IEEE TRANSACTIONS ON

KNOWLEDGE AND DATA ENGINEERING

A publication of the IEEE Computer Society

JULY/AUGUST 2003 VOLUME 15 NUMBER 4 ITKEEH (ISSN 1041-4347)

Arvind Bansal

SPECIAL SECTION ON WWW2002

Guest Editors' Introduction
A.K. Iyengar and D. De Roure 769

Specifying and Enforcing Application-Level Web Security Policies
D. Scott and R. Sharp 771

Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search
T.H. Haveliwala 784

The Yin/Yang Web: A Unified Model for XML Syntax and RDF Semantics
P.F. Patel-Schneider and J. Siméon 797

Scalable Consistency Maintenance in Content Distribution Networks Using Cooperative Leases
A.G. Ninan, P. Kulkarni, P. Shenoy, K. Ramamritham, and R. Tewari 813

Query Expansion by Mining User Logs
H. Cui, J.-R. Wen, J.-Y. Nie, and W.-Y. Ma 829

Managing and Sharing Servents' Reputations in P2P Systems
E. Damiani, S. De Capitani di Vimercati, S. Paraboschi, and P. Samarati 840

Searching with Numbers
R. Agrawal and R. Srikant 855

REGULAR PAPERS

Artificial Intelligence

An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources
Y. Li, Z.A. Bandar, and D. McLean 871

Bioinformatics

Applying Automatically Derived Gene-Groups to Automatically Predict and Refine Metabolic Pathways
A.K. Bansal and C.J. Woolverton 883

Databases

Buffer Queries
E.P.F. Chan 895

Image Representations and Feature Selection for Multimedia Database Search
T. Evgeniou, M. Pontil, C. Papageorgiou, and T. Poggio 911

Temporal Probabilistic Object Bases
V. Biazzo, R. Giugno, T. Lukasiewicz, and V.S. Subrahmanian 921

Data Mining and Knowledge Discovery

Effectively Finding Relevant Web Pages from Linkage Information
J. Hou and Y. Zhang 940

Peculiarity Oriented Multidatabase Mining
N. Zhong, Y.Y.Y. Yao, and M. Ohshima 952

Data Structures and Algorithms

External Sorting: Run Formation Revisted
P. Larson 961

In-Place Reconstruction of Version Differences
R. Burns, L. Stockmeyer, and D.D.E. Long 973

(Contents continued on back cover)

IEEE COMPUTER SOCIETY
http://computer.org
tkde@computer.org

IEEE

Applying Automatically Derived Gene-Groups to Automatically Predict and Refine Metabolic Pathways

Arvind K. Bansal, *Senior Member*, *IEEE*, and Christopher J. Woolverton

Abstract—This paper describes an automated technique to predict integrated pathways and refine existing metabolic pathways using the information of automatically derived, functionally similar gene-groups and orthologs (functionally equivalent genes) derived by the comparison of complete microbial genomes archived in GenBank. The described method integrates automatically derived orthologous and homologous gene-groups (http://www.mcs.kent.edu/~arvind/orthos.html) with the biochemical pathway template available at the KEGG database (http://www.genome.ad.jp), the enzyme information derived from the SwissProt enzyme database (http://expasys.hcuge.ch/), and the Ligand database (http://www.genome.ad.jp). The technique refines existing pathways (based upon the network of reactions of enzymes) by associating corresponding nonenzymatic and regulatory proteins to enzymes and operons and by identifying substituting homologs. The technique is suitable for building and refining integrated pathways using evolutionary diverse organisms. A methodology and the corresponding algorithm are presented. The technique is illustrated by comparing the genomes of *E. coli* and *B. subtilis* with *M. tuberculosis*. The findings about integrated pathways are briefly discussed.

Index Terms—Automation, bacteria, drug-discovery, enzymes, gene-groups, homologs, metabolic pathway, microbes, operons, orthologs, pathogenicity, pathway.

1 Introduction

Understanding microbial machinery has important advantages in bioremediation, management of environmental waste, efficient utilization and generation of energy, and development of more effective antibiotics based upon the regulation of metabolic pathways. Currently, 48 microbial genomes (including many pathogens) have been sequenced and archived, with more than 182 underway. Important aspects of understanding microbial structure and function relationships from DNA sequence data are 1) to derive the metabolic pathways (the network of enzymatic reactions), 2) to understand the regulation of these pathways, and 3) to determine the effect of gene expression on interconnected biochemical reactions. By blocking a part of a metabolic pathway or by regulating its gene-expression, the production of a specific biochemical product can be controlled. The analysis of pathway regulation will help us to understand the dynamic behavior of biochemical metabolism and to study the rate of change of specific biochemical products. Furthermore, the study of the dynamic cellular behaviors dictated by metabolic pathways will facilitate the development of more effective antibacterial drugs with few side effects [19]. However, in order to understand the metabolic mechanisms, distinct biochemical pathways need to be mapped accurately.

- *A.K. Bansal is with the Department of Computer Science, Kent State University, Kent, Ohio 44242. E-mail: arvind@cs.kent.edu, and Intellibio Software and Consultancy Corp., 3109 Killingworth Lane, Twinsburg, OH 44087. E-mail: arvind@intellibiosoft.com.*
- *C.J. Woolverton is with the Department of Biological Sciences, Kent State University, Kent, Ohio 44242. E-mail: cwoolver@kent.edu.*

Manuscript received 10 Jan. 2002; revised 7 Feb. 2002; accepted 21 Feb. 2002. For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number 115677.

Previous techniques to derive metabolic pathways [6], [11], [13], [17], [18], [20] are based upon identifying enzymes in newly sequenced genomes and building a network of enzyme reactions by matching the output of one biochemical reaction with the input of another biochemical reaction. In order to identify the enzymes, two techniques have been used:

1. Newly sequenced genomes are compared against a curated protein database such as SwissProt [2] or Ligand [1] to identify the similar enzymes and their corresponding reactions [6].
2. Newly sequenced genomes are compared against evolutionary close genomes to identify the functionally equivalent enzymes and groups of neighboring genes [20].

Both the schemes have advantages and limitations:

1. The first scheme benefits from a well-curated database (including enzymes other than completed genomes) to identify enzymes. However, the scheme is limited by the fact that it does not take into account the information relating genes being coregulated with a cluster of neighboring genes in the same pathway.
2. The second scheme takes advantage of coregulation of genes clustered together in gene-groups. The advantage of this scheme over the first scheme is that it can also identify, in a restricted sense, some of the regulatory proteins associated with metabolic pathways. However, it is very restrictive because it uses wavelets of gene-groups of adjacent ordered orthologous enzymes (cluster of adjacently occurring

enzymes with best similarity) and it uses the comparison of only evolutionary close genomes, assuming all the gene-groups in the same metabolic pathway should be clustered closely. The use of orthologous gene-groups is quite restricted since enzymes have multiple domains and a coregulated enzyme in a gene-group (having a subset of domains necessary to catalyze the reaction needed in a pathway) will substitute for the putative orthologous (functionally equivalent) enzyme.

In this paper, we relax the above restrictions and use automatically derived gene-groups from a spectrum of genomes to build integrated pathways. We take full advantage of both orthologous (having the same function) and homologous (having similar function) gene-groups. Since the genes in gene-groups have a natural pressure to occur together, it is highly probable that these genes are regulated or expressed together. The use of orthologous and homologous gene-groups is a rich source for identifying similar metabolic pathways as these gene-groups consist of operons (a group of genes involved in a common function within the same metabolic pathway module) and coregulated genes possibly sharing the same control region. This paper also builds upon previous results that evolutionary diverse genomes share a large number of homologous and orthologous gene-groups [4], where the number of gene-groups is also a function of number of genes in each respective genome. This finding has significant impact since a newly sequenced genome can be compared with a large number of genomes to identify almost all its gene-groups. Since our paper uses the network of enzymes developed in Scheme 1, we integrate the advantages of both Schemes 1 and 2, in addition to the advantages gained by relaxing the conditions described above.

We illustrate our techniques using genome comparison of *Escherichia coli* and *Bacillus subtilis* (two genomes extensively investigated in wet laboratories) with *Mycobacterium tuberculosis*. The results presented in this paper show that orthologous gene-groups and a large class of homologous gene-groups obtained from pair-wise genome comparisons map to local sections of metabolic pathways. These local sections of a metabolic pathway become the seed points to trace the metabolic pathway for newly sequenced genomes and these seed points can be explored by the wet-lab scientists to identify new metabolic pathways, especially for classes of organisms living in extreme environmental conditions that have less similarity with previously explored classes of bacteria.

The major contributions of this paper are:

1. It establishes that all the genes of a homologous gene-group belong to the local connected sections of the same metabolic pathway in a genome.
2. Current techniques to reconstruct metabolic pathways have been refined and extended by using a comprehensive set of genomes to identify gene-groups using pair-wise genome comparisons.
3. The current metabolic pathways derived by ortholog analysis (as best homolog) and biochemical reactions have been curated using coregulated homologs since a coregulated enzyme having the necessary domain for catalyzing a reaction substitutes the best homolog.
4. Nonenzymatic genes embedded inside a group of enzymes have been identified. These genes may be involved in the regulation of the gene-group or may be affected in some way by the reactions in the gene-group.

The database of pathway traces and their classification resulting from this research will help the wet lab scientists to identify unannotated enzymes in genomes, by looking for missing enzymes [9] in a pathway trace. Another advantage is that the association of gene-groups with metabolic pathway modules will improve the efficiency of the metabolic pathway deduction.

The paper is organized as follows: Section 2 describes the background, corresponding gene-groups [3], [4], and enzymes and metabolic pathways [2], [11], [18]. Section 3 describes briefly the techniques to derive orthologs and homologous gene-groups [3], [4]. Section 4 describes a technique and algorithms to reconstruct the metabolic pathways integrating homologous gene-groups, chemical reactions of enzymes (as given in the Ligand and SwissProt databases), and the standard template of the metabolic pathway available at the KEGG site. Section 5 describes the results to substantiate the hypothesis (using well-explored genomes in wet labs) that the orthologous and homologous gene-groups mark a large subset of the local sections in a metabolic pathway and can derive unique pathways of *M. tuberculosis.* The last section concludes the paper.

2 Background and Definitions

In this section, we describe background concepts related to genome comparison, orthologs, different classes of gene-groups, enzymes, metabolic pathways, and new definitions related to the automated deduction of metabolic pathways.

2.1 Modeling Genomes and Pair-Wise Genome Comparisons

A genome is modeled as an ordered set of genes. A pair-wise genome comparison is performed by matching the sequences of the corresponding proteins. There may be more than one different subsequence in a protein which is homologous to subsequences of the corresponding proteins in another genome. These subsequences include one or more protein domains.

An *ortholog* is a functional counterpart of a gene in another genome that has arisen from speciation [10]. This paper uses a definition based upon sequence similarity due to the inherent uncertainty in phylogeny resulting from a lateral transfer of genes [3], [4], [12], gene insertions and deletions, gene fusion and splitting [3], and a difference in the evolutionary trees based upon various criteria. A *putative ortholog* is defined as a gene in the second genome such that the corresponding protein has either a unique matching or the best similarity score (above a threshold), with the corresponding protein of a gene in first genome.

A *gene-group* is a cluster of neighboring genes (not necessarily adjacent) with at least two distinct genes that have a natural pressure to occur in close proximity. A *corresponding gene-group* in the second genome is identified

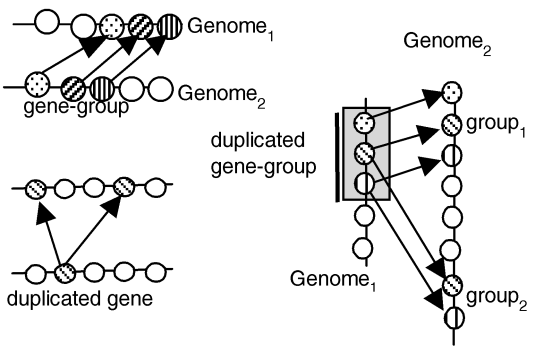

Fig. 1. Gene-groups classification.

by marking the neighboring protein sequences which are homologous to neighboring protein sequences of the first genome. A gene-group is *ordered* if the homologous genes within two corresponding gene-groups occur in the same order. The study of ordered gene-groups is important as it identifies more restricted regulation mechanisms within the cluster of genes in the gene-groups. A *duplicated gene group* in a genome has more than one corresponding disjointed gene groups in the other genome. The presence of duplicated gene-groups may provide fault tolerance to survive extreme environmental conditions. A duplicated gene contains a gene-fragment that has at least two homologous gene-fragments in the other genome such that both fragments are in close proximity (see Fig. 1). The biochemical rationale for this definition of duplicated genes is that duplicated genes may be coregulated or may be involved in the same pathway. Gene duplication may provide fault tolerance against inactivation of one of the genes due to mutation, may be responsible for extra gene-expression, or may provide minor functional variation.

Orthologous gene-groups are comprised of only orthologous genes. The study of orthologous gene-groups is important to annotate the function of gene-groups in genomes and to identify operons (a common functional unit in a pathway involved in a composite reaction or set of coregulated genes).

2.2 Enzymes and Pathways

An enzyme is a protein catalyzing a biochemical reaction. A metabolic pathway is a network of reactions catalyzed by enzymes to transform (break up or synthesize) proteins and other biochemical products. A metabolic pathway is modeled as a directed graph with biochemical products as the nodes and the enzymes as the edges. A metabolic pathway is comprised of enzymes as well as regulatory genes interacting with the corresponding local sections of the pathway.

Complete metabolic pathways have been partitioned into multiple *modules* in the KEGG database (http://www.genome.ad.jp). Some examples of these modules are amino acid biosynthesis, glycolysis, purine and pyrimidine metabolism, folate metabolism, fatty acid biosynthesis, biosynthesis of lipids, the citrate cycle, the CO_2 fixation cycle, and the urea cycle [6], [13], [18].

A reaction-graph models a network of enzyme reactions within a pathway such that enzyme reactions are modeled as directed edges, the substrate (catalyzed by the enzyme reaction) is modeled as a source node, and the product is modeled as a sink node.

A *pathway seed* is a gene-group (identified using pair-wise genome comparison) consisting of at least two enzymes modeling a chain (or coregulated/cotranscribed group) of enzyme reactions in the same pathway or adjacent pathway modules. A pathway seed will be modeled as a set of enzymes.

A *pathway trace* is a set of proteins in a pathway module such that the corresponding enzyme reaction-graph is included within a metabolic pathway module. A pathway trace contains at least one pathway seed. For example, enzymes (*EC 4.1.3.8, EC 1.1.1.86, EC 4.2.1.9, EC 2.6.1.42*) in the metabolic pathway module (see the KEGG database) "valine, leucine, and isoleucine biosynthesis" form a pathway trace.

3 Deriving Orthologs and Gene-Groups

This section briefly describes an algorithm used to identify orthologs, homologous and orthologous gene-groups [3], [4]. The algorithm uses a pair-wise comparison of genomes. To illustrate the examples, we have denoted genes as γ_M^I ($I > 1$ and $M > 1$), where the superscript "I" denotes the genome number and the subscript "M" denotes a generic position of a gene within the genome.

3.1 Identifying Orthologs

Orthologs are identified as the best matching genes in two genomes using pair-wise genome comparison. To identify the best homologs, each genome is modeled as an ordered set of genes and the pair-wise comparison is modeled as a weighted bipartite-graph matching problem [3], [4] where edges show the similarity score between the amino acid sequences of the corresponding genes. The weights of the edges are identified using the Smith-Waterman algorithm [22]. In order to improve the execution efficiency, dissimilar gene-pairs are pruned based on BLAST similarity techniques [1].

After identifying the weights of the edges, the edges are sorted in descending order of the weights. The set of nodes corresponding to the highest weighted edges (best homologs) are collected as putative orthologs. After finding an edge (γ_I^1, γ_J^2) of the highest weight, all the edges involving the nodes γ_I^1 and γ_J^2 are deleted. The process is continued until there are no more edges. The edges starting or ending in genes inside a gene group are biased positively since genes within a gene group are better candidates for preserving a common function within a pathway. Two edges with close weights, if two weights are above a threshold, suggest multiple orthologs or gene-fusion. A detailed algorithm is given in [3].

3.2 Identifying Gene Groups

A set of neighboring genes S_0 for a gene (in Genome_1) that has a corresponding homolog in Genome_2 is marked. Then, a set S_1 (in Genome_2) of homologs for S_0 is marked. Then, the set S_2—a union of all sets of homologs of S_1—in Genome_2 is marked. The presence of two or more common elements in the sets S_0 and S_2 marks the start of a

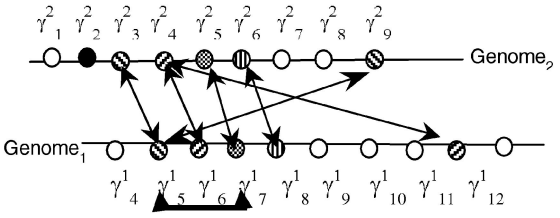

Fig. 2. Identifying corresponding gene-groups.

homologous gene-group. After marking the start of homologous gene-groups, Genome_1 is traversed one node at a time, checking for the presence of an edge in close proximity of the last homologous gene in Genome_2. The method identifies gene groups of any variable size. A detailed algorithm is available in [3].

Example 1. A schematic is given in Fig. 2. We consider the proximity for the neighborhood as 1. The current gene being scanned is gene γ_5^1. The genes γ_3^2 and γ_9^2 are homologs of the gene γ_5^1. The corresponding neighboring set of genes S_0 is $\{\gamma_5^1, \gamma_6^1, \gamma_7^1\}$. The set of corresponding homologs S_1 in Genome_2 is $\{\gamma_3^2, \gamma_4^2, \gamma_5^2, \gamma_9^2\}$. The set of homologs S_2 in Genome_1 corresponding to genes in the set S_1 is $\{\gamma_5^1, \gamma_6^1, \gamma_7^1, \gamma_{12}^1\}$. The set of genes $\{\gamma_5^1, \gamma_6^1, \gamma_7^1\}$ is common to both S_0 and S_2. Since the size of the common set is 3, the gene γ_5^1 becomes the starting point to derive a gene-group. Genome_1 is scanned from the gene γ_5^1 while marking the corresponding neighboring homologs in the Genome_2 and collecting the bag of genes in Genome_1 until no homolog is found in the neighborhood in Genome_2. In this particular case, the gene γ_5^1 is a homolog of the gene γ_3^2, the gene γ_6^1 is a homolog of the gene γ_4^2, the gene γ_7^1 is a homolog of the gene γ_5^2, and the gene γ_8^1 is a homolog of the γ_6^2. The gene-group derived is $\{\gamma_5^1, \gamma_6^1, \gamma_7^1, \gamma_8^1\}$.

4 Identifying Pathway Traces Using Gene-Groups

In this section, we describe a technique and abstract algorithms to identify pathway traces and illustrate the abstract algorithms using examples from *M. tuberculosis*. Algorithms have been described abstractly for the sake of wide readership. Section 4.1 describes a technique, an abstract algorithm, and an illustration to merge adjacent seeds to identify basic pathway traces. Section 4.2 describes a technique and an abstract algorithm to merge pathway-traces identified by single pair-wise genome comparisons and pathway-traces identified by multiple pair-wise genome comparisons.

Genomes were extracted from the "Genbank" using the .gbk file format. The gene-groups and orthologs were obtained from pair-wise comparisons of genomes using Goldie 4.0 (a software library for automated comparison of genomes) [4]. The genome of *E. coli* was compared with the genome of *B. subtilis*, the genome of *M. tuberculosis* was compared with the genome of *E. coli*, and the genome of *M. tuberculosis* was compared with the genome of *B. subtilis*.

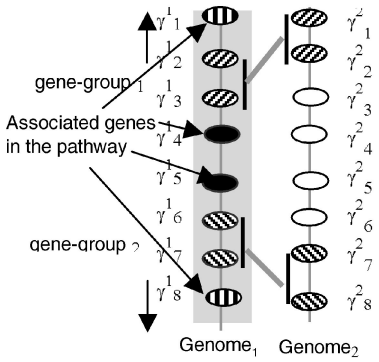

Fig. 3. Identifying a pathway trace.

4.1 Pathway Traces from One Genome-Pair

To identify pathway traces, all the gene-groups identified by the genome-pair comparisons were identified. These gene-groups were sorted in the order of *ordered orthologous gene-groups*, *ordered homologous gene-groups*, *unordered orthologous gene-groups*, *unordered homologous gene-groups*, *gene-groups having multiple duplicated domains*, *duplicated gene-groups*, and *gene-groups with fused domains*. The rationale for this order is that orthologous gene-group correspondences are functionally equivalent and more constrained units and represent strong candidates for operons. Unordered orthologous and homologous groups have putative coregulated domains. Gene-groups with multiple duplicated domains and duplicated gene-groups may vary in the gene-expression due to domain duplication, and may vary in function due to insertion or deletion of one or more genes.

From this set of gene-groups, each gene-group is identified as a seed point iteratively and the corresponding pathway traces are identified. To identify a pathway trace, the neighborhood of the seed point is determined by extending the indices of the two ends of a seed by a user-defined size (in this case 1). If any gene-group resides in the neighborhood and is within the same or adjacent module, then it is merged with the current group and all the embedded genes are included in this larger gene-group. The biochemical rationale for this type of merging results from the instances of insertion and deletion of genes observed in metabolic pathways. Thus, a larger group may appear to split in two groups. This process of merging gene-groups is repeated until there are no more gene-groups in the neighborhood. All the seeds that have been included in the larger group are removed from the set of seeds and the larger seed is inserted in the set of seeds for further processing. The process is repeated until the set of gene-groups is empty. The final set of extended seeds is associated with the corresponding pathway module and this set of pairs of the form (*module-name, pathway-trace*) is passed on to the next stage (see Section 4.2) where pathway traces (within same or adjacent modules) from single pair-wise genome comparisons and multiple genome comparisons are merged together. Some pathway traces occur in two or more adjacent modules since these adjacent modules can be connected through a network of reactions. In such cases, the pathway trace is listed under all the adjacent modules. The overall schematic to identify pathway traces is explained by Fig. 3 and an algorithm is described in Fig. 4.

Algorithm identify pathway-traces;
Input: S_0 — a set of seeds from a comparison of genomes;
Output: A set T of (module-name, set of pathway traces);

{ let set of pathway traces S_4 be an empty set;
S_1 = sort S_0 (the set of seeds) in the order of gene position within the genome being analyzed;
S_2 = the set of seeds after extending the lower and higher boundaries in the set S_1;
while (S_2 is not empty)
{ let s be the current extended seed in S_2;
remove the seed s from S_2;
S_3 = set of adjacent seeds in S_2 having a common gene with the current seed s;
if S_3 is empty then insert the extended seed s in S_4;
else {remove all elements of S_3 from S_2;
s' = extended seed consisting of all the genes in S_3 and the extended seed s in the set S_2;
insert s' in the set S_2;}
}
T = rearrange S_4 as a set of (module-name, set of pathway-traces in the same module);
}

Fig. 4. Identifying pathway traces.

Example 2. In Fig. 3, the gene-group (γ_1^2, γ_2^2) and gene-group (γ_7^2, γ_8^2) in Genome_2 form a neighboring cluster of gene-groups $\{(\gamma_2^1, \gamma_3^1), (\gamma_6^1, \gamma_7^1)\}$ in Genome_1. The set of genes $\{\gamma_4^1, \gamma_5^1\}$ embedded in this cluster and neighboring genes (γ_1^1, γ_8^1) at the end of the cluster are also included. This identifies the cluster $(\gamma_1^1, \gamma_2^1, \gamma_3^1, \gamma_4^1, \gamma_5^1, \gamma_6^1, \gamma_7^1, \gamma_8^1)$ in Genome_1 as a pathway trace.

4.2 Merging Pathway Traces

For a genome, the sets of pathway traces are computed using genome comparisons with other completed genomes. Once all the pathway traces are identified, our computational experiment demonstrates that the pathway traces making up a pathway submodule are scattered on different parts of the genomes. In addition, the traces, identified from different genome comparisons, may vary. Consolidation of all the traces in a module will give a smaller set of merged yet larger pathway traces. There are two types of merging, as follows:

Rule 1. Pathway traces containing one common node are merged. In biochemical terms, a product in one pathway trace is consumed in the second pathway trace or the product may be produced by two different reactions or a substrate may be consumed by two different reactions. This type of merging is common in single genome-pair comparisons when two different pathway traces are obtained from two different seed points occurring at different positions in the genome. For example, in the bottom illustration of Fig. 5, the set of substrates in the first pathway trace is $\{n_1, n_2\}$, the set of products in the first pathway trace is $\{n_2, n_3\}$. The set of substrates in the second pathway trace is $\{n_2\}$ and the set of products in the second pathway trace is $\{n_4\}$. Since n_2 is

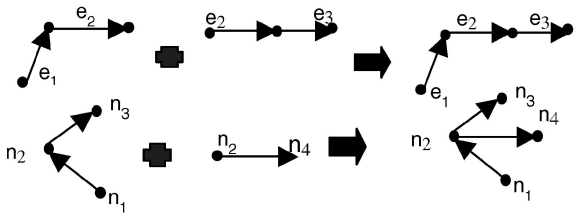

Fig. 5. Merging pathway traces.

shared between the two pathway traces, the two pathway traces are merged.

Rule 2. Pathway traces containing one common edge are merged together. A common edge suggests the same reaction involving the same substrate and the same product catalyzed by the same enzyme. The biochemical rationale is that pathway traces obtained from different genome-pair comparisons vary due to missing genes in different genomes. This difference leads to pathway variations, and only a joint picture obtained from multiple genome comparisons can give a better approximation of the complete pathway of the genome being investigated. For example, In the top illustration of Fig. 5, the first pathway trace $\langle e_1, e_2 \rangle$ merges with the second pathway trace $\langle e_2, e_3 \rangle$ to give a merged pathway trace $\langle e_1, e_2, e_3 \rangle$.

To merge pathway traces, the set of pathway traces derived by the algorithm described in Section 4.1 is sorted and grouped by the same module name. Within each module, the set of pathway traces is sorted in descending order of the number of enzymes contained in the pathway traces. The rationale is that larger pathway traces are likely to merge with a larger number of other pathway traces. Pathway modules are processed one at a time. The largest pathway trace in a given module is removed from this sorted set and processed. Those pathway traces which share an enzyme (or share a compound being consumed or produced) with the current pathway trace are merged with the current pathway trace, the merged pathway traces are removed from the sorted set of pathway traces, and the resulting merged pathway trace is inserted again in the set of the sorted set of pathway traces. In the absence of any pathway trace that could be merged with the current pathway trace, the current pathway trace is added to the set of final pathway traces. The process is repeated until no more pathway traces are left in the sorted set of pathway traces. This process is repeated for every module that contains a pathway trace. An abstract algorithm is given in Fig. 6. The comments are preceded by the symbol "%."

Example 3. This example illustrates the algorithms given in Figs. 4 and 6 using "purine metabolism pathway" in *M. tuberculosis* genome. The seeds derived by the comparisons of *M. tuberculosis* and *E. coli* are given in Table 3 and the seeds derived by the comparison of *M. tuberculosis* and *B. subtilis* are given in Table 4.

Step 1: Identifying Seeds. The algorithms to identify orthologs and corresponding gene-groups [3], [4] derive the pathway seeds for *M. tuberculosis* using comparisons with *E. coli* and *B. subtilis* as shown in Tables 3 and 4. The corresponding set of seeds for "purine metabolism" derived from the comparison with *E. coli* genome is

```
Algorithm merge-pathway-traces
Input: S0 — A set of (module-name, pathway trace);
Output: T2 — A set of merged pathway traces;
{ initialize T1 and T2 to an empty set;
  for (every genome-pair comparison)
  { S1 = rearrange S0 as a set of (module, all pathway traces
         in the module );
    for (every associated module in S1)
    {   S2 = set of pathway traces in the module sorted in
             descending order of size (number of enzymes);
        while (S2 is not empty) % apply Rule 1
        {   pick the largest pathway trace p in S2;
            S3 = set of pathway traces in S2 sharing a
                 common product/substrate with p;
            remove the pathway trace p from S2;
            if (S3 is empty) insert p in T1;
            else{p' = a new merged pathway trace with all the
                      genes in S3 and the pathway trace p;
                 remove the pathway traces in S3 from S2;
                 insert p' in the set S2; }
        } % recycle
    }
  }
  while (T1 is not empty) % Rule 2 for multiple genome-pairs
  {   select the largest pathway trace p in T1;
      S4 = set of pathway traces in T1 sharing a common
           enzyme and the corresponding product substrate
           with p;
      remove p from the set T1;
      if (S4 is empty) insert p in T2;
      else {p' = a new merged pathway trace with all the
                 genes in S4 and the pathway trace p;
            remove S4 from T1;
            insert p' in the set S4; } % recycle
  }
}
```

Fig. 6. Merging pathway traces.

{(*rpoB:2.7.7.6, rpoC:2.7.7.6*), (*gmk: 2.7.4.8, dfp*) (*purE:4.1.1.21, purK:4.1.1.21*) (*recR, Rv3716c, dnaZX:2.7.7.7*)}, and the corresponding set of seeds derived from the comparison with *B. subtilis* genome is {(*rpoB:2.7.7.6, rpoC:2.7.7.6*), (*purB:4.3.2.2, purC:6.3.2.6, purQ:6.3.5.3*), (*purL:6.3.5.3, purF:2.4.2.14, purM:6.3.3.1*), (*purN:2.1.2.2, purH:2.1.2.3*), (*gmk:2.7.4.8, Rv1390:, dfp, priA*), (*ureA:3.5.1.5, ureB:3.5.1.5, ureC:3.5.1.5*), (*Rv3273, purE: 4.1.1.21,purK: 4.1.1.21*), (*hpT: 2.4.2.8, yacA*), (*recR, Rv3716c, dnaZX:2.7.7.7*), (*relA: 2.7.6.5, apt: 2.4.2.7, secF, secD, ruvB, ruvA, Rv2603c*)}.

Step 2: Extending Pathway Seeds to Derive Pathway Traces. This step uses the algorithm as described in Fig. 4. Neither the seeds derived using the comparison with *E. coli* genome nor the seeds derived using the comparison with *B. subtilis* genome can be merged by simple seed extension. Each seed is associated with the corresponding module name in *E. coli* or *B. subtilis* (using KEGG database) and passed to Step 3.

Step 3: Merging Pathway Traces within the Same Genome. An analysis from the KEGG database shows that none of the seeds generated using the *E. coli* genome can be merged together using Rule 1. An analysis of the pathway traces of *B. subtilis* (see KEGG database) demonstrates that the seeds (*gmk:2.7.4.8, Rv1390, dfp, priA*) and (*relA:2.7.6.5, apt:2.4.2.7, secF, secD, ruvB, ruvA, Rv2603c*) share a common product (substrate) "GMP" (Guanosine 5'-phosphate) and are merged to give a trace (*gmk:2.7.4.8, Rv1390, dfp, priA, relA:2.7.6.5, apt:2.4.2.7, secF, secD, ruvB, ruvA, Rv2603c*). The seeds (*purN:2.1.2.2, purH:2.1.2.3*) and (*purB:4.3.2.2, purC:6.3.2.6, purQ:6.3.5.3*) are merged to give a trace (*purN:2.1.2.2, purH:2.1.2.3, purB:4.3.2.2, purC:6.3.2.6, purQ:6.3.5.3*). After the complete merger (using Rule 1), the set of merged pathway traces derived by *B. subtilis* is {(*purN:2.1.2.2, purH:2.1.2.3, purB:4.3.2.2, purC:6.3.2.6, purQ:6.3.5.3, purL:6.3.5.3, purF:2.4.2.14, purM:6.3.3.1, Rv3273, purE: 4.1.1.21, purK: 4.1.1.21, relA:2.7.6.5, apt:2.4.2.7, secF, secD, ruvB, ruvA, Rv2603c, gmk:2.7.4.8, Rv1390:, dfp, priA, hpT: 2.4.2.8, yacA, rpoB:2.7.7.6, rpoC:2.7.7.6*), (*recR, Rv3716c, dnaZX:2.7.7.7*), (*ureA:3.5.1.5, ureB:3.5.1.5, ureC:3.5.1.5*)}.

Step 4: Merging Using Rule 2. The set of merged pathway traces from *E. coli* and *B. subtilis* is merged again if the two pathway traces share at least one enzyme name with same substrate and product. All the pathway traces derived using the genome of *E. coli* are included in one or the other pathway trace derived from the genome of *B. subtilis*. The resulting set of pathway traces is the same as merged pathway traces derived (in Step 3) using the genome of *B. subtilis*.

4.3 Refining Existing Pathways Using Gene-Groups

In this section, we apply the technique (to identify pathway traces) to refine proposed metabolic pathways by integrating pathways with associated nonenzymatic proteins. The technique is based upon mapping the identified pathway traces on the existing metabolic pathway templates derived using a network of reactions as given in the KEGG database (see http://www.genome.ad.jp/kegg-bin/mk_point_html). The refinement was done in three steps as follows:

Step 1. The putative pathway traces are derived by comparing a genome with other known genomes. The traces derived from individual genome comparisons were merged to form a unified set of traces.

Step 2. The standard template of a metabolic pathway (for the genome in question) was procured from the KEGG database (http://www.genome.ad.jp/kegg-bin/mk_point_html). The traces containing the enzymes were mapped using the *EC* numbers of the genes in the gene-groups and the locations of the enzymes near the end-points of the putative pathway trace were detected. Any gene without an enzyme number in the trace was matched against the enzymes in the KEGG pathway using the gene function. The function of the gene was identified from the Genbank annotations, SwissProt enzyme annotations, and Ligand database annotations using the gene-name match (if given). If gene function is matched in the SwissProt enzyme database or Ligand database, then the corresponding *EC* number was assigned to the unannotated gene as a putative enzyme number. If the function did not match exactly, the function of the generic class of the enzymes in the

TABLE 1
Pathway Seeds Using *E. coli* versus *B. subtilis*

Citrate cycle: (sdhA:1.3.99.1, sdhB:1.3.99.1), (sucA:1.2.4.2, sucB:2.3.1.61), (sucC: 6.2.1.5, sucD:6.2.1.5)
Oxidative phosphorylation: (sdhA:1.3.99.1, sdhB:1.3.99.1)
Butanoate metabolism: (sdhA:1.3.99.1, sdhB:1.3.99.1)
Propanoate metabolism: (sucC: 6.2.1.5, sucD:6.2.1.5)
Lysine degradation: (sucA:1.2.4.2, sucB:2.3.6.1)
Urea cycle and metabolism of amino groups: (proB: 2.7.2.11, proA:1.2.1.41), (argC:1.2.1.38, argB:2.7.2.8)
Reductive carboxylate cycle: (*sdhA:1.3.99.1*, *sdhB:1.3.99.1*), (*sucC: 6.2.1.5, sucD: 6.2.1.5*), (***yjcG***, *acs:6.2.1.1*)
Phenylalanine, tyrosine, and tryptophan biosynthesis: (*pheT: 6.1.1.20, pheT: 6.1.1.20*)
Purine metabolism: (*guaA:6.3.4.1*, *guaB:1.1.1.205*), (*purM:6.3.3.1*, *purN:2.1.2.2*), (*prsA:2.7.6.1*, ychB), (*purD:6.3.4.13*, purH:2.1.2.3),
Sulfur metabolism: (*cysC:2.7.1.25*, cysH:2.8.2.-),
Peptideglycan biosynthesis: (***yabB***, ***yabC***, ***ftsI***, murE:6.3.2.13, murD: 6.3.2.9, murF:6.3.2.15, mraY: 2.7.8.13, ***ftsW***, murG:2.4.1.-, murC: 6.3.2.8/6.3.2.13)
Flagellar assembly: (fliE, fliF, fliG, fliI, fliJ, fliM, fliN, flip, fliQ, fliR)
Pentose and glucuronate Interconversions: (araD:5.1.3.4, araA:5.3.1.4, araB:2.7.1.16), (lyxk:2.7.1.16/2.7.1.53, yiaS:5.1.3.4), (***orf.2711***, ygcE:2.7.1.17)
Histidine metabolism: (hisG:2.4.2.17, hisD:1.1.1.23, hisB:4.2.1.19, hisH:2.4.2.-, hisA:5.3.1.16, hisF:2.4.2.-, hisI:3.5.4.19)
Valine, leucine, and isoleucine biosynthesis: (leuC:4.2.1.33, leuB:1.1.1.85, leuA:4.1.13.2)
Folate biosynthesis: (folP:2.5.1.15, hflB:3.4.24.-)
Fatty acid biosynthesis: (***plsX***, fabD:2.3.1.39, fabG:1.1.1.100, ***acpP***)
Terpenoid biosynthesis: (***orf.412***, ispA:2.5.1.10, xseB:3.1.11.6)
Alanine and aspartate metabolism: (argH:4.3.2.1, ***oxyR***),
Cs-branched dibasic acid metabolism: (sucC:6.2.1.5, sucC:6.2.1.5)
Phospholipid degradation: (glpQ:3.1.4.46, ***glpT***)

TABLE 2
A Subset of Pathway Traces in *E. coli*

Citrate cycle: (sdhA:1.3.99.1, sdhB:1.3.99.1, sucC: 6.2.1.5, sucD:6.2.1.5, sucB:2.3.1.61, sucA:1.2.4.2)
Peptideglycan biosynthesis: (***yabB, yabC***, ***ftsI***, murE:6.3.2.13, murF: 6.3.2.15, mraY:2.7.8.13, murD:6.3.2.9, ***ftsW***, murG:2.4.1.-, murC:6.3.2.8, ddlB:6.3.2.4***, ftsQ, ftsA, ftsZ***)
Flagellar assembly: (FliE, FliF, FliG, FliH, FliI, FliJ, FliK, FliL, FliM, FliN, FliO, FliP, FliQ, FliR, ***rcsA, dsrB***)
Urea cycle and metabolism of amino groups: (ppc: 4.1.1.31, argH:4.3.2.1, **oxyR**, argE:3.5.1.16, argC: 1.2.1.38, argB: 2.7.2.8, 2.3.1.1, udhA:1.-.-.-, proB:2.7.2.11, proA:1.2.1.41)
Reductive carboxylate Cycle: (sdhA:1.3.99.1, sdhB:1.3.99.1, sucC: 6.2.1.5, sucD: 6.2.1.5)
Purine metabolism: (prsA:2.7.6.1, upp:2.4.2.9, 6.3.4.13, 2.1.2.2, 6.3.5.3, purM:6.3.3.1, ppk: 2.7.4.1, ppx: 3.6.1.11, ***orf.2454, orf.2455, orf.2456***, ***orf.2457***, guaA:6.3.4.1, guaB:1.2.1.4, xseA:3.1.11.6)
Pentose and glucuronate interconversions: (xylA: 5.3.1.5, xylB: 2.7.1.17, yiaS: 5.1.3.4, lyxK: 2.7.1.5.3, araD: 5.1.3.4, araA: 5.1.3.4, araB: 2.7.1.16)
Histidine metabolism: (hisL, hisG:2.4.2.17, hisI:3.5.4.19, hisA:5.3.1.16, hisH:2.4.2.-, hisB:4.2.1.19, hisC: 2.6.1.9, hisD:1.1.1.23, hisF:2.4.2.-)
Fatty acid biosynthesis: (***plsX***, fabD:2.3.1.39, fabH:2.3.1.41, fabG:1.1.1.100***, acpP***, fabF: 2.3.1.41***, pabC***)

pathway trace were identified and was matched with the function associated with the gene-name. If the generic description matched, then the enzyme in the pathway trace was replaced by the generic enzyme class.

Step 3. After mapping the genes in a pathway trace, the pathway trace was extended by identifying the enzymes in the neighborhood of the pathway trace in the genome and the enzymes in the same or adjacent metabolic pathways. The extended pathway trace in the standard pathway template was matched until the *EC* number of the genes in the pathway trace and the *EC* number in the metabolic pathway module did not match. This extended pathway trace was recorded.

5 Results and Discussion

The predicted gene products can be roughly traced to common metabolic pathways in which known or similar products have been identified. A metabolic pathway can be deduced from linear DNA sequence data using an automated system. The automated system described in this paper compares sequence data between two or more bacterial species and predicts potential pathway processes. Application of the technique by the proposed algorithm used to reconstruct metabolic pathways predicts several gene-groups that can serve as pathway seeds. Association of pathway enzymes or products with the same or similar proteins, derived in wet labs, results in the determination of specific sequences (gene-groups) of known biological function.

Furthermore, the pathway trace analysis is meaningful when comparing predicted enzymes with biochemical functions required in a specific pathway. For example, oxidoreduction, ligation, etc. are required functions in specific pathways. The presence of the gene predicting the enzyme in the pathway trace lends validity to the technique and the algorithm used for data acquisition.

In this section, we first verified the technique by comparing *Escherichia coli* (a well-studied proteobacterium) and *Bacillus subtilis* (a Gram-positive bacterium) genomes. Subsequently, we derived a set of pathway traces for the pathogenic bacterium *Mycobacterium tuberculosis*. To identify the metabolic pathways of *M. tuberculosis*, we compared *E. coli* with *M. tuberculosis* and *B. subtilis* with *M. tuberculosis*. Since there are large numbers of corresponding gene-groups [4], the data are limited to describing a few significant pathway traces identified by gene-group comparisons.

5.1 Hypothesis Verification

The results for verifying the hypothesis (using *E. coli* and *B. subtilis* genome comparison) are summarized in Tables 1 and 2. Table 1 describes a limited representative subset of

TABLE 3
Pathway Seeds Using *E. coli*

Pyruvate metabolism (ptA:2.3.1.8, ackA:2.7.2.1)

Purine metabolism: (rpoB:2.7.7.6, rpoC:2.7.7.6), (gmk: 2.7.4.8, ***dfp***) (purE:4.1.1.21, purK:4.1.1.21) (***recR, Rv3716c***, dnaZX:2.7.7.7)

Pyrimidine metabolism: (rpoB:2.7.7.6, rpoC:2.7.7.6), (pyrB:2.1.3.2 ,pyrC: 3.5.2.3) (***recR, Rv3716c***, dnaZX:2.7.7.7), (gpsI:2.7.7.8, rpsO, truB:4.2.1.70)

Cytrate cycle: (sucC:6.2.1.5, sucD:6.2.1.5)

Glycolysis / gluconeogenesis: (***Rv1021***, eno:4.2.1.11) (***rob***, gpmB:5.4.2.1)

Aminoacyl trna-biosynthesis: (fmt:2.1.2.9, ***fmu***)

Riboflavin metabolism: (***ribG***, ribH:2.5.1.9) (***Rv1410c***, ribC:2.5.1.9), (cobT:2.4.2.21, ***cobS***)

Glyoxylate + Carbon fixation: (gap:1.2.1.-, pgk:2.7.2.3), (***tal,*** tkt:2.2.1.1)

Glycine, serine, and threonine metabolism: (***Rv1516c***, pks5: 1.1.1.103), (thrA:2.7.2.4, thrC: 4.2.99.2, thrB: 2.7.1.39)

t-RNA biosynthesis + protein export: (ileS:6..1.1.5, lspA:3.4.23.36) (***gcvH***, gcvP:1.4.4.2), (pheS:6.1.1.20, pheT:6.1.1.20)

Biotin metabolism : (BioA:2.6.1.62, BioF:2.3.1.47)

Phenylalanine, tyrosine, tryptophan metabolism: (***ppiA***, pabA: 4.1.3.-) (trpE:4.1.3.27, trpC:4.1.1.48, trpB: 4.2.1.20, trpA:4.2.1.20) (aroB: 4.6.1.3, aroK: 2.7.1.71) (***Rv3534C***, Rv3535C:4.1.3.-, ***Rv3536c***)

Urea cycle and metabolism of amino groups: (argC:1.2.1.38, argB:2.7.2.8, argH: 4.3.2.1) (proA: 1.2.1.41, proB:2.7.2.11)

Ribosome related (includes associated protein export): (rpsF, rpsR, rplI) (***secE, nusG***, rplK, rplA, rplJ, rplL) (rpmG, rpmB2), (***yhbZ***, rpmA, rplU), (rplS, ***trmD***, ***rimM***, rpsP, ***ffh***) (rpsI, rplM), (rplQ, rpoA, rpsD, rpsK, rpsM, rpmJ), (***Rv3921c***, rpmH) (rpsL, rpsG, ***fusA, tufA***) (rpsJ, rplC, rplD, rplB, rpsS, rplV, rpsC, rplP, rpmC, rpsQ, rplN, rplX, rplE, rpsN, rpsH, rplF, rplR, rpsE, rpmD, rplO, ***secY***), (***infC***, rpmI, rplT) (***Rv2869c***, ***Rv2870c,*** cdsA:2.7.7.41, ***frr, pyrH:2.7.4.-, tsf***, rpsB)

Folate biosynthesis/Lysine degradation/Ubiquinone biosynthesis: (lipB:6.-.-.-, ***lipA)***

β alanine + Folate biosynthesis: (panD:4.1.1.11, panC:6.3.2.1, folK:2.7.6.3)

Lysine biosynthesis: (***bcp***, Rv2522c:3.5.1.18)

Protein export: (secD, secF)

Nitrogen regulation (two component system): (glnB, ***amtB***)

Thiamine metabolism: (***Rv2971***, thiL:2.7.4.16) (thiE:2.5.1.3, ***thiG, thiC***)

Peptideglycan biosynthesis: (amiA:3.5.1.28, amiB:3.5.1.28)

Sterol + terpenoid biosynthesis: (***Rv3379c***, idsB:2.5.1.10)

Nicotinate and nicotinamide metabolism: (pntAA:1.6.1.1, pntB:1.6.1.1, pntA:1.6.1.1)

Galactose metabolism: (galT:2.7.7.10, galK: 2.7.1.6)

Glyoxylate and dicarboxylate metabolism: (***fdhD***, fdhF: 1.2.1.2)

TABLE 4
Pathway Seeds Using *B. subtilis*

Ribosomal protein related (includes protein export): (rpsF, ***ssb***, rpsR), (nusG, rplK, rplA, rplJ, rplL), (rpsJ, rplC, rplD, rplW, rplB, rpsS, rplV, rpsC, rplP, rpmC, rpsQ, rplN, rplX, rplE, rpsN, rpsH, rplF, rplR, rpsE, rpmD, rplO, ***secY***, adk:2.7.4.3, ***mapx:3.4.11.18***), (***rho***, rpmE, ***prfA, hemK, Rv1301***), : (rplI, dnaB:3.6.1.-), (***InfC,*** rplM, rplT), (***lepA***, rpsT, ***Rv2414c, Rv2415c***), (***obg***, rpmA, rplU), (rpsL, rpsG, ***fusA, tuf***), (***pepR***, gpsI:2.7.7.8, rpsO, ribF:2.7.1.26, truB:4.2.1.70), (rplS, ***trmD:2.1.1.31***, rimM, ***Rv2908c***, rpsP, ffh, ***smc***, ***rnc***), (cdsA: 2.7.7.41***, frr***, ***pyrH: 2.7.4.-***, ***tsf***, rpsB), (rpsI, rplM, truA: 4.2.1.70, rplQ, rpoA: 2.7.7.6, rpsK, rpsM, rpmJ, ***InfA***)

Nitrogen metabolism: (nirB, narK3), (gltD:1.4.1.13, gltB:1.4.1.13)

Pyruvate metabolism (pta:2.3.1.8, ackA:2.7.2.1)

Thiamine metabolism (***lpqL:3.4.21.-,*** thiD:2.7.4.7)

Butanoate metabolism (***aceA***, mmgB:1.1.1.157)

Androgen and estrogen metabolism/aminophohonate metabolism/histidine metabolism/tyrosine metabolism /tryptophan metabolism: (ubiE:2.1.1.-, grcC1:2.5.1.30)

Purine metabolism: (rpoB:2.7.7.6, rpoC:2.7.7.6), (purB:4.3.2.2, purC:6.3.2.6, purQ:6.3.5.3), (purN:2.1.2.2, purH:2.1.2.3), (gmk:2.7.4.8, ***Rv1390***:, ***dfp, priA***), (ureA:3.5.1.5, ureB:3.5.1.5, ureC:3.5.1.5), (***Rv3273***, purE: 4.1.1.21, purK: 4.1.1.21), (hpT:, 2.4.2.8, ***yacA***), (***recR***, ***Rv3716c***, dnaZX:2.7.7.7), (relA: 2.7.6.5, apt: 2.4.2.7, ***secF, secD, ruvB, ruvA, Rv2603c***)

Pyrimidine metabolism: (**rpoB:2.7.7.6, rpoC:2.7.7.6**), (***pyrR***, pyrB:2.1.3.2, pyrC:3.5.2.3, pyrAA:6.3.5.5), (***recR***, ***Rv3716c***, dnaZX:2.7.7.7), (***pepR***, gpsI:2.7.7.8, rpsO, ribF:2.7.1.26, truB:4.2.1.70)

DNA polymerase: (rpoB:2.7.7.6, rpoC:2.7.7.6), (dfrA: 1.5.1.3, thyA: 2.1.1.45), (***recR***, ***Rv3716c***, dnaZX:2.7.7.7)

Pentose and glucuronate interconversions: (fucA:5.1.3.4, xylB:2.7.1.16), (fmt: 2.1.2.9, fmu, rpe:5.1.3.1)

Citrate cycle: (sucC:6.2.1.5, sucD:6.2.1.5), (***Rv1130***, gltA1:4.1.3.7),

Pentose phosphate cycle / glutathione metabolism: (zwf:1.1.1.49, ***gnd2***)
proanoate metabolism/C-branched dibasic acid metabolism/CO2 fixation: (sucC:6.2.1.5, sucD:6.2.1.5),

One Carbon Pool by Folate: (purN:2.1.2.2, purH:2.1.2.3), (dfrA: 1.5.1.3, thyA: 2.1.1.45), (folD: 1.5.1.5, ***Rv3359***)

Amino-acyl trnA biosynthesis (includes protein export): (***Rv1003***, metS:6.1.1.10, Rv1008), (***tsnR***, pheS:6.1.1.20, pheT:6.1.1.20), *(folC:6.3.2.17, valS:6.1.1.19),* (***Rv2553c, Rv2554c***, alas:6.1.1.7),

Valine, leucine, isoleucine biosynthesis: *(folC:6.3.2.17, valS:6.1.1.19), (leuC: 4.2.1.33, leuB: 1.1.1.85)* (ileS, lsp, ***ylyB***)

Porphyrin and Chlorophyll metabolism: (ksgA:2.1.1.-, yabH: 2.7.1.-), (ctaB:2.5.1.-, ***Rv1456c***), (hemyX:1.3.3.4, hemY:4.1.1.37), (cysG2: 2.1.1.107, ***gorA***), (hemA:1.2.1-, hemG: 4.3.1.8, cysG: 4.2.1.75, hemB: 4.2.1.24, hemL: 5.4.3.8)

Glycine, serine, and threonine metabolism: (thrA:1.1.1.3, thrB:2.7.1.39, thrC:4.2.99.2), : (***Rv2869c***, Rv2870c:1.1.1.-, cdsA: 2.7.7.41), (asd: 1.2.1.11, ask:2.7.2.4)

Glutamate metabolism: (murI:5.1.1.3, ***rphA:2.7.7.56***, ***Rv1341***), (gltD: 1.4.1.13, gltB:1.4.1.13)

Glycolysis/ Gluconeogenesis: (gap:1.2.1.12, pgk:2.7.2.3, tpi:5.3.1.1)

Glycoprotein biosynthesis: (ctaB:2.5.1.-, ***Rv1456c***)

Biotin metabolism: (bioA:2.6.1.62, bioF:2.3.1.47, bioD:6.3.3.3),

gene-groups (along with the associated nonenzymes) acting as seeds for pathway traces. Table 2 shows the extended gene-group (along with the associated nonenzymes) in the *E. coli* genome acting as the pathway trace derived from the gene-groups in Table 1.

In the tables, the genes are represented as *gene-name: enzyme classification number*. The genes without enzyme numbers reported indicate that there is no corresponding *EC* number (enzyme classification number) in the KEGG database. An *EC* number without a gene name means that the enzyme is present in the KEGG database. However, the ortholog does not occur within the same gene-group. We mark the genes missing in the KEGG database by bold italicized names and mark the enzymes shared between more than two pathway modules by underlining the enzyme. The results show that the homologous ordered and unordered gene-groups and gene-groups

with duplicated domains completely map on a pathway trace in the same metabolic pathway or adjacent pathway modules. They may also be genes with a similar function or are regulated by the same control region.

For example, *EC 4.2.1.9* and *EC 4.2.1.16* occur together, share the same control region, and have similar function (both are lyases: *EC 4.2.1.9*) and are involved in the "Valine, Leucine, and Isoleucine metabolic pathway." *EC 4.2.1.16* is also involved within the serine pathway. Minor variations occur when the reactions involve two adjacent modules.

Almost all the large pathway traces individually belong to one module. However, one particular enzyme or gene-group may belong to two or more adjacent modules (many times similar). This is due to artificial splitting of metabolic pathways into different modules.

Few genes such as *ppc:4.1.3.1* belong to a different pathway despite being in a gene-group belonging to a different metabolic pathway. A finer, future analysis based on the control regions will remove this discrepancy. Some enzymes, such as *thtR:2.8.1.1*, *upp:2.4.2.9*, *udhA:1.-.-.-*, etc., were missing from the KEGG database and could not be tested despite being in the same gene-group. Due to space limitation, we have not shown the pathways or the genes in duplicated gene-groups. However, the contribution of these gene-groups is very large and can be classified in four different categories as follows:

1. gene-groups that map to a pathway trace are in the same module or in the adjacent modules,
2. a subgroup of a larger gene-group occurring at a different location in the genome,
3. the regulatory and sensor proteins (such as *EC 2.7.3.-*) having multiple occurrences, and
4. rare random duplication bringing enzymes from two metabolic pathways together.

The role of multiple subgroups of the same group of enzymes at different locations in the genome is not clear and may be related to enhanced reaction or production of the correspomding biochemical products. The role of rare random duplication in metabolic pathway is not clear and needs to be investigated further.

Additionally, the data suggest clear redundancy of gene-pairs for gene products that perform equivalent or similar functions. This is evidenced by gene pairs found to be associated with metabolic pathways where amino transferases, ester bond hydrolysis, or carboxylase activities are required in two or more different pathways, for example. The pathway seeds generated from the comparison of *E. coli and B. subtilis* predict specific metabolic events. Extension of these comparisons (Table 2) results in the identification of more conserved sequences for which specific pathway localization can be determined. We present three examples of this below.

Example 4: Peptidoglycan biosynthesis. The bacterial cell wall, or peptidoglycan, is a unique and rigid biopolymer composed of repeating sugar moieties and species-specific (amino acid and polysaccharide) side chains that are linked together. The connected molecules form a three-dimensional matrix that surrounds the cell to create a barrier restricting cell membrane expansion during changes in osmotic pressure. The peptidoglycan precursors are initially synthesized and modified within the cytoplasm from fructose-6-phosphate and are transported through the cell membrane via a 55-carbon isoprenoid lipid carrier, bactoprenol (undecaprenol). Reduction, acetylation aminotransfer, ligation, and other specific modifications of the initial precursors must occur to prepare the N-acetyl-glucosamine-N-acetyl-muramic acid dimers that form the polymer. Additionally, ligases and other enzymes are required to add sidechains to the growing peptidoglycan polymer. Thus, an extensive biosynthetic pathway is required to modify the fructose, add acetyl groups, form the ring structures, and add amino acids, polysaccharides, and peptidoglycan-associated proteins. The synthesis of new cell wall material is required for the overall increase in cell size and in the division of one cell into two new cells during replication. Thus, the synthesis of peptidoglycan must be tightly regulated and coupled to the growth of the bacterial organism to maintain cell integrity. Analysis of the data captured by the automated system above reveals several distinct sequences whose proteins have specific pathway functions. The *mur* genes code for various substrates and enzymes (liagases, reductases, racemases, etc.) for peptidoglycan biosythesis, the *fts* genes code for cell division proteins, the *ddl* genes code peptidoglycan side chain modifications (*ddlB* is a D-ala-D-ala ligase), and *yab* genes code for oxidoreductases and ABC transportor proteins. Thus, the pathway for peptidoglycan synthesis can be more readily defined knowing several of the product sequences.

Example 5: Flagellar Assembly. Prokaryotic flagella are extracellular, protein extensions of the cell whose function is primarily motility. The final structure has three basic parts (basal body, hook, and filament), with precursors synthesized in the cytoplasm for migration out of the cell. Flagellar synthesis, assembly, and function are complicated and dictated by over 30 separate genes. Additionally, integration of the flagellar structure with "motor" proteins and a hydrogen pump system are required for biological function. Analysis of the data captured by the automated system above reveals several distinct sequences whose proteins have specific pathway functions. The *fli* genes code for (mostly structural) flagellar proteins (basal body, hook, filament, motor switch, etc.). The gene *rcsA* codes for a product in the cell capsule surrounding the flagella. The gene *dsrB* (β subunit of dissimilatory sulfite reductase) is probably involved in energy conservation. The automated identification of flagellar genes in several bacteria can be identified in homologous gene-groups and associated with the flagellar biosynthetic pathway. Furthermore, the analysis is logical as it presents the substrate and enzyme families required for flagellar synthesis.

Example 6: Citrate Cycle. The citrate cycle in prokaryotes is a cellular source of carbon skeleton intermediates, NADPH, ATP, and carbon dioxide. However, to function as a catabolic pathway, precursors and enzymes must be available. The formation and breaking of covalent bonds by specific enzymes along the pathway release potential energy via oxidoreduction events until the initial substrate is reduced to carbon dioxide and water. The

catabolism of pyruvate by the citric acid pathway involves oxidoreduction, acetyl group transfer, and ligation of sulfur and carbon groups. Analysis of the data captured by the automated system above reveals several distinct sequences whose proteins have specific pathway functions. The *sdh* genes code for enzymes that facilitate oxidoreduction (especially with sulfide-containing acceptors) and ligation of carbon-sulfur and carbon-nitrogen precursors. Thus, the expected pathway enzymes are predicted by the pathway trace generated by the automated system.

5.2 Predicting Pathways for Genomes

In this section, we apply our technique to identify the pathway seeds and pathway traces of *Mycobacterium tuberculosis*—a deadly pathogen. *Mycobacterium tuberculosis* is the etiologic agent of the ancient, world-wide disease known as tuberculosis. Tuberculosis is primarily a disease of the lungs, but other organs and tissues become infected when the bacteria leave the lungs and disseminate via the bloodstream.

M. tuberculosis is a Gram-positive eubacteria. Cultures of *M. tuberculosis* grow slowly, with a generation time of 18-24 hours. The bacteria are aerobic and produces catalase to detoxify oxyradicals. The cell walls of *Mycobacteria* have a very high glycolipid content of complex fatty acids (e.g., arabinogalactan-lipid complex and mycolic acids). The mycolic acids are composed of 60-90 carbons with aliphatic chains on the α-carbon and a hydroxyl group on the β-carbon. These lipid modifications inhibit normal Gram-staining procedures, requiring harsher techniques, thus resulting in the clinical term "acid-fast bacteria." The penetration of aqueous materials, including soluble antimicrobial compounds, may be prevented by the relatively high lipid content. These lipid modifications may also account for the ability of *M. tuberculosis* to survive within the phagocytes that attempt to kill the invading pathogen. This may be a result of the cell wall lipids, in part, but also of bacterial strategies to counteract the ATPase-mediated acidification processes used by the phagocytes to kill invaders [16].

Other virulence factors have been speculated. Transposon mutagenesis and transfer of tuberculosis genes to *E. coli* have identified potential gene candidates. However, the automated comparison of *M. tuberculosis* genome with other bacterial genomes would quickly identify specific products and pathways responsible for the virulence [8].

The comparison of *M. tuberculosis* with *E. coli* yielded 70 ordered homologous gene groups, 18 unordered homologous gene-groups, 549 duplicated gene-groups, 383 gene-duplications, and 43 neighboring gene fusions. The comparison of *M. tuberculosis* with *B. subtilis* yielded 88 ordered homologous gene-groups, 20 unordered homologous gene-groups, 672 duplicated gene-groups, 568 gene-duplications, and 101 neighboring gene-fusions.

The data show that the major components of the gene-groups are duplicated gene-groups. These gene-groups represent a large amount of the pathway. However, due to space limitations, we use the data related only to ordered and unordered homologous gene-groups (including orthologous gene-groups) derived from the comparison of *M. tuberculosis* with *E. coli* and *B. subtilis*.

Out of 88 ordered and unordered gene-groups found by *M. tuberculosis* and *E. coli* genome comparison, 58 gene-groups were found to be seed-points using the KEGG metabolic pathway database (see Table 3). Out of 108 ordered and unordered gene-groups found by *M. tuberculosis* and *B. subtilis* genome comparison 89 gene-groups were found to be seed-points using the KEGG metabolic pathway database.

TABLE 5
Merged Pathway Traces for *M. tuberculosis*

Purine metabolism: (purD:6.3.4.13, purN:2.1.2.2, purQ:6.3.5.3, purM: 6.3.3.1, ***Rv3273,*** purE:4.1.1.21, PurK: 4.1.1.21, purB:4.3.2.2, purC:6.3.2.6, , apt:2.4.2.7, ***secF, secD, ruvB, ruvA, Rv2603c*** hpt:2.4.2.8, ***yacA***, purH:2.1.2.3/3.5.4.10, ***purA/ 6.3.4.4***, adk:2.7.4.3 (relA: 2.7.6.5, rpoB/rpoC:2.7.7.6, ***recR, Rv3716***, dnaZX:2.7.7.7), (ureA:3.5.1.5, ureB: 3.5.1.5, ureC:3.5.1.5)
Pyrimidine metabolism: (***pyrR***, carA:6.3.5.5, pyrB:2.1.3.2, pyrC:3.5.2.3, **pyrD:1.3.3.1,** pyrE:2.4.2.10, pyrF: 4.1.1.23), (rpoB/rpoC: 2.7.7.6, dnaZX:2.7.7.7, gpsI:2.7.7.8, ***sirR, ltp1, fade21***, rpsO, ribF:2.7.1.26), (truB:4.2.1.70)
gmk: 2.7.4.8 ***dfp, priA***),
Glycine, serine, and threonine metabolism + Lysine biosynthesis: (asd:1.2.1.11, ask: 2.7.2.4, thrA:1.1.1.3, thrC: 4.2.99.2, thrB: 2.7.1.39, ***Rv1516c***, pks5:1.1.1.103, ,)
Phenyl, Tyrosine, Tryptophan biosynthesis: (aroB:4.6.1.3, aroD:1.1.1.25, aroC:4.2.1.10, aroK:2.7.1.71, pepQ:3.4.-.-, **2.5.1.9,** aroF:4.6.1.4, pabA: 4.1.3.-, trpE:4.1.3.27, **2.4.2.18**, **5.3.1.24**, trpC:4.1.1.48, trpB:4.2.1.20, trpA:4.2.1.20, ***lgt***)
Urea Cycle + Amino acid metabolism: (proB:2.7.2.11, proA:1.2.1.41, argB**:**2.7.2.8, argC:1.2.1.38, argD:2.6.1.11, argG:6.3.4.5, argH: 4.3.2.1, argF:2.1.3**.3, *argR,*** argJ:2.3.1.1 /2.3.1.35), (ureA:3.5.1.5, ureB:3.5.1.5, ureC:3.5.1.5)
Porphyrin and Chlorophyll metabolism: (hemA:1.2.1.-, hemL:5.4.3.8, hemB:4.2.1.24, hemG:4.3.1.8, hemY:4.1.1.37, cysG:4.2.1.75, cysG2: 2.1.1.107, ***gorA***, hemyX:1.3.3.4, ksgA: 2.1.1-, ctaB:2.5.1.-, ***Rv1456c***, yabH:2.7.1.-)
Oxidative phosphorylation: (***Rv0967***, ctpV:3.6.1.36, atpH:3.6.1.34, atpA:3.6.1.34, atpG:3.6.1.34, atpD: 3.6.1.34, atpC:3.6.1.34)
Folate Biosynthesis: (lipB:6.-.-.-, ***lipA***, ***3.5.4.16***, recG:3.6.1.-, ***Rv2974c***, ***Rv2975c***, ***moeZ***, folX:4.1.2.25, folK:2.7.6.3, panD:4.1.1.1, panC:6.3.2.1, folP:2.5.1.15, thyA:2.1.1.45, dfrA:1.5.1.3)
Biotin biosynthesis: (***6.2.1.14***, bioF: 2.3.1.47, bioA:2.6.1.62, bioD:6.3.3.3,), (pepQ:3.4.-.-, aroD:4.2.1.10)
One Carbon Pool by folate: (dfrA:1.5.1.3, purH:2.1.2.3, purN:2.1.2.2, thyA:2.1.1.45, fold:1.5.1.5, ***Rv3359***)
Nikotinate and nicotinamide metabolism: (***Rv2420c***, Rv2421c: 2.7.7.18, ***nadA***, nadB:1.4.3.16, nadC:2.4.2.19,pntAA: 1.6.1.1, pntB:1.6.1.1, pntA:1.6.1.1)

Tables 3 and 4 show three types of missing information in enzyme-based pathway databases as follows:

1. The enzymes are part of one gene-group. However, this gene-group has a gene shared between two pathway modules suggesting an intricate relationship between two pathway modules.
2. The genes are associated with a gene-group, but do not have an enzyme number. Such genes are quite large in number. These genes are marked as bold and italicized in Tables 3 and 4. These genes are either not annotated or are regulatory genes involved in the Table 3.

Table 5 shows some of the merged pathway traces from Tables 3 and 4.

In Table 5, the bold enzyme numbers show the missing enzymes (present in the KEGG database) from the pathway trace after merging. The bold and italicized gene-names (present in pathway traces) are missing from the KEGG databases.

These missing genes could be either regulatory proteins or some of them could be a substitution for missing enzymes. The underlined enzymes are intricately linked to another pathway module suggesting a close and intricate relationship between the two modules. For example, *nadB:1.4.3.16* found in gene-group (*nadA, nadB, nad*C) in nikotinate and nikotinamide metabolism is part of alanine and asparatate metabolism. The enzymes missing from the gene-groups but present in KEGG database are either enzymes shuffled from the gene-groups to some other location or are absent in the genome under consideration. For example, *trpD: 2.4.2.18* is present yet shuffled from the main gene-group in *M. tuberculosis* and the gene *trpF: 5.3.1.24* has no homolog in *M. tuberculosis* although present in *B. subtilis*. Such enzymes are possibly substituted by other enzymes. The effect of substitutions will be interesting to study. We have omitted genes marked as hypothetical proteins from Table 5. Some of these hypothetical proteins may be involved in regulation, or may not be annotated.

The subsets of merged and extended pathway traces for *M. tuberculosis* (see Table 5) identify several interrelated metabolic pathways that may explain its virulence. In one example, the aerobic *M. tuberculosis* should use oxidation/reduction enzymes to catalyze proton transport and, thus, the synthesis of ATP via oxidative phosphorylation events. Oxidases and reductases are required for this metabolism to occur. The pathway trace identified several ATP-dependent proton transporting synthetases (*atp genes*:*3.6.1.34*), dehydrogenases (*proA genes*:*1.2.1.*-), and other oxidases (*hemyX genes*:*1.3.3.4*) whose function could direct/regulate ATP synthesis and oxyradical detoxification.

In a second example, *M. tuberculosis* is known for its ability to counteract phagocyte killing via neutalization of the acidic phago-lysosome environment. Speculation indicates that ammonia production by *M. tuberculosis* may be the mechanism by which this occurs [14]. The pathway traces obtained by automatic gene-group comparison (above) identify several interrelated pathways that share in the metabolism of ammonia via the urea cycle (*ure genes: 3.5.1.5*). Additionally, capture of amino groups and their transfer can be mediated through carbamyl phosphate synthetase (*6.3.5.*-) and amino transferases (*5.4.3.*-). These are also predicted by the pathway trace in Table 5. Together, these enzymes participate in the cycling of ammonia and amino groups within *M. tuberculosis* and may contribute to its virulence by significant ammonia production to prevent its destruction by mammalian phagocytes.

6 Conclusion

In this paper, we have described a novel graph-based scheme [5] and algorithms to automate and refine the reconstruction of metabolic pathways of newly sequenced genomes using the automatically derived orthologous and homologous gene-groups [3], [4]. We also analyze a subset of derived metabolic pathways of *Mycoplasma tuberculosis*—a deadly pathogen. The technique is based upon the marking of local sections in metabolic pathways as pathway seeds, extending the pathway seeds to pathway traces, and merging pathway traces obtained from multiple genome comparisons.

Unlike previous schemes [20] that compare evolutionarily similar species, this scheme uses the findings of Bansal [4] (that the number of gene-groups is largely a function of genome size) to compare genomes from different genome families.

This gene-group based analysis also identifies nonenzymatic genes affected by enzymatic reactions in a section of metabolic pathway. The scheme can also predict the *EC* number for missing genes in the large pathway traces and the interrelationship of different pathway modules. An interesting finding is that gene-groups containing regulatory proteins, sensor proteins, and ABC transporters [15], [21] are frequently duplicated. The predicted pathway substrates and enzymes are consistent with expected gene products in each of the respective analyses, providing a high degree of validity for the automated system.

References

[1] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman, "Basic Alignment Search Tools," *J. Molecular Biology,* vol. 215, pp. 403-410, 1990.

[2] A. Bairoch, "The ENZYME Database in 2000," *Nucleic Acids Research,* pp. 304-305, 2000.

[3] A.K. Bansal, P. Bork, and P.J. Stuckey, "Automated Pair-Wise Comparisons of Microbial Genomes," *Math. Modeling and Scientific Computing,* vol. 9, no. 1, pp. 1-23, 1998.

[4] A.K. Bansal, "An Automated Comparative Analysis of 17 Complete Microbial Genomes," *Bioinformatics,* vol. 15, no. 11, pp. 900-908, 1999.

[5] A.K. Bansal, "A Framework of Automated Reconstruction of Microbial Metabolic Pathways," *Proc. IEEE Int'l Conf. Bioinformatics and Biomedical Eng.,* pp. 184-190, 2000.

[6] H. Bono, H. Ogata, S. Goto, and M. Kanehisa, "Reconstruction of Amino Acid Biosynthesis Pathways from the Complete Genome Sequence," *Genome Research,* vol. 8, no. 3, pp. 203-210, 1998.

[7] P. Bork, T. Dandekar, Y. Diaz-Lazcoz, F. Eisenhaber, M.A. Huynen, and Y. Yuan, "Predicting Function: From Gene to Genomes and Back," *J. Molecular Biology,* vol. 283, pp. 707-725, 1998.

[8] R. Brosch, S.V. Gordon, and A. Pym et al., "Comparative Genomics of the Mycobacteria," *J. Medical Microbiolgy,* vol. 290, no. 2, pp. 143-152, 2000.

[9] S.J. Cordwell, "Microbial Genomes and Missing Enzymes: Redefining Biochemical Pathways," *Archives Microbiology,* vol. 172, no. 5, pp. 269-279, 1999.

[10] W.M. Fitch, "Distinguishing Homologous from Analogous Proteins," *Systematic Zoology,* vol. 19, pp. 99-113, 1970.

[11] S. Goto, T. Nishioka, and M. Kanehisa, "LIGAND: Chemical Database for Enzyme Reactions," *Bioinformatics,* vol. 14, no. 7, pp. 591-599, 1998.

[12] M.A. Huynen and P. Bork, "Measuring Genome Evolution," *Proc. Nat'l Academy of Science,* vol. 95, pp. 5849-5856, 1998.

[13] P.D. Karp, M. Krummenacker, S. Paley, and J. Wagg, "Integrated Pathway-Genome Databases and Their Role in Drug Discovery," *Trends Biotechnology,* vol. 17, no. 7, pp. 275-281, 1999.

[14] A. Kowl, A. Chodias, and M. Treder et. al., "Cloning and Characterization of Secretory Tyrosine Phosphatases of Mycobacterium Tuberculosis," *J. Bacteriology,* vol. 182, no. 19, pp. 5425-5432, 2000.

[15] K.J. Linton and C.F. Higgins, "The *Escherichia coli* ATP-Binding Cassette (ABC) Proteins," *Molecular Microbiology,* vol. 28, no. 1, pp. 5-13, 1998.

[16] A.A. Salyers and D.D. Whitt, *Bacterial Pathogenesis: A Molecular Approach.* ASM Press, 1994.

[17] E. Selkov, N. Maltsev, G.J. Olsen, R. Overbeek, and W.B. Whitman, "A Reconstruction of the Metabolism of *Methanococcusjannaschi* from Sequence Data," *Gene*, vol. 197, nos. 1-2, pp. 11-26, 1997.
[18] E. Selkov Jr., Y. Grechkin, N. Mikhailova, and E. Selkov, "MPW: The Metabolic Pathways Database," *Nucleic Acids Research*, vol. 26, no. 1, pp. 43-45, 1998.
[19] S. Schuster, T. Dandekar, and D.A. Fell, "Detection of Elementary Flux Modes in Biochemical Networks: A Promising Tool for Pathway Analysis and Metabolic Engineering," *Trends Biotechnology*, vol. 17, no. 2, pp. 53-60, 1999.
[20] R.L. Tatusov, M. Mushegian, P. Bork, N. Brown, W.S. Hayes, M. Borodovsky, K.E. Rudd, and E.V. Koonin, "Metabolism and Evolution of *Haemophilius Influenzae* Deduced From a Whole-Genome Comparison with *Escherichia Coli*," *Current Biology*, vol. 6, pp. 279-291, 1996.
[21] K. Tomi and M. Kanehisa, "A Comparative Analysis of ABC Transporters in Complete Microbial Genomes," *Genome Research*, vol. 8, pp. 1048-1059, 1998.
[22] M.S. Waterman, *Introduction to Computational Biology: Maps, Sequence, and Genomes*. Chapman & Hall, 1995.

Arvind K. Bansal received the BTech (electrical engineering) and MTech (computer science) degree from the Indian Institute of Techology at Kanpur in 1979 and 1983. He received the PhD (computer science) degree from Case Western Reserve University in 1988. He is an associate professor in the Department of Computer Science at Kent State University, Kent, Ohio. His research interests and research contributions are in bioinformatics, distributed artificial intelligence, logic programming, multimedia Information retrieval, and distributed multimedia environments and languages. He is a senior member of the IEEE, a member of the ACM, a member of the New York Academy of Sciences, and a member of the American Association of Advancement of Science.

Christopher J. Woolverton received the MS and PhD (medical microbiology) degrees from West Virginia University, Morgantown, West Virginia in 1984 and 1986, respectively. He received his postdoctoral training (immunology) from the University of North Carolina at Chapel Hill. He is an associate professor in the Department of Biological Sciences at Kent State University, Kent, Ohio. His research interests and contributions are in the areas of microbial virulence factors, antimicrobial chemotherapy, biosensors, and bioinformatics. He is a member of the American Society for Microbiology, Sigma Xi, and the New York Academy of Sciences.