Nucleic Acids Research, 2000, Vol. 28, No. 20 **4029–4036**

Automatic detection of conserved gene clusters in multiple genomes by graph comparison and P-quasi grouping

Wataru Fujibuchi, Hiroyuki Ogata, Hideo Matsuda[1] and Minoru Kanehisa*

Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan and [1]Graduate School of Engineering Science, Osaka University, Toyonaka, Osaka 560-8531, Japan

Received May 16, 2000; Revised July 10, 2000; Accepted August 4, 2000

ABSTRACT

We previously reported two graph algorithms for analysis of genomic information: a graph comparison algorithm to detect locally similar regions called correlated clusters and an algorithm to find a graph feature called P-quasi complete linkage. Based on these algorithms we have developed an automatic procedure to detect conserved gene clusters and align orthologous gene orders in multiple genomes. In the first step, the graph comparison is applied to pairwise genome comparisons, where the genome is considered as a one-dimensionally connected graph with genes as its nodes, and correlated clusters of genes that share sequence similarities are identified. In the next step, the P-quasi complete linkage analysis is applied to grouping of related clusters and conserved gene clusters in multiple genomes are identified. In the last step, orthologous relations of genes are established among each conserved cluster. We analyzed 17 completely sequenced microbial genomes and obtained 2313 clusters when the completeness parameter *P* was 40%. About one quarter contained at least two genes that appeared in the metabolic and regulatory pathways in the KEGG database. This collection of conserved gene clusters is used to refine and augment ortholog group tables in KEGG and also to define ortholog identifiers as an extension of EC numbers.

INTRODUCTION

With the availability of complete genome sequences for an increasing number of organisms whole genome comparison has become a powerful method for understanding genome structure, function and evolution. In the traditional sequence comparison, gene functions can be predicted by sequence similarity by establishing orthologous relations to well-characterized genes in other organisms. In whole genome comparison, additional clues for functional implications may be obtained by examining positional coupling of genes on the chromosome by establishing conservation of gene orders and gene clusters. Suppose, for example, that gene A in one organism is functionally well characterized and gene A′ in another organism is predicted as orthologous to A. Suppose also that there is gene B immediately adjacent to gene A but its function is unknown. If its ortholog B′ is adjacent to A′ or if the positional coupling of A–B is conserved among relatively distant species, there is a good chance that genes A and B are functionally coupled.

A number of reports have already been made concerning the conserved clusters of genes by comparative analysis of complete genome sequences. Generally speaking, even between phylogenetically close species, such as between *Escherichia coli* and *Haemophilus influenzae*, there is a considerable amount of juxtaposition of genes, but at the same time there is a tendency for short-range conservation of gene clusters (1–3). The conserved clusters are likely to represent functionally coupled genes, such as those forming operon structures for co-expression and/or those encoding physically interacting protein subunits (4–6). In addition to such an ancient evolutionary origin, a multicistronic gene cluster sometimes results from horizontal transfer between species (7,8). Furthermore, multiple genes in a bacterial operon tend to be fused into a single gene encoding a multi-domain protein in eukaryotic genomes (9,10). We have also utilized the information about sequence similarity and positional correlation of genes for functional prediction of ABC transporters (11) and other membrane proteins (12), as well as for metabolic pathway reconstruction from complete genome sequences (13).

The primary information in the genome is the sequence of nucleotides and the resulting sequence of amino acids, but at a higher level of abstraction the genome can be viewed as a sequence of genes. The local alignment of nucleotide sequences or amino acid sequences is based on an optimization procedure to detect locally similar subsequences using a measure of similarity defined for pairs of nucleotides or pairs of amino acids. In contrast, the local alignment of genomes is used to detect locally conserved gene clusters using a measure of gene similarity, which may be defined by the sequence

*To whom correspondence should be addressed. Tel: +81 774 38 3270; Fax: +81 774 38 3269; Email: kanehisa@kuicr.kyoto-u.ac.jp
Present addresses:
Wataru Fujibuchi, National Center for Biotechnology Information, National Institutes of Health, Building 38A, Room B2N14, Bethesda, MD 20894, USA
Hiroyuki Ogata, Information Génétique et Structurale, CNRS-UMR 1889, 31 Chemin Joseph Aiguier, 13402 Marseille Cedex 20, France

similarity scores of gene pairs. While in the sequence alignment problem the order of nucleotides or amino acids may not be changed, genome alignment allows rearrangements of corresponding genes. For example, when (A,A′), (B,B′), (C,C′) are similar gene pairs, the genome alignment A–B–C and A′–C′–B′ is a perfect match containing a reversal. Thus, local genome alignment requires detecting clusters of similar genes that are localized in two genomes. In the accompanying paper (14) we developed a graph comparison algorithm to detect correlated clusters of corresponding nodes, which is here applied to dealing with the local genome alignment problem.

The purpose of the present analysis is two-fold. First, we attempt to develop an automatic procedure to construct a 'multiple genome alignment' for conserved gene clusters. Second, we report the construction of ortholog group tables in KEGG (15) and the assignment of ortholog identifiers as an extension of EC (enzyme commssion) numbers.

MATERIALS AND METHODS

The open reading frame (ORF) data of 17 organisms

The ORF data for 17 complete genomes were prepared from the KEGG (Kyoto Encyclopedia of Genes and Genomes) GENES database (http://www.genome.ad.jp/kegg/kegg2.html), which is compiled from the complete sequence data originally released in GenBank. These comprise 12 bacterial genomes: *E.coli* (16), *H.influenzae* (17), *Helicobacter pylori* (18), *Bacillus subtilis* (19), *Mycoplasma genitalium* (20), *Mycoplasma pneumoniae* (21), *Mycobacterium tuberculosis* (22), *Chlamydia trachomatis* (23), *Borrelia burgdorferi* (24), *Treponema pallidum* (25), *Synechocystis* PCC6803 (26) and *Aquifex aeolicus* (27); four archaeal genomes: *Methanococcus jannashii* (28), *Methanobacterium thermoautotrophicum* (29), *Archaeoglobus fulgidus* (30) and *Pyrococcus horikoshii* (31); one eukaryotic genome: *Saccharomyces cerevisiae* (32). The KEGG/GENES database contains its own gene annotations together with those annotated by SWISS-PROT, GenBank and the original genome project teams.

In the present analysis we do not distinguish the strands on which genes are located. Thus, the sequential order of genes is defined by the smaller numbers of the gene positions in the GenBank database, namely the first nucleotide positions of the genes in one strand and the last nucleotide positions of the genes on the complementary strand. We consider only protein coding genes, excluding tRNA, rRNA and other RNA genes.

Pairwise comparison of genomes

A schematic view of the entire procedure of our automated analysis is shown in Figure 1. There are three steps to compile a multiple alignment of significant gene clusters among multiple genomes: (i) extraction of conserved (correlated) gene clusters in two genomes by the pairwise graph comparison algorithm; (ii) identification of related gene clusters in multiple genomes by P-quasi complete linkage analysis; and (iii) identification of orthologous, paralogous and fused genes in each gene cluster by the P-quasi and COG grouping methods to generate gene cluster tables.

Here we define a conserved gene cluster as a group of homologous genes that are located at contiguous positions in the genomes of multiple organisms. In order to detect such

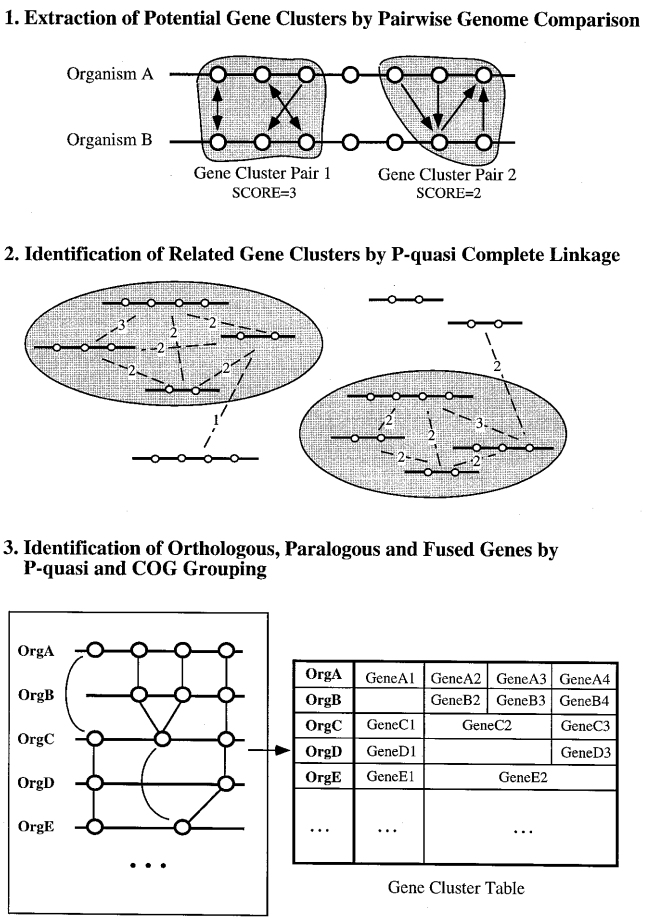

Figure 1. A schematic view of the entire procedure to extract conserved gene clusters in multiple genomes. (Step 1) A gene cluster pair is a group of related gene pairs that are located at contiguous positions in two genomes. An arrow indicates the best hit or the bi-directional best hit relation by SSEARCH. The similarity score of each gene cluster pair is defined by the smaller number of related genes (linked by arrows) in one genome. Thus, multiple links to the same node are counted just once. (Step 2) The cluster pairs are grouped by the P-quasi complete linkage method. The numbers indicate the scores retained from Step 1. (Step 3) Once a group of related gene clusters is obtained, the second P-quasi and the COG methods are used to establish the relationships of individual genes, including gene orders, orthologs, paralogs and fused genes.

conserved gene clusters in two genomes, we first make a similarity matrix of genes, which contains elements for all possible pairwise comparisons of protein coding genes. The genes are ordered on each axis of the matrix according to the positions in each genome. The similarity of two genes is defined by an amino acid sequence comparison using the SSEARCH program based on the Smith–Waterman algorithm (33). When the optimized (opt) score of SSEARCH is 100 or more, the matrix element is 1; otherwise the element is 0. Thus, the similarity matrix can also be viewed as a dot matrix for comparison of gene orders in two genomes. A diagonal stretch of ones in the matrix corresponds to a conserved gene cluster. More generally, a conserved gene cluster appears on the dot matrix as a local cluster of ones because it can contain rearrangements of genes, such as inversions, transpositions, fusions and fissions, as well as gaps (unpaired genes).

We apply the graph comparison algorithm (14) and detect conserved (correlated) gene clusters containing at least two homologous gene pairs, allowing rearrangements of gene positions and allowing gap lengths of up to two in each genome. The opt score of 100 in SSEARCH is a relatively low threshold value, which produces a number of spurious hits. In order to screen out the noise of such random occurrences of small clusters, we count only those clusters that contain at least two best hits. Here the best hit is the highest scoring gene pair when one gene in one genome is compared against all genes in the other genome.

Identification of related gene clusters in multiple genomes

Once all pairwise comparisons of 17 genomes, including self-comparisons, are made, a set of conserved gene cluster pairs is obtained together with a similarity score for each cluster pair. We define the similarity score between two gene clusters as the number of best hit gene pairs where multiple pairs involving the same node are combined in order to avoid counting paralogs more than once (see Fig. 1). Apparently this set contains the same gene clusters that are shared partially or entirely in multiple genomes when pairwise similarity links are combined. We thus perform grouping of gene clusters based on linkage of similar cluster pairs and compile groups of conserved gene clusters among multiple organisms by a clustering algorithm.

Generally speaking there are two representative clustering algorithms: single linkage and complete linkage. When single linkage is applied to our problem, many gene clusters tend to be merged into a small number of large groups. In the worst case, two unrelated clusters may be merged in the same group, which makes it difficult to obtain a multiple alignment of gene orders. On the other hand, complete linkage identifies only small groups of uniform gene cluster organizations, which is not suitable for detecting variations among species. In order to produce moderately conserved gene clusters among various species, we introduce the P-quasi complete linkage method (34), which satisfies the condition that any member in one group has linkages (similarity links) to $\geq P\%$ of all the members within the group. When P is 100% each member is linked to all other members in the group, which is equivalent to normal complete linkage. In contrast, single linkage requires just one link to any member in the group and P is virtually 0% when the number of members in the group is large.

For the threshold of similarity links we consider that two gene clusters are linked when two or more best hit gene pairs are shared, thus avoiding weak cluster connections by only one gene pair relationship. We have tried various percentage values for the completeness parameter P, namely $P = 0$ (single linkage), 20, 40, 60, 80 and 100% (complete linkage). It was extremely time and memory consuming to compute P-quasi completeness of gene cluster pairs for all the 17 organisms; it took between 3 and 7 days, depending on the parameter P, on a SGI Origin 2000 with 10 parallel CPUs.

Multiple genome alignment within a conserved gene cluster

The last step is to identify orthologous genes within individual gene clusters and to make rough drafts of multiple genome alignments, which are then manually edited to produce KEGG ortholog group tables. Again, we use the P-quasi complete linkage method to establish groups of homologous genes and at the same time to detect fused or split genes. This is based on the following criteria.

- Groups of homologous genes are defined by the P-quasi complete linkage method within each gene cluster with an opt score of 100 for the SSEARCH cut-off value and a completeness parameter P of 20%.
- When one gene in a genome corresponds to more than one homologous gene in another genome, it is considered to be a fused gene if all the following four conditions are met:

(i) the gene has two or more homologs (opt score $\geq$ 100) in another genome;

(ii) some of the homologs have no similarity (opt score < 100) with each other;

(iii) the sum of the lengths of the non-similar homologs is not much greater (<50 amino acids) than the gene length;

(iv) the similarity regions of the gene to the non-similar homologs have small overlaps (up to 10 amino acids).

Otherwise, the homologs are assumed to be paralogs.

- Paralogs are further divided into different ortholog groups by the COG (Cluster of Orthologous Groups of proteins) triangle method (35) of linkages based on best hits.

RESULTS

Conserved gene clusters in pairwise genome comparisons

An example of the pairwise genome comparison is shown in Figure 2, which is a dot plot representation of the similarity matrix for all the protein coding genes between *M.pneumoniae* and *M.genitalium* (Fig. 2a) and between *C.trachomatis* and *M.genitalium* (Fig. 2b). Each dot corresponds to a SSEARCH score of 100 or more for the pairwise comparison of protein coding genes at the amino acid sequence level. Note that diagonal stretches are often interrupted in a comparison of the two closely related mycoplasmas because of the existence of RNA genes, which are not excluded from the axes and whose similarities are not examined.

Figure 2c and d shows the results of applying the graph comparison algorithm to detect correlated clusters. Many dots that exist in the SSEARCH result are screened out and only a relatively small number of stretches are found for conserved gene clusters in the two genomes. It is remarkable that the conserved clusters of the two mycoplasmas are located almost co-linearly on the two genomes, with one large chromosomal translocation (21) and some additions and amplifications of genes in *M.pneumoniae* (36). However, genomic rearrangements are usually so extensive that such distinct co-linearity is not found except for very closely related species, as exemplified in Figure 2d. Here the longest stretch is a ribosomal protein cluster, which is known to be one of the most conserved gene clusters in all bacterial and archaeal genomes (see the KEGG ortholog group tables listed in Table 1).

In the KEGG system we maintain precomputed SSEARCH scores for all pairwise comparisons of completely sequenced genomes. The dot plot analysis such as shown in Figure 2 can be performed by the genome comparison tool (available at http://www.genome.ad.jp/kegg-bin/mk_genome_cmp_java).

In addition, a list of conserved gene clusters can be generated for any pair of genomes by the graph comparison algorithm (available at http://www.genome.ad.jp/kegg-bin/genome_cmp).

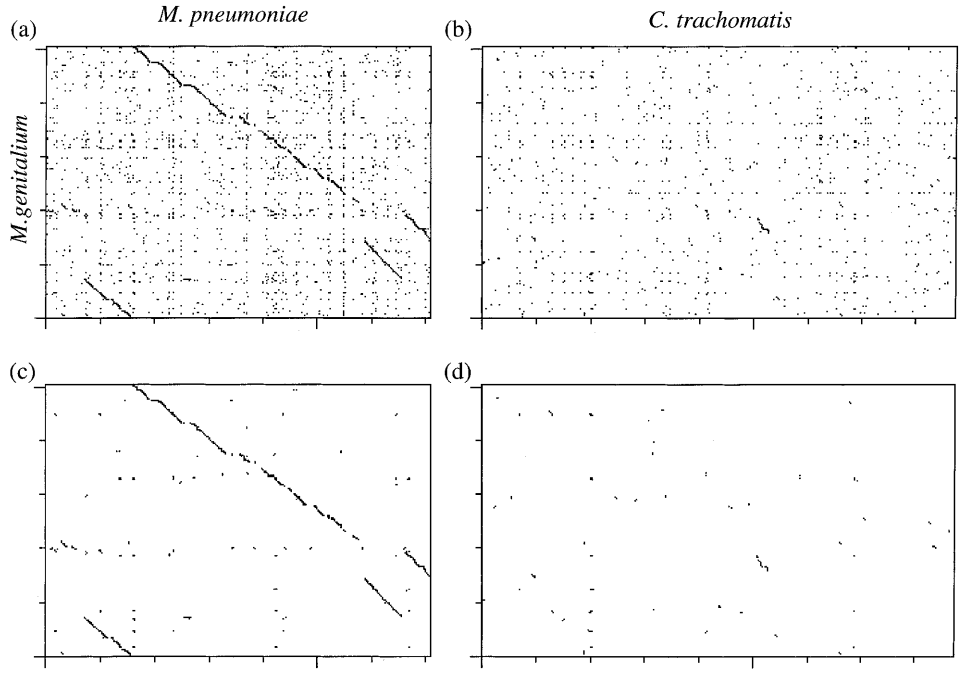

Figure 2. The dot plot matrices representing the sequence similarity results by SSEARCH (upper) and the conserved cluster search results by our algorithm (lower) for pairwise comparisons of all protein coding genes between *M.genitalium* and *M.pneumoniae* (left) and between *C.trachomatis* and *M.genitalium* (right).

Table 1. Selected ortholog group tables in KEGG

Ortholog group	URL
Ribosomal protein cluster	http://www.genome.ad.jp/kegg/ortholog/tab01030.html
ATP synthase	http://www.genome.ad.jp/kegg/ortholog/tab03110.html
Tryptophan biosynthesis	http://www.genome.ad.jp/kegg/ortholog/tab00400.html
Glycolysis	http://www.genome.ad.jp/kegg/ortholog/tab00010.html
RNA polymerase	http://www.genome.ad.jp/kegg/ortholog/tab03020.html
NADH dehydrogenase	http://www.genome.ad.jp/kegg/ortholog/tab03100.html
Pyruvate oxidoreductase	http://www.genome.ad.jp/kegg/ortholog/tab03120.html
ABC transporters	http://www.genome.ad.jp/kegg/ortholog/tab02010.html

Fraction of genes in the conserved gene clusters

The fraction of genes in conserved gene clusters is obviously dependent on the closeness of the species being compared. Figure 3 shows variations of such fractions for all pairwise genome comparisons in our dataset excluding *S.cerevisiae*. The vertical axis shows the ratio of the number of genes in the clusters to the total number of genes in the genome. Shaded boxes and open triangles represent, respectively, the genomes that have larger and smaller numbers of genes in the pairwise genome comparison. The horizontal axis shows the phylogenetic distance between the two genomes measured as the percentage difference in small rRNA sequences according to the Ribosomal Database Project II site (http://www.cme.msu.edu/rdp/).

Not surprisingly, the ratio of clustered genes decreases as the phylogenetic distance between two species increases. The ratio plateaus at a constant level of ~8% when the phylogenetic distance reaches 30%. It is apparent that many clusters are not significant when the species being compared are too closely related. Thus, we introduce a cut-off value for selecting significant gene clusters that are conserved despite extensive rearrangements of the entire genomes. In this study we extracted only those clusters that were found in species with a phylogenetic distance of 20% or more. This condition was highly useful in eliminating apparent clusters that were found only between close species, for example between *M.genitalium* and *M.pneumoniae* or between *E.coli* and *H.influenzae*.

Conserved gene clusters in multiple genomes

The gene clusters that are conserved among multiple genomes are obtained by merging related gene clusters identified in the pairwise genome comparisons using P-quasi complete linkage analysis. Obviously the number and the size of such clusters as

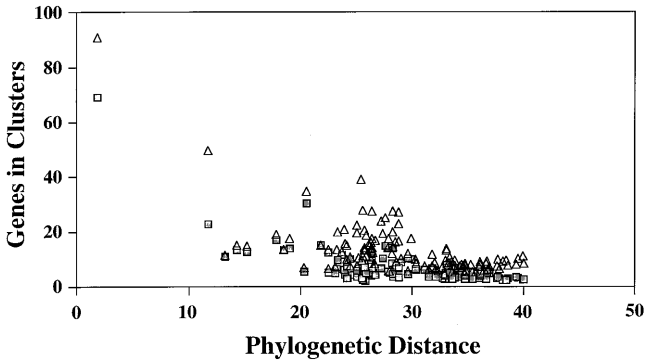

Figure 3. The percentage of genes in the conserved clusters relative to the total number of genes in the genome when two genomes are compared. The percentages for the larger (shaded boxes) and the smaller genome (open triangles) in the pairwise comparison are plotted against the phylogenetic distance between the two genomes according to the percent difference in small rRNA sequences.

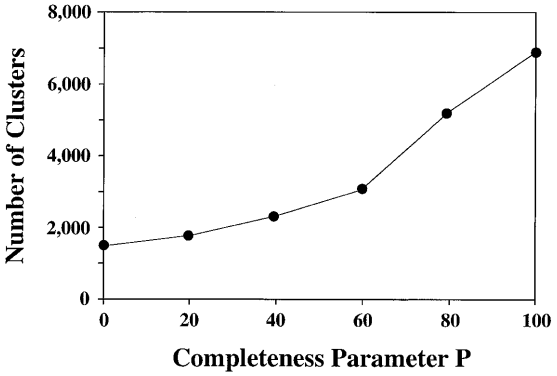

Figure 4. The number of groups formed by merging related clusters is plotted against the completeness parameter *P* in a P-quasi complete linkage analysis. The parameter values of 100 and 0 correspond, respectively, to complete linkage and single linkage.

well as their quality are dependent on the parameter *P*, the extent of completeness for linkage. Figure 4 shows the number of conserved clusters in multiple genomes plotted against *P*, where *P* = 100% corresponds to complete linkage and *P* = 0% corresponds to single linkage. The number of clusters increases as the extent of completeness becomes higher, from a minimum of 1462 in single linkage to a maximum of 6825 in complete linkage.

Although it was not possible to determine an appropriate value for *P* from Figure 4 alone, the break at 60% appeared to be a highest limit for the extent of completeness. Thus, using values of 40 and 20% we examined how well known gene clusters, including those shown in Table 1, could be reproduced by our method. Generally speaking, the optimal value varied for different gene clusters. For example, the value of 40% was most suitable for F1-F0 ATP synthase and pyruvate oxidoreductase, 20% was better for NADH dehydrogense and peptide ABC transporters and either value was appropriate for the largest ribosomal protein gene cluster and the tryptophan operon. The manner of cluster conservation seemed to be dependent, at least, on the conservation of individual genes (in terms of amino acid sequence) and the number of paralogous genes outside the cluster. In the following analysis we chose the value of 40% as one of the acceptable values. We note that in practice our computational results are used only as rough approximations for the KEGG ortholog group tables, which are refined by human experts with additional analyses.

Gene cluster tables

The results of identifying orthologous gene relations within each conserved cluster are represented in the form of a gene cluster table (Fig. 1), which may be considered as a multiple genome alignment. A complete collection of gene cluster tables obtained with *P* = 40 and 20% for the 17 genomes is available at http://kanehisa.kuicr.kyoto-u.ac.jp/Paper/gclust/

Figure 5a shows an example of a gene cluster table, that for the *trp* gene cluster for the tryptophan biosynthesis pathway, which was obtained with *P* = 40%. The *trp* gene clusters in nine of 13 organisms containing this pathway were identified, with the exceptions *Synechocystis*, *A.aeolicus* and *S.cerevisiae*, whose genes were dispersed in the genome. Figure 5b is the corresponding ortholog group table for tryptophan biosynthesis in KEGG (Table 1), which was originally constructed from FRECs (functionally related enzyme clusters) analysis using graph comparison of genomes and metabolic pathways (14) and manually refined with additional analyses. While the current automatic method performed relatively well, it obviously did not detect orthologous genes that either did not belong to gene clusters or had only weak sequence similarities.

The total of 2313 clusters obtained with *P* = 40% (Fig. 4) contained 370 clusters that were considered not to be significant because the phylogenetic distance was below the 20% threshold. When the rest were compared with the KEGG metabolic and regulatory pathway maps, ~27% of the significant clusters contained at least two genes that also appeared on the KEGG pathways. In some cases genes that were adjacent in the genome, i.e. were in the gene cluster, were not adjacent in the metabolic pathways. A case in point is the two glycolytic enzyme genes (Table 1) 6-phosphofructokinase (EC 2.7.1.11) and pyruvate kinase (EC 2.7.1.40), which form a conserved gene cluster in several Gram-positive bacteria, but which are distantly placed (six steps apart) in the glycolysis pathway. This gene cluster is known to be critical in regulating the overall direction of glycolysis or gluconeogenesis. In our previous FRECs analysis (14) this particular gene cluster was not detected because the separation in the pathway was above the threshold of allowable gaps.

Fused genes

As can be seen in Figure 5, our method correctly identifies fused genes and their orthologous relationship to other genes. In fact, we assign orthologs in two steps: first by defining gene clusters where just groups of possible orthologs are identified, and second by generating gene cluster tables where more precise relations are established. Both steps utilize sequence similarities in pairwise genome comparisons, especially best hits and bi-directional best hits, and positional correlations on the genomes. In the second step we also examine the possibility of fused genes according to the criteria given in Materials and Methods. A number of cases are known where multiple

(a) Computationally generated gene cluster table

organ-ism	map00400 (10.3)	map00400 (10.3)	map00400 (5.9)	map00400 (4.2)	map00130 map00400 (2.9)	map00130 map00400 (2.8)	map00130 map00400 (5.9)
Eco	b1260	b1261	b1262		b1263		b1264
Hin		HI1388	HI1389.1		HI1388	HI1389	HI1387
	HI1432	HI1431					
Hpy	HP1277	HP1278	HP1279		HP1281		HP1282
Bsu	trpA	trpB	trpC	trpF		trpD	trpE
Mtu	Rv1613	Rv1612	Rv1611				Rv1609
Ctr	CT171	CT170					
Mja	MJ1038	MJ1037					
Mth	MTH1660	MTH1659	MTH1657	MTH1658	MTH1656	MTH1661	MTH1655
Afu	AF1599	AF1600	AF1604	AF1601	AF1602	AF1604	AF1603

(b) KEGG ortholog group table

organ-ism	EC 4.2.1.20		EC 5.3.1.24	EC 4.1.1.48	EC 2.4.2.18	EC 4.1.3.27	
Eco	b1260(trpA)	b1261(trpB)	b1262(trpC)		b1263(trpD)		b1264(trpE)
Hin	HI1432	HI1431	HI1389.1		HI1389	HI1388 HI1171	HI1387
Hpy	HP1277	HP1278	HP1279		(HP1280)	HP1281	HP1282
Bsu	trpA	trpB	trpF	trpC	trpD	pabA(trpG)	trpE
Mtu	Rv1613	Rv1612		Rv1611	Rv2192c		Rv1609 Rv2386c
Ctr	CT171	CT170	CT327				
Mja	MJ1038	MJ1037	MJ0451	MJ0918	MJ0234	MJ0238	MJ1075
Mth	MTH1660	MTH1659 MTH1476	MTH1658	MTH1657	MTH1661	MTH1656	MTH1655
Afu	AF1599	AF1600 AF1240	AF1601	AF1604		AF1602	AF1603

Figure 5. The gene cluster corresponding to the *trp* operon for tryptophan biosynthesis. (**a**) The gene cluster table computationally generated with $P = 40\%$ and (**b**) the manually refined table as represented in the KEGG ortholog group table. The columns in these tables represent groups of orthologous genes, which are annotated with the KEGG pathway map numbers and similarity weights in (a) and with the EC numbers in (b). The shading in (b) denotes possible operon structures, which is better viewed by the coloring at the KEGG web site (http://www.genome.ad.jp/kegg/ortholog/tab00400.html). The gene names in parentheses are alternative names, except for HP1280 which contains a frameshift (no amino acid sequence). Eco, *Escherichia coli*; Hin, *Haemophilus influenzae*; Hpy, *Helicobacter pylori*; Bsu, *Bacillus subtilis*; Mtu, *Mycobacterium tuberculosis*; Ctr, *Chlamydia trachomatis*; Mja, *Methanococcus jannashii*; Mth, *Methanobacterium thermoautotrophicum*; Afu, *Archaeoglobus fulgidus*.

proteins encoded in operons in some organisms are fused into single proteins in other organisms. A few examples from Table 1 are bacterial RNA polymerase subunits β and β′, which are fused in *H.pylori*, NADH dehydrogenase chains C and D, which are fused in *E.coli*, and pyruvate:ferredoxin oxidoreductase α and β subunits, which are fused in *E.coli* and *Synechocystis*. The fusion (or fission) patterns are somewhat more complicated for the genes of the tryptophan biosynthesis pathway shown in Figure 5; there are three fusion events: *trpC* and *trpF* in three proteobacteria, *trpC* and *trpD* in *A.fulgidus* and *trpD* and *trpG* in *E.coli*.

DISCUSSION

The sequence similarity found in database searches is often the basis of assigning gene functions. However, one of the major problems in interpreting the search results is that there is no predefined threshold of percent identity or similarity score that assures functional identity. In fact, it is not simply the score that matters, but many other additional factors need to be integrated to come up with a conclusion, such as the class of protein, the location of similar segments, the types of amino acids conserved, the number of paralogs found and the evolutionary distance between the species being compared. The complete genome sequence provides additional constraints for better interpretation of sequence similarity relations. For example, when a pair of genes taken from different genomes exhibits a bi-directional best hit relation, it is a good indication of functional orthology, even if the sequence similarity score is relatively low. Here bi-directional best hit means that when one gene (or one set of paralogous genes) in one genome is searched against all genes in the other genome, the other gene (or the other set of paralogous genes) is the best hit, and vice versa. The COG database (35,37) applies this criterion to three distant species and attempts to classify all proteins in the completely sequenced genomes.

In contrast, the collection of KEGG ortholog group tables is not a functional categorization of individual proteins. Rather, it is intended to provide more specific functional information about cellular roles; how proteins interact with each other to form a pathway or a molecular complex. As reported in this paper, this type of higher level functional information is sometimes encoded in the genome as a gene cluster, which is conserved among different species and which can be detected by whole genome comparisons of both positional and sequence information for genes. From a practical point of view of systematically uncovering such functional information, we have developed an automated procedure to detect and align conserved gene clusters in multiple genomes based on two graph algorithms. One is the algorithm to compare two graphs and identify locally similar subgraphs called correlated clusters (14); the other is the algorithm to find a graph feature called the P-quasi complete subgraph (34). A complete graph is a fully connected graph with each node linked to all other nodes, while in a P-quasi complete graph each node is linked to at least $P\%$ of all other nodes. In these graph analyses edges are unweighted, as well as undirected. Thus, sequence similarity is either present or absent according to a threshold similarity score and also whether it is a best hit or not.

Most methods in computational molecular biology explicitly utilize numerical scores in order to determine biological meanings of sequence similarities, expression similarities and other relations. For example, in hierarchical cluster analysis the linkage of two clusters is determined by the shortest distance (single linkage), the longest distance (complete linkage), or an average distance among all pairwise distances (similarity scores) between members of the respective clusters. Numerical scores are definitely useful when the computational analysis involves only a specific type of data. In general, however, biological interpretation is an integrated process which requires concurrent evaluation of different types of data. As we have seen, the existence of a best hit or the presence of a positional correlation can sometimes be more meaningful than how high the similarity score is. We believe that it is better to convert numerical scores into only ones and zeros in order to evaluate different types of data in a common framework.

This abstraction is also relevant to our graph-based approach. We are interested in understanding higher level cellular functions that result from networks of interacting proteins, rather than functions of single proteins, from the information in the genome. We consider a huge graph of genes or proteins as nodes which are linked by different types of edges, including experimentally determined protein–protein interactions, computationally derived similarity relations, positional correlations of genes in the genome and many other relations (38). We are trying to correlate structural features of this graph, without considering weights for edges, to higher level functions. Our graph-based approach, which is also a logic-based approach, is an attempt to integrate different types of data and knowledge towards automating, at least in part, human reasoning steps.

At present there are about 70 ortholog group tables available in KEGG. The metabolic pathway portion of the tables was originally organized using the results of aligning genes from different organisms against a conserved portion of the metabolic pathway in a FRECs analysis (14). The present study has expanded the original collection in several respects. It has refined the existing tables by distinguishing different subunits with the same EC (enzyme commission) number in an enzyme complex and clarifying complicated relations involving fused genes (see Table 1 for examples). It has also been helpful in identifying new gene clusters that consist of non-enzyme proteins. This has led us to introduce what we call ortholog identifiers for categorization of functional homologs in KEGG. We have been using the EC numbers for functional identification of enzymes and enzyme genes, especially for mapping genes in the genome onto gene products (enzymes) in the pathway. Starting with KEGG release 14.0 in April 2000, this mapping is to be done using the ortholog identifiers instead of the EC numbers. Table 2 shows a few examples of ortholog identifiers which are, at the moment, not associated with any hierarchical classification such as the EC numbering system. Because of the ortholog identifiers it is now possible to computerize and utilize networks of interacting molecules, including both metabolic pathways and regulatory pathways, in a uniform way.

Table 2. Examples of ortholog identifiers in KEGG

Ortholog identifier	Definition	EC number	Pathway	Gene
E5.4.2.1	phosphoglycerate mutase	5.4.2.1	map00010	eco:b0755, eco:b4395, ...
E4.2.1.20A	tryptophan synthase alpha chain	4.2.1.20	map00400	eco:b1260, hin:HI1432, ...
E4.2.1.20B	tryptophan synthase beta chain			eco:b1261, hin:HI1431, ...
RP-S2	small subunit ribosomal protein S2	none	map03010b	eco:b0169, hin:HI0913, ...
			map03010a	mja:MJ0982, mth:MTH44, ...

ACKNOWLEDGEMENTS

We thank Drs Kenta Nakai and Atsushi Ogiwara for helpful discussions. This work was supported by a Grant-in-Aid for Scientific Research on the Priority Area 'Genome Science' from the Ministry of Education, Science, Sports and Culture of Japan and the Genome Frontier Project 'Genetic and Molecular Networks' from the Science and Technology Agency of Japan. The computational resource was provided by the Supercomputer Laboratory, Institute for Chemical Research, Kyoto University.

REFERENCES

1. Tatusov,R.L., Mushegian,A.R., Bork,P., Brown,N.P., Hayes,W.S., Borodovsky,M., Rudd,K.E. and Koonin,V. (1996) *Curr. Biol.*, **6**, 279–291.
2. Watanabe,H., Mori,H., Itoh,T. and Gojobori,T. (1997) *J. Mol. Evol.*, **44**, S57–S64.
3. Siefert,J.L., Martin,K.A., Abdi,F., Widger,W.R. and Fox,G.E. (1997) *J. Mol. Evol.*, **45**, 467–472.
4. Tamames,J., Casari,G., Ouzounis,C. and Valencia,A. (1997) *J. Mol. Evol.*, **44**, 66–73.
5. Dandekar,T., Snel,B., Huynen,M. and Bork,P. (1998) *Trends Biochem. Sci.*, **23**, 324–328.
6. Overbeek,R., Fonstein,M., D'Souza,M., Pusch,G.D. and Maltsev,N. (1999) *Proc. Natl Acad. Sci. USA*, **96**, 2896–2901.
7. Lawrence,J.G. and Roth,J.R. (1996) *Genetics*, **143**, 1843–1860.
8. Xu,Y., Murray,B.E. and Weinstock,G.M. (1998) *Infect. Immun.*, **66**, 4313–4323.
9. Davidson,J.N. and Peterson,M.L. (1997) *Trends Genet.*, **13**, 281–285.
10. Marcotte,E.M., Pellegrini,M., Ng,H.L., Rice,D.W., Yeates,T.O. and Eisenberg,D. (1999) *Science*, **285**, 751–753.
11. Tomii,K. and Kanehisa,M. (1998) *Genome Res.*, **8**, 1048–1059.
12. Kihara,D. and Kanehisa,M. (2000) *Genome Res.*, **10**, 731–743.
13. Ogata,H., Goto,S., Sato,K., Fujibuchi,W., Bono,H. and Kanehisa,M. (1999) *Nucleic Acids Res.*, **27**, 29–34.
14. Ogata,H., Fujibuchi,W., Goto,S. and Kanehisa,M. (2000) *Nucleic Acids Res.*, **28**, 4021–4028.
15. Kanehisa,M. and Goto,S. (2000) *Nucleic Acids Res.*, **28**, 27–30.
16. Blattner,F.R., Plunkett,G. III, Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides.J., Glasner,J.D., Rode,C.K., Mayhew,G.F. *et al.* (1997) *Science*, **277**, 1453–1474.
17. Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J.-F., Dougherty,B.A., Merrick,J.M. *et al.* (1995) *Science*, **269**, 496–512.
18. Tomb,J.-F., White,O., Kerlavage,A.R., Clayton,R.A., Sutton,G.G., Fleischmann,R.D., Ketchum,K.A., Klenk,H.P., Gill,S. and Dougherty,B.A. *et al.* (1997) *Nature*, **388**, 539–547.
19. Kunst,F., Ogasawara,N., Moszer,I., Albertin,A.M., Alloni,G., Azevedo,V., Bertero,M.G., Bessieres,P., Bolotin,A., Borchert,S. *et al.* (1997) *Nature*, **390**, 249–256.
20. Fraser,C.M., Gocayne,J.D., White,O., Adams,M.D., Clayton,R.A., Fleischmann,R.D., Bult,C.J., Kerlavage,A.R., Sutton,G.G., Kelley,J.M. *et al.* (1995) *Science*, **270**, 397–403.
21. Himmelreich,R., Hilbert,H., Plagens,H., Pirkl,E., Li,B.-C. and Herrmann,R. (1996) *Nucleic Acids Res.*, **24**, 4420–4449.
22. Cole,S.T., Brosch,R., Parkhill,J., Garnier,T., Churcher,C., Harris,D., Gordon,S.V., Eiglmeier,K., Gas,S., Barry,C.E. *et al.* (1998) *Nature*, **393**, 537–544.
23. Stephens,R.S., Kalman,S., Lammel,C., Fan,J., Marathe,R., Aravind,L., Mitchell,W., Olinger,L., Tatusov,R.L., Zhao,Q. *et al.* (1998) *Science*, **282**, 754–759.
24. Fraser,C.M., Casjens,S., Huang,W.M., Sutton,G.G., Clayton,R., Lathigra,R.,White,O., Ketchum,K.A., Dodson,R., Hickey,E.K. *et al.* (1997) *Nature*, **390**, 580–586.
25. Fraser,C.M., Norris,S.J., Weinstock,G.M., White,O., Sutton,G.G., Dodson,R., Gwinn,M., Hickey,E.K., Clayton,R., Ketchum,K.A. *et al.* (1998) *Science*, **281**, 375–388.
26. Kaneko,T., Sato,S., Kotani,H., Tanaka,A., Asamizu,E., Nakamura,Y., Miyajima,N., Hirosawa,M., Sugiura,M., Sasamoto,S. *et al.* (1996) *DNA Res.*, **3**, 109–136.
27. Deckert,G., Warren,P.V., Gaasterland,T., Young,W.G., Lenox,A.L., Graham,D.E., Overbeek,R., Snead,M.A., Keller,M., Aujay,M. *et al.* (1998) *Nature*, **392**, 353–358.
28. Bult,C.J., White,O., Olsen,G.J., Zhou,L., Fleischmann,R.D., Sutton,G., Blake,J.A., FitzGerald,L.M., Clayton,R.A., Gocayne,J.D. *et al.* (1996) *Science*, **273**, 1058–1073.
29. Smith,D.R., Doucette-Stamm,L.A., Deloughery,C., Lee,H., Dubois,J., Aldredge,T., Bashirzadeh,R., Blakely,D., Cook,R., Gilbert,K. *et al.* (1997) *J. Bacteriol.*, **179**, 7135–7155.

30. Klenk,H.P., Clayton,R.A., Tomb,J.-F., White,O., Nelson,K.E., Ketchum,K.A., Dodson,R.J., Gwinn,M., Hickey,E.K., Peterson,J.D. *et al.* (1997) *Nature*, **390**, 364–370.
31. Kawarabayasi,Y., Sawada,M., Horikawa,H., Haikawa,Y., Hino,Y., Yamamoto,S., Sekine,M., Baba,S., Kosugi,H., Hosoyama,A. *et al.* (1998) *DNA Res.*, **5**, 55–76.
32. Goffeau,A., Aert,R., Agostini-Carbone,M.L., Ahmed,A., Aigle,M., Alberghina,L., Albermann,K., Albers,M., Aldea,M., Alexandraki,D. *et al.* (1997) *Nature*, **387** (suppl.), 1–105.
33. Smith,T.F. and Waterman,M.S. (1983) *J. Mol. Biol.*, **147**, 195–197.
34. Matsuda,H., Ishihara,T. and Hashimoto,A. (1999) *Theor. Comput. Sci.*, **210**, 305–325.
35. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) *Science*, **278**, 631–637.
36. Himmelreich,R., Plagens,H., Hilbert,H., Reiner,B. and Herrmann,R. (1997) *Nucleic Acids Res.*, **25**, 701–712.
37. Tatusov,R.L., Galperin,M.Y., Natale,D.A. and Koonin,E.V. (2000) *Nucleic Acids Res.*, **28**, 33–36.
38. Kanehisa,M. (2000) *Post-genome Informatics*. Oxford University Press, Oxford, UK.