BIOINFORMATICS

Vol. 15 no. 11 1999
Pages 900–908

An automated comparative analysis of 17 complete microbial genomes

Arvind K. Bansal

Department of Mathematics and Computer Science, Kent State University, Kent, OH 44242, USA

Abstract

Motivation: *As sequenced genomes become larger and sequencing becomes faster, there is a need to develop accurate automated genome comparison techniques and databases to facilitate derivation of genome functionality; identification of enzymes, putative operons and metabolic pathways; and to derive phylogenetic classification of microbes.*

Results: *This paper extends an automated pair-wise genome comparison technique (Bansal* et al.*, Math. Model. Sci. Comput., **9**, 1–23, 1998, Bansal and Bork, in* First International Workshop of Declarative Languages*, Springer, pp. 275–289, 1999) used to identify orthologs and gene groups to derive orthologous genes in a group of genomes and to identify genes with conserved functionality. Seventeen microbial genomes archived at ftp://ncbi.nlm.nih.gov/genbank/genomes have been compared using the automated technique. Data related to orthologs, gene groups, gene duplication, gene fusion, orthologs with conserved functionality, and genes specifically orthologous to* Escherichia coli *and pathogens has been presented and analyzed.*

Availability: *A prototype database is available at ftp://www.mcs.kent.edu/~arvind/intellibio/orthos.html. The software is free for academic research under an academic license. The detailed database for every microbial genome in NCBI is commercially available through intellibio software and consultancy corporation (Web site: http://www.mcs.kent.edu/~arvind/intellibio.html).*

Contact: *arvind@mcs.kent.edu*

Introduction

Microbes (bacteria and archaea) serve as model organisms for understanding basic metabolic functions. Microbes are also important targets in biotechnology, disease treatment and ecology. The comparison of genomes forms an important technique in identifying the functionality of individual genes (Tatusov *et al.*, 1997; Bansal *et al.*, 1998; Bork *et al.*, 1998), which is essential for the identification of functionality unique to a group of genomes and for mapping the metabolic pathways (Selkov *et al.*, 1997; Tatusov *et al.*, 1996) of the organisms. Gene function, identification of genes specific to pathogens, and the study of metabolic pathways and their variations will facilitate the discovery of the causes of diseases.

The first microbial genome was completely sequenced in 1995 (Fleischmann *et al.*, 1995). Currently, 25 completed genomes – 23 microbes (references to the papers related to genome sequencing are available at ftp://ncbi.nlm.nih.gov/genbank/genomes/bacteria in the submitted genome files), *Caenorhabditis elegans* and *Saccharomyces cerevisiae* – are archived at NCBI (ftp://ncbi.nlm.nih.gov/genbank/genomes) and many are underway. A major portion of the human genome will be sequenced by the year 2000.

As genomes become available at a faster rate, accurate automated techniques will become the first necessary step for cross-species comparison. However, the cross-species comparisons to identify orthologs (exact functional counterparts of genes in different genomes) must be done carefully since similarity-based comparisons (Altschul *et al.*, 1990) identify homologs (similar genes derived from some common ancestral gene). Homologs also includes paralogs, which may have a different functionality due to gene duplications. Putative orthologs are identified efficiently by filtering out dissimilar protein sequnces using BLAST, aligning the similar pairs of protein sequences using the Smith–Waterman algorithm (Waterman, 1995), and then using a variation of weighted bipartite graph matching technique (Bansal *et al.*, 1998; Bansal and Bork, 1999) to find the best matches.

A database of putative orthologs and gene groups will facilitate the identification of putative functionality of genes and gene groups in newly sequenced genomes. The comparison of a union of sets of orthologs in the complete set of genomes against a newly sequenced genome, using the bi-partite graph-matching technique (Bansal *et al.*, 1998), will identify a major part of functionality of newly sequenced genome in a single pass.

This paper compares a set of 17 microbial genomes: *Aquifex aeolicus*, *Archaeoglobus fulgidus*, *Borrelia burgdorferi*, *Bacillus subtilis*, *Chlamydia trachomatis*, *Escherichia coli*, *Haemophilus influenzae*, *Helicobacter pylori*, *Mycoplasma genitalium*, *Methanococcus*

jannaschii, *Mycoplasma pneumoniae*, *Methanobacterium thermoautotrophicum*, *Mycobacterium tuberculosis*, *Pyrococcus horikoshii*, *Rickettsia prowazekii*, *Synechocystis* sp. PCC6803 and *Treponema pallidum*. The paper identifies the orthologs, orthologous gene groups (gene groups composed of orthologs), putative gene fusions (genes homologous to a fusion of two subsequences of adjacent genes) and gene groups which have multiple homologous correspondences in other genomes.

The results can be immediately used to identify enzymes and the functionality of genes in newly sequenced genomes; to curate gene name and functionality in genome databases; to perform multiple sequence alignments of orthologs; to facilitate identification of the operons in the genomes using the orthologous gene groups; and to provide new insight in phylogenetic classification of microbes. The results in this paper, augmented with secondary structure information, are already being used to understand the regulation mechanism in operons involving ribosomal proteins (Vitreschak *et al.*, 1999).

The technique presented in this paper uses comparison of protein sequences for the corresponding genes in two different genomes. The automated comparison was performed using the prototype software library Goldie 2.0 (Genome Ortholog Detection and Inference Engine), a significant enhancement of Goldie 1.0 (Bansal *et al.*, 1998; Bansal and Bork, 1999). The software needs only gbk-format files from NCBI, and is portable across different Unix-based architectures which support Sicstus Prolog (http://www.sics.se), the WU-BLASTP package (http://blast.wustl.edu) and the Smith–Waterman software library (http://www-hto.usc.edu/software/seqaln/).

Definitions

Unfortuntely, the genome organization is quite complex. In order to understand the definitions of different gene groups involved in useful phenomena, some mathematical notations have been used in this paper. A subset is denoted by $\subset$, non-inclusion in a set is denoted by $\not\subset$, and length of a subsequence s is denoted by $|s|$. The inclusion of the subsequence in a sequence is denoted by the mathematical symbol for '$\subseteq$'. Set difference is denoted by '$-$', and a small number is denoted by the Greek symbol ε.

Modeling genomes

A genome Γ is modeled as an ordered set of genes $\langle \gamma_1, \gamma_2, \ldots, \gamma_N \rangle$ where N is the number of genes in the genome. The set of protein sequences corresponding to the protein coding regions in a genome Γ is modeled as $\langle \pi_1, \pi_2, \ldots, \pi_N \rangle$ where π_I is the protein sequence corresponding to the protein coding region of the gene γ_I. A subsequence of protein sequence π_I is denoted as δ_I. There may be more than one different subsequences in a protein which are homologous to protein sequences corresponding to different genes in another genome. These subsequences include one or more protein domains. However, the explicit knowledge of protein domains is absent during the comparison. For the sake of convenience, the comparison of genome Γ_1 with another genome Γ_2 will be denoted as $\Gamma_1 \mapsto \Gamma_2$, and the alignment of two largest protein subsequences δ_I and δ_J with the best alignment score will be denoted as $\delta_I \leftrightarrow \delta_J$.

Orthologs

An *ortholog* is an exact functional counterpart of a gene in another genome that has arisen from speciation (Fitch, 1970). However, uncertainty is inherent in phylogeny due to lateral transfer of genes (Huynen and Bork, 1998), gene insertions and deletions, gene fusion and splitting (Bansal *et al.*, 1998), and difference in the evolutionary trees based upon different criteria (Olsen *et al.*, 1994; Gruber and Bryant, 1997; Snel *et al.*, 1998). This paper uses a definition based upon sequence similarity to define putative orthologs. *Putative ortholog* is defined as a gene γ_{2J}, in a genome Γ_2 such that it has the best similarity score (above a threshold) with another gene γ_{1I} in a genome Γ_1 during the pair-wise comparison of genomes Γ_1 and Γ_2. A *gene function is conserved* if a gene is orthologous in the genomes of two or more major genome families: proteobacteria, gram-positive or archaea.

Gene groups

A *gene group* Σ is a cluster of neighboring genes $\langle \gamma_I, \gamma_J, \gamma_K \ldots \rangle$ with at least two distinct genes which have a natural pressure to occur in close proximity. Close proximity of two gene positions indexed by I and J is defined as $0 < I - \mathrm{c} < J < I + \mathrm{c} <$ genome size and $0 < J - \mathrm{c} < I < J + \mathrm{c} <$ genome size where c is a small constant experimentally limited by 12 (Bansal *et al.*, 1998). The study of gene groups is important since

1. Gene groups are starting points to identify operons in newly sequenced genomes.
2. Gene groups can be used to quantify the amount of duplication in newly sequenced genomes.
3. The study of variations – insertions, deletions, and change in the order – in these gene groups will facilitate the study of variations in metabolic pathways in two genomes.
4. The number of orthologous gene groups can be used as a measure of similarity between two genome functions.

A *gene group* $\langle \gamma_{2M}, \gamma_{2N}, \gamma_{2P} \ldots \rangle (M \neq N \neq P$; and M, N, and P are in close proximity) in the genome

Γ_2 is identified by marking the corresponding protein sequences $\langle \pi_{2M}, \pi_{2N}, \pi_{2P} \ldots \rangle$ in Γ_2 to an ordered bag of protein sequence $\langle \pi_{1I}, \pi_{1J}, \pi_{1K} \ldots \rangle (I \leq J \leq K)$ in Γ_1 corresponding to the gene group $\langle \gamma_{1I}, \gamma_{1J}, \gamma_{1K} \ldots \rangle$ in the genome Γ_1 such that $(\delta_{2M} \subseteq \pi_{2M}) \leftrightarrow (\delta_{1I} \subseteq \pi_{1I})$, $(\delta_{2N} \subseteq \pi_{2N}) \leftrightarrow (\delta_{1J} \subseteq \pi_{1J})$, and $(\delta_{2P} \subseteq \pi_{2P}) \leftrightarrow (\delta_{1K}) \subseteq \pi_{1K})$. A gene group $\langle \gamma_{2M}, \gamma_{2N}, \gamma_{2P} \ldots \rangle$ is ordered if

$$M < N < P \text{ when } I < J < K, \text{ or} \quad (1)$$
$$M > N > P \text{ when } I > J > K \quad (2)$$

The study of *ordered gene groups* is an important starting point to identify and study the operons and common sub-units in metabolic pathways of genomes.

A gene group $\langle \gamma_{2M}, \gamma_{2N}, \gamma_{2P} \ldots \rangle$ is *unordered* if one or more of the following conditions are satisfied:

$$(M > N \text{ when } I < J) \text{ or } (N > P \text{ when } J < K) \quad (3)$$
$$(M < N \text{ when } I > J) \text{ or } (N < P \text{ when } J > K) \quad (4)$$
$$(M \neq N \text{ and } I = J) \text{ or } (N \neq P \text{ and } J = K) \quad (5)$$

The third condition is possible when there are multiple protein subsequences (corresponding to a single gene) which are homologous to protein subsequences corresponding to two or more different genes. The study of unordered groups is important:

1. to understand the mechanism of variations in metabolic pathways of two different genomes, and
2. to understand the evolution based upon variations in metabolic pathways of different genomes

Orthologous gene groups comprise only orthologous genes. The study of orthologous gene groups is important since it annotates the function of gene groups in newly sequenced genomes. The number of orthologous gene groups also provides an important measure of the functional similarity of two genomes. Experimental data shows that a large percentage of orthologous gene groups are ordered. However, there are many unordered orthologous gene groups caused by the permutation of gene order as described in conditions (3) and (4) in the definition of ordered groups. For example, the gene group (*Bs:bioB*, *Bs:bioD*, *Bs:bioF*, *Bs:bioA*, *Bs:bioW*) is orthologous to the gene group (*Mj:MJ1296*, *Mj:MJ1299*, *Mj:MJ1298*, *Mj:MJ1300*, *Mj:MJ1297*) where *Bs* denotes the genome *B.subtilis*, and *MJ* denotes the genome *M.jannaschii*. At this point the biological significance of permutation of the gene order (in a gene group) on the functionality of gene groups is unknown limiting the further refinement of this definition.

A *multigene group* Σ_{1I} is an unordered gene group in a genome Γ_1 satisfying following two conditions:

1. Σ_{1I} has a corresponding gene group Σ_{2M} in the genome Γ_2, and
2. Σ_{2M} (or whose proper subset) has at least one more corresponding multigene group $\Sigma_{1J} (I \neq J)$ in the genome Γ_1 such that two Σ_{1I} and Σ_{1J} are disjoint – do not have a common gene.

The study of multigene groups is important since understanding the duplication of gene groups will facilitate the understanding of variation and functional similarity in different metabolic pathways comprised of these gene groups. The data presented in this paper suggests that the duplication of gene groups plays a major role in the change in the functionality of genomes.

A *duplicated gene* is a single gene γ_{1I} in Γ_1 (not inside any gene group) whose corresponding protein sequence π_{1I} is homologous to two or more protein sequences $\pi_{2M}, \pi_{2N} \ldots (M \neq N)$ in Γ_2 such that the corresponding genes γ_{2M} and γ_{2N} are adjacent. The data presented in this paper suggests that duplication of single genes is also a common phenomenon, and plays a major role in the change in functionality of genomes.

A fused gene group has a gene γ_{1I} in the genome Γ_1 such that two (or more) non-overlapping (Bansal *et al.*, 1998) protein subsequences $\delta_{1M}, \delta_{1N} \subseteq \pi_{1I} (\delta_{1M} \not\subset \delta_{1N}$ and $|\delta_{1M} - \delta_{1N}| \leq \in)$ align with the protein subsequences $\delta_{2U} \subseteq \pi_{2J}$ and $\delta_{2V} \subseteq \pi_{2K} (J \neq K$, and J and K are in close proximity). The data suggests that fused genes are another important mechanism of change in function of microbial genomes.

Note that the definitions of orthologous groups, multigene group, duplicated genes, and fused genes are mutually exclusive as follows:

1. Orthologous gene groups have one-to-one mapping between the corresponding gene groups while multigene groups have many-to-one mapping of the corresponding gene group in another genome. Additionally, multigene groups do not contain orthologs.
2. Multigene groups result from the duplication of gene groups, and duplicated genes involve the duplication of single genes not inside any gene group.
3. Fused genes involve non-overlapping protein subsequences, while duplicate genes involve duplication of the same protein sequence.

Nomenclature

This paper uses NCBI nomenclature for gene names and abbreviates *A.aeolicus* to *Aa*, *A.fulgidus* to *Af*, *B.burgdorferi* to *Bb*, *B.subtilis* to *Bs*, *C.trachomatis* to *Ct*, *E.coli* to *Ec*, *H.influenzae* to *Hi*, *H.pylori* to *Hp*,

M.genitalium to *Mg*, *M.jannaschii* to *Mj*, *M.pneumoniae* to *Mp*, *M.thermoautotrophicum* to *Mt*, *M.tuberculosis* to *Tb*, *R.prowazekii* to *Rp*, *P.horikoshii* to *Ph*, *T.pallidum* to *Tp*, and *Synechocystis* sp. PCC6803 to *Sy*.

When there is no ambiguity, *E.coli* name is used. To avoid ambiguity in the presence of genes of multiple genomes, a protein sequence (or the corresponding gene) from a specific genome is referred to as *Genome name:gene name*, an aligned protein subsequence is referred to *as Genome name:gene name:alignment_start..alignment_stop*. A gene without a known functionality has been referred to as *Genome name:orf.index*, where *index* is the ordering of the protein coding region (annotated with 'CDS' in gbk files) within the genome.

Methods

This section briefly describes the algorithm used to identify orthologs and orthologous gene groups (Bansal *et al.*, 1998; Bansal and Bork, 1999), and extends the algorithm to identify orthologs in a group of genomes.

Identifying orthologs

The pair-wise comparison of two genomes is modeled as a weighted bipartite graph-matching problem (Papadimitrou and Steiglitz, 1982). The weights of the edges are identified using the Smith–Waterman algorithm and PAM120 matrix. The gene corresponding to the nodes of the best matching edges of the bipartite graphs are taken as orthologs, and are deleted from the further consideration. A scheme based upon weighting edges using BLAST scores will be faster, but less accurate (Brenner *et al.*, 1998).

Since naive bipartite matching will have $N \times M$ gene pairs, where N and M are the number of genes in two respective genomes, the total cost of comparing two genomes using the Smith–Waterman alignment will be $N \times M \times O(K^2)$ where K is the average number of characters in a gene representation. In order to improve the execution efficiency, the number of gene pairs are pruned based on BLAST similarity matching techniques. Only those gene pairs are used which have a *high-score* value above a certain threshold – 50 for evolutionary close genome families and 30 for distant families such as gram-negative and archaea, or gram-positive and archaea – and a *chance* value threshold – 1.0×10^{-5} for evolutionarily close genome families and 1.0×10^{-3} for distant genomes such as gram-negative and archaea or gram-positive and archaea). The rationale is as follows:

1. The ortholog similarity statistically erodes by 15–20 % for evolutionary distant families such as proteobacteria and archaea (Bansal *et al.*, 1998; Brenner *et al.*, 1998).
2. For evolutionary close genomes lowering the threshold may result in spurious data.

Gene pair filtering after the BLAST phase ensures (in most of the cases) approximately five edges incident upon the same node. This makes the cost of alignment $N \times O(M \times K) + \mathrm{c} \times N \times O(K^2)$ where c is a small constant, N is the number of genes in Γ_1, and M is the number of genes in Γ_2. The first term is the comparison cost in the BLAST phase, and the second term is the cost in the Smith–Waterman phase. After identifying the weights of the edges using the Smith–Waterman alignment, a variation of weighted bipartite graph-matching sorts the edges in the descending order. The set of nodes corresponding to the highest weighted edges are collected as putative orthologs. After finding an edge (π_{1I}, π_{2j}) of the highest weight, all the edges involving the nodes π_{1I} and π_{2j} are deleted. The process is continued until there are no more edges. The edges starting or ending in genes inside a gene group are positively biased as genes within a gene group are better candidates for preserving a common functionality. Two edges with close weights, if two weights are above a threshold, suggest multiple orthologs or gene fusion, and need further analysis to identify the fused genes. Multiple edges with close weights (below a threshold) suggest the presence of paralogs.

Identifying gene groups

First, gene groups are identified. A neighboring group S_0 for a gene in Γ_1 is marked. A *neigboring group* is a group of genes in close proximity. Then a set S_1 (in Γ_2) of homologs for S_0 is marked. Then the set S_2 – a union of all the sets of homologs of S_1 – in Γ_1 is marked. A non-empty intersection of the sets S_0 and S_2, with more than one element in the intersection, marks the presence of the start of a homologous gene group. After marking the start of homologous gene groups, the genome Γ_1 is traversed one node at a time, checking for the presence of an edge in the close proximity of the last homologous gene in the genome Γ_2. The method identifies gene groups of any variable size.

Orthologs in groups of genomes

Multiple genome pairs were compared against a common representative genome in the family and against *E.coli* and *B.subtilis*. *Escherichia coli* was chosen for proteobacteria, *B.subtilis* was chosen for gram-positive bacteria, and *M.jannaschii* was chosen for an archaea. The rationale for selection was:

1. *E.coli* and *B.subtilis* are thoroughly explored genes in the wet labs, and one of the largest ones in their respective families;

2. many of the pathogens are either proteobacteria or gram-positive bacteria; and

3. my experiments reveal that there are many genes which are absent in at least one genome of the same family, but have orthologs outside the family.

The intersection of sets of the orthologs obtained from pair-wise comparison of genomes against a common genome determined the orthologs within the set of genomes. The sets of conserved genes were obtained by first comparing all the genomes against *E.coli* and then comparing all the genomes against *B.subtilis*. The two sets were slightly different for the following reasons:

1. *E.coli* or *B.subtilis* alone do not share all the orthologs with other genomes.

2. Experimental results showed that the comparison against one common genome misses some orthologs. The discrepancy is due to the combination of the approximation involved in string-matching algorithms and the BLAST cutoff of the high-score and chance value.

For each ortholog in the set of orthologs derived using *E.coli* as a common genome, the corresponding entry was identified in the set of orthologs derived using *B.subtilis* as a common genome. The union of set of genomes for these two entries derived the group of genomes containing the ortholog. The process was repeated for every ortholog. The genes with most conserved functionality were identified by marking orthologs in 14 or more genomes. The rationale is that a gene orthologous to 14 of the given 17 genomes (containing four archaea genomes) is orthologous in proteobacteria, gram-positive and archaea.

Results and discussion

To compare genome groups, the set of genomes was divided into four categories: proteobacteria, spirochetes, gram-positive and archaea (Olsen *et al.*, 1994). The category of proteobacteria contains *E.coli*, *H.influenzae*, *H.pylori* and *R.prowazekii*; spirochetes contains *B.burgdorferi* and *T.pallidum*; gram-positive contains *B.subtilis*, *M.genitalium*, *M.pneumoniae* and *M.tuberculosis*; and archaea contains *A.fulgidus*, *M.jannaschii*, *M.thermoautotrophicum* and *P.horikoshii*. Three microbes, *C.trachmomatis*, *Synechocystis* sp. PCC6803 and *A.aeolicus*, were only compared with *E.coli* and *B.subtilis*.

New annotations

Escherichia coli contains 42 genes of unknown and undocumented functionality (marked as ORF in NCBI databank), which have an ortholog in *B.subtilis* (using the same NCBI data-bank) with known functionality. Some of the examples are (*Ec:orf.263*, *Bs:xynB*), (*Ec:orf.317*, *Bs:adhA*), (*Ec:orf.325*, *Bs:mmgD*), (*Ec:orf.490*, *Bs:wapa*), (*Ec:orf.533*, *Bs:ebrB*), and (*Ec:orf.588*, *Bs:cstA*). Similarly, 504 genes of *B.subtilis* with unknown and undocumented functionality in the NCBI database have orthologs with documented functionality in *E.coli* within the NCBI database. Similarly, other genomes have been annotated.

Analysis of orthologs

Table 1 summarizes the result about the orthologs between various genome pairs. The data shows the number of orthologs/percentage of orthologs with respect to the number of genes in the first genome/percentage of orthologs with respect to the number of genes in the second genome. A complete list of orthologs for *E.coli* vs. *B.subtilis* containing detailed information is available at ftp://www.mcs.kent.edu/~arvind/intellibio/ortho.

The results show that the genomes within the same family have a large percentage of orthologs, such as *Ec-Hi*, *Mg-Mp*, *Mg-Bs*, *Mp-Bs*, etc. However, cross-family comparisons also reveal important data. For example, *Ec-Bs*, *Ec-Mg*, *Ec-Bs*, *Ec-Tb*, *Ec-Bb*, *Bs-Bb*, *Ec-Aa* have a significant number of orthologous genes. The genome comparison of *Ec* vs. *Bs* shows that orthologous genes in two genomes (from two different families) match well with the name and enzyme (if present) annotations at NCBI. The results also point out missing enzymes and gene functionality in these two genomes. Ninety-seven percent of the genes in *M.genitalium* have an ortholog in *M.pneumoniae*.

Study of gene groups

Table 2 shows the number of different types of gene groups in genome pairs Γ_1–Γ_2. The first column shows gene pairs, the second column shows ordered gene groups/orthologous gene groups, the third column shows multigene groups in Γ_1/number of fused genes in Γ_1/duplicated genes in Γ_1, and the fourth column shows multigene groups in Γ_2/fused genes in Γ_2/duplicated genes in Γ_2 . The following observations were made:

1. The number of gene groups within the same family had higher percentages in terms of the size of the smaller genome.

2. There are some genome pairs found between proteo- and gram-positive bacteria such as *Ec-Bs*, *Ec-Tb*, *Bs-Hi*, that have a large number of orthologous gene groups. The number of orthologous gene groups paired between *Ec-Hp* and *Bs-Hp*, *Ec-Ct* and *Bs-Ct*, *Ec-Tp* and *Bs-Tp* are comparable.

3. Duplication is a common means of change in

Table 1. Orthologs in genome pairs

Pairs	Orthologs	Pairs	Orthologs
Ec-Hi	1226/28%/71%	*Bs-Ec*	1276/31%/29%
Ec-Rp	516/12%/61%	*Bs-Hi*	802/19%/46%
Ec-Hp	694/16%/44%	*Bs-Rp*	440/10%/52%
Ec-Ct	469/10%/52%	*Bs-Hp*	645/15%/41%
Ec-Bb	411/9%/44%	*Bs-Ct*	454/11%/50%
Ec-Tp	467/10%/45%	*Bs-Bb*	444/10%/52%
Ec-Tb	1016/23%/25%	*Bs-Tp*	471/11%/45%
Ec-Mp	271/6%/40%	*Bs-Tb*	1005/24%/25%
Ec-Mg	246/5%/52%	*Bs-Mp*	326/7%/48%
Ec-Bs	1276/29%/31%	*Bs-Mg*	314/7%/66%
Ec-Sy	983/22%/31%	*Bs-Sy*	906/22%/28%
Ec-Aa	810/18%/53%	*Bs-Aa*	761/18%/50%
Ec-Mt	568/13%/30%	*Bs-Mt*	595/14%/31%
Ec-Mj	522/12%/31%	*Bs-Mj*	562/13%/33%
Ec-Ph	536/12%/27%	*Bs-Ph*	575/14%/29%
Ec-Af	637/14%/26%	*Bs-Af*	669/16%/27%
Hi-Rp	431/25%/51%		
Hi-Hp	574/33%/36%	*Af-Mj*	790/32%/47%
Rp-Hp	394/47%/25%	*Af-Mt*	812/33%/43%
Bb-Tp	441/51%/42%	*Af-Ph*	714/29%/36%
Mg-Mp	454/97%/66%	*Mj-Mt*	860/51%/49%
Mg-Tb	233/49%/5%	*Mj-Ph*	651/38%/32%
Mp-Tb	252/37%/6%	*Mt-Ph*	621/33%/31%

Table 2. Gene groups in genome pairs

Pair	Gene group data		
Ec-Hi	272/269	410/9/294	326/9/327
Ec-Rp	98/69	140/6/29	85/6/170
Ec-Hp	96/70	148/1/83	98/2/164
Ec-Ct	73/52	102/1/148	69/1/130
Ec-Bb	77/55	129/4/86	96/4/113
Ec-Tp	57/50	123/0/17	101/0/156
Ec-Tb	120/107	507/15/267	489/16/336
Ec-Mp	36/33	85/3/282	64/3/180
Ec-Mg	35/28	84/5/194	68/5/142
Ec-Bs	172/166	1073/22/874	1053/24/703
Ec-Sy	86/46	578/9/397	571/10/642
Ec-Aa	93/58	225/13/73	194/14/175
Ec-Mt	78/36	255/7/100	204/8/206
Ec-Mj	59/23	174/5/86	130/5/193
Ec-Ph	64/34	287/5/458	250/7/349
Ec-Af	80/42	521/12/491	487/14/410
Bs-Ec	172/166	1053/24/703	1073/22/874
Bs-Hi	135/105	306/3/30	259/3/453
Bs-Rp	87/52	136/13/31	91/13/180
Bs-Hp	86/60	158/1/87	116/1/231
Bs-Ct	69/63	98/0/140	69/0/156
Bs-Bb	66/61	143/2/97	132/2/144
Bs-Tp	67/49	96/0/31	76/0/205
Bs-Tb	145/134	609/15/396	597/16/527
Bs-Mp	58/56	97/3/301	79/3/257
Bs-Mg	54/62	102/5/233	87/5/223
Bs-Sy	92/48	518/8/406	507/8/636
Bs-Aa	114/58	161/6/26	117/6/191
Bs-Mt	64/35	188/10/96	172/10/218
Bs-Mj	60/31	186/11/65	164/11/224
Bs-Ph	69/35	304/5/404	267/5/407
Bs-Af	86/53	446/18/629	451/18/531
Hi-Rp	56/63	66/4/13	59/4/87
Hi-Hp	65/57	76/0/45	66/0/86
Hp-Rp	56/34	38/2/13	19/3/20
Bb-Tp	64/74	34/0/16	38/0/37
Tb-Mg	35/32	34/0/94	28/0/37
Tb-Mp	33/35	42/0/141	35/0/57
Mg-Mp	8/19	55/5/76	64/8/70
Mt-Af	104/90	174/11/174	182/10/121
Mj-Af	98/76	152/7/168	182/7/54
Ph-Af	78/70	193/9/230	184/9/219
Mt-Mj	103/93	150/7/48	137/6/69
Mt-Ph	73/61	94/2/104	101/3/39
Mj-Ph	69/51	131/3/113	117/3/61

genome functionality. Both multigene groups and duplicated genes are large in number. The number of multigene groups is mainly a function of the genome size. To a lesser extent, the number of multigene groups is positively related to genomes being in the same family.

4. Fused genes are present in genome pairs across various families, such as the genome pairs *Ec-Tb*, *Ec-Bs*, *Ec-Aa*, and *Ec-Af*, *Ec-Sy*, *Bs-Rp*, *Bs-Tb*, and *Bs-Af*. The fused genes do not have any correlation with genomes being in the same family.

Analysis of fused genes

A close analysis of fused genes in the genome pairs *Ec-Bs* reveals that two different subsequences of adjacent genes join to form a composite gene with a combined functionality. The functionality of many of these fused genes are unknown in the NCBI database. For example, the proteins *Ec:yadG*, *Ec:sapF*, *Ec:modC*, *Ec:orf.1455*, *Ec:orf.1882*, *Ec:hisP* are homologous to the fusion of proteins *Bs:yvrO:18..173* and *Bs:fhuC:140..240*. Since *modC* and *fhuC* and *hisP* are transport ATP binding proteins, the data suggest that *Ec:yadG*, *Ec:orf.1455* and *Ec:orf.1882* are also possible transport binding proteins. Similarly, the protein *Ec:mrdA* (function undocumented in the NCBI database) is homologous to the fusion of the adjacent proteins *Bs:pbpB:194..292* and *Bs:spoVD:45..167*. Both *Bs:pbpB* and *Bs:spovD* are penicillin binding proteins, which suggest that mrdA is a pencillin binding protein. However, *Bs:spovD* is orthologous to *Ec:ftsI*, and *Ec:mrdA* is orthologous to *Bs:pbpC* – a pencillin binding protein. The protein *Ec:orf.805* is homologous to the fusion of two adjacent proteins *Bs:appD* and *Bs:appF*. The enzyme *Ec:fabG* – 3-oxoacyl-[acyl-carrier protein] reduc-

tase – is homologous to the fusion of *Bs:yusR* (function not documented) and *Bs:yusS* (function not documented). However, *Ec:fabG* is orthologous to *Bs:fabG*. The protein *Ec:yehX*, a transport binding protein, is homologous to the fusion of *Bs:cydC* and *Bs:cydD*. Both *Bs:cydC* and *Bs:cydD* are ABC membrane transporters (ATP-binding proteins). An aerobic respiration control sensor protein *Ec:arcB* (*Ec:arcB:286..501*, *Ec:arcB:525..638*) is homologous to (*Bs:phoR:354..571*, *Bs:phoP:4..115*) and (*Bs:yycG:371..598*, *Bs:yycF:4..114*). *phoR* is a two-component sensor histadine kinase, and *phoP* is a two-component sensor regulator. The genes *Ec:arcB*, *Bs:phoR* and *Bs:phoP* are not orthologous to any other gene. It appears to be a case of gene fusion with *Ec:arcB* having a combined functionality.

The data suggests that gene fusion is a possible mechanism for the formation of new genes with a composite function. However, the presence of an ortholog (at a site different than the component gene fragments) of some fused genes in the same genome containing component homologous subsequences complicates the matter, and needs an explanation.

Multigene groups and duplicated genes

The data from pair-wise genome comparisons suggest that multigene groups are mainly a function of the number of genes in a genome. Larger genomes have more genes and multigene groups. For example, in the *Ec-Hi* comparison, *Hi* has 326 multigene groups while *E.coli* has 410 multigene groups. In the *Bs-Mp* comparison, *Bs* has 97 multigene groups compared to 79 multigene groups in *Mp*. In the *Bs-Ph* comparison, *Bs* has 304 multigene groups compared to 267 in *Ph*.

The number of multigene groups also depends, to a lesser extent, on the genome being in the same family. Comparison of *Ec-Hi* and *Bs-Hi* shows that *E.coli* has 410 multigene groups compared to 306 multigene groups in *B.subtilis*. A similar trend is visible in other genome pairs.

The number of duplicated genes seems to be independent of the number of genes in a genome. Some genomes have more duplicated genes with respect to one genome than other genomes. The comparisons of the pairs *Ec-Rp*, *Bs-Rp*, *Hi-Rp* show that *Ec*, *Bs* and *Hi* have few duplicates of genes in *Rp*. The comparisons of the genome pairs *Ec-Af*, *Ec-Ph*, *Bs-Af*, *Bs-Ph* show larger number of duplicated genes than *Af* or *Ph* being compared with other archaea genomes.

Conserved gene functions

Table 3 shows the orthologs occurring in maximum number of genomes. The names are given in the form *E.coli gene name*/*B.subtilis gene name*. In case *E.coli* and *B.subtilis* genes have the same name, only one name has been used. The result reveals that many orthologs with conserved functionality are related to the mechanism of transcription and translation. Some of the ribosomal proteins have no orthologs in some of the genomes: *rpmF* has 9, *rpmE* has 9, *rpsU* has 9, *rplY* has 10, *rpsF* has 10, *rpsT* has 11, *rpmB* has 9, and *rpmJ* has 11. Ribosomal proteins *rpsF*, *rpsP*, *rpsR*, *rpsT*, *rpsU*, *rplI*, *rplP*, *rplQ*, *rplS*, *rplT*, *rplU*, *rplY*, *rpmA*, *rpmB*, *rpmD*, *rpmE*, *rpmF*, *rpmG*, *rpmH*, *rpmI* and *rpmJ* have no orthologs in archaea microbes *M.jannaschii*, *M.thermoautotrophicum*, *A.fulgidus* and *P.horikoshii*.

Table 3. Genes with conserved functions

Genome count	Gene count	*E.coli* gene name/*B.subtilis* gene name for different gene names, otherwise *E.coli* gene name
17	57	*alaS*, *argS*, *dnaX*, *ffh*, *ftsy*, *fusA/fus*, *gltX*, *glyA*, *hflB*, *hisS*, *ileS*, *infB*, *ksgA*, *map*, *metG*, *mopA/groEL*, *nusA*, *pheS*, *prlA/secY*, *recA*, *rplA*, *rplB*, *rplC*, *rplE*, *rplF*, *rplK*, *rplM*, *rplN*, *rplO*, *rplR*, *rplX*, *rplV*, *rpsB*, *rpsC*, *rpsD*, *rpsE*, *rpsG*, *rpsH*, *rpsI*, *rpsJ*, *rpsK*, *rpsL*, *rpsM*, *rpsQ*, *rpsS*, *rpoB*, *rpoC*, *secD/secF*, *tufB/tufA*, *topA*, *trpS*, *truA*, *uvrB*, *ychf/yyaF*, *ycfH/yabD*, *ygjD/ydiE*, *yhbZ/obg*
16	16	*cysS*, *eno*, *ftsZ*, *hflb*, *lon*, *mesJ/yacA*, *mrsA/ybbT*, *pepP/yqhT*, *pgk*, *pheT*, *rplD*, *rpsO*, *serS*, *ychF/yyAF*, *ycfH/yabD*, *yfjB/yjbN*
15	12	*adk*, *ndk*, *nth*, *orf.174/yluA*, *prsA/prs*, *pyrG/ctrA*, *pyrH/smbA*, *tpiA/tpi*, *tmk*, *mesJ,/yacA*, *ycfF/hit,-/yerN*

Table 4. Orthologs in various groups of genomes

Groups	Orthologs	Groups	Orthologs
Proteobacteria		Archaea	
Ec-Hi-Rp-Hp	258	*Mj-Mt-Af-Ph*	358
Ec-Hi-Rp	370	*Mj-Mt-Ph*	439
Ec-Hp-Rp	297	*Mj-Af-Mt*	553
Ec-Hi-Hp	450	*Mj-Af-Ph*	459
Hi-Hp-Rp	282		
Gram-positive		Other	
Bs-Mg-Mp-Tb	215	*Ec-Hi-Rp-Hp-Bs*	225
Bs-Mg-Mp	286	*Bs-Mg-Mp-Tb-Ec*	181
Bs-Mg-Tb	228	*Mj-Mt-Af-Ph-Ec-Bs*	124
Bs-Mp-Tb	228	*Mj-Mt-Af-Ph-Ec*	139
Tb-Mg-Mp	214	*Mj-Mt-Af-Ph-Bs*	150

Orthologs in genome groups

Table 4 shows a comparative analysis of orthologs in different groups of genomes. The result shows that a large percentage of orthologous groups are ordered. However,

Table 5. Orthologs specific to pathogens and *E.coli*

Group of genomes: set of genes	
Ec-Bb-Hi-Hp-Rp-Tp:hflC	*Ec-Bb-Ct-Hi-Tb:recB*
Ec-Bb-Hi-Rp-Tp: hflK	*Ec-Ct-Hi-Rp-Tb:orf.1343*
Ec-Bb-Hi-Tp: hrpA	*Ec-Bb-Hi-Tb: (fba,recC)*,
Ec-Hi-Rp-Tb: (hscA, yhjE)	*Ec-Bb-Rp-Tb:orf.2034*
Ec-Bb-Hi-Tp: hupA,	*Ec-Bb-Ct-Hi: ydeA*,
Ec-Ct-Hi-Rp: ydeA,	*Ec-Ct-Hi-Rp: (yigN,ccmA)*,
Ec-Ct-Hi-Hp: pal,	*Ec-Bb-Tb-Tp: thiZ*,
Ec-Ct-Hi-Tb:pbpG,	*Ec-Hi-Hp-Tb: fic*,
Ec-Bb-Ct-Tp: orf.2559	*Ec-Mg-Mp:potI*,
Ec-Mp-Tb: yhfV	*Ec-Hi-Mp: orf.255*,
Ec-Tb-Tp:orf.2363	*Ec-Hi-Tp: (mreD,secE)*
Ec-Rp-Tb: (*yhcM*, *gppA*, *smtA*, *orf.1269*)	
Ec-Hp-Tb: (*asnA*, *hdhA*, *add*), *Ec-Ct-Hi*:(*orf.1597*, *orf.1602*, *trpR*)	
Ec-Bb-Hi: (*pepD*, *orf.1839*, *fucP*, *rfaF*, *orf.3153*, *yibQ*)	
Ec-Hi-Rp: (*ampG*, *yraP*, *vacJ*, *secB*, *orf.634*, *dsbB*, *yfhE*, *bolA*, *cyaY*, *orf.2833*)	
Ec-Hi-Hp: (*bisZ*, *sdaC*, *mdaB*, *ykgB*, *phnA*, *orf.2936*, *yibN*, *pnuC*)	
Ec-Hi-Tb: (*aceE*, *glnE*, *plsB*, *glnD*, *dld*, *tesB*, *tesB*, *orf.606*, *nadR menE*, *yjeR*, *menC*, *tag*, *yijC*, *ccmB*, *orf.669*, *frdC*, *yibN*, *frdD*, *orf.2751*)	

there are unordered orthologous groups. The significance of alteration of gene order on the overall functionality of the gene group is still not clear. This limitation can only be answered by wet lab experiments. Many orthologs of *Ec-Hi-Hp* (450 orthologs) are not shared by *Rp* (*Ec-Hi-Hp-Rp* has 258 orthologs). There are 215 orthologs in the gram-positive family, of which 181 also occur in *E.coli*. Similarly, there are almost 258 orthologs in *EC-Hi-Hp-Rp*. 225 of these orthologs also occur in *B.subtilis*. Archaea genomes share a high percentage of orthologs among themselves. The percentage of orthologs in archaea shared with proteobacteria and gram-positive is significantly less. There are 358 orthologs in *Mt-Mj-Af-Ph*. However, only 139 of these are present in *E.coli* and 150 are present in *B.subtilis*. *Pyrococcus horikoshii* is somewhat separated from *M.jannaschii*, *M.thermoautotrophicum* and *A.fulgidus*: *Mt-Mj-Af* has 553 orthologs which reduces to 358 orthologs for *Mt-Mj-Af-Ph*.

Specific orthologs

Table 5 shows a possible list of genes which are specific only to a set of pathogens (*B.burgdorferi*, *C.trachomatis*, *H.influenzae*, *H.pylori*, *M.genitalium*, *M.pneumoniae*, *M.tuberculosis*, *R.prowazekii*) and *E.coli*. Note that a gene function specific in a more restricted set of genomes is also specific in a set which includes it. For example, *recB* is specific to *Ec-Bb-Ct-Hi-Tb*, which implies that *recB* is also specific to *Ec-Hi-Tb*. The specific genes have been listed as *group of genomes: set of genes specific to the group of genomes*.

Conclusion

The paper described an automated scheme for the identification of orthologs, orthologous gene groups, multigene groups, duplicated genes, groups of genomes containing orthologs, and genes with conserved functionality in 17 microbial genomes from different families.

The results show that the relative percentage of orthologous genes (compared to genome size) is higher within the same genome family. The result on duplication shows that duplication is a major mechanism for the change in the functionality of genomes.

The result shows that two adjacent genes may fuse together to give a new gene with combined functionality. Archaea share a much smaller number of orthologs with genomes in other families. Many of the genes with conserved functionality are related to mechanism of transcription and translation. The comparison of *E.coli* vs. pathogen genomes shows the presence of orthologs specific to *E.coli* and a set of pathogens which are absent in other genomes. An analysis of orthologs specific to group of pathogens, missing genes corresponding to conserved functions, or variations in genes involved in conserved functions may give clues to genes involved in bacterial diseases.

Acknowledgements

This research was supported in part by a grant by the Research Council at Kent State University and an Ohio Board of Regents Ph.D. enhancement grant. The author acknowledges Warren Gish for providing WU-BLASTP software and Paul Hardy for the portable Smith–Waterman library, and the Liquid Crystal Institute and the OCARnet project at Kent State University for the use of high performance computers. The author acknowledges the researchers whose papers on genome sequencing could not be included due to space limitations. The author thanks two anonymous referees for their thorough and insightful comments.

References

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D. (1990) Basic alignment search tools. *J. Mol. Biol.*, **215**, 403–410.

Bansal,A.K., Bork,P. and Stuckey,P.J. (1998) Automated pair-wise comparisons of microbial genomes. *Math. Model. Sci. Comput.*, **9**, 1–23.

Bansal,A.K. and Bork,P. (1999) Applying logic programming to derive novel functional information of genomes. *First International Workshop of Declarative Languages* Springer Verlag, San Antonio, **LNCS 1551**, pp. 275–289.

Bork,P., Dandekar,T., Diaz-Lazcoz,Y., Eisenhaber,F., Huynen,M.A. and Yuan,Y. (1998) Predicting function: from gene to genomes and back. *J. Mol. Biol.*, **283**, 707–725.

Brenner,S.E., Chothia,C. and Hubbard,T.J.P. (1998) Assessing sequence comparison methods with reliable structurally distant

evolutionary relationships. *Proc. Natl. Acad. Sci. USA*, **95**, 6073–6078.

Fitch,W.M. (1970) Distinguishing homologous from analogous proteins. *System. Zool.*, **19**, 99–113.

Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R. *et al.* (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, 496–512.

Gruber,T.M. and Bryant,D.A. (1997) Molecular systematic studies of eubacteria, using σ^{70} type sigma factors of group 1 and group 2. *J. Bacteriol.*, **179**, 1734–1747.

Huynen,M.A. and Bork,P. (1998) Measuring genome evolution. *Proc. Natl. Acad. Sci. USA*, **95**, 5849–5856.

Olsen,J., Woese,C.R. and Overbeek,R. (1994) The winds of evolutionary change: breathing new life intomicrobiology. *J. Bacteriol.*, **176**, 1–6.

Papadimitrou,C.H. and Steiglitz,K (1982) *Combinatorial Optimization: Algorithm andComplexity*. Prentice Hall.

Selkov,E., Maltsev,N., Olsen,G.J., Overbeek,R. and Whitman,W.B. (1997) A reconstruction of the me tabolism of *Metanococcus jannaschi* from sequence data. *Gene*, **197**, GC11–26.

Snel,B., Bork,P. and Huynen,M.A. (1998) Genome phylogeny based on gene content. *Nat. Genet.*, **21**, 108–110.

Tatusov,R.L., Mushegian,M., Bork,P., Brown,N.P., Hayes,W.S., Borodovsky,M., Rudd,K.E. and Koonin,E.V. (1996) Metabolism and evolution of *Haemophilius influenzae* deduced from a whole genome comparison with *Escherichia coli*. *Curr. Biol.*, **6**, 279–291.

Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.

Vitreschak,A., Bansal,A.K., Titov,I.I. and Gelfand,M.S. (1999) Conserved RNA structures regulation initiation of translation of *Escherichia coli* and *Haemophilus influenzae* ribosomal protein operons. *Biofizika*, **44**, 601–610.

Waterman,M.S. (1995) *Introduction to Computational Biology: Maps, Sequences, and Genomes*. Chapman & Hall, London.