**ORI-GENE: gene classification based on the evolutionary tree**

Hideaki Mizuno\(^1,\,3,\,4,\,*\), Yoshimasa Tanaka\(^1\), Kenta Nakai\(^2\) and Akinori Sarai\(^3,\,*\)

\(^1\)Institute of Biological Sciences, University of Tsukuba, 1-1-1 Ten-nodai, Tsukuba, Ibaraki 305-8572, \(^2\)Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639 and \(^3\)Tsukuba Life Science Center, The Institute of Physical & Chemical Research (RIKEN), 3-1-1 Koyadai, Tsukuba, Ibaraki 305-0074, Japan

Received on January 19, 2000; revised on August 19, 2000; accepted on October 2, 2000

**ABSTRACT**

**Motivation:** Genome projects have produced large amounts of data on the sequences of new genes whose functions are as yet unknown. The functions of new genes are usually inferred by comparing their sequences with those of known genes, but evaluation of the sequence homology of individual genes does not make the most of the available sequence information. Therefore, new methods and tools for extracting more biological information from homology searches would be advantageous.

**Results:** We have developed a computational tool, ORI-GENE, to analyze the results of sequence homology searches from the perspective of the evolution of selected sets of new genes. ORI-GENE has a graphical interface and accomplishes two important tasks: first, based on the output of homology searches, it identifies species with similar genes and displays their pattern of distribution on the phylogenetic tree. This function enables one to infer the way in which a given gene may have propagated among species over time. Second, from the distribution patterns, it predicts the point at which a given gene may have been first acquired (i.e. its ‘origin’), then classifies the gene on that basis. Because it makes use of available evolutionary information to show the way in which genes cluster among species, ORI-GENE should be an effective tool for the screening and classification of new genes revealed by genome analysis.

**Availability:** ORI-GENE is retrievable via the Internet at: http://www.rtc.riken.go.jp/jouhou/ORI-GENE.

**Contact:** hideaki.mizuno@roche.com; sarai@rtc.riken.go.jp

\(*\)To whom correspondence should be addressed.

\(^4\)Present address: Nippon Roche Research Center, 200 Kajiwara, Kamakura, Kanagawa 249-8530, Japan.

**INTRODUCTION**

The first complete genome of a bacterium, *Haemophilus influenzae*, was published in 1995 (Fleischmann et al., 1995). Since then, the genomes of 30 other bacteria, 8 archaea, and the unicellular eukaryote *Saccharomyces cerevisiae* (Goffeau et al., 1996) have been sequenced completely. Also, the multicellular eukaryote *Caenorhabditis elegans* (C.elegans Sequencing Consortium, 1998) and *Drosophila melanogaster* (Adams et al., 2000) have been sequenced to near completion. Currently, ongoing projects are endeavoring to sequence the genomes of diverse organisms, ranging from bacteria to *Homo sapiens*. As a consequence of this nucleotide sequence analysis, a large number of genes have been discovered whose functions remain unknown, and various computational methods and tools have proven essential for their classification and annotation, and for predicting the functions of the newly discovered genes.

Typically, gene sequencing and identification are followed by a homology search using programs such as BLAST (Altschul et al., 1990) and FASTA (Pearson and Lipman, 1988). These analyses are usually based on the assumption that genes with similar sequences have similar functions, and are mostly used to find genes with similarities to others whose functions have already been experimentally shown. With this approach, however, queries showing no similarity to known genes yield no useful information. Consequently, analysis relying on a similarity to known genes alone does not take maximum advantage of the information made available by homology searches. For instance, similarities among unknown genes provide information about the presence of those genes in different species, whereas the absence of similarity raises the possibility that the gene in question is unique to a particular species. This ‘*in silico*’ hybridization would enable one to detect the presence of selected genes among
species and to determine the gene’s pattern of distribution on the phylogenetic tree. Unfortunately, because there was generally an insufficient amount of available sequence data, such analyses had been rarely carried out. Now, although large amounts of sequence data are available from the complete/partial genome sequences of many diverse organisms and from millions of ESTs resulting from myriad cDNA sequencing projects, there is still the problem that there are few computational tools available for in silico hybridization. Some programs such as BLATAx (Koonin et al., 1996) and SEALS (Walker and Koonin, 1997), which can filter BLAST output according to user-specified taxa, may serve for this purpose. However, these tools use rather complicated command-line operation and some skill is required for operating them and interpreting their text-based output. Thus, we have developed a more user-friendly bioinformatic tool with graphical user interface, ORI-GENE, which accomplishes two key tasks: (1) it displays gene distribution patterns on the phylogenetic tree; and (2) it classifies genes based on evolutionary hypotheses. Here, we describe how this tool can be used to classify genes based on their likely evolution.

SYSTEMS AND METHODS

Overview

ORI-GENE is a tool with which to analyze the distributions of genes among species on the basis of the output of homology searches. It enables investigators to visualize gene distribution patterns on the phylogenetic tree, which is useful in the context of two scenarios. The first is when a gene of interest is distributed only in a particular clade, making it likely that unique characteristics of the clade are dictated by the gene. From that perspective, detection of the gene’s presence in a species might enable inference of the gene’s specificity. The second scenario is when transmission of a gene usually occurs from an ancestral line to a descendant. In that case, if the gene distribution pattern were projected on the phylogenetic tree, we should be able to readily infer when the gene was acquired—i.e. its ‘origin’—and how it was propagated. To display patterns of gene distribution, ORI-GENE performs the following (Figure 1a):

(1) it scans each hit in the homology search output;
(2) it identifies the source species of hit genes from their accession numbers;
(3) it projects the identified species on the phylogenetic tree.

The second function of ORI-GENE is classification of genes based on evolutionary hypotheses. When multiple genes are given, ORI-GENE automatically predicts their origin from the distribution pattern, and classifies them based on their origin. This classification offers insight into important evolutionary events, e.g. transformation from unicellular to multicellular organisms and from invertebrate to vertebrate. To classify genes through identification of their origin, ORI-GENE performs the following (Figure 1b):

(4) it selects species having sequence similarity above a threshold in the distribution pattern;
(5) it traces the pattern back to the first branching point (origin) on the phylogenetic tree;
(6) it classifies the gene based on its origin.

Phylogenetic tree of species

The phylogenetic tree on which identified species are projected was retrieved from the NCBI taxonomy database (Leipe and Soussov, 1995). This taxonomy represents our current knowledge on the evolution, and contains the names of all the organisms represented in the genetic databases (56 000 species, as of January 1999), and describes their relationship by hierarchical classification (e.g. kingdom, phylum, class, etc.). From the data, categories whose taxonomic position are unclear—i.e. ‘Viroids’, ‘Viruses’, ‘Other’ (artificial sequences), ‘Unclassified’, ‘environmental samples’ and ‘Unidentified’—were excluded because they cause confusion during the projection. The tree is based on this taxonomy source. Thus each branching point in the tree (Figures 1-4) represents a taxonomic rank and each branch length is proportional to the number of ranks between the branching points.

Homology search

Homology searches were performed using BLAST2 algorithm (Altschul et al., 1997), which allowed gapped alignment without filtering. TBLASTN program was used to search the GenBank database with BLOSUM62 matrix. To detect the presence of a particular gene among species, all hits were output. To be certain that sufficient output was obtained, option -v, which defines the number of one-line sequence descriptions, was set to 5000. The threshold above which the hit genes are considered for the following analyses can be controlled manually (see Section Implementation below).

Implementation

ORI-GENE is a stand-alone tool written in the C programming language for the Macintosh computer. It requires 50 MB of disk space for installation and 2 MB of memory for execution.

ORI-GENE employs the mouse-based graphical user interfaces shown in Figure 2. The ‘Tree landscape viewer’ (Figure 2a) displays an overview of the phylogenetic
Fig. 1. (a) Flowchart for the visualization of gene distribution patterns on phylogenetic trees: (1) scanning through the hit list in the homology search output; (2) identification of the source species for each hit from the accession number; and (3) projection of the identified species on the phylogenetic tree. (b) Procedure for gene classification through identification of its hypothetical ‘origin’; (4) selecting species having sequence similarity above a preset threshold in the distribution pattern; (5) tracing back to their first break point (origin) on the phylogenetic tree; and (6) gene classification based on the origin.

tree of species, upon which gene distribution patterns are projected. In the figure, projected nodes are colored according to the sequence similarity of the hits. ‘Tree explorer’ (Figure 2b) shows an enlargement of the region corresponding to the box in Figure 2a. As with the NCBI taxonomy web browser, users can explore the taxonomy by pointing and clicking on selected objects. For example, by clicking on the arrowheads at the branching points of lineages, one can ‘open’ (descend) or ‘close’ (ascend) the taxonomy. The ‘Gene description browser’ (Figure 2c) is an interface enabling one to browse gene descriptions of the homology search output. Within the interface, each description is colored according to its sequence similarity. The ‘Threshold operator’ (Figure 2d) provides the threshold above which the hit genes are considered for the following analyses. With this interface, users can set the threshold. The selection is immediately reflected by the identification of the origin. The default threshold score (score of the highest-scoring segment pair) is set to 200. The ‘Cluster list browser’ (Figure 2e) is used for origin-based gene classification. Each line on the list shows the origins and the number of genes classified into respective clusters. Clicking on a line causes a pull down menu to appear showing a list of genes. The distribution patterns can be displayed by selecting a menu item (gene).

RESULTS AND DISCUSSION

Displaying patterns of gene distribution

Visualization of gene distribution patterns on the evolutionary tree gives us perspective on gene propagation.
Fig. 2. ‘Tree landscape viewer’ (a) displays an overview of the phylogenetic tree of the species from NCBI. ‘Tree explorer’ (b) shows an enlarged region corresponding to the box in panel ‘a’ and provides the ability to descend or ascend the nodes. ‘Gene description browser’ (c) is an interface allowing users to browse the gene descriptions in the homology search output. ‘Threshold operator’ (d) provides graphical interface to set the threshold. ‘Cluster list browser’ (e) shows each origin-based cluster and the number of genes included within it.

Fig. 3. Examples of visualization of distribution patterns. The tubulin beta chain of H.sapiens (sp:TBB1_HUMAN) (a) and the ribulose 1,5-bisphosphate carboxylase small subunit (rbcS) of Synechocystis sp. (sp:RBS_SYNY3) (b) were projected on the phylogenetic tree.
Fig. 4. Gene classification of 6241 *S.cerevisiae* ORFs. Many origins (branching points) were found to exist between the root and *S.cerevisiae*. Identification of the ORF origins revealed their organization into ten clusters. Each cluster and the number of ORFs included within it are shown.

Here we present two examples. We first analyzed human beta tubulin, a constituent of microtubules (sp:TBB1_HUMAN). Microtubules perform a variety of essential functions in eukaryotic cells, including regulation of chromosome movement and vesicular traffic, and maintenance of cell shape and morphogenesis. Although microtubule-like structures have been found in prokaryotic cells, no report of bacterial ‘microtubules’ conforms to the standard biochemical, morphological, and sequence definition (Bermudes et al., 1994). Thus, the distribution of beta tubulin would be expected to be limited to eukaryotes. In accordance with this expectation, ORI-GENE projected a distribution pattern for beta tubulin and similar genes in which they were distributed broadly among eukaryotes, but were absent from bacteria and archaea (Figure 3a).

We also examined the distribution of ribulose 1,5-bisphosphate carboxylase (RuBisCO) small subunit (rbcS) from the photosynthetic bacteria *Synechocystis* sp. (sp:RBS_SYNY3). During photosynthesis, RuBisCO catalyzes a key reaction in the Calvin reductive pentose phosphate cycle. In this case, the distribution pattern revealed the discontinuous presence of the rbcS gene in some blocks of photosynthetic species, cyanobacteria, proteobacteria, plantae (Figure 3b). This prediction is consistent with reports suggesting that RuBisCO propagated with rampant horizontal transfer among cyanobacteria, proteobacteria and plastid (Delwiche and Palmer, 1996; Paoli et al., 1998). In addition, rbcS genes might have propagated in plastid-containing eukaryotes as a consequence of multiple symbiotic events (Gray, 1989). Thus, the pattern of rbcS distribution revealed by ORI-GENE appropriately exhibited aspects of propagation mediated by multiple events.

**Gene classification based on origin**

When the origins of 6241 ORFs of yeast *S.cerevisiae* were identified and classified, the ORFs arranged themselves
into ten clusters (Figure 4). The ‘S.cerevisiae’ cluster contained the most genes (3817 ORFs), followed by the ‘root’ (581 ORFs) and ‘Fungi/Metazoa group’ (500 ORFs) clusters. We then compared the results of this classification with the yeast functional catalogue in the MIPS database (Mewes et al., 1998), which lists related yeast ORFs with well-characterized functions (Figure 5).

In the ‘root’ cluster, gene categories for metabolism and energy production yielded particularly high ratios. In contrast, categories for signal transduction, intracellular transportation and cell growth + cell division + DNA synthesis produced high ratios in the ‘Eukaryota’ cluster. In the ‘S.cerevisiae’ cluster, the highest ratio was for unclassified genes.

Gene classification based on origin reflects evolutionary records, and the characteristics of the classified genes may have some correlation with their origins. For example, the ORFs in the ‘root’ origin cluster (581 ORFs) are likely to be related to fundamental activity of life such as metabolism and energy production. Thus, origin-based gene classification may be capable of inferring characteristics of genes. It is notable that the only requirement for origin identification was species information obtained from the accession number. Even when a query showed similarity only to a poorly annotated sequence, such as an EST, the gene classification could be made and the origin of the gene identified.

ORI-GENE thus offers a new tool for the screening of genes among different species. For example, finding multicellular-organism-specific genes from the gene set of the nematode C.elegans is interesting because such genes can serve as experimental targets revealing the mechanism of development and differentiation. In addition, from
the viewpoint of drug design, finding species-specific genes, e.g. genes shared only among bacteria, would be useful because such genes would be good ‘drug target’ candidates, enabling bacteria to be selectively killed without affecting their human hosts.

**Current limitations and prospects**

At present the methodology used in ORI-GENE is limited by three factors. (1) It may be biased by the rate of sequencing. Since not all species have been completely sequenced, investigators must judge whether in a given distribution pattern, a gene is absent or merely unsequenced. For instance, if homologous genes are distributed among very different species, it may indicate horizontal gene transfer. However, the distribution patterns may be due to the incomplete sequencing. (2) There is ambiguity in the phylogenetic tree. The NCBI taxonomy approximates current phylogenetic knowledge; nonetheless, there may be ambiguous classifications that disagree with the latest data. For example, according to the NCBI taxonomy, the branching of Dictyostelium discoideum, one of the most sequenced organisms, antedates the divergence of Animalia, Fungi, and Plantae. However, recent reports strongly suggest that D. discoideum is closely related to Animalia and Fungi, and that Plantae is an outgroup of the Animalia–Fungi clade (Kuma et al., 1995; Baldauf and Doolittle, 1997). Such ambiguity in the phylogenetic tree could influence interpretation of the distribution pattern. (3) Appropriate similarity thresholds vary. The similarity thresholds depend on the rate of change in the amino acid sequence and vary from gene to gene. There will be no universal threshold that is appropriate for all cases. Using our method of classification, some genes, having evolved with unusual speed, will show an obviously incorrect origin.

The first and second points mentioned above reflect the incompleteness of our knowledge. But since the amount of available sequence data is increasing rapidly, leading to further phylogenetic analysis and improved taxonomy data, we anticipate this situation will be rectified in due time. As new information becomes available, we will concurrently update the ORI-GENE data. The third problem is more complex. Given the variability of the rate of gene evolution, there is no universal threshold score applicable to all purposes. ORI-GENE provides an interface (Figure 2d) that enables users to adjust the threshold and to recalculate the distribution pattern. Despite these caveats, the present results, obtained using currently available data, seem reasonable and suggest that ORI-GENE is an effective tool for the classification and screening of new genes. Further, as more and more genomes are sequenced completely, we would expect the utility of ORI-GENE to increase substantially.

**ACKNOWLEDGEMENTS**

The authors are grateful to Drs Takahiro Morio, Masanori Arita, Hidemasa Bono, Atsushi Kurabayashi, Mitsuteru Nakao and Takashi Gojobori for valuable comments on this research.

**REFERENCES**


