MBGD and eggNOG groups ###
    {
        SELECT ?mbgd_grp ?eggnog_grp (COUNT(DISTINCT ?common_seq) AS ?n_common_seq)
        WHERE {
            {
                SELECT DISTINCT ?mbgd_grp ?eggnog_grp ?common_seq
                WHERE {
                    ?mbgd_grp orth:member ?mbgd_mem .
                    ?mbgd_mem orth:gene/orth:crossReference ?common_seq ;
                              mbgd:domainBegin ?mbgd_begin ;
                              mbgd:domainEnd ?mbgd_end .
                    ?eggnog_grp orth:member ?eggnog_mem .
                    ?eggnog_mem orth:protein/orth:crossReference ?common_seq ;
                              mbgd:domainBegin ?eggnog_begin ;
                              mbgd:domainEnd ?eggnog_end .
                    BIND (IF(?mbgd_begin < ?eggnog_begin, ?mbgd_begin, ?eggnog_begin) AS ?total_begin)
                    BIND (IF(?mbgd_begin < ?eggnog_begin, ?eggnog_begin, ?mbgd_begin) AS ?overlap_begin)
                    BIND (IF(?mbgd_end < ?eggnog_end, ?eggnog_end, ?mbgd_end) AS ?total_end)
                    BIND (IF(?mbgd_end < ?eggnog_end, ?mbgd_end, ?eggnog_end) AS ?overlap_end)
                    BIND ((?total_end - ?total_begin + 1) AS ?total_len)
                    BIND ((?overlap_end - ?overlap_begin + 1) AS ?overlap_len)
                    BIND ((xsd:decimal(?overlap_len) / ?total_len) AS ?r_overlap)
                    FILTER (?r_overlap > 0.5)
                }
            }
        }
    }

    ### Determine the relationship between MBGD and eggNOG groups ###
    BIND ((xsd:decimal(?n_common_seq) / ?n_mbgd) AS ?r1)
    BIND ((xsd:decimal(?n_common_seq) / ?n_eggnog) AS ?r2)
    BIND (((2*?r1*?r2)/(?r1+?r2)) AS ?f_measure)
    BIND (IF(?f_measure > 0.7, "equiv", IF(?r1 > 0.7, "super", IF(?r2 > 0.7, "sub", "others"))) AS ?group_relation)
}
ORDER BY DESC(?f_measure)
```

**Figure 19. A SPARQL query to compare ortholog groups obtained from different databases.**

### 3.3.5 Gene functions and phylogenetic patterns

As the third application, we showed queries that find relationships between gene functions and taxonomy of organisms by tracing the linkages in RDF (Figure 20). As an ortholog group is connected to a set of organisms as well as a set of functional categories through its members in RDF (Figure 20A), it can link between a gene function and a set of organisms having that function. If a functional category (GO term) is specified, we can obtain genes assigned that functional category and then the ortholog groups containing them. For example, the query shown in Figure 20B searches for MBGD clusters that include members assigned a specific GO term, GO:0009288 (bacterial-type flagellum) in *Escherichia coli* K-12 MG1655 (NCBI taxonomy ID 511145). The obtained list of MBGD clusters is shown in Figure 20D. One of the clusters (in this example, cluster 12897, namely fliG) is specified in the query shown in Figure 20C, which searches for organisms included in this cluster and returns the phylogenetic pattern summarized at a given taxonomic rank (in this example, phylum) using the hierarchical taxonomic classification.

Using the R environment (see 3.2 Methods), the SPARQL queries in Figure 20B and C were sequentially executed to obtain the phylogenetic pattern of the clusters, and then the results were visualized as a heat map (Figure 20E, see Figure 21 for the R source code). Out of the 26 clusters, 19 showed relatively wide organismal distribution ranging in at least 16 phyla of bacteria, whereas the other 7 clusters distributed in smaller ranges, including at least *Proteobacteria* that *E. coli* belongs to. More specifically, among the former 19 clusters with overall similarity in distributions, slight differences were observed. The differences in the phylogenetic pattern could reflect species- or taxon-specific functions of the bacterial flagellum genes, although they basically have known functions of bacterial motility. Specifically, clusters 14760 (flgI) and 14931 (flgH) tend to be missing in the phyla including gram-positive bacteria (marked by + in Figure 5E). Here, flgI products constitute the P (peptidoglycan) ring in the peptidoglycan layer and flgH products constitute the L (lipopolysaccharide) ring in the outer membrane. The molecular characteristics of these gene products correlate with the phylogenetic pattern that these genes are missing in the phyla including gram-positive bacteria that have a thick peptidoglycan layer and lack an outer membrane [66]. Notably, 4 clusters (fliR, fliN, fliQ and fliH) contain genes from *Chlamydiae*, whereas other clusters do not. Considering that *Chlamydia* are non-motile bacteria, this result suggests that these genes could be related to functions other than motility. In fact, these genes are known components of the type III secretion system, which delivers effectors into eukaryotic cells and is evolutionarily related to the bacterial flagellum [67,68].

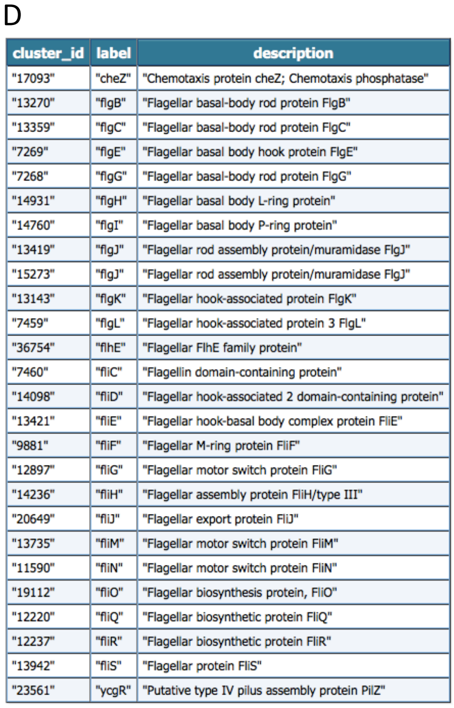


B

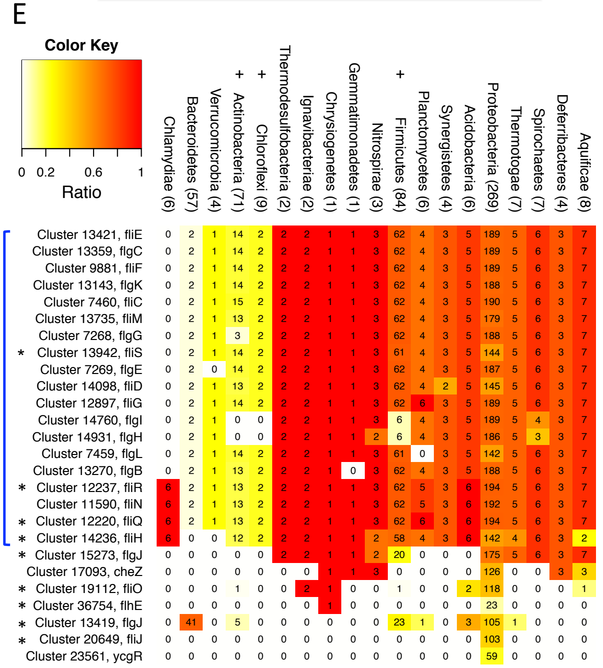
```
define input:inference "mbgd"

SELECT DISTINCT ?cluster_id ?label ?description
WHERE {
    ?group a orth:OrthologGroup ;
           orth:inDataset mbgdr:2014-01_default ;
           dct:identifier ?cluster_id ;
           rdfs:label ?label ;
           dct:description ?description ;
           orth:member/orth:gene ?gene .
    ?gene orth:organism/orth:taxon tax:511145 ;
           mbgd:uniprot/goa:goa go:0009288 .
}
ORDER BY ?label
```

C

```
SELECT ?phylum ?num_all
 (COUNT(DISTINCT ?organism) AS ?num)
 ((xsd:decimal(?num) / ?num_all) AS ?ratio)
WHERE {
    ?group a orth:OrthologGroup ;
           orth:inDataset mbgdr:2014-01_default ;
           dct:identifier "12897";
           orth:member ?member .
    ?member orth:organism ?organism .
    ?organism orth:taxon ?taxid .
    ?taxid rdfs:subClassOf* ?taxon .
    ?taxon up:rank up:Phylum ;
           up:scientificName ?phylum .
    {
        SELECT ?taxon (COUNT(?org) AS ?num_all)
        WHERE {
            mbgdr:2014-01_default orth:organism ?org .
            ?org orth:taxon ?tax .
            ?tax rdfs:subClassOf* ?taxon .
        }
    }
}
ORDER BY ?phylum
```



| cluster_id | label | description |
|---|---|---|
| "17093" | "cheZ" | "Chemotaxis protein cheZ; Chemotaxis phosphatase" |
| "13270" | "flgB" | "Flagellar basal-body rod protein FlgB" |
| "13359" | "flgC" | "Flagellar basal-body rod protein FlgC" |
| "7269" | "flgE" | "Flagellar basal body hook protein FlgE" |
| "7268" | "flgG" | "Flagellar basal-body rod protein FlgG" |
| "14931" | "flgH" | "Flagellar basal body L-ring protein" |
| "14760" | "flgI" | "Flagellar basal body P-ring protein" |
| "13419" | "flgJ" | "Flagellar rod assembly protein/muramidase FlgJ" |
| "15273" | "flgJ" | "Flagellar rod assembly protein/muramidase FlgJ" |
| "13143" | "flgK" | "Flagellar hook-associated protein FlgK" |
| "7459" | "flgL" | "Flagellar hook-associated protein 3 FlgL" |
| "36754" | "flhE" | "Flagellar FlhE family protein" |
| "7460" | "fliC" | "Flagellin domain-containing protein" |
| "14098" | "fliD" | "Flagellar hook-associated 2 domain-containing protein" |
| "13421" | "fliE" | "Flagellar hook-basal body complex protein FliE" |
| "9881" | "fliF" | "Flagellar M-ring protein FliF" |
| "12897" | "fliG" | "Flagellar motor switch protein FliG" |
| "14236" | "fliH" | "Flagellar assembly protein FliH/type III" |
| "20649" | "fliJ" | "Flagellar export protein FliJ" |
| "13735" | "fliM" | "Flagellar motor switch protein FliM" |
| "11590" | "fliN" | "Flagellar motor switch protein FliN" |
| "19112" | "fliO" | "Flagellar biosynthesis protein, FliO" |
| "12220" | "fliQ" | "Flagellar biosynthetic protein FliQ" |
| "12237" | "fliR" | "Flagellar biosynthetic protein FliR" |
| "13942" | "fliS" | "Flagellar protein FliS" |
| "23561" | "ycgR" | "Putative type IV pilus assembly protein PilZ" |



| | Chlamydiae (6) | Bacteroidetes (57) | Verrucomicrobia (4) | + Actinobacteria (71) | + Chloroflexi (9) | Thermodesulfobacteria (2) | Ignavibacteriae (2) | Chrysiogenetes (1) | Gemmatimonadetes (1) | Nitrospirae (3) | + Firmicutes (84) | Planctomycetes (6) | Synergistetes (4) | Acidobacteria (6) | Proteobacteria (269) | Thermotogae (7) | Spirochaetes (7) | Deferribacteres (4) | Aquificae (8) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Cluster 13421, fliE | 0 | 2 | 1 | 14 | 2 | 2 | 2 | 1 | 1 | 3 | 62 | 4 | 3 | 5 | 189 | 5 | 6 | 3 | 7 |
| Cluster 13359, flgC | 0 | 2 | 1 | 14 | 2 | 2 | 2 | 1 | 1 | 3 | 62 | 4 | 3 | 5 | 189 | 5 | 6 | 3 | 7 |
| Cluster 9881, fliF | 0 | 2 | 1 | 14 | 2 | 2 | 2 | 1 | 1 | 3 | 62 | 4 | 3 | 5 | 189 | 5 | 6 | 3 | 7 |
| Cluster 13143, flgK | 0 | 2 | 1 | 14 | 2 | 2 | 2 | 1 | 1 | 3 | 62 | 4 | 3 | 5 | 188 | 5 | 6 | 3 | 7 |
| Cluster 7460, fliC | 0 | 2 | 1 | 15 | 2 | 2 | 2 | 1 | 1 | 3 | 62 | 4 | 3 | 5 | 190 | 5 | 6 | 3 | 7 |
| Cluster 13735, fliM | 0 | 2 | 1 | 13 | 2 | 2 | 2 | 1 | 1 | 3 | 62 | 4 | 3 | 5 | 179 | 5 | 6 | 3 | 7 |
| Cluster 7268, flgG | 0 | 2 | 1 | 3 | 2 | 2 | 2 | 1 | 1 | 3 | 62 | 4 | 3 | 5 | 188 | 5 | 6 | 3 | 7 |
| * Cluster 13942, fliS | 0 | 2 | 1 | 14 | 2 | 2 | 2 | 1 | 1 | 3 | 61 | 4 | 3 | 5 | 144 | 5 | 6 | 3 | 7 |
| Cluster 7269, flgE | 0 | 2 | 0 | 14 | 2 | 2 | 2 | 1 | 1 | 3 | 62 | 4 | 3 | 5 | 187 | 5 | 6 | 3 | 7 |
| Cluster 14098, fliD | 0 | 2 | 1 | 13 | 2 | 2 | 2 | 1 | 1 | 3 | 62 | 4 | 2 | 5 | 145 | 5 | 6 | 3 | 7 |
| Cluster 12897, fliG | 0 | 2 | 1 | 14 | 2 | 2 | 2 | 1 | 1 | 3 | 62 | 6 | 3 | 5 | 189 | 5 | 6 | 3 | 7 |
| Cluster 14760, flgI | 0 | 2 | 1 | 0 | 0 | 2 | 2 | 1 | 1 | 3 | 6 | 4 | 3 | 5 | 189 | 5 | 4 | 3 | 7 |
| Cluster 14931, flgH | 0 | 2 | 1 | 0 | 0 | 2 | 2 | 1 | 1 | 2 | 6 | 4 | 3 | 5 | 186 | 5 | 3 | 3 | 7 |
| Cluster 7459, flgL | 0 | 2 | 1 | 14 | 2 | 2 | 2 | 1 | 1 | 3 | 61 | 0 | 3 | 5 | 142 | 5 | 6 | 3 | 7 |
| Cluster 13270, flgB | 0 | 2 | 1 | 13 | 2 | 2 | 2 | 1 | 0 | 3 | 62 | 4 | 3 | 5 | 188 | 5 | 6 | 3 | 7 |
| * Cluster 12237, fliR | 6 | 2 | 1 | 13 | 2 | 2 | 2 | 1 | 1 | 3 | 62 | 5 | 3 | 6 | 194 | 5 | 6 | 3 | 7 |
| Cluster 11590, fliN | 6 | 2 | 1 | 13 | 2 | 2 | 2 | 1 | 1 | 3 | 62 | 5 | 3 | 6 | 192 | 5 | 6 | 3 | 7 |
| * Cluster 12220, fliQ | 6 | 2 | 1 | 13 | 2 | 2 | 2 | 1 | 1 | 3 | 62 | 6 | 3 | 6 | 194 | 5 | 6 | 3 | 7 |
| * Cluster 14236, fliH | 6 | 0 | 0 | 12 | 2 | 2 | 2 | 1 | 1 | 2 | 58 | 4 | 3 | 6 | 142 | 4 | 6 | 3 | 2 |
| * Cluster 15273, flgJ | 0 | 0 | 0 | 0 | 0 | 2 | 2 | 1 | 1 | 2 | 20 | 0 | 0 | 0 | 175 | 5 | 6 | 3 | 7 |
| Cluster 17093, cheZ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 3 | 0 | 0 | 0 | 0 | 126 | 0 | 0 | 3 | 3 |
| * Cluster 19112, fliO | 0 | 0 | 0 | 1 | 0 | 0 | 2 | 1 | 0 | 0 | 1 | 0 | 0 | 2 | 118 | 0 | 0 | 0 | 1 |
| * Cluster 36754, flhE | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 23 | 0 | 0 | 0 | 0 |
| * Cluster 13419, flgJ | 0 | 41 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 23 | 1 | 0 | 3 | 105 | 1 | 0 | 0 | 0 |
| * Cluster 20649, fliJ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 103 | 0 | 0 | 0 | 0 |
| Cluster 23561, ycgR | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 59 | 0 | 0 | 0 | 0 |

**Figure 20. Retrieval of phylogenetic patterns of orthologs related to a specific function.**

(A) Schematic diagram of the RDF graph structure related to the queries in B and C. (B) SPARQL query to get MBGD clusters including members related to the GO term GO:0009288 (bacterial-type flagellum). (C) SPARQL query to obtain organisms that contain members of an ortholog group. (D) Search results of the query shown in B. (E) The queries shown in B and C were executed and the results were visualized using R. The number of target organisms in each phylum is shown in parenthesis. After obtaining the output from R, the phyla containing gram-positive bacteria (+) and genes functioning in the flagellar export system (*) are marked, and the blue line was added to represent clusters with relatively wide organismal distribution (in at least 16 phyla).

```
library(SPARQL)
library(gplots)

endpoint <- "http://sparql.nibb.ac.jp/sparql"
query.file1 <- "example3-1.rq"
query.file2 <- "example3-2.rq"

read.query <- function(query.file, arg="") {
  query.lines <- scan(query.file, what="char", sep="\n")
  query.lines <- gsub(”\\$ARG", arg, query.lines) # replacing a placeholders in the query
  query.string <- paste(query.lines, collapse="\n")
  return(query.string)
}

### Execute SPARQL query 1 ###
query1.result <- SPARQL(endpoint, read.query(query.file1))
cluster.list <- query1.result$results$cluster_id
label.list <- query1.result$results$label
n_clusters <- length(cluster.list)

### Execute SPARQL query 2 ###
result.list <- c()
phylum.list <- c()
for (i in 1:n_clusters) {
  query2.result <- SPARQL(endpoint, read.query(query.file2, arg=cluster.list[i]))
  result.list[[i]] <- query2.result
  phylum.list <- c(phylum.list, query2.result$results$phylum)
}
phylum.list <- sort(unique(phylum.list))
n_phylum <- length(phylum.list)

### Summarization ###
ratio.mat <- matrix(1:(n_clusters * n_phylum), n_clusters, n_phylum)
num.mat <- matrix(1:(n_clusters * n_phylum), n_clusters, n_phylum)
num.list <- c()
for (i in 1:n_clusters) {
  result <- result.list[[i]]$results
  for (j in 1:n_phylum) {
    ratio <- result$ratio[result$phylum==phylum.list[j]]
    num <- result$num[result$phylum==phylum.list[j]]
    num_all <- result$num_all[result$phylum==phylum.list[j]]
    ratio.mat[i,j] <- ifelse(length(ratio)==1, ratio, 0)
    num.mat[i,j] <- ifelse(length(num)==1, num, 0)
    if (length(num_all)==1) {
      num.list[j] <- num_all
    }
  }
}

### Visualization ###
heatmap.2(ratio.mat, col=rev(heat.colors(256)), cellnote=num.mat, notecex=0.8, notecol="black",
          dendrogram="none", trace="none", density.info="none", key.xlab="Ratio",
          labRow=paste("Cluster ", cluster.list, ", ", label.list, sep=""), offsetRow=-27, adjRow=c(1,NA),
          labCol=paste(phylum.list, " (", num.list, ")", sep=""), srtCol=270, offsetCol=-45, adjCol=c(1,NA))
```

**Figure 21. R source code for execution of SPARQL queries and visualization of the results.**

This R source code was used to execute SPARQL queries shown in Figure 20A,C and to output the heat map shown in Figure 20E. Note that the query shown in Figure 20A was saved as a file “example3-1.rq”, and Figure 20B as a file “example3-2.rq” in which the cluster ID “12897” was replaced by a placeholder “$ARG”.

### 3.3.6 Performance of the RDF store

The performances of the existing programs implementing the Semantic Web technology have recently been improved. These improvements include updates of Virtuoso. However, the performances need to be further improved. With respect to hardware performance, high-speed drives such as solid-state drives can enhance the database performance. We measured the time required for loading and querying to the Virtuoso installed on a solid-state drive. As a result, the loading of the dataset "MBGD chromosomes and plasmids" in Table 2 (6,796,757 triples) took 28.4 s, indicating that the loading speed was 14.4 million triples per minute. The execution time required for the queries is shown in Table 3.

**Table 7. Time required for executing SPARQL queries.**

| Query | Time (s) |
|---|---:|
| Figure 3A | 1.2 |
| Figure 4A | 3.3 |
| Figure 5A | 1.6 |
| Figure 5B | 19.1 |

## 3.4 Discussion

In the study presented in this chapter, together with my collaborators I developed an RDF model for integrating the genomic data of multiple organisms using the ortholog information as a hub. The RDF model is constructed based on the OrthO, which is a compact ontology for general use but is also designed to be extensible to cover database-specific concepts. We demonstrated the usefulness of our model in the integrative analysis of multiple genomes by describing the MBGD database in RDF and linking the data to various other resources. With respect to database integration, this study addressed several aspects. One aspect is to create linkages between various biological resources on the basis of the connective nature of RDF. Another is to use the characteristics of orthologs as a hub of links between organisms. The third is to utilize ontologies that have the ability to unite different terminologies.

One of the main advantages of using the Semantic Web technology instead of conventional technologies is that databases get highly connectable to other resources, both locally and globally on the Web. More specifically, improved local connectivity means that when several graphs are imported into an RDF store, a merged graph is automatically generated, and the concatenation of the RDF files immediately produces a valid merged RDF file (in the case of specific formats such as Turtle and N-triples [69]). Moreover, improved global connectivity means that the RDF data are accessible through SPARQL across the Web, which will ultimately transform the Web into a big database cooperatively constructed by developers worldwide. The Semantic Web includes another form of global access that does not require the SPARQL endpoint; direct access to a resource URI can be used as an easy way of connecting the resource information if it returns data. Our database not only accepts access to the SPARQL endpoint but also provides an easy access without SPARQL by specifying URIs on the web browsers (see 3.2.5 Browsing the RDF).

Database integration consequently enables the comparison of corresponding data from different data sources that otherwise could not be easily compared. In this study, we compared the ortholog groups in different classification systems (i.e., MBGD and eggNOG) by comparing the members in each group. Finding the corresponding groups (or subgroups) through such a comparison of their members is not limited to the case of ortholog group comparison but instead is a general issue for other types of groupings such as functional categories. In general, comparing members between different grouping systems produces new relations between the groups (or between the concepts behind each group, such as GO terms). Even if different grouping systems use specific types of resource IDs, we can identify the corresponding genes in RDF as those having links to a common reference through *orth:crossReference* or its subproperties.

Ortholog information, by nature, has a hub structure that connects corresponding genes between organisms. Besides, the evolutionary relationships between orthologs or organisms are often represented in tree structures. Thus, the representation of ortholog information and their evolutionary relationship using graphs is a straightforward approach. In fact, we demonstrated that the RDF version of ortholog and taxonomy information is useful for analyzing the genomic contents of multiple organisms (Figure 20). The use of recursive property paths of SPARQL 1.1 functionalities (such as *rdfs:subClassOf** and *orth:member+*) is suitable for the retrieval of data by traversing a hierarchical structure and enhances the usefulness of the RDF version of the ortholog database.

Because of the existence of various formats for ortholog information, there is a need for standards in the orthology field. OrthoXML was developed as the first standard format of orthology information. In comparison to XML, the RDF representation has several merits. It enables data merging by just concatenating the collected files and is searchable by complex queries. Besides, RDF has a high flexibility and extensibility. For example, some ortholog databases such as MBGD and eggNOG include positional information within genes for which sequence homology is detected. OrthoXML is not suitable for expressing the concept of domain-level orthology. Although the OrthO basically conforms to OrthoXML but it can be extended to represent database-specific concepts such as domain-level orthology. Thus, the RDF representation using the OrthO is a good candidate of a more general model for describing ortholog information.

Although only a universal terminology is ideal in terms of worldwide database integration, each research group may propose their own terminology depending on their specific scopes, resulting in the existence of equivalent concepts in different terminologies. However, the problem of different terminologies could be solved by translating them. Here, we created an association between OrthO, OGO, SO, and SIO. Besides, each database has its specific concepts and it will be convenient for the maintainer to treat the database by designing the terminologies that reflect the concepts. In this study, we defined the MBGD-specific terms under the general terms of OrthO. Thus, the general terminology worked as an abstraction layer over the database-specific terms. Searching the database using the general terminology as abstraction enables the search against similar data in different databases, thus enabling crossover search and integrative analysis. The Semantic Web, with this abstraction mechanism, reduces the burden of data integration even though individual database developers implement their databases differently according to their own philosophies, which is quite common in cutting-edge scientific fields.

## 3.5 Conclusions

Together with my collaborators, I developed a general RDF model for describing ortholog information on the basis of an ontology OrthO. The model enables the integration of functional information for multiple organisms. Besides, the ortholog information from different data sources can be compared using the OrthO as a shared ontology. By representing the data in this RDF model, the ortholog database can work as a hub structure for biological databases in the Semantic Web, and it will contribute to knowledge discovery through integrative data analysis.

# Chapter 4 Functional analysis of orthologous gene conservation

## 4.1 Background

There were many pioneering works on the molecular evolution of mammalian protein sequences [19], which were followed by large-scale comparative analyses between species. Wolfe and Sharp [20] analyzed a collection of 363 mouse and rat orthologous gene pairs, and Murphy [21] examined 615 pairs of orthologous genes between human and rodents. Makalowski et al. [22] performed a comparative analysis for 1,196 cDNA pairs between human and rodents. These studies revealed that the evolutionary rates of protein sequences depend on the protein functions. For example, ribosomal proteins and Ras-like GTPases are highly conserved [22], while proteins for antimicrobial host defenses are highly divergent [21].

On the other hand, comparisons of upstream non-coding sequences have been conducted to investigate the regulatory sequences. The complete sequences of mammalian genomes [23-25] facilitated large-scale comparisons of non-coding sequences, which provided insights about regulatory sequences. Iwama and Gojobori [26] compared the upstream sequences of 3,750 human-mouse orthologous gene pairs and found that transcription factor genes, particularly those related to developmental processes, show high upstream sequence conservation. Lee et al. [27] also reported that genes involved in adaptive processes tend to have highly conserved upstream regions in mammalian genomes. Choi et al. [28] investigated the levels of non-coding conservation, focusing on tissue-specific genes.

While many efforts have been made to examine protein sequence conservation or regulatory sequence conservation, the relationships between them are poorly understood. Although several researchers have addressed a similar issue, where the relationship between protein evolution and regulatory evolution was examined based on microarray expression data [70-75], there is a discrepancy among their conclusions. Some of the researchers concluded that these two kinds of evolution are decoupled [70,73], while others claimed that there was indeed a correlation between them [71,72,74,75]. Since a substantial amount of the regulatory information is embedded in the promoter region, which is located proximal to the transcriptional start site, examining the protein sequence evolution in relation to the promoter sequence is an alternative approach to address this problem. Castillo-Davis et al. [76] made the first investigation of the relationship between protein and cis-regulatory sequence evolution using nematode genomes, and observed a weak correlation. As a step to broaden our understanding of genome evolution and function, it seems important to examine these sequences in mammalian genomes, and to analyze them in detail to dissect the relationship. However, such a sequence level analysis has not been carried out for mammals. One of the main problems is the precise determination of the TSS, which is indispensable for identifying reliable promoter regions.

Experimentally validated TSS information can provide a basis for a reliable promoter analysis. Based on large-scale collections of full-length cDNAs [77-80], Database of Transcriptional Start Sites (DBTSS) [81-83] was constructed and enabled the reliable identification, annotation and analysis of promoter regions [84-86]. Since abundant TSS data for human and mouse were integrated into DBTSS, large-scale cross-species comparisons of promoter regions became possible [87,88].

Along with my collaborators, I compared promoter sequences as well as protein sequences for 6,901 human and mouse orthologous genes, aiming at two points [29]. First, we carried out a comprehensive comparison of human and mouse promoter sequences, to examine the relationship between promoter conservation and gene function. Second, we tried to elucidate what kinds of relationships exist between promoter conservation and protein conservation in mammals. In the second part, we not only examined the extent of correlation between them, but also investigated the relationship in further detail, by decomposing it based on the functional categories of genes. The results revealed that there seem to be nonparallel components between protein and promoter sequence evolution.

## 4.2 Methods

### 4.2.1 Sequence comparison

From DBTSS, we obtained human and mouse orthologous gene pairs with experimentally validated TSS information. The definition of an orthologous relationship is based on HomoloGene [89]. One-to-multi orthologous relationships were removed, resulting in 8,429 one-to-one orthologous gene pairs. Since the TSSs for a given gene are not fixed but vary on the chromosome, a representative TSS was defined for each gene, as described in Yamashita et al. [84]. Based on the positions of representative TSSs, sequences from -1000 to +200 were defined as putative promoter sequences. Promoters of orthologous gene pairs were aligned by the local alignment program *water* from the EMBOSS package [90]. In addition, promoter pairs to be used as a negative control were created by shuffling the original pairings, and were aligned similarly. The protein sequences of orthologous gene pairs were obtained from the NCBI reference sequence (RefSeq) database [91]. They were also aligned with *water*. For additional analyses by global alignments, *needle* from the EMBOSS package was used. Furthermore, we confirmed the results after eliminating coding sequences contained in promoter sequences, as follows. The coding sequences downstream of the TSSs were removed by restricting the promoter sequences from -1000 to -1 of the TSSs. In addition, since 16% of the shortened sequences (1,101 out of 6,901) still contained coding sequences, I used the other 5,800 sequences for the additional analyses.

To display the distributions of the alignment scores, they were transformed by common logarithmic transformation, and then the densities were estimated by R with the Gaussian kernel and a band width of 0.5. For protein sequences, protein diversity, instead of identity, was subjected to the logarithmic transformation. In addition, to avoid zero before the logarithmic transformation, a small number was added. Thus, 105 – *identity* was subjected to the logarithmic transformation. This transformation is similar to that described in a previous study on protein evolutionary rates [92].

### 4.2.2 Functional annotation of genes

Annotations of genes were based on the gene ontology (GO) [93]. The GO annotations for the human and mouse genes were obtained from the gene2go file at NCBI [45]. In this study, to summarize the attributes of the genes, we developed a slimmed-down version of the GO vocabulary (GO slim), as follows. A set of high level terms was selected to cover most aspects of each of the three ontologies (52 terms for biological process, 22 terms for cellular component and 26 terms for molecular function; for the complete list of selected GO terms, see Table 8). Basically, GO terms containing over 100 genes were selected, although well-known cellular components with smaller number of genes, such as C:lysosome and C:peroxisome, were also included. Overly general terms, such as C:cell, P:physiological process and F:binding, were removed, because their biological interpretation seems uninformative. Each GO term was mapped to the GO slim terms using map2slim.pl from the go-perl package [33]. Note that several GO slim terms can be assigned to a single gene; that is, the GO slim terms are not mutually exclusive. In the following sections of this chapter, the GO slim terms are referred to as “GO term” for short.

### 4.2.3 Significance test of conservation

We tested whether the alignment scores (or percentage identities) of a set of genes associated with a given GO term are significantly high or low by a Wilcoxon rank sum test. The *P*-values were used to calculate the false discovery rate (FDR) [94]. It should be noted that the genes used as a control group of a term are those that are not associated with the term, but with other terms. For example, in the case of “transcription” of biological process, 640 genes are associated with the term among 6,901 genes. Of the 6,261 genes that are not associated with “transcription”, 2,116 genes are missing terms of biological processes. Since these “uncharacterized” genes had low sequence conservation tendencies, we eliminated them from the control gene set. The resulting control set in the case of “transcription” is thus composed of 4,145 genes.

**Table 8. List of 100 GO terms selected for the functional analysis.**

| GO ID | ontology | GO term |
|---|---|---|
| GO:0000902 | biological process | cellular morphogenesis |
| GO:0005975 | biological process | carbohydrate metabolism |
| GO:0006066 | biological process | alcohol metabolism |
| GO:0006118 | biological process | electron transport |
| GO:0006259 | biological process | DNA metabolism |
| GO:0006350 | biological process | transcription |
| GO:0006396 | biological process | RNA processing |
| GO:0006412 | biological process | protein biosynthesis |
| GO:0006457 | biological process | protein folding |
| GO:0006461 | biological process | protein complex assembly |
| GO:0006468 | biological process | protein amino acid phosphorylation |
| GO:0006508 | biological process | proteolysis |
| GO:0006512 | biological process | ubiquitin cycle |
| GO:0006520 | biological process | amino acid metabolism |
| GO:0006629 | biological process | lipid metabolism |
| GO:0006811 | biological process | ion transport |
| GO:0006915 | biological process | apoptosis |
| GO:0006928 | biological process | cell motility |
| GO:0006950 | biological process | response to stress |
| GO:0006955 | biological process | immune response |
| GO:0007010 | biological process | cytoskeleton organization and biogenesis |
| GO:0007049 | biological process | cell cycle |
| GO:0007155 | biological process | cell adhesion |
| GO:0007165 | biological process | signal transduction |
| GO:0007166 | biological process | cell surface receptor linked signal transduction |
| GO:0007186 | biological process | G-protein coupled receptor protein signaling pathway |
| GO:0007242 | biological process | intracellular signaling cascade |
| GO:0007243 | biological process | protein kinase cascade |
| GO:0007264 | biological process | small GTPase mediated signal transduction |
| GO:0007267 | biological process | cell-cell signaling |
| GO:0007275 | biological process | development |
| GO:0007399 | biological process | nervous system development |
| GO:0007600 | biological process | sensory perception |
| GO:0008283 | biological process | cell proliferation |
| GO:0008610 | biological process | lipid biosynthesis |
| GO:0009056 | biological process | catabolism |
| GO:0009607 | biological process | response to biotic stimulus |
| GO:0009628 | biological process | response to abiotic stimulus |
| GO:0009653 | biological process | morphogenesis |
| GO:0009892 | biological process | negative regulation of metabolism |
| GO:0016192 | biological process | vesicle-mediated transport |
| GO:0019752 | biological process | carboxylic acid metabolism |
| GO:0030154 | biological process | cell differentiation |
| GO:0042221 | biological process | response to chemical stimulus |
| GO:0045045 | biological process | secretory pathway |
| GO:0045449 | biological process | regulation of transcription |
| GO:0046907 | biological process | intracellular transport |
| GO:0048513 | biological process | organ development |
| GO:0048518 | biological process | positive regulation of biological process |
| GO:0048519 | biological process | negative regulation of biological process |

| | | |
|---|---|---|
| GO:0051186 | biological process | cofactor metabolism |
| GO:0051276 | biological process | chromosome organization and biogenesis |
| GO:0000151 | cellular component | ubiquitin ligase complex |
| GO:0005615 | cellular component | extracellular space |
| GO:0005654 | cellular component | nucleoplasm |
| GO:0005681 | cellular component | spliceosome complex |
| GO:0005694 | cellular component | chromosome |
| GO:0005730 | cellular component | nucleolus |
| GO:0005739 | cellular component | mitochondrion |
| GO:0005764 | cellular component | lysosome |
| GO:0005768 | cellular component | endosome |
| GO:0005777 | cellular component | peroxisome |
| GO:0005783 | cellular component | endoplasmic reticulum |
| GO:0005794 | cellular component | Golgi apparatus |
| GO:0005829 | cellular component | cytosol |
| GO:0005840 | cellular component | ribosome |
| GO:0005886 | cellular component | plasma membrane |
| GO:0012505 | cellular component | endomembrane system |
| GO:0015629 | cellular component | actin cytoskeleton |
| GO:0015630 | cellular component | microtubule cytoskeleton |
| GO:0031012 | cellular component | extracellular matrix |
| GO:0031090 | cellular component | organelle membrane |
| GO:0031967 | cellular component | organelle envelope |
| GO:0031982 | cellular component | vesicle |
| GO:0000287 | molecular function | magnesium ion binding |
| GO:0003677 | molecular function | DNA binding |
| GO:0003700 | molecular function | transcription factor activity |
| GO:0003723 | molecular function | RNA binding |
| GO:0003735 | molecular function | structural constituent of ribosome |
| GO:0003924 | molecular function | GTPase activity |
| GO:0004518 | molecular function | nuclease activity |
| GO:0004672 | molecular function | protein kinase activity |
| GO:0004842 | molecular function | ubiquitin-protein ligase activity |
| GO:0004872 | molecular function | receptor activity |
| GO:0005102 | molecular function | receptor binding |
| GO:0005198 | molecular function | structural molecule activity |
| GO:0005216 | molecular function | ion channel activity |
| GO:0005386 | molecular function | carrier activity |
| GO:0005506 | molecular function | iron ion binding |
| GO:0005509 | molecular function | calcium ion binding |
| GO:0005524 | molecular function | ATP binding |
| GO:0005525 | molecular function | GTP binding |
| GO:0008092 | molecular function | cytoskeletal protein binding |
| GO:0008233 | molecular function | peptidase activity |
| GO:0008270 | molecular function | zinc ion binding |
| GO:0015075 | molecular function | ion transporter activity |
| GO:0016491 | molecular function | oxidoreductase activity |
| GO:0016887 | molecular function | ATPase activity |
| GO:0030234 | molecular function | enzyme regulator activity |
| GO:0042578 | molecular function | phosphoric ester hydrolase activity |

## 4.3 Results

### 4.3.1 Comparison of promoter regions between human and mouse

We analyzed sequence conservation of 8,429 promoter pairs of one-to-one orthologous genes between human and mouse. These pairs were compared by using the local alignment program *water* from the EMBOSS package [90]. The resulting distributions of the alignment scores are shown in Figure 22. The distribution has two peaks: a major peak around 1000, and a minor peak a little lower than 100. The minor peak corresponds to the negative control distribution created from randomly shuffled promoter pairs (depicted with a dashed line), indicating the presence of non-orthologous promoters that are not evolutionally related to each other (for an explanation of this phenomenon, see 4.4 Discussion). The apparent separation of the major and minor peaks indicates that we can discriminate orthologous promoters from non-orthologous ones by examining the local alignment scores. For the following analyses, we used the 6,901 promoter pairs with alignment scores ≥ 200 (82% of the initial data set) to eliminate non-orthologous pairs. The threshold of 200 was chosen so that the proportion of non-orthologous pairs with scores over the threshold was low enough: 200 is the 1.5 percentile of the negative control distribution, and the height of the minor peak is 0.16 times that of the negative control, and thus the proportion of non-orthologous pairs with scores ≥ 200 is estimated to be 0.24% (see Figure 23). It was possible that the offset of representative TSSs between human and mouse could bias the alignment scores. We evaluated this effect by estimating the offset from the differences in the local alignment end positions and shifting the mouse promoter as much as the offset. As a result of the promoter alignment with the offset correction, we confirmed that the bias was very small (data not shown). Therefore, we retained the original approach.

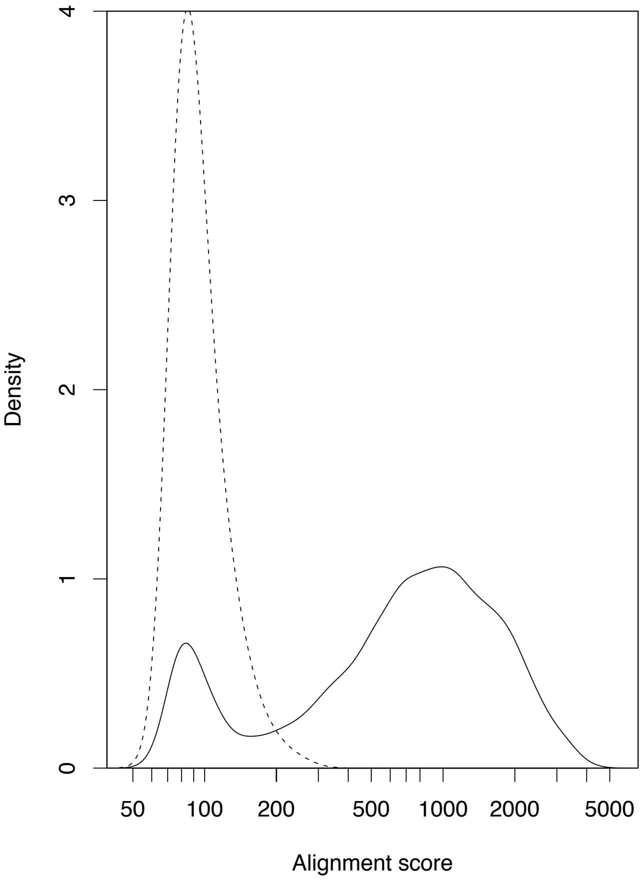


**Figure 22. Distribution of alignment scores of human and mouse promoters.**

The distribution for the orthologous gene pairs is depicted by the solid line, and the distribution for the negative control pairs is shown by the dashed line. The x-axis is shown in a logarithmic scale.

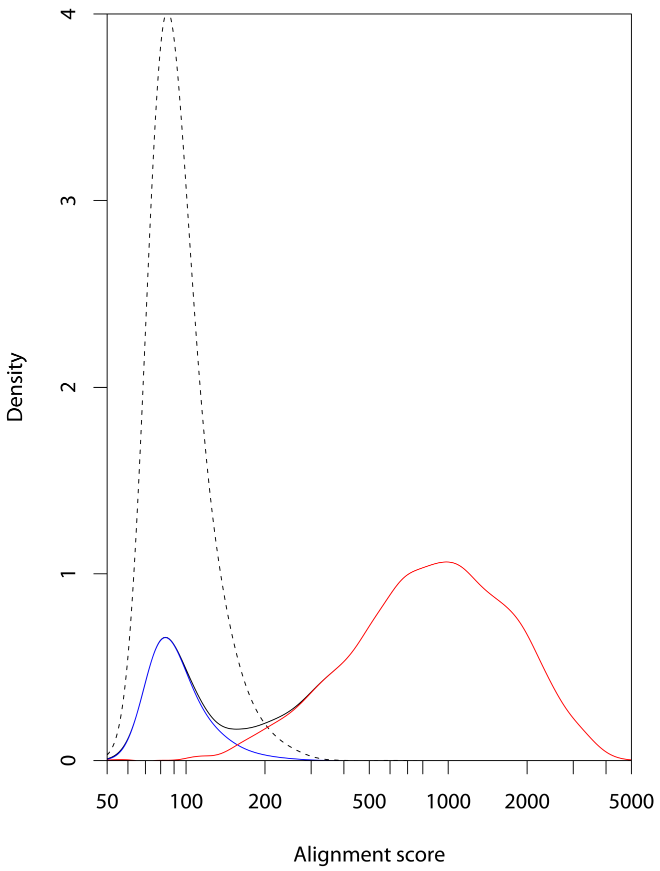


**Figure 23. Estimated distributions of orthologous and non-orthologous promoter pairs.**

- - - - Negative control

—— Estimated proportion of non-orthologous promoters

—— Estimated proportion of orthologous promoters

### 4.3.2 Promoter conservation

On the basis of the promoter sequence comparison between human and mouse for the 6,901 genes, we investigated the relationship between gene function and promoter conservation. Annotations of genes were made by associating human genes with GO terms. To this end, we developed a slimmed-down version of the GO vocabulary, containing 52 terms for biological process (P), 22 for cellular component (C) and 26 for molecular function (F) (Table 8, see 4.2 Methods for details). I tested whether the alignment scores for a set of genes associated with a GO term are significantly high or low by a Wilcoxon rank sum test. The resulting GO terms with high promoter conservation are listed in Table 9, and those with low conservation are in Table 10 (only terms with FDR < 0.05 are in the tables; for the complete list of results, see Table 11). Figure 24 shows the distributions of the alignment scores for several GO terms with significant tendencies (all of the distributions for the GO terms listed in Table 9 and 10 are shown in Figure 25 and 26, respectively). When we tried the global alignment score, we obtained quite similar tendencies (data not shown). We also confirmed that eliminating the coding sequences from the promoter dataset does not significantly influence the observed tendencies (data not shown, see 4.2 Methods for details).

In Table 9, we confirmed that the most significant terms are P:development and P:regulation of transcription [26,27]. Furthermore, an overall observation of the table revealed that the terms with high promoter conservation are related to signaling events inside as well as outside of the cell (P:cell-cell signaling, P:cell surface receptor linked signal transduction, P:ion transport, and P:intracellular signaling cascade). On the other hand, Table 10 covers a wide range of metabolism (P:lipid metabolism, P:carbohydrate metabolism, P:protein biosynthesis, P:proteolysis, P:electron transport, F:oxidoreductase activity, F:nuclease activity). Table 10 also contains cellular components, such as C:mitochondrion, C:lysosome, C:ribosome and C:peroxisome, which correspond to the metabolism-related terms.

**Table 9. GO categories with high promoter conservation.**

Terms of biological process are labeled as P, cellular component as C, molecular function as F.

| GO term | Number of genes | *P*-value | FDR |
|---|---|---|---|
| P:development | 649 | 0 | 0 |
| P:regulation of transcription | 602 | 1.67E-15 | 8.33E-14 |
| F:transcription factor activity | 263 | 3.44E-15 | 1.15E-13 |
| P:transcription | 640 | 4.11E-14 | 1.03E-12 |
| P:nervous system development | 154 | 1.99E-10 | 3.98E-09 |
| P:organ development | 213 | 2.30E-10 | 3.83E-09 |
| P:signal transduction | 994 | 5.19E-10 | 7.41E-09 |
| F:DNA binding | 628 | 3.19E-08 | 3.99E-07 |
| P:morphogenesis | 212 | 9.78E-08 | 1.09E-06 |
| P:cell surface receptor linked signal transduction | 363 | 2.23E-06 | 2.23E-05 |
| P:negative regulation of metabolism | 107 | 1.02E-05 | 9.27E-05 |
| F:receptor binding | 221 | 1.90E-05 | 0.000159 |
| P:cell-cell signaling | 176 | 2.27E-05 | 0.000175 |
| F:cytoskeletal protein binding | 137 | 4.97E-05 | 0.000355 |
| P:negative regulation of biological process | 327 | 6.87E-05 | 0.000458 |
| F:ion channel activity | 98 | 9.87E-05 | 0.000617 |
| C:extracellular matrix | 111 | 0.000119 | 0.000698 |
| C:actin cytoskeleton | 85 | 0.000164 | 0.000909 |
| P:cell differentiation | 173 | 0.000179 | 0.000944 |
| P:cell adhesion | 242 | 0.000182 | 0.000912 |
| P:cellular morphogenesis | 111 | 0.000607 | 0.002892 |
| F:ion transporter activity | 237 | 0.001493 | 0.006785 |
| P:protein amino acid phosphorylation | 213 | 0.001593 | 0.006928 |
| P:ion transport | 239 | 0.001825 | 0.007603 |
| F:protein kinase activity | 220 | 0.002033 | 0.008132 |
| P:intracellular signaling cascade | 431 | 0.006872 | 0.026430 |
| P:chromosome organization and biogenesis | 105 | 0.007832 | 0.029007 |
| C:plasma membrane | 608 | 0.008026 | 0.028664 |
| F:GTPase activity | 88 | 0.011437 | 0.039436 |
| P:cytoskeleton organization and biogenesis | 155 | 0.011868 | 0.039561 |
| P:small GTPase mediated signal transduction | 126 | 0.012373 | 0.039914 |

**Table 10. GO categories with low promoter conservation.**

Terms of biological process are labeled as P, cellular component as C, molecular function as F.

| GO term | Number of genes | *P*-value | FDR |
|---|---|---|---|
| C:mitochondrion | 398 | 5.31E-09 | 5.31E-07 |
| F:oxidoreductase activity | 309 | 2.07E-08 | 1.03E-06 |
| C:lysosome | 77 | 9.94E-08 | 3.31E-06 |
| C:ribosome | 114 | 7.54E-07 | 1.89E-05 |
| P:lipid metabolism | 260 | 1.04E-06 | 2.08E-05 |
| P:carboxylic acid metabolism | 225 | 4.43E-06 | 7.38E-05 |
| F:structural constituent of ribosome | 130 | 5.76E-06 | 8.22E-05 |
| P:amino acid metabolism | 112 | 0.000102 | 0.001277 |
| P:electron transport | 151 | 0.000236 | 0.002623 |
| P:catabolism | 260 | 0.000251 | 0.002509 |
| P:carbohydrate metabolism | 220 | 0.000278 | 0.002531 |
| C:peroxisome | 49 | 0.000623 | 0.005192 |
| P:protein biosynthesis | 283 | 0.00063 | 0.004849 |
| F:nuclease activity | 60 | 0.000772 | 0.005518 |
| P:response to biotic stimulus | 318 | 0.000893 | 0.005956 |
| C:nucleolus | 63 | 0.004455 | 0.027845 |
| P:immune response | 270 | 0.005437 | 0.031984 |
| F:iron ion binding | 111 | 0.0055 | 0.030554 |
| F:peptidase activity | 227 | 0.005592 | 0.029433 |
| P:proteolysis | 259 | 0.006844 | 0.034218 |

## Table 11. Statistical significance of promoter conservation for 100 GO terms.

| GO term | number of samples | number of control | *P*-value for high conservation | FDR for high conservation | *P*-value for low conservation | FDR for low conservation |
|---|---|---|---|---|---|---|
| P:development | 649 | 4136 | 0 | 0 | 1.000000 | 1.000000 |
| P:regulation of transcription | 602 | 4183 | 1.67E-15 | 8.33E-14 | 1.000000 | 1.010101 |
| F:transcription factor activity | 263 | 4760 | 3.44E-15 | 1.15E-13 | 1.000000 | 1.020408 |
| P:transcription | 640 | 4145 | 4.11E-14 | 1.03E-12 | 1.000000 | 1.030928 |
| P:nervous system development | 154 | 4631 | 1.99E-10 | 3.98E-09 | 1.000000 | 1.041667 |
| P:organ development | 213 | 4572 | 2.30E-10 | 3.83E-09 | 1.000000 | 1.052632 |
| P:signal transduction | 994 | 3791 | 5.19E-10 | 7.41E-09 | 1.000000 | 1.063830 |
| F:DNA binding | 628 | 4395 | 3.19E-08 | 3.99E-07 | 1.000000 | 1.075269 |
| P:morphogenesis | 212 | 4573 | 9.78E-08 | 1.09E-06 | 1.000000 | 1.086957 |
| P:cell surface receptor linked signal transduction | 363 | 4422 | 2.23E-06 | 2.23E-05 | 0.999998 | 1.098899 |
| P:negative regulation of metabolism | 107 | 4678 | 1.02E-05 | 9.27E-05 | 0.999990 | 1.111100 |
| F:receptor binding | 221 | 4802 | 1.90E-05 | 0.000159 | 0.999981 | 1.123574 |
| P:cell-cell signaling | 176 | 4609 | 2.27E-05 | 0.000175 | 0.999977 | 1.136338 |
| F:cytoskeletal protein binding | 137 | 4886 | 4.97E-05 | 0.000355 | 0.999950 | 1.149368 |
| P:negative regulation of biological process | 327 | 4458 | 6.87E-05 | 0.000458 | 0.999931 | 1.162711 |
| F:ion channel activity | 98 | 4925 | 9.87E-05 | 0.000617 | 0.999901 | 1.176355 |
| C:extracellular matrix | 111 | 4401 | 0.000119 | 0.000698 | 0.999881 | 1.190335 |
| C:actin cytoskeleton | 85 | 4427 | 0.000164 | 0.000909 | 0.999837 | 1.204622 |
| P:cell differentiation | 173 | 4612 | 0.000179 | 0.000944 | 0.999821 | 1.219294 |
| P:cell adhesion | 242 | 4543 | 0.000182 | 0.000912 | 0.999818 | 1.234343 |
| P:cellular morphogenesis | 111 | 4674 | 0.000607 | 0.002892 | 0.999393 | 1.249241 |
| F:ion transporter activity | 237 | 4786 | 0.001493 | 0.006785 | 0.998508 | 1.263934 |
| P:protein amino acid phosphorylation | 213 | 4572 | 0.001593 | 0.006928 | 0.998407 | 1.280009 |
| P:ion transport | 239 | 4546 | 0.001825 | 0.007603 | 0.998176 | 1.296332 |
| F:protein kinase activity | 220 | 4803 | 0.002033 | 0.008132 | 0.997967 | 1.313115 |
| P:intracellular signaling cascade | 431 | 4354 | 0.006872 | 0.026430 | 0.993129 | 1.324172 |
| P:chromosome organization and biogenesis | 105 | 4680 | 0.007832 | 0.029007 | 0.992170 | 1.340770 |
| C:plasma membrane | 608 | 3904 | 0.008026 | 0.028664 | 0.991975 | 1.358870 |
| F:GTPase activity | 88 | 4935 | 0.011437 | 0.039436 | 0.988566 | 1.373008 |
| P:cytoskeleton organization and biogenesis | 155 | 4630 | 0.011868 | 0.039561 | 0.988133 | 1.391737 |
| P:small GTPase mediated signal transduction | 126 | 4659 | 0.012373 | 0.039914 | 0.987629 | 1.410898 |
| P:cell proliferation | 258 | 4527 | 0.016427 | 0.051335 | 0.983575 | 1.425471 |
| F:GTP binding | 160 | 4863 | 0.026124 | 0.079164 | 0.973879 | 1.432175 |
| F:calcium ion binding | 280 | 4743 | 0.058031 | 0.170681 | 0.941974 | 1.405931 |
| P:cell motility | 105 | 4680 | 0.061020 | 0.174343 | 0.938989 | 1.422710 |
| F:receptor activity | 391 | 4632 | 0.078737 | 0.218713 | 0.921269 | 1.417337 |
| F:structural molecule activity | 307 | 4716 | 0.079056 | 0.213666 | 0.920950 | 1.438984 |
| P:G-protein coupled receptor protein signaling pathway | 153 | 4632 | 0.108505 | 0.285540 | 0.891506 | 1.415089 |
| F:phosphoric ester hydrolase activity | 113 | 4910 | 0.138352 | 0.354748 | 0.861663 | 1.389779 |
| F:enzyme regulator activity | 238 | 4785 | 0.148055 | 0.370138 | 0.851956 | 1.396648 |
| P:protein complex assembly | 122 | 4663 | 0.154119 | 0.375899 | 0.845897 | 1.409829 |
| P:vesicle-mediated transport | 190 | 4595 | 0.166697 | 0.396897 | 0.833317 | 1.412401 |
| P:protein kinase cascade | 118 | 4667 | 0.173566 | 0.403642 | 0.826451 | 1.424916 |
| C:nucleoplasm | 107 | 4405 | 0.198841 | 0.451912 | 0.801180 | 1.405578 |
| C:Golgi apparatus | 216 | 4296 | 0.202850 | 0.450778 | 0.797165 | 1.423509 |
| F:zinc ion binding | 600 | 4423 | 0.232496 | 0.505425 | 0.767514 | 1.395479 |
| F:carrier activity | 149 | 4874 | 0.239577 | 0.509738 | 0.760441 | 1.408224 |
| P:intracellular transport | 350 | 4435 | 0.240786 | 0.501638 | 0.759226 | 1.432503 |
| P:sensory perception | 111 | 4674 | 0.293922 | 0.599840 | 0.706102 | 1.357889 |
| P:cell cycle | 340 | 4445 | 0.297867 | 0.595733 | 0.702148 | 1.376760 |

| | | | | | | |
|---|---|---|---|---|---|---|
| F:RNA binding | 290 | 4733 | 0.320681 | 0.628785 | 0.679334 | 1.358669 |
| P:positive regulation of biological process | 275 | 4510 | 0.345462 | 0.664349 | 0.654555 | 1.335827 |
| F:ATP binding | 520 | 4503 | 0.376693 | 0.710741 | 0.623319 | 1.298582 |
| F:ubiquitin-protein ligase activity | 144 | 4879 | 0.426143 | 0.789154 | 0.573880 | 1.221021 |
| P:DNA metabolism | 266 | 4519 | 0.454829 | 0.826962 | 0.545189 | 1.185194 |
| P:apoptosis | 244 | 4541 | 0.532993 | 0.951773 | 0.467026 | 1.037835 |
| C:microtubule cytoskeleton | 115 | 4397 | 0.540403 | 0.948076 | 0.459626 | 1.044604 |
| P:protein folding | 121 | 4664 | 0.586660 | 1.011483 | 0.413366 | 0.961316 |
| C:endoplasmic reticulum | 268 | 4244 | 0.591106 | 1.001875 | 0.408913 | 0.973602 |
| C:extracellular space | 179 | 4333 | 0.608820 | 1.014700 | 0.391202 | 0.954152 |
| P:response to chemical stimulus | 129 | 4656 | 0.645196 | 1.057698 | 0.354828 | 0.887070 |
| F:ATPase activity | 130 | 4893 | 0.647215 | 1.043895 | 0.352808 | 0.904636 |
| P:ubiquitin cycle | 235 | 4550 | 0.654462 | 1.038829 | 0.345556 | 0.909357 |
| C:endomembrane system | 163 | 4349 | 0.667170 | 1.042454 | 0.332852 | 0.899600 |
| C:chromosome | 108 | 4404 | 0.692915 | 1.066024 | 0.307111 | 0.853086 |
| C:organelle envelope | 148 | 4364 | 0.696927 | 1.055950 | 0.303096 | 0.865987 |
| P:response to abiotic stimulus | 148 | 4637 | 0.720140 | 1.074836 | 0.279881 | 0.823178 |
| C:vesicle | 86 | 4426 | 0.734101 | 1.079560 | 0.265927 | 0.805838 |
| P:secretory pathway | 102 | 4683 | 0.777510 | 1.126826 | 0.222512 | 0.695349 |
| C:endosome | 33 | 4479 | 0.801091 | 1.144416 | 0.198946 | 0.641762 |
| C:ubiquitin ligase complex | 122 | 4390 | 0.815112 | 1.148045 | 0.184907 | 0.616357 |
| C:organelle membrane | 242 | 4270 | 0.832397 | 1.156107 | 0.167616 | 0.577986 |
| P:cofactor metabolism | 100 | 4685 | 0.901092 | 1.234373 | 0.098920 | 0.353287 |
| C:cytosol | 171 | 4341 | 0.924121 | 1.248812 | 0.075887 | 0.281064 |
| C:spliceosome complex | 37 | 4475 | 0.931290 | 1.241720 | 0.068727 | 0.264335 |
| F:magnesium ion binding | 135 | 4888 | 0.946019 | 1.244762 | 0.053988 | 0.215951 |
| P:alcohol metabolism | 133 | 4652 | 0.977205 | 1.269098 | 0.022798 | 0.094993 |
| P:lipid biosynthesis | 101 | 4684 | 0.978371 | 1.254321 | 0.021633 | 0.094058 |
| P:RNA processing | 198 | 4587 | 0.979718 | 1.240149 | 0.020285 | 0.092203 |
| P:response to stress | 446 | 4339 | 0.984169 | 1.230211 | 0.015833 | 0.075393 |
| P:proteolysis | 259 | 4526 | 0.993157 | 1.226120 | 0.006844 | 0.034218 |
| F:peptidase activity | 227 | 4796 | 0.994409 | 1.212693 | 0.005592 | 0.029433 |
| F:iron ion binding | 111 | 4912 | 0.994501 | 1.198194 | 0.005500 | 0.030554 |
| P:immune response | 270 | 4515 | 0.994563 | 1.184004 | 0.005437 | 0.031984 |
| C:nucleolus | 63 | 4449 | 0.995546 | 1.171231 | 0.004455 | 0.027845 |
| P:response to biotic stimulus | 318 | 4467 | 0.999107 | 1.161752 | 0.000893 | 0.005956 |
| F:nuclease activity | 60 | 4963 | 0.999228 | 1.148538 | 0.000772 | 0.005518 |
| P:protein biosynthesis | 283 | 4502 | 0.999370 | 1.135647 | 0.000630 | 0.004849 |
| C:peroxisome | 49 | 4463 | 0.999377 | 1.122896 | 0.000623 | 0.005192 |
| P:carbohydrate metabolism | 220 | 4565 | 0.999722 | 1.110802 | 0.000278 | 0.002531 |
| P:catabolism | 260 | 4525 | 0.999749 | 1.098625 | 0.000251 | 0.002509 |
| P:electron transport | 151 | 4634 | 0.999764 | 1.086700 | 0.000236 | 0.002623 |
| P:amino acid metabolism | 112 | 4673 | 0.999898 | 1.075159 | 0.000102 | 0.001277 |
| F:structural constituent of ribosome | 130 | 4893 | 0.999994 | 1.063824 | 5.76E-06 | 8.22E-05 |
| P:carboxylic acid metabolism | 225 | 4560 | 0.999996 | 1.052627 | 4.43E-06 | 7.38E-05 |
| P:lipid metabolism | 260 | 4525 | 0.999999 | 1.041666 | 1.04E-06 | 2.08E-05 |
| C:ribosome | 114 | 4398 | 0.999999 | 1.030927 | 7.54E-07 | 1.89E-05 |
| C:lysosome | 77 | 4435 | 1.000000 | 1.020408 | 9.94E-08 | 3.31E-06 |
| F:oxidoreductase activity | 309 | 4714 | 1.000000 | 1.010101 | 2.07E-08 | 1.03E-06 |
| C:mitochondrion | 398 | 4114 | 1.000000 | 1.000000 | 5.31E-09 | 5.31E-07 |

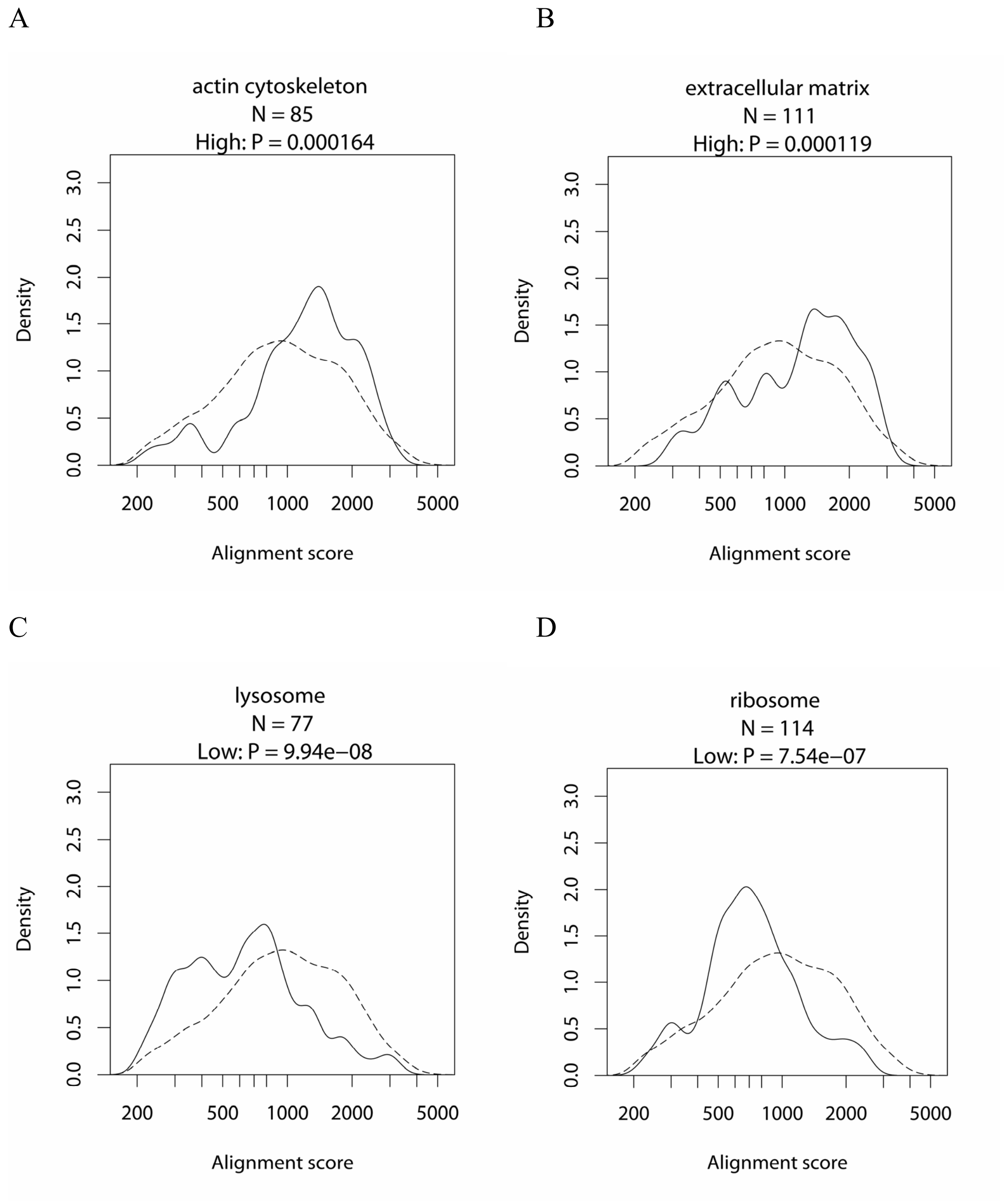
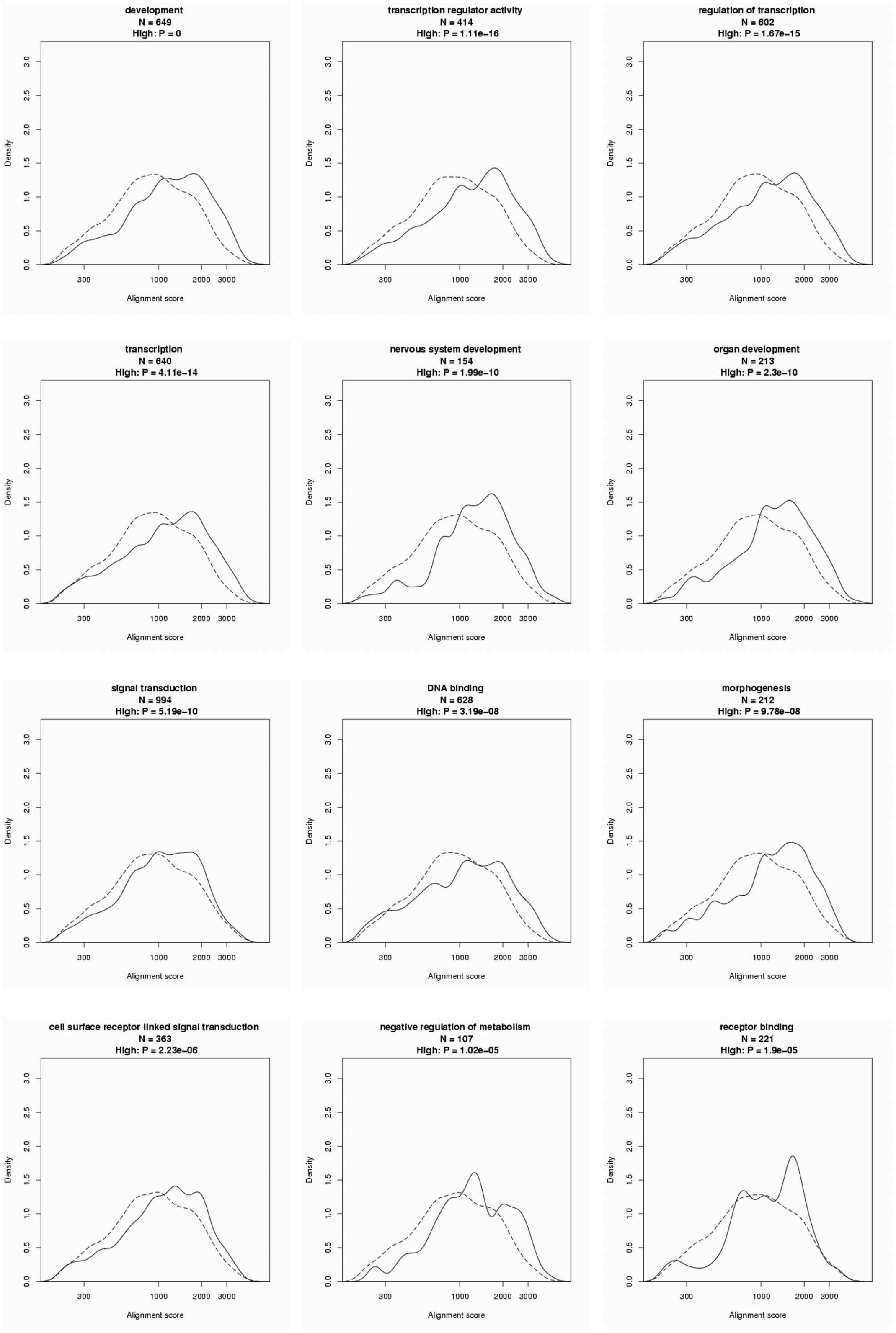
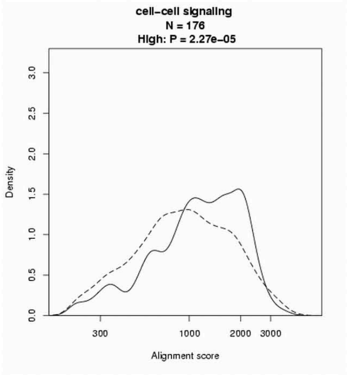


**Figure 24. Distribution of alignment scores of promoters for specific genes.**

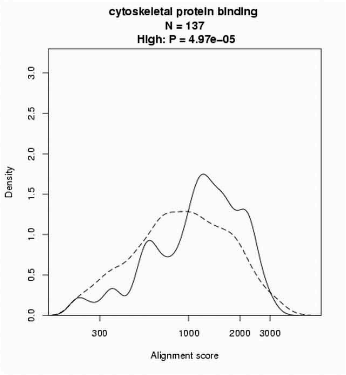
For the high conservation tendency, actin cytoskeleton (A) and extracellular matrix (B), for the low conservation tendency, lysosome (C) and ribosome (D). For each of A-D, the solid line shows the distribution of the alignment scores for genes with the specific GO term, and the dashed line shows the distribution for the control gene set (see 4.2 Methods for details).

**Figure 25. GO categories with high promoter conservation.**

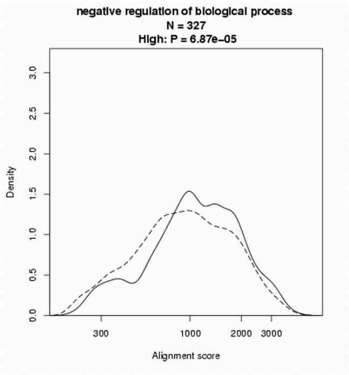




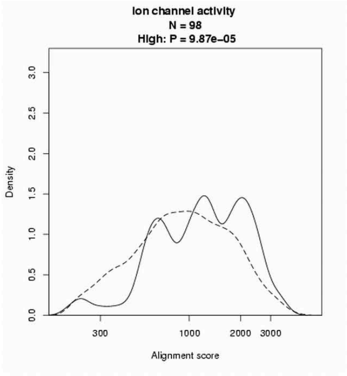
cell−cell signaling
N = 176
High: P = 2.27e−05



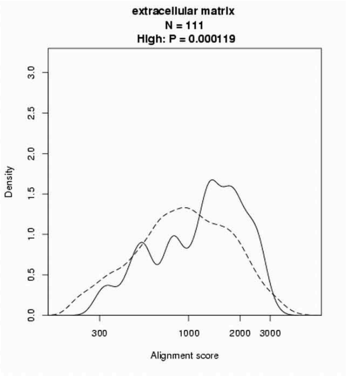
cytoskeletal protein binding
N = 137
High: P = 4.97e−05



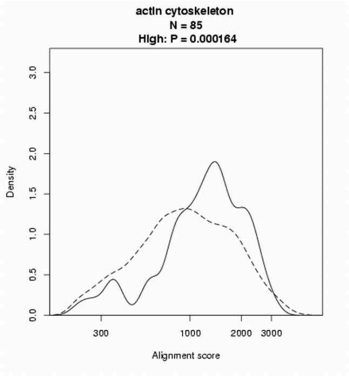
negative regulation of biological process
N = 327
High: P = 6.87e−05



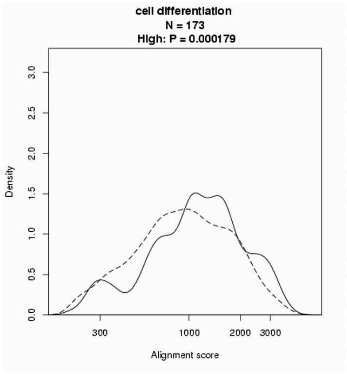
ion channel activity
N = 98
High: P = 9.87e−05



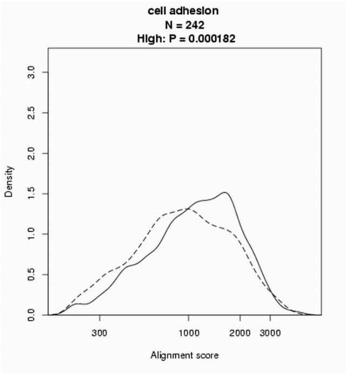
extracellular matrix
N = 111
High: P = 0.000119



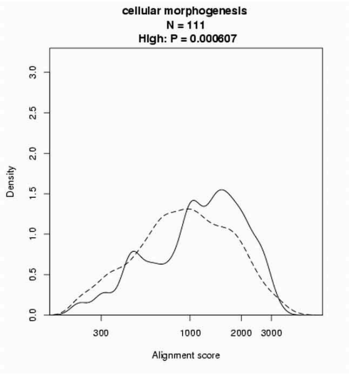
actin cytoskeleton
N = 85
High: P = 0.000164



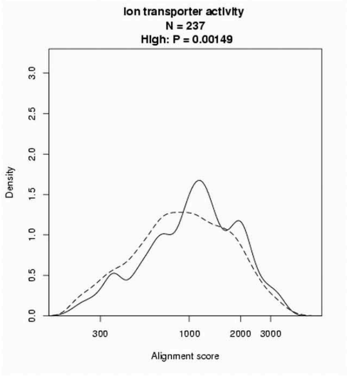
cell differentiation
N = 173
High: P = 0.000179



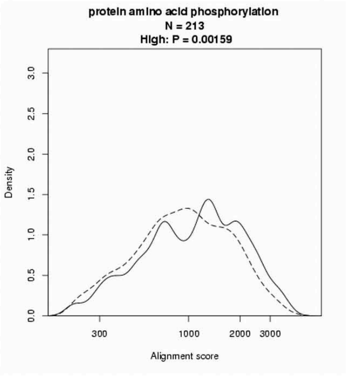
cell adhesion
N = 242
High: P = 0.000182



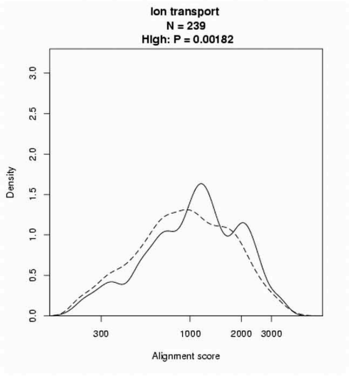
cellular morphogenesis
N = 111
High: P = 0.000607



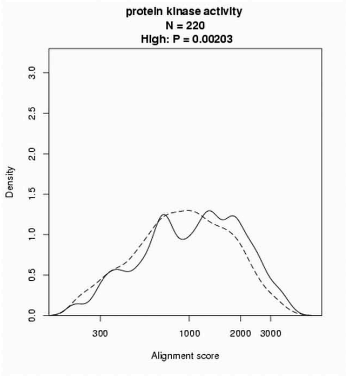
ion transporter activity
N = 237
High: P = 0.00149

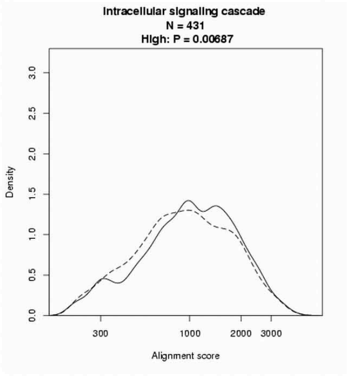


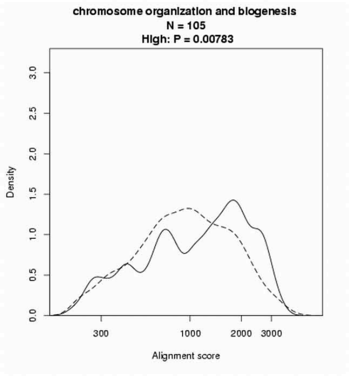
protein amino acid phosphorylation
N = 213
High: P = 0.00159

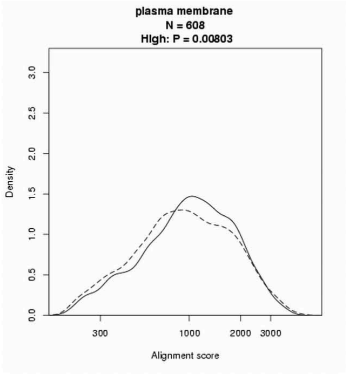


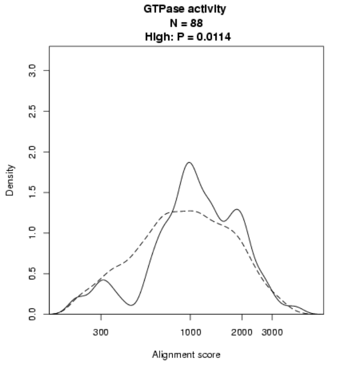
ion transport
N = 239
High: P = 0.00182

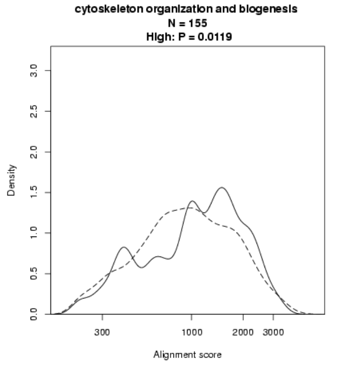

protein kinase activity
N = 220
High: P = 0.00203
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
300
1000
2000
3000
Alignment score

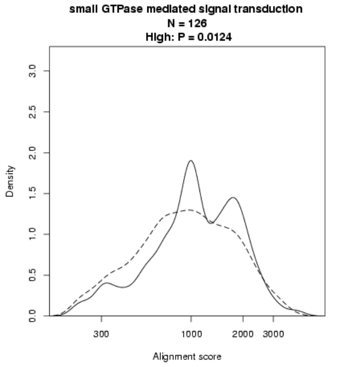

intracellular signaling cascade
N = 431
High: P = 0.00687
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
300
1000
2000
3000
Alignment score


chromosome organization and biogenesis
N = 105
High: P = 0.00783
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
300
1000
2000
3000
Alignment score

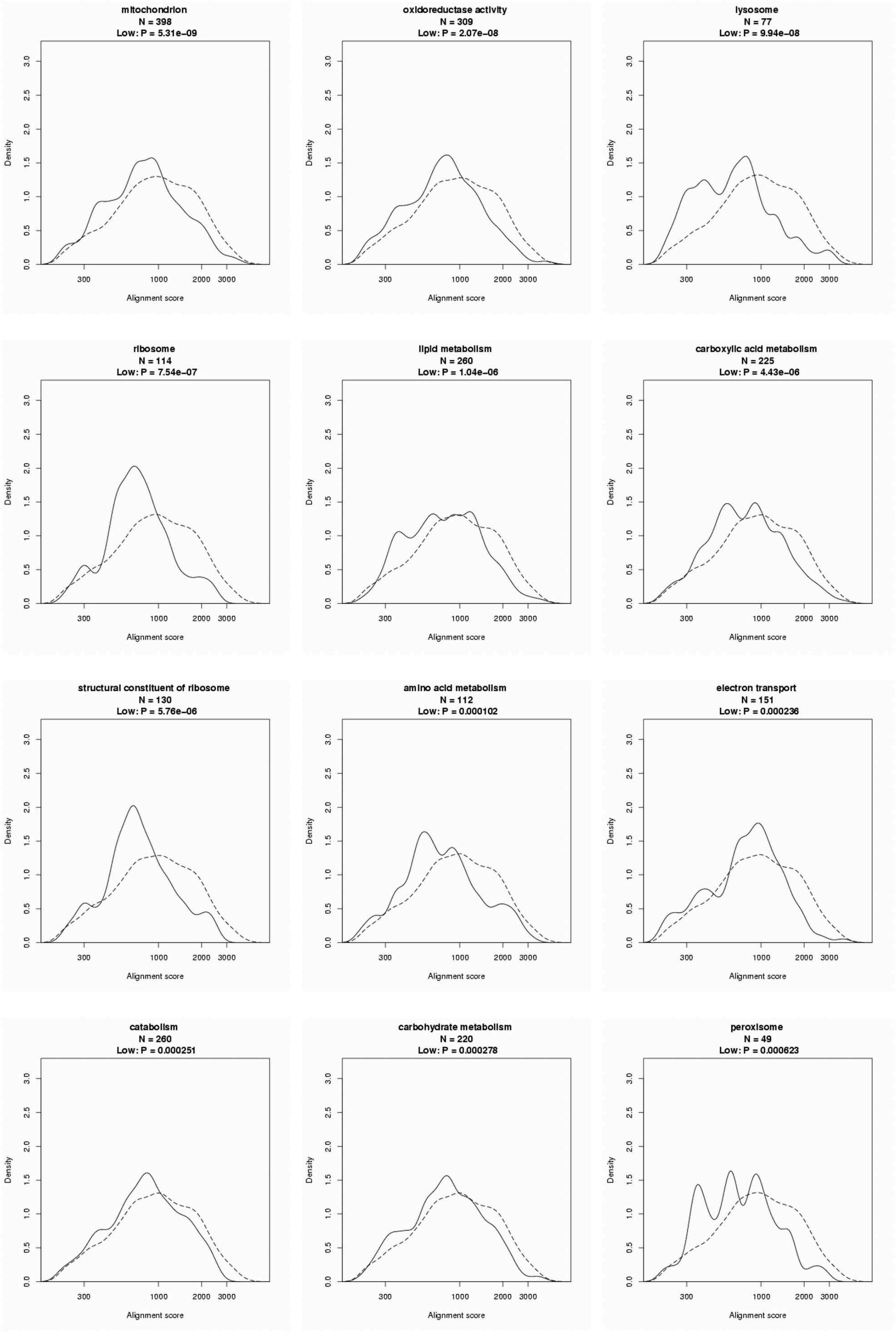

plasma membrane
N = 608
High: P = 0.00803
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
300
1000
2000
3000
Alignment score

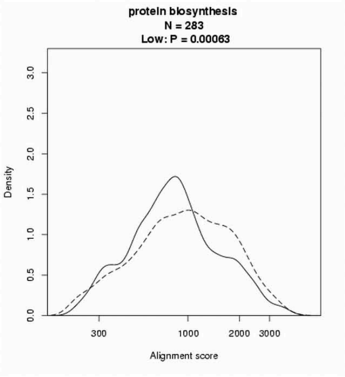

GTPase activity
N = 88
High: P = 0.0114
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
300
1000
2000
3000
Alignment score

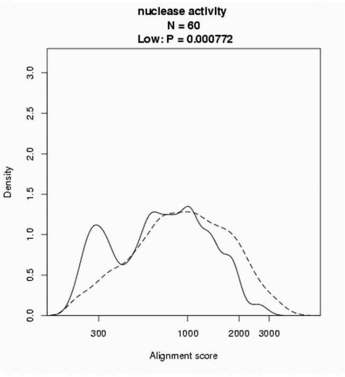

cytoskeleton organization and biogenesis
N = 155
High: P = 0.0119
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
300
1000
2000
3000
Alignment score

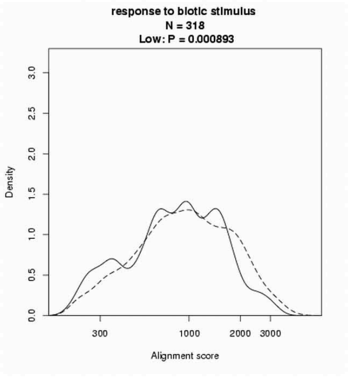

small GTPase mediated signal transduction
N = 126
High: P = 0.0124
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
300
1000
2000
3000
Alignment score

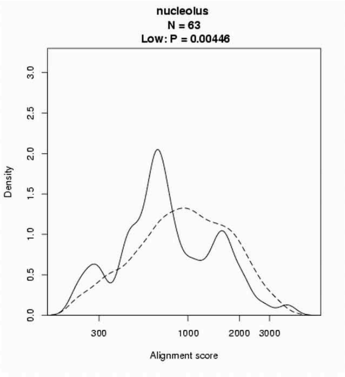
## Figure 26. GO categories with low promoter conservation.

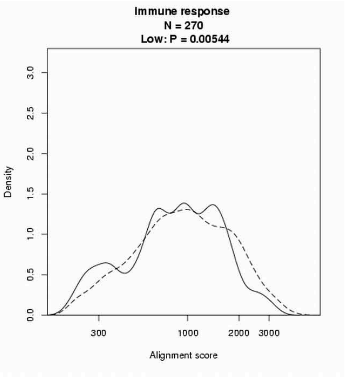


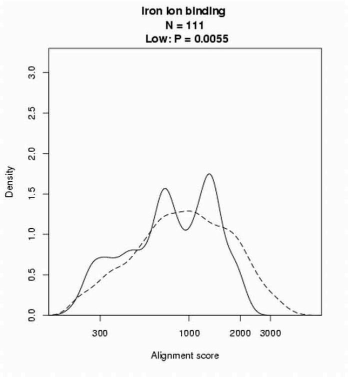

protein biosynthesis
N = 283
Low: P = 0.00063
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
300
1000
2000
3000
Alignment score

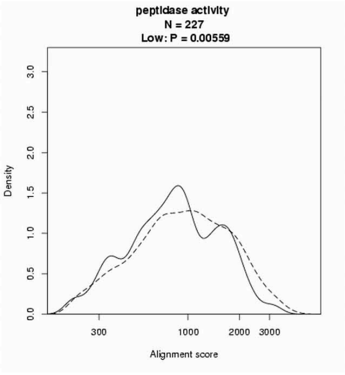

nuclease activity
N = 60
Low: P = 0.000772
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
300
1000
2000
3000
Alignment score

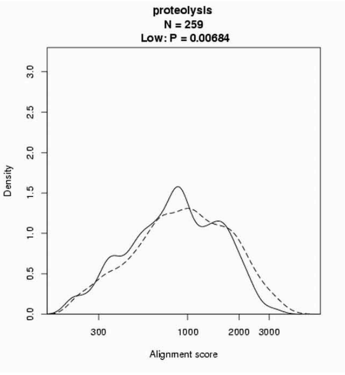

response to biotic stimulus
N = 318
Low: P = 0.000893
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
300
1000
2000
3000
Alignment score


nucleolus
N = 63
Low: P = 0.00446
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
300
1000
2000
3000
Alignment score


immune response
N = 270
Low: P = 0.00544
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
300
1000
2000
3000
Alignment score


iron ion binding
N = 111
Low: P = 0.0055
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
300
1000
2000
3000
Alignment score


peptidase activity
N = 227
Low: P = 0.00559
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
300
1000
2000
3000
Alignment score


proteolysis
N = 259
Low: P = 0.00684
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
300
1000
2000
3000
Alignment score

### 4.3.3 Protein conservation

The protein conservation tendencies were examined in a similar manner to those of the promoter conservation, using protein sequences obtained from the RefSeq database. Since the alignment score largely depends on the protein length, we used the percentage identity for protein sequences, instead of the alignment scores. GO terms showing high protein conservation are listed in Table 12, and those with low conservation are in Table 13 (only terms with a FDR < 0.05; for the complete list of results, see Table 14). Figure 27 shows the distributions of conservation levels for several GO terms with significant tendencies (all of the distributions for the GO terms in Table 12 and 13 are shown in Figure 28 and 29, respetively). When we tried global alignment, we obtained quite similar tendencies (data not shown), which is reasonable, given that the coverages of the local alignments were mostly over 95% (data not shown).

Table 12 includes well-known categories for high protein conservation: actins [19], ribosomal proteins, Ras-like GTPases [22] and RNA processing [95], and for low protein conservation, P:immune response [21]. By looking over Table 12, we realized that the categories are composed of a series of processes required for gene expression; from intracellular signaling cascade and regulation of transcription, to RNA processing, protein biosynthesis and intracellular transport. We also find C:cytosol and C:nucleoplasm, where the above-mentioned processes take place, and C:actin cytoskeleton, which is known to be involved in transcription [96]. On the other hand, in Table 13, the terms with low conservation are related to extracellular regions or cell surface (C:extracellular space, C:extracellular matrix, C:plasma membrane, F:receptor activity, F:receptor binding, P:cell-cell signaling or P:cell adhesion) or to membrane-bounded organelles (C:lysosome, C:mitochondrion or C:peroxisome). Other terms, such as F:oxidoreductase activity, F:peptidase activity, F:nuclease activity, P:electron tansport and P:proteolysis, correspond to the functions of these cellular components.

**Table 12. GO categories with high protein conservation.**

Terms of biological process are labeled as P, cellular component as C, molecular function as F.

| GO term | Number of genes | *P*-value | FDR |
|---|---|---|---|
| F:GTPase activity | 88 | 0 | 0 |
| F:GTP binding | 160 | 0 | 0 |
| P:intracellular transport | 350 | 0 | 0 |
| P:small GTPase mediated signal transduction | 126 | 1.11E-16 | 2.78E-15 |
| F:RNA binding | 290 | 1.33E-15 | 2.66E-14 |
| C:cytosol | 171 | 3.70E-11 | 6.17E-10 |
| P:RNA processing | 198 | 3.25E-10 | 4.64E-09 |
| C:Golgi apparatus | 216 | 5.63E-10 | 7.03E-09 |
| P:intracellular signaling cascade | 431 | 2.57E-09 | 2.85E-08 |
| C:spliceosome complex | 37 | 6.46E-09 | 6.46E-08 |
| P:transcription | 640 | 1.76E-08 | 1.60E-07 |
| P:regulation of transcription | 602 | 2.02E-08 | 1.68E-07 |
| F:ATP binding | 520 | 2.85E-08 | 2.19E-07 |
| C:actin cytoskeleton | 85 | 5.37E-08 | 3.83E-07 |
| P:vesicle-mediated transport | 190 | 7.02E-08 | 4.68E-07 |
| P:cytoskeleton organization and biogenesis | 155 | 9.26E-08 | 5.78E-07 |
| F:cytoskeletal protein binding | 137 | 1.44E-07 | 8.47E-07 |
| P:secretory pathway | 102 | 7.91E-07 | 4.39E-06 |
| C:nucleoplasm | 107 | 1.22E-06 | 6.40E-06 |
| C:ribosome | 114 | 1.36E-06 | 6.81E-06 |
| P:protein biosynthesis | 283 | 1.56E-06 | 7.45E-06 |
| P:ubiquitin cycle | 235 | 2.86E-06 | 1.30E-05 |
| F:ion channel activity | 98 | 7.08E-05 | 0.000308 |
| P:protein amino acid phosphorylation | 213 | 0.000101 | 0.000420 |
| F:ATPase activity | 130 | 0.000143 | 0.000572 |
| C:endomembrane system | 163 | 0.000154 | 0.000591 |
| F:protein kinase activity | 220 | 0.000178 | 0.000659 |
| P:nervous system development | 154 | 0.000293 | 0.001045 |
| F:transcription factor activity | 263 | 0.000465 | 0.001603 |
| C:microtubule cytoskeleton | 115 | 0.000595 | 0.001982 |
| C:vesicle | 86 | 0.000732 | 0.002362 |
| F:structural molecule activity | 307 | 0.000801 | 0.002505 |
| F:structural constituent of ribosome | 130 | 0.000843 | 0.002554 |
| F:ubiquitin-protein ligase activity | 144 | 0.001969 | 0.005792 |
| C:organelle membrane | 242 | 0.004122 | 0.011778 |
| P:cell cycle | 340 | 0.006445 | 0.017904 |
| F:ion transporter activity | 237 | 0.012295 | 0.033229 |

**Table 13. GO categories with low protein conservation.**

Terms of biological process are labeled as P, cellular component as C, molecular function as F.

| GO term | Number of genes | *P*-value | FDR |
|---|---|---|---|
| P:response to biotic stimulus | 318 | 4.08E-49 | 4.08E-47 |
| P:immune response | 270 | 1.16E-44 | 5.79E-43 |
| C:extracellular space | 179 | 3.49E-37 | 1.16E-35 |
| P:response to stress | 446 | 5.05E-26 | 1.26E-24 |
| F:oxidoreductase activity | 309 | 2.35E-12 | 4.70E-11 |
| F:receptor activity | 391 | 1.11E-11 | 1.85E-10 |
| F:receptor binding | 221 | 2.15E-11 | 3.07E-10 |
| P:lipid metabolism | 260 | 5.95E-11 | 7.43E-10 |
| P:electron transport | 151 | 7.64E-10 | 8.49E-09 |
| C:lysosome | 77 | 6.38E-08 | 6.38E-07 |
| F:peptidase activity | 227 | 6.15E-07 | 5.59E-06 |
| P:cell proliferation | 258 | 1.65E-06 | 1.38E-05 |
| P:cell adhesion | 242 | 2.16E-06 | 1.66E-05 |
| C:mitochondrion | 398 | 3.00E-05 | 0.000214 |
| P:proteolysis | 259 | 4.53E-05 | 0.000302 |
| C:extracellular matrix | 111 | 5.52E-05 | 0.000345 |
| C:peroxisome | 49 | 8.02E-05 | 0.000472 |
| F:nuclease activity | 60 | 8.50E-05 | 0.000472 |
| C:plasma membrane | 608 | 0.000291 | 0.001533 |
| P:apoptosis | 244 | 0.001373 | 0.006866 |
| P:carboxylic acid metabolism | 225 | 0.002527 | 0.012032 |
| P:response to abiotic stimulus | 148 | 0.004444 | 0.020198 |
| P:positive regulation of biological process | 275 | 0.004599 | 0.019996 |
| P:response to chemical stimulus | 129 | 0.004626 | 0.019273 |
| P:lipid biosynthesis | 101 | 0.005576 | 0.022302 |
| P:cell-cell signaling | 176 | 0.006021 | 0.023159 |
| P:sensory perception | 111 | 0.008042 | 0.029784 |

**Table 14. Statistical significance of protein conservation for 100 GO terms.**

| GO term | number of samples | number of control | *P*-value for high conservation | FDR for high conservation | *P*-value for low conservation | FDR for low conservation |
|---|---|---|---|---|---|---|
| F:GTPase activity | 88 | 4935 | 0 | 0 | 1.000000 | 1.000000 |
| F:GTP binding | 160 | 4863 | 0 | 0 | 1.000000 | 1.010101 |
| P:intracellular transport | 350 | 4435 | 0 | 0 | 1.000000 | 1.020408 |
| P:small GTPase mediated signal transduction | 126 | 4659 | 1.11E-16 | 2.78E-15 | 1.000000 | 1.030928 |
| F:RNA binding | 290 | 4733 | 1.33E-15 | 2.66E-14 | 1.000000 | 1.041667 |
| C:cytosol | 171 | 4341 | 3.70E-11 | 6.17E-10 | 1.000000 | 1.052632 |
| P:RNA processing | 198 | 4587 | 3.25E-10 | 4.64E-09 | 1.000000 | 1.063830 |
| C:Golgi apparatus | 216 | 4296 | 5.63E-10 | 7.03E-09 | 1.000000 | 1.075269 |
| P:intracellular signaling cascade | 431 | 4354 | 2.57E-09 | 2.85E-08 | 1.000000 | 1.086957 |
| C:spliceosome complex | 37 | 4475 | 6.46E-09 | 6.46E-08 | 1.000000 | 1.098901 |
| P:transcription | 640 | 4145 | 1.76E-08 | 1.60E-07 | 1.000000 | 1.111111 |
| P:regulation of transcription | 602 | 4183 | 2.02E-08 | 1.68E-07 | 1.000000 | 1.123596 |
| F:ATP binding | 520 | 4503 | 2.85E-08 | 2.19E-07 | 1.000000 | 1.136364 |
| C:actin cytoskeleton | 85 | 4427 | 5.37E-08 | 3.83E-07 | 1.000000 | 1.149425 |
| P:vesicle-mediated transport | 190 | 4595 | 7.02E-08 | 4.68E-07 | 1.000000 | 1.162791 |
| P:cytoskeleton organization and biogenesis | 155 | 4630 | 9.26E-08 | 5.78E-07 | 1.000000 | 1.176471 |
| F:cytoskeletal protein binding | 137 | 4886 | 1.44E-07 | 8.47E-07 | 1.000000 | 1.190476 |
| P:secretory pathway | 102 | 4683 | 7.91E-07 | 4.39E-06 | 0.999999 | 1.204818 |
| C:nucleoplasm | 107 | 4405 | 1.22E-06 | 6.40E-06 | 0.999999 | 1.219511 |
| C:ribosome | 114 | 4398 | 1.36E-06 | 6.81E-06 | 0.999999 | 1.234566 |
| P:protein biosynthesis | 283 | 4502 | 1.56E-06 | 7.45E-06 | 0.999998 | 1.249998 |
| P:ubiquitin cycle | 235 | 4550 | 2.86E-06 | 1.30E-05 | 0.999997 | 1.265819 |
| F:ion channel activity | 98 | 4925 | 7.08E-05 | 0.000308 | 0.999929 | 1.281961 |
| P:protein amino acid phosphorylation | 213 | 4572 | 0.000101 | 0.000420 | 0.999899 | 1.298570 |
| F:ATPase activity | 130 | 4893 | 0.000143 | 0.000572 | 0.999857 | 1.315601 |
| C:endomembrane system | 163 | 4349 | 0.000154 | 0.000591 | 0.999846 | 1.333128 |
| F:protein kinase activity | 220 | 4803 | 0.000178 | 0.000659 | 0.999822 | 1.351111 |
| P:nervous system development | 154 | 4631 | 0.000293 | 0.001045 | 0.999707 | 1.369462 |
| F:transcription factor activity | 263 | 4760 | 0.000465 | 0.001603 | 0.999535 | 1.388243 |
| C:microtubule cytoskeleton | 115 | 4397 | 0.000595 | 0.001982 | 0.999405 | 1.407613 |
| C:vesicle | 86 | 4426 | 0.000732 | 0.002362 | 0.999268 | 1.427526 |
| F:structural molecule activity | 307 | 4716 | 0.000801 | 0.002505 | 0.999199 | 1.448114 |
| F:structural constituent of ribosome | 130 | 4893 | 0.000843 | 0.002554 | 0.999157 | 1.469349 |
| F:ubiquitin-protein ligase activity | 144 | 4879 | 0.001969 | 0.005792 | 0.998031 | 1.489599 |
| C:organelle membrane | 242 | 4270 | 0.004122 | 0.011778 | 0.995878 | 1.508907 |
| P:cell cycle | 340 | 4445 | 0.006445 | 0.017904 | 0.993555 | 1.528547 |
| F:ion transporter activity | 237 | 4786 | 0.012295 | 0.033229 | 0.987707 | 1.543292 |
| P:ion transport | 239 | 4546 | 0.030566 | 0.080436 | 0.969438 | 1.538790 |
| P:protein folding | 121 | 4664 | 0.034044 | 0.087293 | 0.965961 | 1.558002 |
| P:chromosome organization and biogenesis | 105 | 4680 | 0.036151 | 0.090378 | 0.963855 | 1.580090 |
| P:negative regulation of metabolism | 107 | 4678 | 0.036606 | 0.089282 | 0.963400 | 1.605667 |
| C:organelle envelope | 148 | 4364 | 0.041055 | 0.097751 | 0.958950 | 1.625339 |
| C:ubiquitin ligase complex | 122 | 4390 | 0.049298 | 0.114646 | 0.950710 | 1.639154 |
| C:nucleolus | 63 | 4449 | 0.052885 | 0.120193 | 0.947125 | 1.661624 |
| P:signal transduction | 994 | 3791 | 0.059572 | 0.132383 | 0.940431 | 1.679340 |
| P:development | 649 | 4136 | 0.068289 | 0.148455 | 0.931715 | 1.694027 |
| C:endosome | 33 | 4479 | 0.072924 | 0.155158 | 0.927094 | 1.716841 |
| P:cellular morphogenesis | 111 | 4674 | 0.096274 | 0.200570 | 0.903738 | 1.705166 |
| F:DNA binding | 628 | 4395 | 0.127413 | 0.260027 | 0.872593 | 1.678064 |
| F:magnesium ion binding | 135 | 4888 | 0.178351 | 0.356703 | 0.821664 | 1.611106 |

| | | | | | | |
|---|---|---|---|---|---|---|
| C:endoplasmic reticulum | 268 | 4244 | 0.191479 | 0.375448 | 0.808535 | 1.617069 |
| P:protein complex assembly | 122 | 4663 | 0.220653 | 0.424333 | 0.779367 | 1.590544 |
| F:carrier activity | 149 | 4874 | 0.234886 | 0.443181 | 0.765132 | 1.594024 |
| P:cofactor metabolism | 100 | 4685 | 0.290470 | 0.537908 | 0.709555 | 1.509691 |
| F:phosphoric ester hydrolase activity | 113 | 4910 | 0.322610 | 0.586564 | 0.677414 | 1.472638 |
| P:catabolism | 260 | 4525 | 0.424096 | 0.757314 | 0.575922 | 1.279827 |
| C:chromosome | 108 | 4404 | 0.488428 | 0.856891 | 0.511602 | 1.162732 |
| P:alcohol metabolism | 133 | 4652 | 0.577437 | 0.995582 | 0.422588 | 0.982762 |
| P:protein kinase cascade | 118 | 4667 | 0.594004 | 1.006786 | 0.406023 | 0.966720 |
| P:cell differentiation | 173 | 4612 | 0.687263 | 1.145439 | 0.312757 | 0.762821 |
| F:calcium ion binding | 280 | 4743 | 0.760671 | 1.247001 | 0.239342 | 0.598356 |
| P:amino acid metabolism | 112 | 4673 | 0.782782 | 1.262552 | 0.217238 | 0.557021 |
| P:organ development | 213 | 4572 | 0.818392 | 1.299035 | 0.181621 | 0.477951 |
| P:morphogenesis | 212 | 4573 | 0.820673 | 1.282302 | 0.179340 | 0.484703 |
| P:carbohydrate metabolism | 220 | 4565 | 0.835908 | 1.286012 | 0.164104 | 0.455846 |
| P:G-protein coupled receptor protein signaling pathway | 153 | 4632 | 0.846529 | 1.282620 | 0.153485 | 0.438528 |
| F:zinc ion binding | 600 | 4423 | 0.849586 | 1.268038 | 0.150421 | 0.442416 |
| P:negative regulation of biological process | 327 | 4458 | 0.883863 | 1.299798 | 0.116146 | 0.351956 |
| P:cell surface receptor linked signal transduction | 363 | 4422 | 0.916573 | 1.328367 | 0.083433 | 0.260729 |
| P:DNA metabolism | 266 | 4519 | 0.922990 | 1.318558 | 0.077016 | 0.248440 |
| P:cell motility | 105 | 4680 | 0.951250 | 1.339789 | 0.048757 | 0.162524 |
| F:iron ion binding | 111 | 4912 | 0.978445 | 1.358951 | 0.021559 | 0.074341 |
| F:enzyme regulator activity | 238 | 4785 | 0.985550 | 1.350068 | 0.014452 | 0.051615 |
| P:sensory perception | 111 | 4674 | 0.991960 | 1.340486 | 0.008042 | 0.029784 |
| P:cell-cell signaling | 176 | 4609 | 0.993980 | 1.325306 | 0.006021 | 0.023159 |
| P:lipid biosynthesis | 101 | 4684 | 0.994426 | 1.308455 | 0.005576 | 0.022302 |
| P:response to chemical stimulus | 129 | 4656 | 0.995375 | 1.292695 | 0.004626 | 0.019273 |
| P:positive regulation of biological process | 275 | 4510 | 0.995402 | 1.276156 | 0.004599 | 0.019996 |
| P:response to abiotic stimulus | 148 | 4637 | 0.995557 | 1.260199 | 0.004444 | 0.020198 |
| P:carboxylic acid metabolism | 225 | 4560 | 0.997474 | 1.246842 | 0.002527 | 0.012032 |
| P:apoptosis | 244 | 4541 | 0.998627 | 1.232873 | 0.001373 | 0.006866 |
| C:plasma membrane | 608 | 3904 | 0.999709 | 1.219157 | 0.000291 | 0.001533 |
| F:nuclease activity | 60 | 4963 | 0.999915 | 1.204717 | 8.50E-05 | 0.000472 |
| C:peroxisome | 49 | 4463 | 0.999920 | 1.190381 | 8.02E-05 | 0.000472 |
| C:extracellular matrix | 111 | 4401 | 0.999945 | 1.176406 | 5.52E-05 | 0.000345 |
| P:proteolysis | 259 | 4526 | 0.999955 | 1.162738 | 4.53E-05 | 0.000302 |
| C:mitochondrion | 398 | 4114 | 0.999970 | 1.149391 | 3.00E-05 | 0.000214 |
| P:cell adhesion | 242 | 4543 | 0.999998 | 1.136361 | 2.16E-06 | 1.66E-05 |
| P:cell proliferation | 258 | 4527 | 0.999998 | 1.123594 | 1.65E-06 | 1.38E-05 |
| F:peptidase activity | 227 | 4796 | 0.999999 | 1.111110 | 6.15E-07 | 5.59E-06 |
| C:lysosome | 77 | 4435 | 1.000000 | 1.098901 | 6.38E-08 | 6.38E-07 |
| P:electron transport | 151 | 4634 | 1.000000 | 1.086957 | 7.64E-10 | 8.49E-09 |
| P:lipid metabolism | 260 | 4525 | 1.000000 | 1.075269 | 5.95E-11 | 7.43E-10 |
| F:receptor binding | 221 | 4802 | 1.000000 | 1.063830 | 2.15E-11 | 3.07E-10 |
| F:receptor activity | 391 | 4632 | 1.000000 | 1.052632 | 1.11E-11 | 1.85E-10 |
| F:oxidoreductase activity | 309 | 4714 | 1.000000 | 1.041667 | 2.35E-12 | 4.70E-11 |
| P:response to stress | 446 | 4339 | 1.000000 | 1.030928 | 5.05E-26 | 1.26E-24 |
| C:extracellular space | 179 | 4333 | 1.000000 | 1.020408 | 3.49E-37 | 1.16E-35 |
| P:immune response | 270 | 4515 | 1.000000 | 1.010101 | 1.16E-44 | 5.79E-43 |
| P:response to biotic stimulus | 318 | 4467 | 1.000000 | 1.000000 | 4.08E-49 | 4.08E-47 |

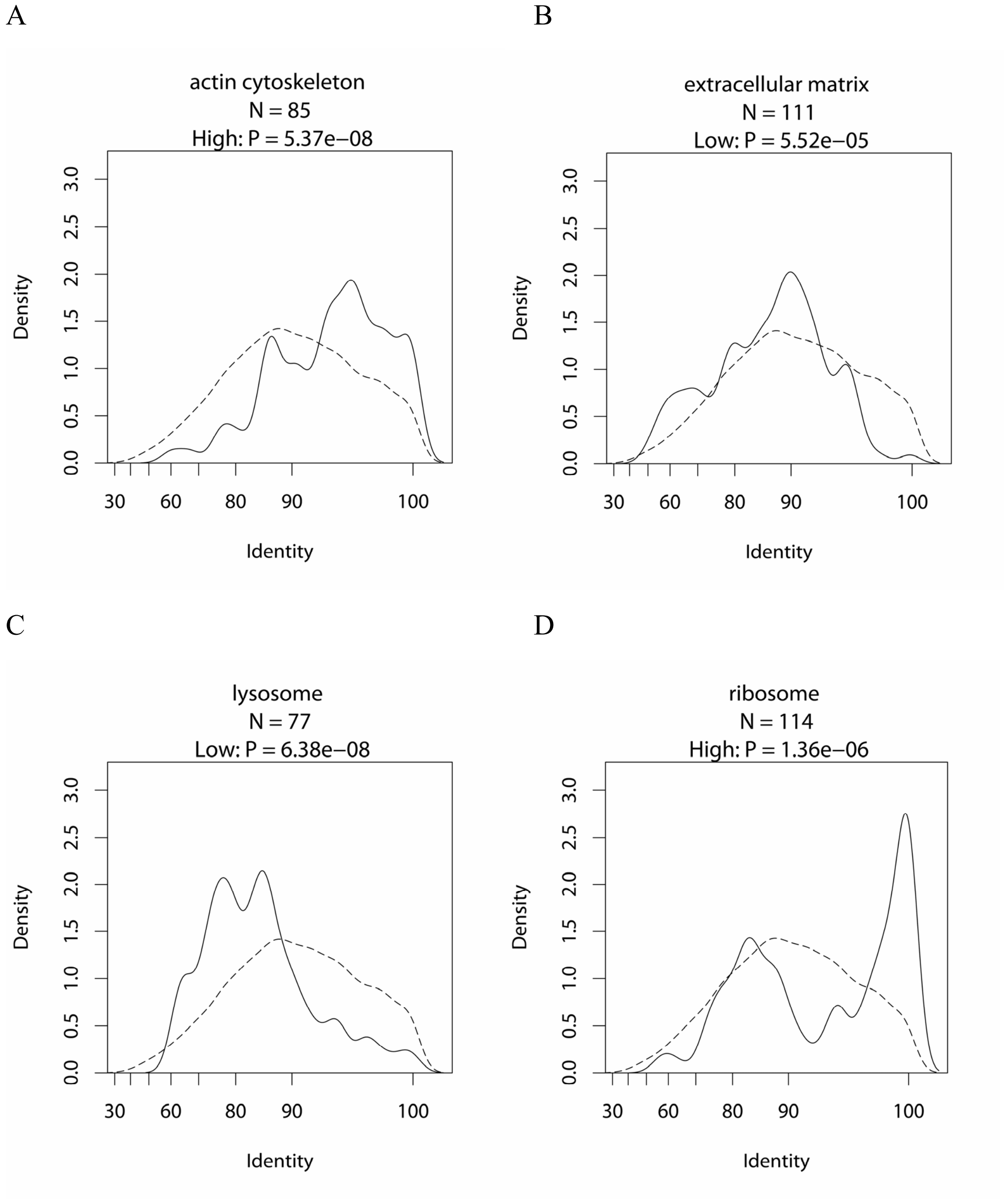
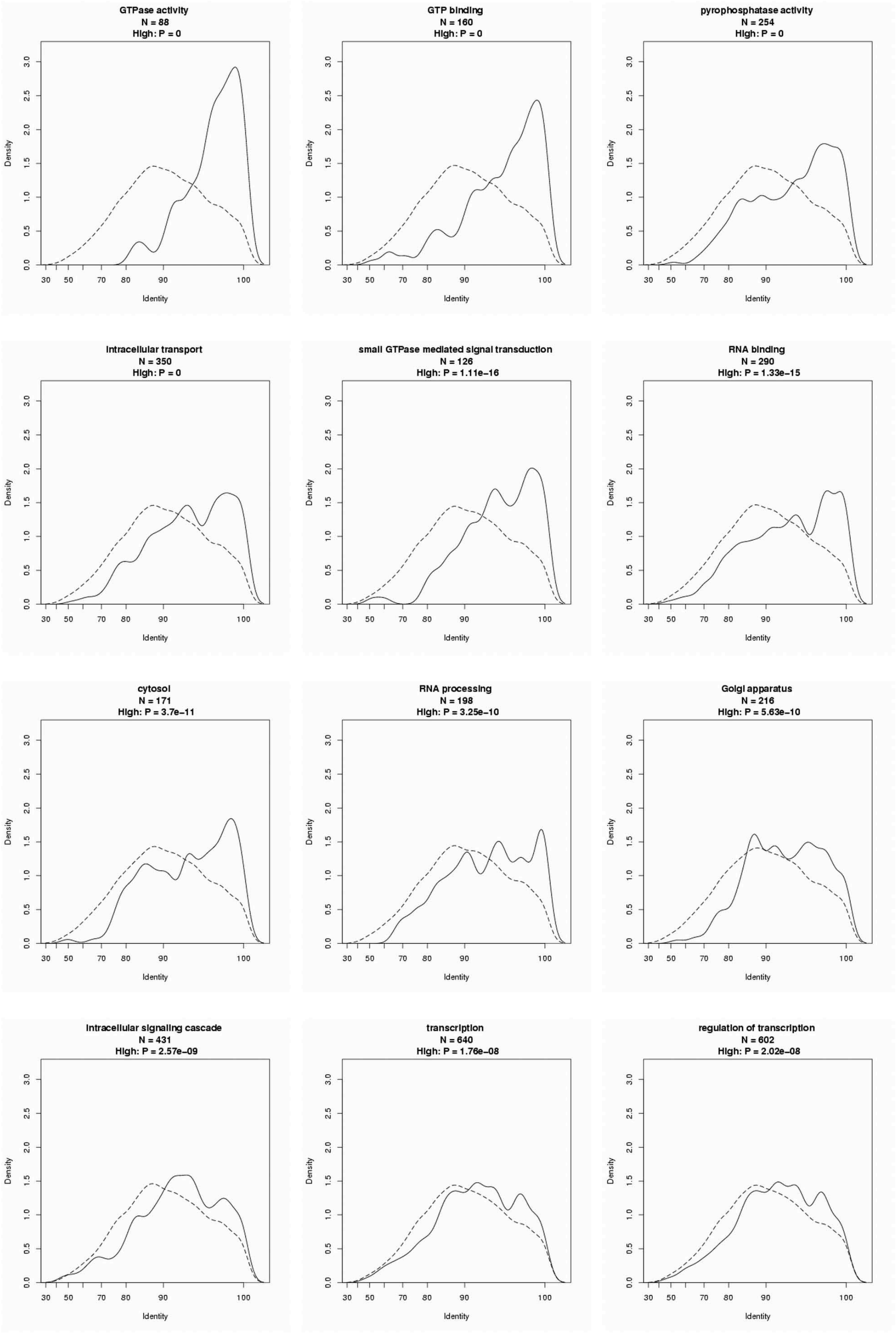
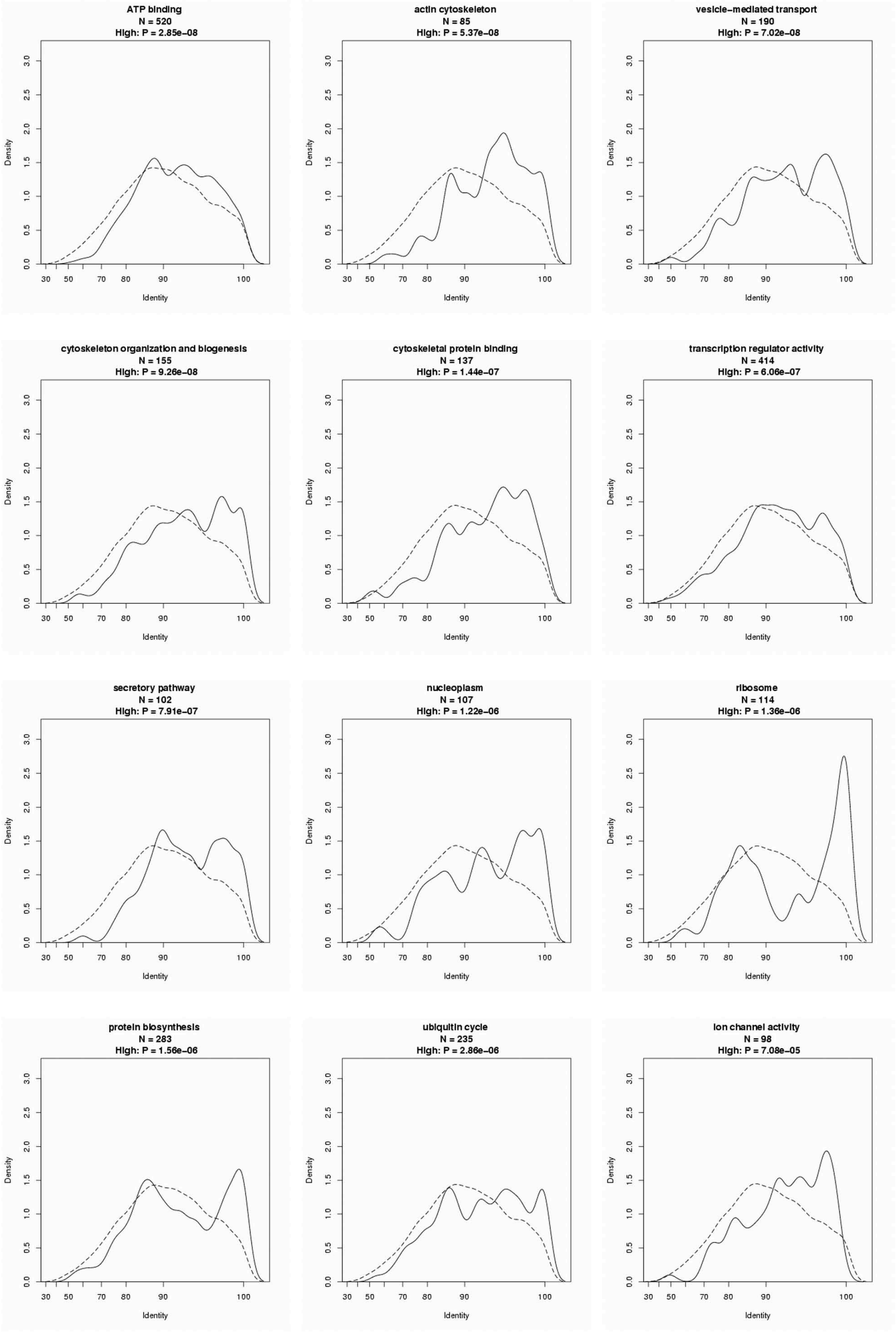


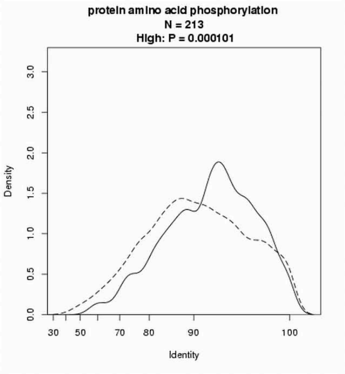
**Figure 27. Distribution of percentage identities of human and mouse protein sequences.**

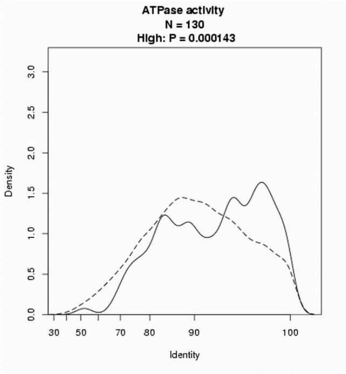
For the high conservation tendency, actin cytoskeleton (A) and ribosome (D), for the low conservation tendency, extracellular matrix (B) and lysosome (C). For each of A-D, the solid line shows the distribution of the identities for genes with the specific GO term, and the dashed line shows the distribution for the control gene set.

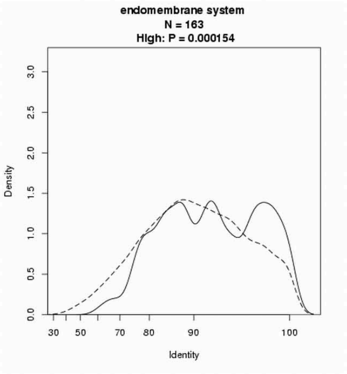
## Figure 28. GO categories with high protein conservation.

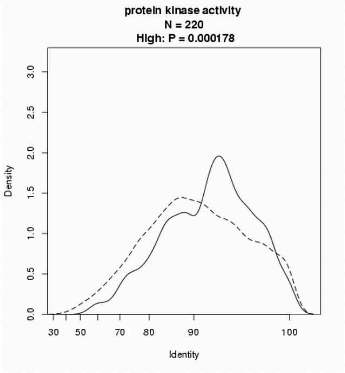


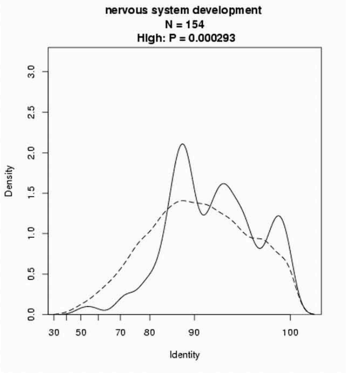

ATP binding
N = 520
High: P = 2.85e−08
actin cytoskeleton
N = 85
High: P = 5.37e−08
vesicle−mediated transport
N = 190
High: P = 7.02e−08
cytoskeleton organization and biogenesis
N = 155
High: P = 9.26e−08
cytoskeletal protein binding
N = 137
High: P = 1.44e−07
transcription regulator activity
N = 414
High: P = 6.06e−07
secretory pathway
N = 102
High: P = 7.91e−07
nucleoplasm
N = 107
High: P = 1.22e−06
ribosome
N = 114
High: P = 1.36e−06
protein biosynthesis
N = 283
High: P = 1.56e−06
ubiquitin cycle
N = 235
High: P = 2.86e−06
ion channel activity
N = 98
High: P = 7.08e−05
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Identity
30
50
70
80
90
100

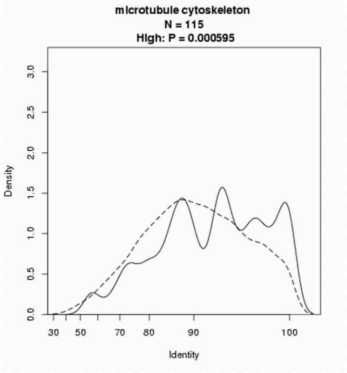

protein amino acid phosphorylation
N = 213
High: P = 0.000101
Density
0.0 0.5 1.0 1.5 2.0 2.5 3.0
30 50 70 80 90 100
Identity

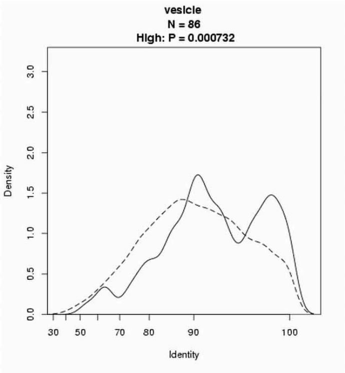

ATPase activity
N = 130
High: P = 0.000143
Density
0.0 0.5 1.0 1.5 2.0 2.5 3.0
30 50 70 80 90 100
Identity

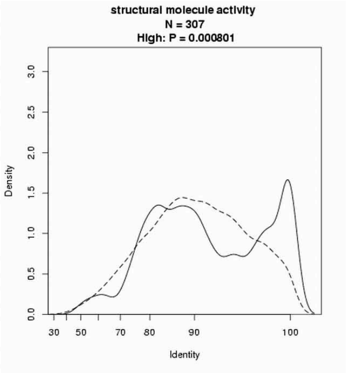

endomembrane system
N = 163
High: P = 0.000154
Density
0.0 0.5 1.0 1.5 2.0 2.5 3.0
30 50 70 80 90 100
Identity

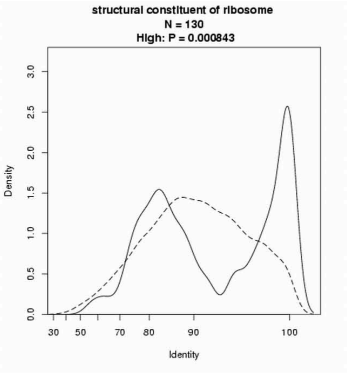

protein kinase activity
N = 220
High: P = 0.000178
Density
0.0 0.5 1.0 1.5 2.0 2.5 3.0
30 50 70 80 90 100
Identity

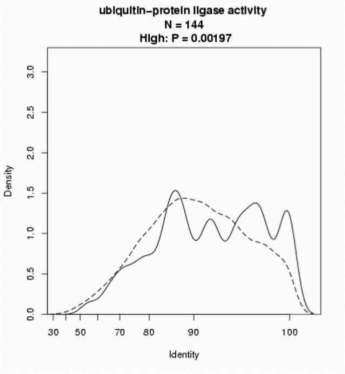

nervous system development
N = 154
High: P = 0.000293
Density
0.0 0.5 1.0 1.5 2.0 2.5 3.0
30 50 70 80 90 100
Identity

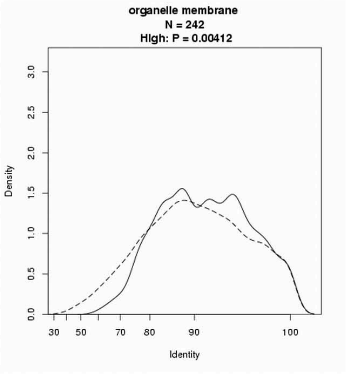

microtubule cytoskeleton
N = 115
High: P = 0.000595
Density
0.0 0.5 1.0 1.5 2.0 2.5 3.0
30 50 70 80 90 100
Identity

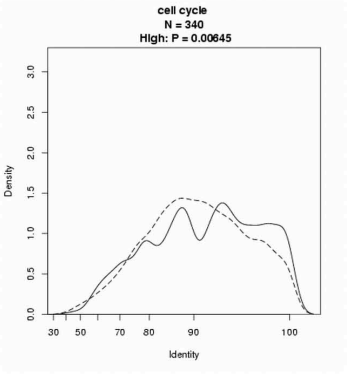

vesicle
N = 86
High: P = 0.000732
Density
0.0 0.5 1.0 1.5 2.0 2.5 3.0
30 50 70 80 90 100
Identity


structural molecule activity
N = 307
High: P = 0.000801
Density
0.0 0.5 1.0 1.5 2.0 2.5 3.0
30 50 70 80 90 100
Identity


structural constituent of ribosome
N = 130
High: P = 0.000843
Density
0.0 0.5 1.0 1.5 2.0 2.5 3.0
30 50 70 80 90 100
Identity


ubiquitin-protein ligase activity
N = 144
High: P = 0.00197
Density
0.0 0.5 1.0 1.5 2.0 2.5 3.0
30 50 70 80 90 100
Identity


organelle membrane
N = 242
High: P = 0.00412
Density
0.0 0.5 1.0 1.5 2.0 2.5 3.0
30 50 70 80 90 100
Identity


cell cycle
N = 340
High: P = 0.00645
Density
0.0 0.5 1.0 1.5 2.0 2.5 3.0
30 50 70 80 90 100
Identity

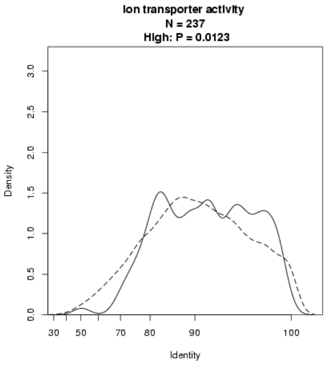
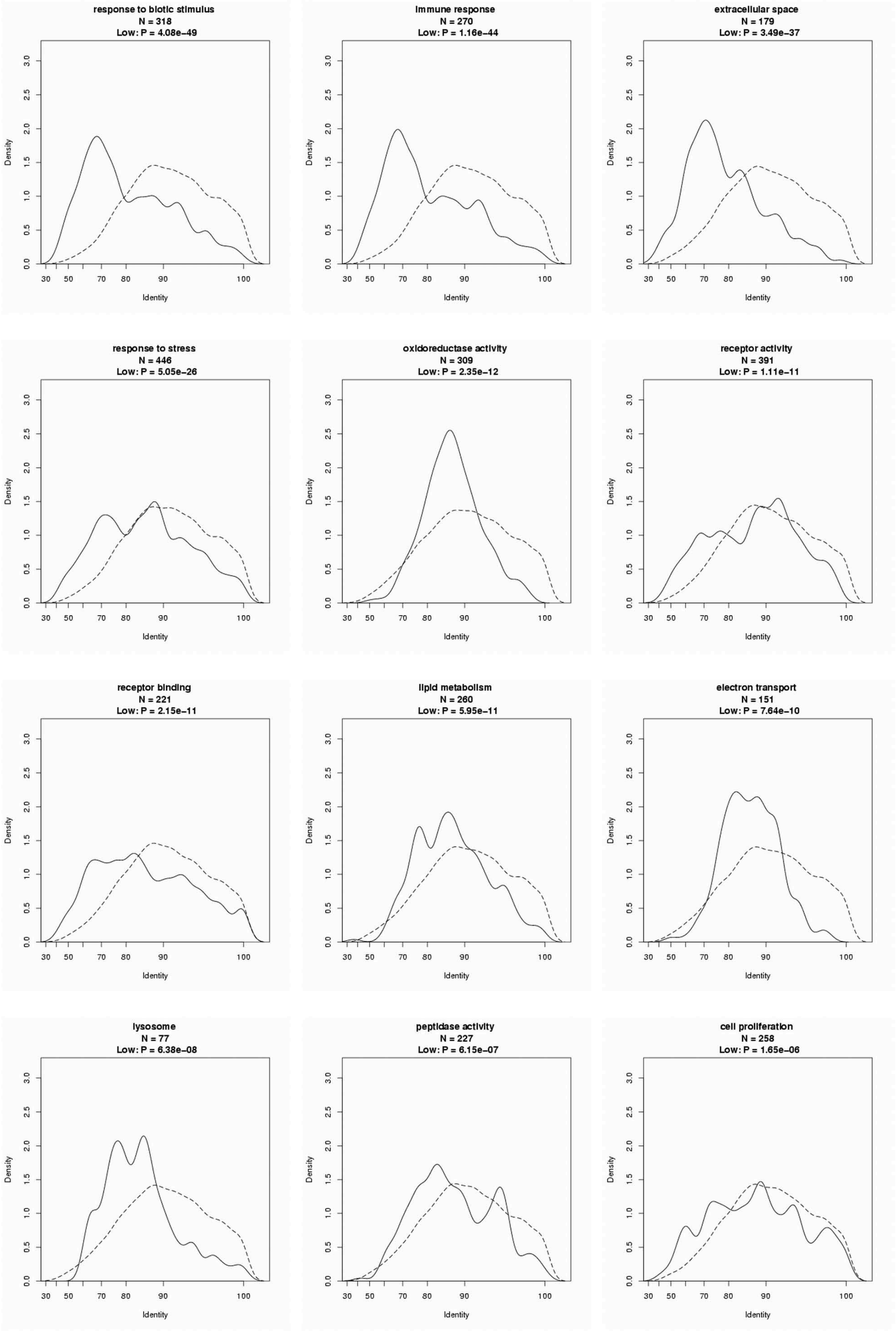
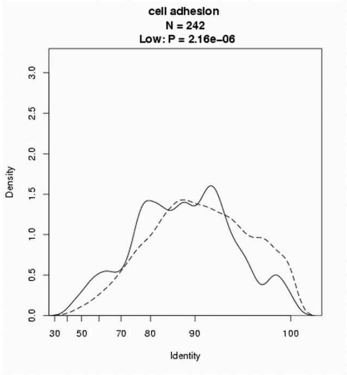
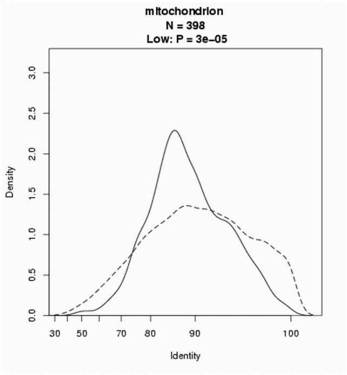
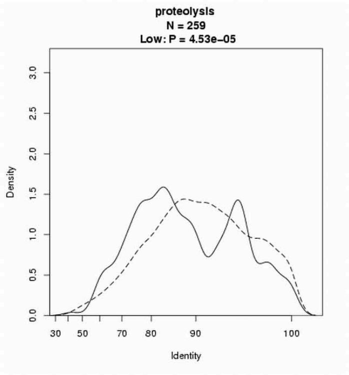

Ion transporter activity
N = 237
High: P = 0.0123
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
30
50
70
80
90
100
Identity

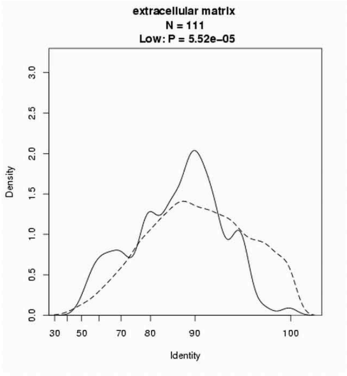
Figure 29. GO categories with low protein conservation.

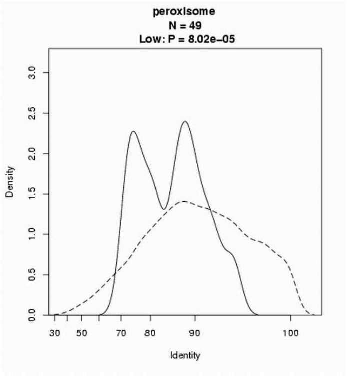


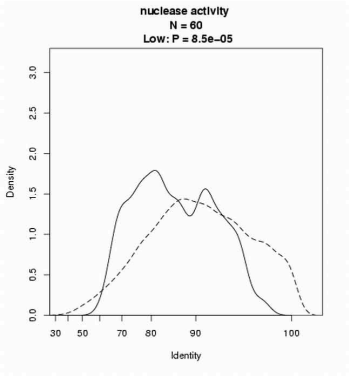

cell adhesion
N = 242
Low: P = 2.16e−06
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
30
50
70
80
90
100
Identity

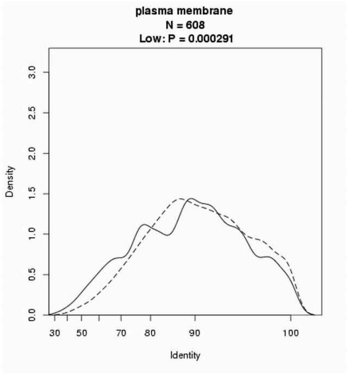

mitochondrion
N = 398
Low: P = 3e−05
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
30
50
70
80
90
100
Identity

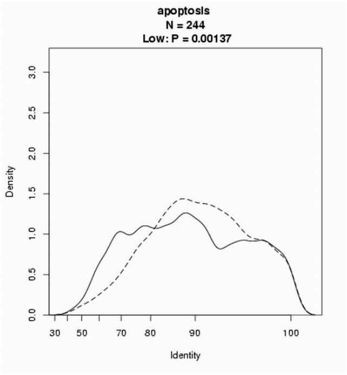

proteolysis
N = 259
Low: P = 4.53e−05
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
30
50
70
80
90
100
Identity

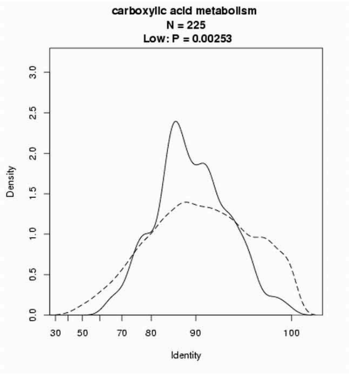

extracellular matrix
N = 111
Low: P = 5.52e−05
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
30
50
70
80
90
100
Identity

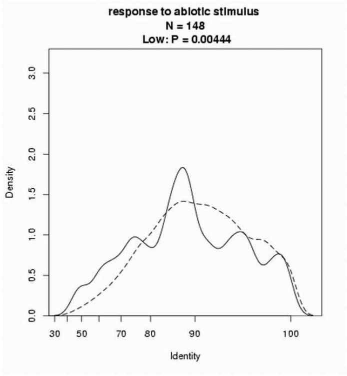

peroxisome
N = 49
Low: P = 8.02e−05
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
30
50
70
80
90
100
Identity

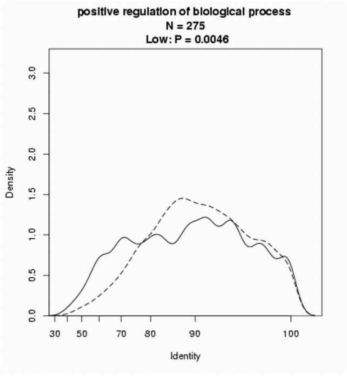

nuclease activity
N = 60
Low: P = 8.5e−05
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
30
50
70
80
90
100
Identity

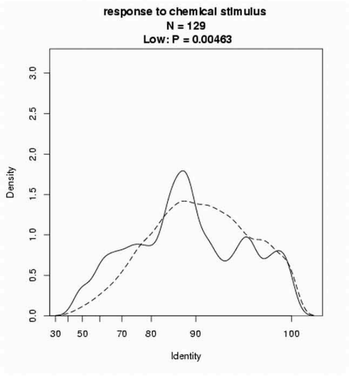

plasma membrane
N = 608
Low: P = 0.000291
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
30
50
70
80
90
100
Identity

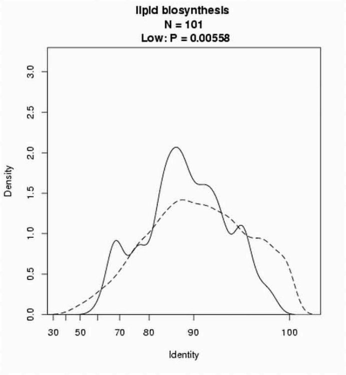

apoptosis
N = 244
Low: P = 0.00137
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
30
50
70
80
90
100
Identity

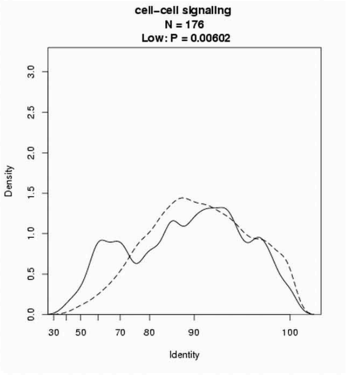

carboxylic acid metabolism
N = 225
Low: P = 0.00253
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
30
50
70
80
90
100
Identity

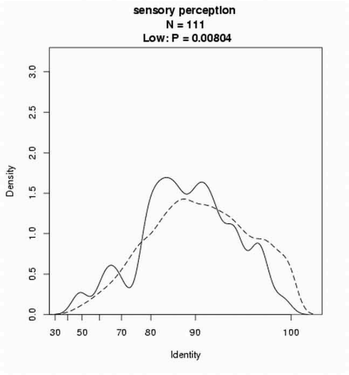

response to abiotic stimulus
N = 148
Low: P = 0.00444
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
30
50
70
80
90
100
Identity


positive regulation of biological process
N = 275
Low: P = 0.0046
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
30
50
70
80
90
100
Identity


response to chemical stimulus
N = 129
Low: P = 0.00463
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
30
50
70
80
90
100
Identity


lipid biosynthesis
N = 101
Low: P = 0.00558
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
30
50
70
80
90
100
Identity


cell−cell signaling
N = 176
Low: P = 0.00602
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
30
50
70
80
90
100
Identity


sensory perception
N = 111
Low: P = 0.00804
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
30
50
70
80
90
100
Identity

### 4.3.4 Promoter conservation and protein conservation

To examine the relationship between promoter conservation and protein conservation, we calculated the correlation coefficient of promoter conservation (raw alignment score obtained by *water*) and protein conservation (percentage identity obtained by *water*). This correlation was very weak (the Kendall's rank correlation is 0.193, Figure 30), suggesting that the promoter and protein sequences are under different types of selective pressure. We further investigated the relationship between protein and promoter conservation in detail, by decomposing it based on GO categories. From Table 9, 10, 12, and 13, the terms that have significant conservation tendencies for both protein sequences and promoter sequences were extracted and compiled as a 2 by 2 cross table (Table 15). Although the results using mouse gene annotations were mostly consistent with the result based on human genes, P:cell-cell signaling showed high protein conservation based on mouse annotation (Figure 31B) whereas low protein conservation on human annotation (Figure 31A). An examination of the contents of the two gene sets revealed that the observed difference seems to be derived from the different GO annotation status between human and mouse. Specifically, 151 genes out of 176 are annotated as P:cell-cell signaling only in human, and these genes seems to contribute to the low protein conservation tendency (Figure 31C).

Table 15 illustrates the relationship between protein conservation and promoter conservation, on the functional category basis. GO terms in the upper right cell, which have high conservation for both protein and promoter sequences, are related to transcription regulation or intracellular signaling. In contrast, the membrane-bounded organelles engaged in metabolism are in the lower left cell, showing low conservation for both protein and promoter. Interestingly, several terms are in the upper left and lower right cell, indicating opposite characteristics for protein and promoter conservation. For example, although genes related to signaling events showed high promoter conservation, they do not always have high protein conservation, but can even have low protein conservation; P:cell-cell signaling shows low protein conservation, while F:regulation of transcription shows high protein conservation. An analogous situation can be seen in the case of genes with low promoter conservation; among metabolism-related terms, C:ribosome shows high protein conservation, while C:mitochondrion shows low protein conservation. These results illustrate that there seems to be a nonparallel component in protein and promoter sequence evolution.

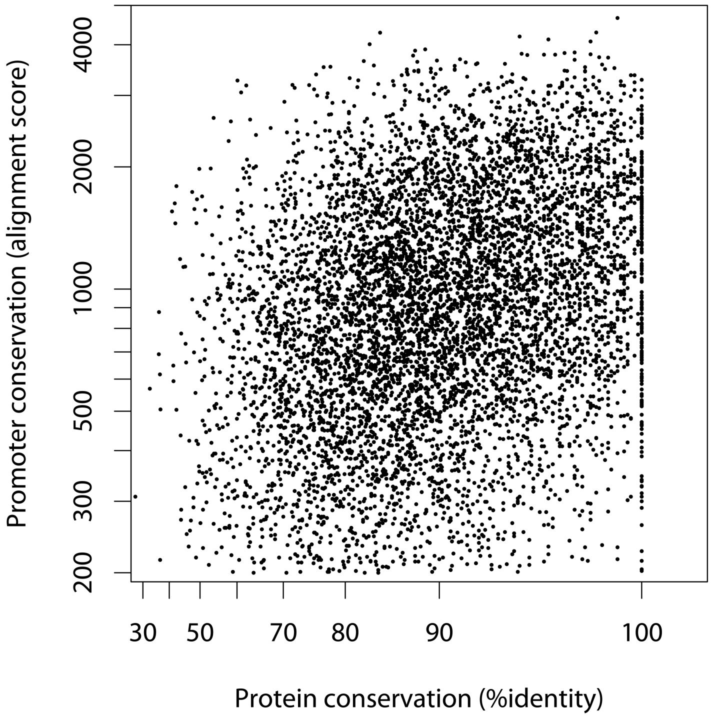


**Figure 30. Correlation between protein conservation and promoter conservation.**

**Table 15. Summary of GO categories that show significant conservation tendencies for both protein and promoter sequences.**

| | | | |
|---|---|---|---|
| **Promoter conservation** | **High** | F:receptor binding (221)<br>P:cell-cell signaling (176)<br>C:extracellular matrix (111)<br>P:cell adhesion (242)<br>C:plasma membrane (608) | P:regulation of transcription (602)<br>F:transcription factor activity (263)<br>P:transcription (640)<br>P:nervous system development (154)<br>F:cytoskeletal protein binding (137)<br>F:ion channel activity (98)<br>C:actin cytoskeleton (85)<br>P:protein amino acid phosphorylation (213)<br>F:ion transporter activity<br>F:protein kinase activity (220)<br>P:intracellular signaling cascade (431)<br>F:GTPase activity<br>P:cytoskeleton organization and biogenesis<br>P:small GTPase mediated signal transduction |
| | **Low** | P:proteolysis (259)<br>F:peptidase activity (227)<br>P:immune response (270)<br>P:response to biotic stimulus (318)<br>F:nuclease activity (60)<br>C:peroxisome (49)<br>P:electron transport (151)<br>P:carboxylic acid metabolism (225)<br>P:lipid metabolism (260)<br>C:lysosome (77)<br>F:oxidoreductase activity (309)<br>C:mitochondrion (398) | P:protein biosynthesis (283)<br>F:structural constituent of ribosome (130)<br>C:ribosome (114) |
| | | **Low** | **High** |
| | | **Protein conservation** | |

In each cell, the GO categories are ordered by promoter conservation. The number of genes for each term is shown in parentheses. GO annotations associated with human genes were used to make this table.

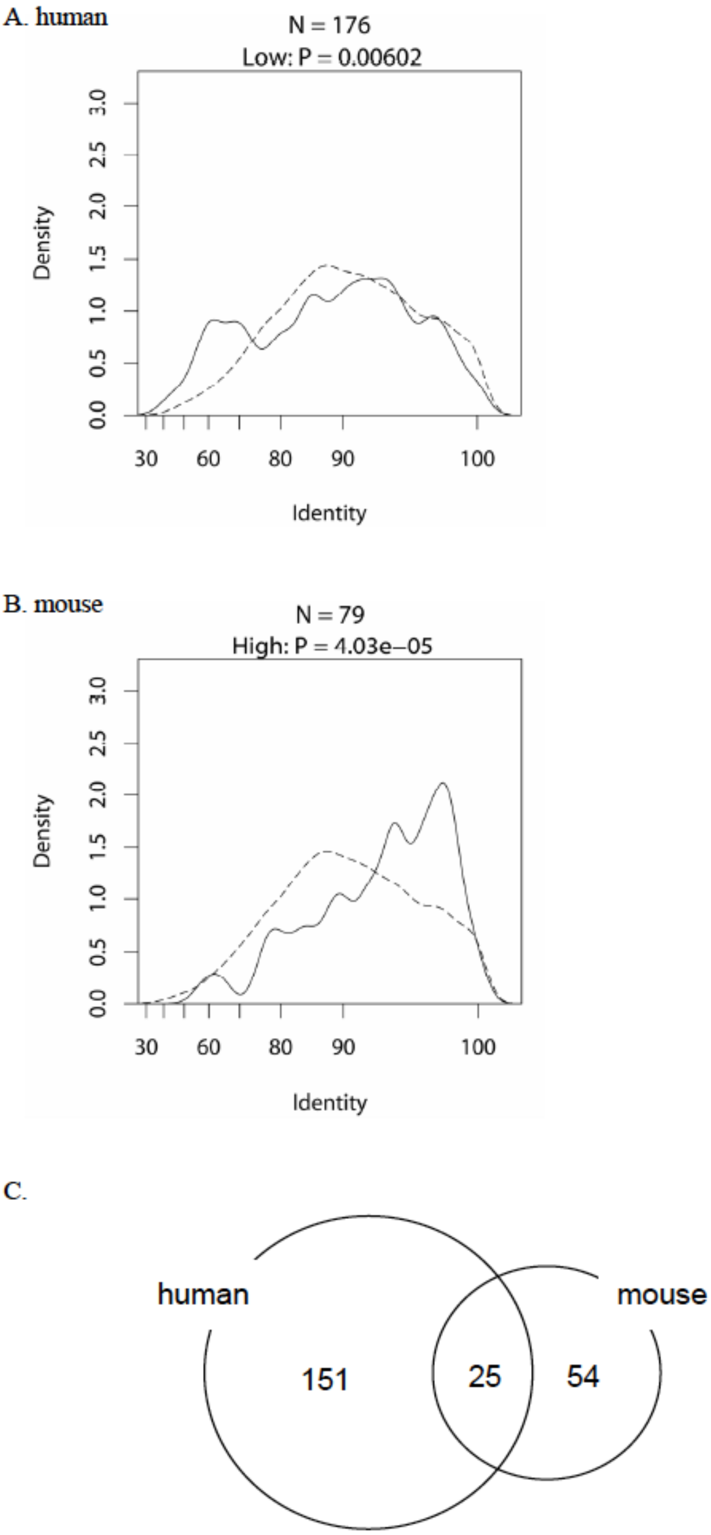


**Figure 31. Protein conservation of human and mouse "cell-cell signaling" genes.**

### 4.3.5 Example: ribosomal proteins

Unlike other categories, C:ribosome shows a bimodal distribution of protein conservation (Figure 27B); one is around 100% identity, and the other ranges from 70% to 90%. Consistently, several categories related to C:ribosome (P:protein biosynthesis and F:structural constituent of ribosome) also show bimodal distributions (Figure 28). This result could be due to different evolutionary rates between cytoplasmic and mitochondrial ribosomal protein [97]. Therefore, we checked the annotations for the genes in the C:ribosome category, using the NCBI RefSeq database [91]. In fact, the peak with high protein conservation is substantially composed of cytoplasmic ribosomal proteins, while the peak with lower protein conservation mainly comprises nuclear-encoded mitochondrial ribosomal proteins (Figure 32). Notably, the general protein conservation tendency described in previous sections holds here: proteins in the cytosol show high protein conservation, while proteins in membrane-bounded organelles, such as mitochondria, have low protein conservation.

Besides the protein conservation, we examined the promoter conservation tendency for the two subsets of the C:ribosome category, cytoplasmic and mitochondrial ribosomal proteins. In contrast to the protein conservation, we could not observe a significant difference in the conservation levels between these two subgroups (*P*-value = 0.34 by Wilcoxon rank sum test; see Figure 32). Apparently, the protein conservation is drastically different between cytoplasmic and mitochondrial ribosomal proteins, whereas the distribution of promoter conservation is quite similar. This result underscores the decoupled property of protein and promoter sequence evolution.

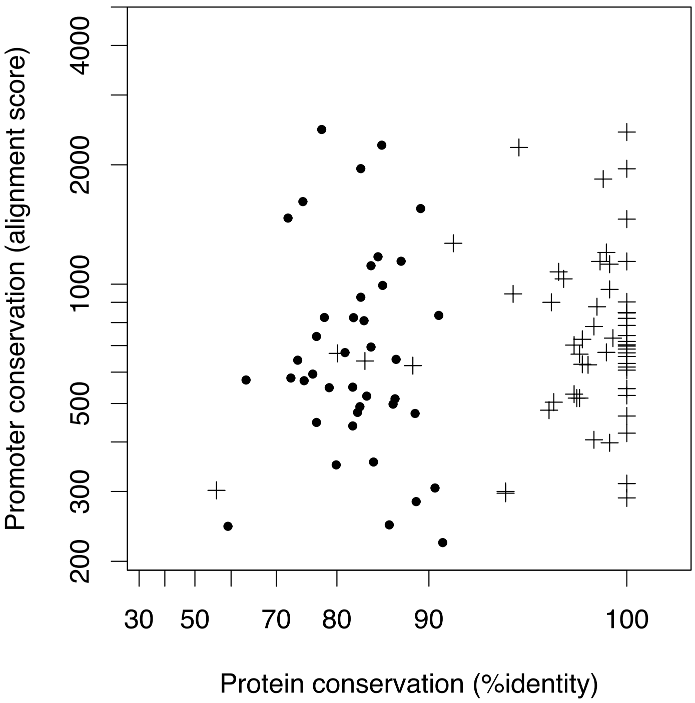


**Figure 32. Protein conservation and promoter conservation for two subsets of ribosomal proteins.** Crosses represent cytoplasmic ribosomal proteins (58 genes). Dots represent mitochondrial ribosomal proteins (41 genes). The conspicuous outlier corresponding to (56, 302), "ribosomal L1 domain containing 1", does not seem to be an actual ribosomal protein and might have been erroneously annotated by an electronic procedure.

## 4.4 Discussion

When we conducted a comprehensive comparison of promoter sequences for human and mouse orthologous genes, we noted that the promoter pairs of orthologous genes contained non-orthologous promoters. The source of these non-orthologous promoters could be the potential false pairings in the orthologous table. Another possible reason is the presence of alternative promoters [98,99], which can result in the failure to select the corresponding TSSs between human and mouse. The other possible cause is the existence of species-specific promoters; for example, our group recently reported that there are human promoter sequences whose counterparts are completely missing in the mouse genomic sequences [100]. Nevertheless, despite these problems that may cause mis-pairing of non-orthologous promoters, as much as 82% of the promoter pairs were shown to be evolutionally related in the data set. Although the dynamic aspects of TSSs, such as TSS diversification ad TSS turn over, have been highlighted recently [98,99,101,102], our results show that the representative TSS for each gene has been generally sustained during the evolution of the human and mouse lineages.

We focused on gene pairs with promoters that appeared to be truly evolutionally related, and examined the relationship between promoter conservation and gene function. We found that the terms with high promoter conservation are related to signaling events inside as well as outside of the cell. Considering that the promoter conservation levels reflect the regulatory information contained in the sequence, the results suggest that these genes require more regulatory information embedded in the promoter. It is reasonable to suppose that more regulatory information enables more sophisticated changes of expression levels, thereby allowing these proteins to work effectively as signaling molecules. On the other hand, genes involved in metabolism, which showed low promoter conservation, may require relatively less regulatory information in their promoter sequences. Consistently, a recent study revealed that housekeeping genes tend to show reduced upstream sequence conservation [103]. Specifically, in relation to ribosomal proteins, Perry et al. [104] pointed out that most of their promoters contain transposable elements, resulting in a low conservation. The reduced regulatory information in the promoters of ribosomal proteins might be compensated by the translational regulation mechanism directed by the 5' terminal oligopyrimidine sequence in their mRNAs [105].

Related discussions on regulatory sequence conservation have been made for specific categories of genes. Iwama and Gojobori [26] found that transcription factor genes, particularly those related to developmental processes, show high upstream sequence conservation, suggesting that these genes form highly connected regulatory networks. Lee et al. [27] reported that genes involved in adaptive processes tend to have highly conserved upstream regions in mammalian genomes, and also suggested the complex combinatorial circuitry of their transcriptional regulation. There have been other approaches based on whole genome comparisons, where highly conserved non-coding regions were found to be associated with developmental genes [95,106,107]. However, as Lee et al. suggested [27], most of these regions are far from genes and have little overlap with promoter regions. Thus, it seems that these regions are conserved independently from the promoter regions.

The conserved elements in the promoter may be either very short, or spread over a much longer region than the 1,200 bases. In both cases, our measures will report poor conservation when there is just a right amount of it. The local alignment score we used to measure promoter conservation can be roughly considered as a combination of identity and alignment length. Identity reflects the rates of substitutions and indels, and length reflects larger insertions, such as transposon insertions. When we examined the promoter conservation tendency for each GO term, by using alignment length or percentage identity as a measure of conservation, the tendencies were consistent with each other (Figure 33). Thus, the evolutionary pressures of each functional category on alignment length and identity work in the same direction.

When we investigated the relationship between protein conservation and promoter conservation in mammals, we observed a very weak correlation between them. This suggests that substantial portions of the evolutionary changes of promoter and protein sequences are under different types of selective pressures. This observation is consistent with the nematode [76] and yeast [108] cases, and thus the very weak correlation between protein and promoter conservation might be universal from unicellular organisms to higher vertebrates.

In order to understand the relationship of protein and promoter sequence conservation in terms of gene functions, we examined it by a decomposition based on GO categories. When we dissected not only promoter conservation but also protein conservation, different trends were observed for proteins and promoters. As for proteins, high conservations were observed for terms related to a wide range of gene expression processes that occur in the cytosol and the nucleoplasm, while low conservations were observed for terms related to extracellular regions, cell surface and membrane-bounded organelles (such as mitochondrion, peroxisome and lysosome). Although the results for the membrane-bounded organelles seem surprising, considering that they often carry out basic, conserved metabolic process, they can also be considered as being topologically “outside” of the cell, given that they are on the opposite side of the membrane from the cytosol. The problem of the determinant of the protein evolutionary rate [92,109] needs to be solved to fully clarify the phenomenon. Nevertheless, our observation provides the trends of the protein sequence evolution in terms of functional categories. Comparing these trends with those of promoters, we found that these two kinds of trends are nonparallel: protein conservation depends on whether they are on the cytosolic side or not, while promoter conservation seems to depend on whether the gene is related to signaling or metabolism. Specifically, cytoplasmic ribosomal proteins, which exist in the cytosol and are engaged in metabolism, have high protein conservation in spite of low promoter conservation. On the other hand, cell-cell signaling genes, which act outside or at the surface of the cell to convey signals, show low protein conservation in spite of high promoter conservation. These terms may provide evidence that decoupled properties exist between protein and promoter sequence evolution.

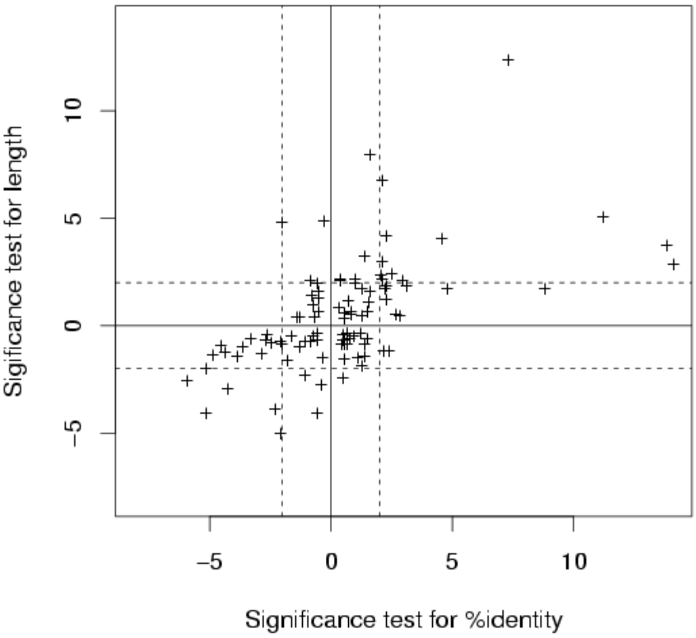


**Figure 33. Promoter conservation tendency for each GO category based on alignment length and percentage identity.**

Each axis is based on the common logarithm of *P*-values of significance tests for each GO. Plus means high identity and minus means low identity. Dashed lines correspond to a *P*-value of 0.01.

## 4.5 Conclusions

In this study described in this chapter, my collaborators and I examined the relationship between protein conservation and promoter conservation in detail, by decomposing it based on functional categories. The results show the relation of gene function to protein conservation and promoter conservation, and revealed that there seem to be nonparallel components between protein and promoter sequence evolution. This study will provide a basis to understand the evolution of mammalian genes and their regulation. Further efforts are now being made to construct reliable promoter sequences based on full-length cDNAs. Future analyses of multiple species will clarify the evolutionary mechanisms of the coding and regulatory sequences more precisely.

# Chapter 5 Concluding remarks

As concluding remarks, I will summarize the studies presented in this thesis and provide the future directions based on these studies. In this thesis, I describe the following three computational approaches for analyzing and utilizing the gene orthology: identification of orthology at the domain level; construction of a database of orthology; and comparative analysis of protein-coding and promoter sequences based on orthology.

Firstly, we developed a method for improving domain-level ortholog classification on the basis of the optimization of a score and demonstrated the effectiveness of the method using the manually curated reference databases. For this purpose, we designed a score for evaluating ortholog clusters at the domain level using multiple alignments and demonstrated that the method contributes to the improvement of the clustering. This method will enhance the reliability of ortholog databases and thereby contribute to comparative analyses using them.

Secondly, I developed a general RDF model for describing ortholog information on the basis of an ontology OrthO. The model enables the integration of functional information for multiple organisms. Furthermore, the ortholog information from different data sources can be compared using the OrthO as a shared ontology. By representing the data in this RDF model, the ortholog database can work as a hub structure for biological databases in the Semantic Web, and it will contribute to knowledge discovery through integrative data analysis.

I also examined the relationship between protein conservation and promoter conservation in detail by decomposing it based on functional categories. Our results show the relation of gene function to protein conservation and promoter conservation, revealing that there seem to be nonparallel components between protein and promoter sequence evolution. I believe that this study will provide a basis to understand the evolution of mammalian genes and their regulation. Further efforts are now being made to construct reliable promoter sequences based on full-length cDNAs. Future analyses of multiple species will more precisely clarify the evolutionary mechanisms of the coding and regulatory sequences.

In the first study of domain-level orthology, we mainly focused on microbial genomes, where gene fusion events prevail but the domain architectures are relatively simple. In higher organisms with more complex domain architectures, the detection of domain-level orthology will be more challenging. Although the pipeline probably needs to be changed in order to deal with more ambiguous domain boundaries, the basic strategy based on the score optimization will be an effective approach. As for the second study of the database construction, the data model was designed to be general; thus the target species are not limited. The model even accepts orthology data from other ortholog databases. Therefore, it is possible to construct a meta-database of orthologs based on the framework presented here. The meta-database will contribute not only to the interoperability of the ortholog database but also to the assessment of the reliability of various ortholog databases. In this study, the orthology of the protein-coding region is considered. However, interestingly, the data model can deal with other regions, such as regulatory regions. It is thus possible to extend the database to protein-coding and promoter sequences of various organisms. In the third topic of this thesis, the comparative analysis was limited to human and mouse. However, the extension of the ortholog database mentioned above will enable more comprehensive analysis, contributing to a better understanding of the genome evolution of a broader range of organisms.

# Acknowledgements

I wish to express my gratitude to Professor Kenta Nakai, my supervisor in graduate school, for his continuous advice and encouragement during and after the graduate program. I am grateful to Dr. Ikuo Uchiyama for his insightful discussions and comments throughout the collaborative research studies. I would like to thank Professor Kengo Kinoshita and Dr. Riu Yamashita for their fruitful discussions and encouragement regarding my research. I would like to thank Hiroyo Nishide for her dedication to the computer system support and service development. I would also like to thank the organizers and participants of the BioHackathon series of meetings for sharing their experiences in the Semantic Web technologies.

# References

[1] Fang, G., Bhardwaj, N., Robilotto, R. and Gerstein, M.B. (2010) Getting started in gene orthology and functional analysis. *PLoS Comput. Biol.* **6**(3):e1000703.

[2] Fitch, W.M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.* **19**(2):99-113.

[3] Sonnhammer, E.L. and Koonin, E.V. (2002) Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet.* **18**(12):619-620.

[4] Koonin, E.V. (2005) Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.* **39**:309-338.

[5] Tatusov, R.L., Koonin, E.V. and Lipman, D.J. (1997) A genomic perspective on protein families. *Science* **278**(5338):631-637.

[6] Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. and Yeates, T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U. S. A.* **96**(8):4285-4288.

[7] Pagani, I., Liolios, K., Jansson, J., Chen, I.M., Smirnova, T., Nosrat, B., Markowitz, V.M. and Kyrpides, N.C. (2012) The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.* **40**(Database issue):D571-579.

[8] Uchiyama, I. (2006) Hierarchical clustering algorithm for comprehensive orthologous-domain classification in multiple genomes. *Nucleic Acids Res.* **34**(2):647-658.

[9] Chiba, H. and Uchiyama, I. (2014) Improvement of domain-level ortholog clustering by optimizing domain-specific sum-of-pairs score. *BMC Bioinformatics* **15**(1):148.

[10] Berners-Lee, T. and Hendler, J. (2001) Publishing on the semantic web. *Nature* **410**(6832):1023-1024.

[11] Katayama, T., Wilkinson, M.D., Micklem, G., Kawashima, S., Yamaguchi, A., Nakao, M., Yamamoto, Y., Okamoto, S., Oouchida, K., Chun, H.W. *et al.* (2013) The 3rd DBCLS BioHackathon: improving life science data integration with Semantic Web technologies. *J. Biomed. Semantics* **4**(1):6.

[12] Chen, H., Yu, T. and Chen, J.Y. (2013) Semantic Web meets Integrative Biology: a survey. *Brief. Bioinform.* **14**(1):109-125.

[13] Katayama, T., Wilkinson, M.D., Aoki-Kinoshita, K.F., Kawashima, S., Yamamoto, Y., Yamaguchi, A., Okamoto, S., Kawano, S., Kim, J.D., Wang, Y. *et al.* (2014) BioHackathon

series in 2011 and 2012: penetration of ontology and linked data in life science domains. *J. Biomed. Semantics* **5**(1):5.

[14] Chiba, H., Nishide, H. and Uchiyama, I. (2015) Construction of an ortholog database using the semantic web technology for integrative analysis of genomic data. *PLoS ONE* **10**(4):e0122802.

[15] Uchiyama, I., Mihara, M., Nishide, H. and Chiba, H. (2013) MBGD update 2013: the microbial genome database for exploring the diversity of microbial world. *Nucleic Acids Res.* **41**(Database issue):D631-635.

[16] O'Brien, K.P., Remm, M. and Sonnhammer, E.L. (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.* **33**(Database issue):D476-480.

[17] Ureta-Vidal, A., Ettwiller, L. and Birney, E. (2003) Comparative genomics: genome-wide analysis in metazoan eukaryotes. *Nat. Rev. Genet.* **4**(4):251-262.

[18] Batzoglou, S., Pachter, L., Mesirov, J.P., Berger, B. and Lander, E.S. (2000) Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res.* **10**(7):950-958.

[19] Nei, M. (1987) *Molecular Evolutionary Genetics*. Columbia University Press, New York.

[20] Wolfe, K.H. and Sharp, P.M. (1993) Mammalian gene evolution: nucleotide sequence divergence between mouse and rat. *J. Mol. Evol.* **37**(4):441-456.

[21] Murphy, P.M. (1993) Molecular mimicry and the generation of host defense protein diversity. *Cell* **72**(6):823-826.

[22] Makalowski, W., Zhang, J. and Boguski, M.S. (1996) Comparative analysis of 1196 orthologous mouse and human full-length mRNA and protein sequences. *Genome Res.* **6**(9):846-857.

[23] Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* **409**(6822):860-921.

[24] Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A. *et al.* (2001) The sequence of the human genome. *Science* **291**(5507):1304-1351.

[25] Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**(6915):520-562.

[26] Iwama, H. and Gojobori, T. (2004) Highly conserved upstream sequences for transcription factor genes and implications for the regulatory network. *Proc. Natl. Acad. Sci. U. S. A.* **101**(49):17156-17161.

[27] Lee, S., Kohane, I. and Kasif, S. (2005) Genes involved in complex adaptive processes tend to have highly conserved upstream regions in mammalian genomes. *BMC Genomics* **6**:168.

[28] Choi, S.S., Bush, E.C. and Lahn, B.T. (2006) Different classes of tissue-specific genes show different levels of noncoding conservation. *Genomics* **87**(3):433-436.

[29] Chiba, H., Yamashita, R., Kinoshita, K. and Nakai, K. (2008) Weak correlation between sequence conservation in promoter regions and in protein-coding regions of human-mouse orthologous gene pairs. *BMC Genomics* **9**:152.

[30] Kuzniar, A., van Ham, R.C., Pongor, S. and Leunissen, J.A. (2008) The quest for orthologs: finding the corresponding gene across genomes. *Trends Genet.* **24**(11):539-551.

[31] Kristensen, D.M., Wolf, Y.I., Mushegian, A.R. and Koonin, E.V. (2011) Computational methods for Gene Orthology inference. *Brief. Bioinform.* **12**(5):379-391.

[32] Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**:41.

[33] Jensen, L.J., Julien, P., Kuhn, M., von Mering, C., Muller, J., Doerks, T. and Bork, P. (2008) eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res.* **36**(Database issue):D250-254.

[34] Chen, F., Mackey, A.J., Stoeckert, C.J., Jr. and Roos, D.S. (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.* **34**(Database issue):D363-368.

[35] Altenhoff, A.M., Schneider, A., Gonnet, G.H. and Dessimoz, C. (2011) OMA 2011: orthology inference among 1000 complete genomes. *Nucleic Acids Res.* **39**(Database issue):D289-294.

[36] Storm, C.E. and Sonnhammer, E.L. (2002) Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics* **18**(1):92-99.

[37] Dufayard, J.F., Duret, L., Penel, S., Gouy, M., Rechenmann, F. and Perriere, G. (2005) Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics* **21**(11):2596-2603.

[38] Huerta-Cepas, J., Bueno, A., Dopazo, J. and Gabaldon, T. (2008) PhylomeDB: a database for genome-wide collections of gene phylogenies. *Nucleic Acids Res.* **36**(Database issue):D491-496.

[39] Gray, G.S. and Fitch, W.M. (1983) Evolution of antibiotic resistance genes: the DNA sequence of a kanamycin resistance gene from Staphylococcus aureus. *Mol. Biol. Evol.* **1**(1):57-66.

[40] Storm, C.E. and Sonnhammer, E.L. (2003) Comprehensive analysis of orthologous protein domains using the HOPS database. *Genome Res.* **13**(10):2353-2362.

[41] Tatusov, R.L., Galperin, M.Y., Natale, D.A. and Koonin, E.V. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28**(1):33-36.

[42] Haft, D.H., Selengut, J.D., Richter, R.A., Harkins, D., Basu, M.K. and Beck, E. (2013) TIGRFAMs and Genome Properties in 2013. *Nucleic Acids Res.* **41**(Database issue):D387-395.

[43] Wang, L. and Jiang, T. (1994) On the complexity of multiple sequence alignment. *J. Comput. Biol.* **1**(4):337-348.

[44] Thompson, J.D., Plewniak, F. and Poch, O. (1999) A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res.* **27**(13):2682-2690.

[45] Powell, S., Szklarczyk, D., Trachana, K., Roth, A., Kuhn, M., Muller, J., Arnold, R., Rattei, T., Letunic, I., Doerks, T. *et al.* (2012) eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res.* **40**(Database issue):D284-289.

[46] Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J. *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**:539.

[47] Price, M.N., Dehal, P.S. and Arkin, A.P. (2010) FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**(3):e9490.

[48] Eddy, S.R. (2009) A new generation of homology search tools based on probabilistic inference. *Genome Inform.* **23**(1):205-211.

[49] Unger, R., Uliel, S. and Havlin, S. (2003) Scaling law in sizes of protein sequence families: from super-families to orphan genes. *Proteins* **51**(4):569-576.

[50] Uchiyama, I., Higuchi, T. and Kawai, M. (2010) MBGD update 2010: toward a comprehensive resource for exploring microbial genome diversity. *Nucleic Acids Res.* **38**(Database issue):D361-365.

[51] Tocchini-Valentini, G.D., Fruscoloni, P. and Tocchini-Valentini, G.P. (2005) Structure, function, and evolution of the tRNA endonucleases of Archaea: an example of subfunctionalization. *Proc. Natl. Acad. Sci. U. S. A.* **102**(25):8933-8938.

[52] Dessimoz, C., Boeckmann, B., Roth, A.C. and Gonnet, G.H. (2006) Detecting non-orthology in the COGs database and other approaches grouping orthologs using genome-specific best hits. *Nucleic Acids Res.* **34**(11):3309-3316.

[53] Koonin, E.V., Aravind, L. and Kondrashov, A.S. (2000) The impact of comparative genomics on our understanding of evolution. *Cell* **101**(6):573-576.

[54] Uchiyama, I., Mihara, M., Nishide, H. and Chiba, H. (2015) MBGD update 2015: microbial genome database for flexible ortholog analysis utilizing a diverse set of genomic data. *Nucleic Acids Res.* **43**(Database issue):D270-276.

[55] Gennari, J.H., Musen, M.A., Fergerson, R.W., Grosso, W.E., Crubezy, M., Eriksson, H., Noy, N.F. and Tu, S.W. (2003) The evolution of Protege: an environment for knowledge-based systems development. *Int. J. Hum. Comput. Stud.* **58**(1):89-123.

[56] RDF 1.1 Turtle. [http://www.w3.org/TR/turtle]

[57] Schmitt, T., Messina, D.N., Schreiber, F. and Sonnhammer, E.L. (2011) Letter to the editor: SeqXML and OrthoXML: standards for sequence and orthology information. *Brief. Bioinform.* **12**(5):485-488.

[58] Erling, O. and Mikhailov, I. (2007) RDF Support in the Virtuoso DBMS. *Proceedings of the 1st Conference on Social Semantic Web (CSSW)*:59-68.

[59] Sonnhammer, E.L., Gabaldon, T., Sousa da Silva, A.W., Martin, M., Robinson-Rechavi, M., Boeckmann, B., Thomas, P.D., Dessimoz, C. and Quest for Orthologs, c. (2014) Big data and other challenges in the quest for orthologs. *Bioinformatics* **30**(21):2993-2998.

[60] Minarro-Gimenez, J.A., Madrid, M. and Fernandez-Breis, J.T. (2009) OGO: an ontological approach for integrating knowledge about orthology. *BMC Bioinformatics* **10**(Suppl 10):S13.

[61] van der Heijden, R.T., Snel, B., van Noort, V. and Huynen, M.A. (2007) Orthology prediction at scalable resolution by phylogenetic tree analysis. *BMC Bioinformatics* **8**:83.

[62] Eilbeck, K., Lewis, S.E., Mungall, C.J., Yandell, M., Stein, L., Durbin, R. and Ashburner, M. (2005) The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol.* **6**(5):R44.

[63] Dumontier, M., Baker, C.J., Baran, J., Callahan, A., Chepelev, L., Cruz-Toledo, J., Del Rio, N.R., Duck, G., Furlong, L.I., Keath, N. *et al.* (2014) The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery. *J. Biomed. Semantics* **5**(1):14.

[64] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**(1):25-29.

[65] Dimmer, E.C., Huntley, R.P., Alam-Faruque, Y., Sawford, T., O'Donovan, C., Martin, M.J., Bely, B., Browne, P., Mun Chan, W., Eberhardt, R. *et al.* (2012) The UniProt-GO Annotation database in 2011. *Nucleic Acids Res.* **40**(Database issue):D565-570.

[66] Terashima, H., Kojima, S. and Homma, M. (2008) Flagellar motility in bacteria structure and function of flagellar motor. *Int. Rev. Cell Mol. Biol.* **270**:39-85.

[67] Kim, J.F. (2001) Revisiting the chlamydial type III protein secretion system: clues to the origin of type III protein secretion. *Trends Genet.* **17**(2):65-69.

[68] Abby, S.S. and Rocha, E.P. (2012) The non-flagellar type III secretion system evolved from the bacterial flagellum and diversified into host-cell adapted systems. *PLoS Genet.* **8**(9):e1002983.

[69] RDF 1.1 N-Triples. [http://www.w3.org/TR/n-triples/]

[70] Wagner, A. (2000) Decoupled evolution of coding region and mRNA expression patterns after gene duplication: implications for the neutralist-selectionist debate. *Proc. Natl. Acad. Sci. U. S. A.* **97**(12):6579-6584.

[71] Gu, Z., Nicolae, D., Lu, H.H. and Li, W.H. (2002) Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends Genet.* **18**(12):609-613.

[72] Makova, K.D. and Li, W.H. (2003) Divergence in the spatial pattern of gene expression between human duplicate genes. *Genome Res.* **13**(7):1638-1645.

[73] Jordan, I.K., Marino-Ramirez, L., Wolf, Y.I. and Koonin, E.V. (2004) Conservation and coevolution in the scale-free human gene coexpression network. *Mol. Biol. Evol.* **21**(11):2058-2070.

[74] Nuzhdin, S.V., Wayne, M.L., Harmon, K.L. and McIntyre, L.M. (2004) Common pattern of evolution of gene expression level and protein sequence in Drosophila. *Mol. Biol. Evol.* **21**(7):1308-1317.

[75] Khaitovich, P., Hellmann, I., Enard, W., Nowick, K., Leinweber, M., Franz, H., Weiss, G., Lachmann, M. and Paabo, S. (2005) Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science* **309**(5742):1850-1854.

[76] Castillo-Davis, C.I., Hartl, D.L. and Achaz, G. (2004) cis-Regulatory and protein evolution in orthologous and duplicate genes. *Genome Res.* **14**(8):1530-1536.

[77] Suzuki, Y. and Sugano, S. (2003) Construction of a full-length enriched and a 5'-end enriched cDNA library using the oligo-capping method. *Methods Mol. Biol.* **221**:73-91.

[78] Ota, T., Suzuki, Y., Nishikawa, T., Otsuki, T., Sugiyama, T., Irie, R., Wakamatsu, A., Hayashi, K., Sato, H., Nagai, K. *et al.* (2004) Complete sequencing and characterization of 21,243 full-length human cDNAs. *Nat. Genet.* **36**(1):40-45.

[79] Carninci, P. and Hayashizaki, Y. (1999) High-efficiency full-length cDNA cloning. *Methods Enzymol.* **303**:19-44.

[80] Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H. *et al.* (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**(6915):563-573.

[81] Suzuki, Y., Yamashita, R., Nakai, K. and Sugano, S. (2002) DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs. *Nucleic Acids Res.* **30**(1):328-331.

[82] Suzuki, Y., Yamashita, R., Sugano, S. and Nakai, K. (2004) DBTSS, DataBase of Transcriptional Start Sites: progress report 2004. *Nucleic Acids Res.* **32**(Database issue):D78-81.

[83] Yamashita, R., Suzuki, Y., Wakaguri, H., Tsuritani, K., Nakai, K. and Sugano, S. (2006) DBTSS: DataBase of Human Transcription Start Sites, progress report 2006. *Nucleic Acids Res.* **34**(Database issue):D86-89.

[84] Yamashita, R., Suzuki, Y., Sugano, S. and Nakai, K. (2005) Genome-wide analysis reveals strong correlation between CpG islands with nearby transcription start sites of genes and their tissue specificity. *Gene* **350**(2):129-136.

[85] Sun, H., Palaniswamy, S.K., Pohar, T.T., Jin, V.X., Huang, T.H. and Davuluri, R.V. (2006) MPromDb: an integrated resource for annotation and visualization of mammalian gene promoters and ChIP-chip experimental data. *Nucleic Acids Res.* **34**(Database issue):D98-103.

[86] Jin, V.X., Singer, G.A., Agosto-Perez, F.J., Liyanarachchi, S. and Davuluri, R.V. (2006) Genome-wide analysis of core promoter elements from conserved human and mouse orthologous pairs. *BMC Bioinformatics* **7**:114.

[87] Suzuki, Y., Yamashita, R., Shirota, M., Sakakibara, Y., Chiba, J., Mizushima-Sugano, J., Nakai, K. and Sugano, S. (2004) Sequence comparison of human and mouse genes reveals a homologous block structure in the promoter regions. *Genome Res.* **14**(9):1711-1718.

[88] Palaniswamy, S.K., Jin, V.X., Sun, H. and Davuluri, R.V. (2005) OMGProm: a database of orthologous mammalian gene promoters. *Bioinformatics* **21**(6):835-836.

[89] Maglott, D., Ostell, J., Pruitt, K.D. and Tatusova, T. (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* **33**(Database issue):D54-58.

[90] Rice, P., Longden, I. and Bleasby, A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**(6):276-277.

[91] Pruitt, K.D., Tatusova, T. and Maglott, D.R. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **33**(Database issue):D501-504.

[92] Wall, D.P., Hirsh, A.E., Fraser, H.B., Kumm, J., Giaever, G., Eisen, M.B. and Feldman, M.W. (2005) Functional genomic analysis of the rates of protein evolution. *Proc. Natl. Acad. Sci. U. S. A.* **102**(15):5483-5488.

[93] Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32**(Database issue):D258-261.

[94] Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. Ser. B. (Stat. Method.)* **57**(1):289-300.

[95] Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S. and Haussler, D. (2004) Ultraconserved elements in the human genome. *Science* **304**(5675):1321-1325.

[96] Obrdlik, A., Kukalev, A. and Percipalle, P. (2007) The function of actin in gene transcription. *Histol. Histopathol.* **22**(9):1051-1055.

[97] Pietromonaco, S.F., Hessler, R.A. and O'Brien, T.W. (1986) Evolution of proteins in mammalian cytoplasmic and mitochondrial ribosomes. *J. Mol. Evol.* **24**(1-2):110-117.

[98] Landry, J.R., Mager, D.L. and Wilhelm, B.T. (2003) Complex controls: the role of alternative promoters in mammalian genomes. *Trends Genet.* **19**(11):640-648.

[99] Kimura, K., Wakamatsu, A., Suzuki, Y., Ota, T., Nishikawa, T., Yamashita, R., Yamamoto, J., Sekine, M., Tsuritani, K., Wakaguri, H. *et al.* (2006) Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes. *Genome Res.* **16**(1):55-65.

[100] Tsuritani, K., Irie, T., Yamashita, R., Sakakibara, Y., Wakaguri, H., Kanai, A., Mizushima-Sugano, J., Sugano, S., Nakai, K. and Suzuki, Y. (2007) Distinct class of putative "non-conserved" promoters in humans: Comparative studies of alternative promoters of human and mouse genes. *Genome Res.* **17**(7):1005-1014.

[101] Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C. *et al.* (2005) The transcriptional landscape of the mammalian genome. *Science* **309**(5740):1559-1563.

[102] Frith, M.C., Ponjavic, J., Fredman, D., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y. and Sandelin, A. (2006) Evolutionary turnover of mammalian transcription start sites. *Genome Res.* **16**(6):713-722.

[103] Farre, D., Bellora, N., Mularoni, L., Messeguer, X. and Alba, M.M. (2007) Housekeeping genes tend to show reduced upstream sequence conservation. *Genome Biol.* **8**(7):R140.

[104] Perry, R.P. (2005) The architecture of mammalian ribosomal protein promoters. *BMC Evol. Biol.* **5**(1):15.

[105] Yoshihama, M., Uechi, T., Asakawa, S., Kawasaki, K., Kato, S., Higa, S., Maeda, N., Minoshima, S., Tanaka, T., Shimizu, N. *et al.* (2002) The human ribosomal protein genes: sequencing and comparative analysis of 73 genes. *Genome Res.* **12**(3):379-390.

[106] Sandelin, A., Bailey, P., Bruce, S., Engstrom, P.G., Klos, J.M., Wasserman, W.W., Ericson, J. and Lenhard, B. (2004) Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics* **5**(1):99.

[107] Woolfe, A., Goodson, M., Goode, D.K., Snell, P., McEwen, G.K., Vavouri, T., Smith, S.F., North, P., Callaway, H., Kelly, K. *et al.* (2005) Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.* **3**(1):e7.

[108] Chin, C.S., Chuang, J.H. and Li, H. (2005) Genome-wide regulatory complexity in yeast promoters: separation of functionally conserved and neutral sequence. *Genome Res.* **15**(2):205-213.

[109] Pal, C., Papp, B. and Lercher, M.J. (2006) An integrated view of protein evolution. *Nat. Rev. Genet.* **7**(5):337-348.