COXPRESdb: a database to compare gene coexpression in seven model animals

Takeshi Obayashi and Kengo Kinoshita*

Graduate School of Information Science, Tohoku University, 6-3-09, Aramaki-Aza-Aoba, Aoba-ku, Sendai 980-8679, Japan

Received September 15, 2010; Revised October 23, 2010; Accepted October 25, 2010

ABSTRACT

Publicly available databases of coexpressed gene sets are a valuable resource for a wide variety of experimental studies, including gene targeting for functional identification, and for investigations of regulatory mechanisms or protein–protein interaction networks. Although coexpressed gene databases are becoming more and more popular in the field of plant biology, those with animal data are rather limited, possibly due to the lower reliability of the coexpression data. The original COXPRESdb (coexpressed gene database) (http://coxpresdb.jp) represented the coexpression relationship for human and mouse. Here, we report updates of this database that especially focus on the enhancement of the reliability of gene coexpression data in animals. For this purpose, we implemented a new comparable coexpression measure, Mutual Rank, included five other animal species, rat, chicken, zebrafish, fly and nematoda, to assess the conservation of coexpression, and added different layers of omics data into the integrated network of genes. Comparison of coexpression is a key concept to enhance the reliability of gene coexpression, and the integration of different information can reduce the noise inherent in the information. With the functions for gene network representation, COXPRESdb can help researchers to clarify the functional and regulatory networks of genes in a broad array of animal species.

INTRODUCTION

Genes involved in related biological pathways are usually expressed cooperatively for their functions, and thus information on their coexpression is a key to understand the biological systems at the molecular level (1). Coexpression data have been utilized in a wide variety of experimental designs, such as gene targeting, regulatory investigations and/or identification of potential partners in protein–protein interactions (PPIs) (2). A reliable estimation of coexpressed gene relationships requires large amounts of gene expression data from DNA microarray experiments, which are now available in various public repositories (3,4) for several species. Using these large public data sources, a number of coexpression databases have been constructed, and are widely used. However, most of the databases have been constructed for plant researchers (2), and thus the use of coexpression data in animal research is rather limited.

One possible reason for the limited use of coexpression data in animals is that such coexpression data showed lower performance in identifying the functionally related genes for mammals than for Arabidopsis (5). Among the many possible reasons for this, one reason is the lower variety of microarray samples in animals. Although more data are usually available for animals than for plants, the expression data for animals are quite biased, due to the inclusion of medical samples, such as cancer cells. On the contrary, in the field of Arabidopsis, systematic collections were promoted by the AtGenExpress international collaboration (6–8), and a much broader variety of samples were obtained. The tissue organization and the regulatory mechanisms in animals are far more complicated than those in plants, and thus the number of samples required for the estimation of reliable coexpression will be much larger. In addition, the higher frequency of alternative splicing in mammals hinders a precise evaluation of the strength of gene coexpression.

Our coexpression database, COXPRESdb, was originally constructed in 2007 to use coexpression data for human and mouse, where network representations of gene coexpression data were provided to show the relationships among coexpression modules, in addition to the gene-to-gene relationships (9). In this article, we describe new features of the COXPRESdb. The main improvement is the enhancement of the relevance of coexpression with the limited variety of expression data available in public databases for animals. For this purpose, the evolutionary
conservation was analyzed, and different data were integrated. The former method is one of the most powerful approaches to evaluate the relevance of gene coexpression (10,11), since if similar coexpression relations are observed for the orthologous genes in different species, then the possibility of experimental/technical artifacts is considered to be reduced. To implement this idea, we first developed a new coexpression measure to directly compare coexpression strengths in different species. Then, five other species, rat, chicken, zebrafish, fly and nematoda, were added. These coexpression data are provided in a comparative gene list representation.

In addition, the protein–protein interaction (PPI) and KEGG pathways in the coexpressed gene networks were added to the interaction networks, because integrating different layers of information is considered to enhance the reliability of the regulatory relationship. Details of the new features are described in the following sections, along with examples of the comparative coexpressed gene list (Figures 2 and 3) and the integrated gene network (Figure 4). The history of COXPRESdb, as well as other miscellaneous updates, is summarized in Table 1.

### COMPARATIVE COEXPRESSED GENE LIST

#### Gene coexpression for different species

A comparison of the coexpression data for different species is one of the approaches to enhance the relevance of the gene coexpression data. Figure 1 shows a schematic diagram to evaluate the reliabilities of the coexpression of a pair of genes and the expression data of a guide gene. (A) To investigate the reliability of the coexpression of a pair of genes in a species, a comparison with the coexpression of a pair of orthologous genes in other species can be used. If the coexpression relationship is also observed in other species, then the coexpression in the original species is reliable. (B) To investigate the reliability of the expression pattern of a guide gene, coexpressed gene lists are compared. If most of the coexpression in the coexpressed gene list is observed in other species, then the expression pattern of the guide gene in the original species is reliable. This reliability of the gene expression pattern can be used to investigate that of coexpression, because the coexpression between a pair of genes with a reliable expression pattern is credible.

### Table 1. Evolution of COXPRESdb versions over the last 3 years

<table>
<thead>
<tr>
<th>Version</th>
<th>Date</th>
<th>Human</th>
<th>Mouse</th>
<th>Rat</th>
<th>Chicken</th>
<th>Zebrafish</th>
<th>Fly</th>
<th>Nematoda</th>
<th>Other annotations</th>
</tr>
</thead>
<tbody>
<tr>
<td>ver.4.1</td>
<td>2010-09-12</td>
<td>†</td>
<td>†</td>
<td>†</td>
<td>†</td>
<td>†</td>
<td>†</td>
<td>†</td>
<td>Updated</td>
</tr>
<tr>
<td>ver.4.0</td>
<td>2009-08-25</td>
<td>†</td>
<td>†</td>
<td></td>
<td>c2.0</td>
<td>c1.0</td>
<td>c1.0</td>
<td>c1.0</td>
<td>Updated</td>
</tr>
<tr>
<td>ver.3.3</td>
<td>2008-11-12</td>
<td>†</td>
<td>†</td>
<td>†</td>
<td>†</td>
<td></td>
<td></td>
<td></td>
<td>Updated</td>
</tr>
<tr>
<td>ver.3.2</td>
<td>2008-05-29</td>
<td>†</td>
<td></td>
<td>†</td>
<td>†</td>
<td></td>
<td></td>
<td></td>
<td>Updated</td>
</tr>
<tr>
<td>ver.3.1</td>
<td>2008-08-04</td>
<td>c3.1</td>
<td>c2.1</td>
<td>c1.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Updated</td>
</tr>
<tr>
<td>ver.3.0</td>
<td>2008-03-17</td>
<td>c3.0</td>
<td>†</td>
<td></td>
<td>c1.0</td>
<td></td>
<td></td>
<td></td>
<td>Updated</td>
</tr>
<tr>
<td>ver.2.0</td>
<td>2007-06-22</td>
<td>c2.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Updated</td>
</tr>
</tbody>
</table>

†, same condition as previous version. Details of the updates of ‘Other annotations’ are available at: http://coxpresdb.jp/versions.shtml. Date are written as year-month-day format.

![Figure 1](https://academic.oup.com/nar/article-abstract/39/suppl_1/D1016/2507295)
diagram of a comparison of the gene coexpression to evaluate the relevance of the data. When coexpression is observed in a pair of genes in a single species, the reliability of the coexpression may be weak, due to experimental and technical artifacts, such as cross hybridization of probes, secondary-structure formation of the probes to prevent hybridization with the target RNA, and/or inappropriate data treatment. However, if we also find coexpressed gene pairs in orthologs at the same time, then the relevance of the coexpression becomes much higher, because it is less likely that invalid expression data have been obtained multiple times for the related genes. In addition, gene coexpression found in multiple species will be more biologically relevant, because important regulatory mechanisms are conserved during evolution. Actual examples of conserved coexpression are shown in the following section.

Several reports have addressed the effectiveness of conserved coexpression to identify functionally related genes (10,11). These studies focused on orthologous gene clusters represented by the metagene or KOG gene, rather than the gene itself, and coexpression between any pair of gene clusters was calculated using order statistics (10) or ‘between-species averaging’ (11). These studies clearly showed improved performance in predicting gene functions. However, any gene has the possibility of becoming a key to investigate a phenomenon of interest, regardless of the existence of orthologs. Therefore, instead of making orthologous gene clusters, we simply added the coexpression data of other species to the coexpressed gene list of interest, to confirm the coexpression relationships of a pair of genes and the expression data of a guide gene.

The original version of COXPRESdb already contained coexpression data for two species: human and mouse (9). To enhance this approach, we added five more model species in the current version: rat, chicken, zebrafish, fly and nematoda, with the same calculation protocols as those used for human and mouse (9). The details for the raw data are summarized in Table 2.

### Measure of gene coexpression

To evaluate the strength of coexpression, the Pearson and Spearman correlation coefficients are widely used, but we found that these measures were not suitable for direct comparisons among different species and that the correlation rank-based measure, mutual rank (MR, which is calculated as the geometric mean of the correlation rank of gene A to gene B and of gene B to gene A), is more suitable to compare the coexpression data in multiple species on average (5). In addition, we have performed several successful case studies using MR coexpression data to identify new gene functions in the applications for Arabidopsis (12). Therefore, we adopted MR as the coexpression measure in the COXPRESdb, to compare the coexpression strengths among multiple species.

To further investigate the relevance of gene coexpression between two genes of interest, we also provided a function to check the ‘stability of coexpression’, which is the degree of change in the coexpression when major microarray samples are subtracted (13), by using the ‘coexpression viewer’ tool in the ‘Draw box’. More stable coexpression suggests stronger functional relationships between the genes.

### An example of a comparative coexpressed gene table

The table on the left in Figure 2 shows the human coexpressed gene list from SLC39A7. The top line is the guide gene itself, and the following lines indicate the strongly coexpressed genes ordered by their coexpression strength, as measured by MR (smaller MR indicates stronger coexpression), while the five columns on the right indicate the gene coexpression strengths in other species. Note that the chicken ortholog was not available for this guide gene, and thus it was omitted from this table. On the second line, the mouse coexpression column (Mmu MR column in the table) shows the coexpression between the mouse orthologs to human SLC39A7 and that to COPG. Its coexpression strength is 38.7, which is strong enough, because the MR value distribution is 1 for the total number of genes in the species for the coexpression analysis (about 20 000 for human, mouse and rat). We previously showed that MR values less than 5000 represent significant coexpression (12), but in view of this observation, weak coexpression was judged by an MR >1000 and is shown in a faint color. The coexpressed genes from human SLC39A7 are fairly well conserved in other species, indicating that most of the coexpression in this list as well as the expression pattern of SLC39A7 are reliable.

Note that in some cases, such as the 4th line for the PP1B gene in Caenorhabditis elegans (Cel MR column), there are multiple orthologous genes in a species, and the coexpression values are shown in parallel in a single cell (8.9 and 17.3).

### Table 2. Coexpression data information in COXPRESdb

<table>
<thead>
<tr>
<th>Species</th>
<th>Abbreviation</th>
<th>Version</th>
<th>Number of genes</th>
<th>Number of experiments</th>
<th>Number of microarray slides</th>
<th>Microarray platform</th>
<th>Last update</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>Hsa</td>
<td>c3.1</td>
<td>19 777</td>
<td>123</td>
<td>4401</td>
<td>GPL570</td>
<td>2008-04-08</td>
</tr>
<tr>
<td>Mouse</td>
<td>Mmu</td>
<td>c2.1</td>
<td>21 036</td>
<td>154</td>
<td>2226</td>
<td>GPL1261</td>
<td>2009-04-08</td>
</tr>
<tr>
<td>Rat</td>
<td>Rno</td>
<td>c2.0</td>
<td>13 751</td>
<td>164</td>
<td>3526</td>
<td>GPL1355</td>
<td>2009-08-25</td>
</tr>
<tr>
<td>Chicken</td>
<td>Gga</td>
<td>c1.0</td>
<td>13 757</td>
<td>22</td>
<td>352</td>
<td>GPL3213</td>
<td>2009-08-25</td>
</tr>
<tr>
<td>Zebrafish</td>
<td>Dre</td>
<td>c1.0</td>
<td>10 112</td>
<td>41</td>
<td>590</td>
<td>GPL1319</td>
<td>2009-08-25</td>
</tr>
<tr>
<td>Fly</td>
<td>Dme</td>
<td>c1.0</td>
<td>12 626</td>
<td>52</td>
<td>1022</td>
<td>GPL1322</td>
<td>2009-08-25</td>
</tr>
<tr>
<td>Nematoda</td>
<td>Cel</td>
<td>c1.0</td>
<td>17 256</td>
<td>35</td>
<td>514</td>
<td>GPL200</td>
<td>2009-08-25</td>
</tr>
</tbody>
</table>

Species abbreviations in COXPRESdb are used in accordance with those in GenomeNet (http://www.genome.jp). Date are written as year-month-day format.
In this comparative gene list, two genes were found to be unsupported by other species. One is the fifth strongest coexpressed gene, YIPF3; and the other is the 24th strongest coexpressed gene, YIPF2, which are marked with red circles in Figure 2. There are two possible explanations for the unsupported coexpression: (i) these coexpressions with the guide gene are human specific, and (ii) the expression patterns of these two genes are affected by some technical problems. To assess these two possibilities, coexpressed gene lists from these two genes can be used. As shown in the center table of Figure 2, the coexpressed genes from YIPF3 were mostly conserved, while those from YIPF2 were not supported by mouse and rat (right table of Figure 2), indicating that the expression pattern of YIPF2 was less reliable than that of YIPF3, and suggesting that the coexpression between SLC38A7 and YIPF3 in the table on the left is human specific.

In addition to evaluating the reliabilities of coexpression between a pair of genes and the expression pattern of the guide gene, a coexpression comparison can be used to identify ‘functional orthologs’ from traditional sequence orthologs. Figure 3 shows a comparative coexpressed gene list from a mouse guide gene. For this mouse guide gene, there are two human orthologs. Both of the human genes show strong coexpression with the orthologous genes of the coexpressed genes in mouse. However, for almost all of the coexpressed gene relationships, the gene on the left side (PRSS1) shows stronger coexpression, suggesting that PRSS1 is more likely to be a functional ortholog, rather than PRSS3, which might have acquired some other cellular functions after gene duplication.

INTEGRATING DIFFERENT INFORMATION INTO THE GENE COEXPRESSION NETWORK

In addition to the conserved coexpression, two types of omics data were mapped to the gene coexpression network: the known pathways and the PPIs. The known pathway information on gene coexpression networks is shown on the nodes, which are marked with the KEGG pathway annotation (14) (Figure 4). For each gene network, up to five KEGG pathways with larger numbers of genes within the pathway are selected. For example, the gene network for human CD3D has eight genes for ‘T cell receptor signaling pathway’ (red circles in Figure 4). The same marks in the table, placed just
in other species, we also used IntAct data for PPI in the current version of COXPRESdb. Number of networks with PPI and with KEGG annotation is summarized in Table 3.

In addition to the pre-calculated gene networks for each gene, each GO functional group and each tissue, we also provide a ‘NetworkDrawer’ tool, to draw the gene network for the user-defined set of genes. This tool detects genes displaying coexpression and/or PPI with the query genes, according to the user-defined criteria. The result networks are depicted in the Pajak (15) and Cytoscape (16) formats, in addition to the default PNG figure.

**OTHER IMPROVEMENTS**

**New top page design**

The top page of COXPRESdb was revised to intuitively show the four major functions of our database. COXPRESdb has eleven tools, in addition to the pre-calculated pages. To easily access these tools, we classified them into two categories, ‘Search box’ and ‘Draw box’. The Search box is composed of tools to search genes by various aspects, such as coexpression and PPI, and it outputs a gene table. On the other hand, the tools in Draw box can draw various pictures, such as a gene network, a hierarchical gene cluster and a detailed view of gene coexpression. There are two additional boxes, ‘Browse’ and ‘Bulk download’. The Browse box has a link to the gene networks for 49 tissues. These gene networks are huge, and therefore, these networks are provided as a Google Map interface. The Bulk download box is used for downloading coexpression data and protein subcellular localization predicted by WoLF PSORT (17). These data are now available under the Creative Commons Attribution license.

A simple gene search is available at the search window just below the scenic picture. The user can search genes and GO terms using keywords, gene aliases, Entrez gene IDs or GO IDs. The default searches for all of them. Users can access the coexpressed gene list and network for any single guide gene in human, mouse and rat.

**Version numbering**

COXPRESdb is composed of our own coexpression data and integrated public annotations. The coexpression data are updated yearly, and the public annotations are updated every few months (Table 1). Each data update is assigned a new version number, to prevent confusion about the data update and as a reference for users’ publications. The major version number essentially corresponds to the coexpression data update, whereas the minor version number corresponds to the public data update. Version histories can be checked at: http://coxpresdb.jp/versions.shtml. Note that we maintain previous versions of the gene coexpression data, but we do not store any other annotations.
ACKNOWLEDGEMENTS

The super-computing resource was provided by Human Genome Center, Institute of Medical Science, The University of Tokyo.

FUNDING

Grant-in-Aid for Scientific Research (No. 50397048, to T.O.); Grant-in-Aid for Innovative Areas “HD physiology” (No. 22136005, to K.K.). Funding for open access charge: MEXT of Japan.

Conflict of interest statement. None declared.

REFERENCES


