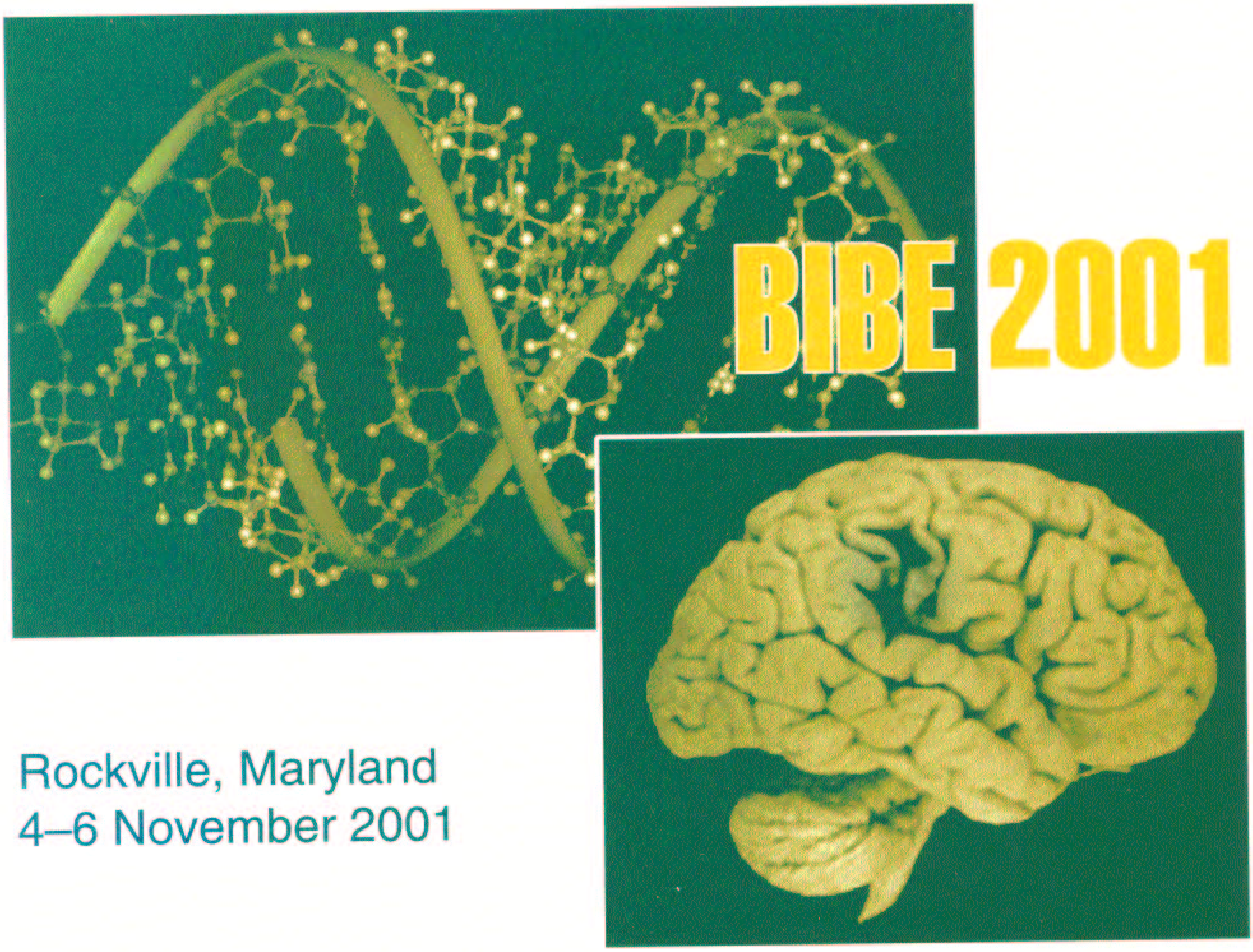
Proceedings

IEEE International Symposium on

Bio-Informatics and Biomedical Engineering



BIBE 2001

Rockville, Maryland
4–6 November 2001

Sponsors:
IEEE Computer Society
IEEE Computer Society Task Force on Virtual Intelligence (TFVI)

IEEE
COMPUTER
SOCIETY

Proceedings of the 2nd IEEE International Symposium on Bio-informatics and Biomedical Engineering, Maryland, USA, November 2001, IEEE Computer Society Order Number PRO1423, ISBN 0-7695-1423-5, Library of Congress Control Number 20001096085

Integrating Co-regulated Gene-groups and Pair-wise Genome Comparisons to Automate Reconstruction of Microbial Pathways

Arvind K. Bansal
Department of Computer Science
Kent State University, Kent, OH 44242, USA
http://www.cs.kent.edu/~arvind e-mail:arvind@cs.kent.edu
&
Intellibio Software and Consultancy Corporation
3109 Killingworth Lane, Twinsburg, OH 44087, USA
e-mail:arvind@intellibiosoft.com

Abstract

This paper extends previously described automated techniques by automatically integrating the information about automatically derived co-transcribed gene-groups, functionally similar gene-groups derived using automated pair-wise genome comparisons and automatically derived orthologs (functionally equivalent genes) to derive microbial metabolic pathways. The method integrates automatically derived co-transcribed gene-groups with orthologous and homologous gene-groups (http://www.mcs.kent.edu/~arvind/orthos.html), the biochemical pathway template available at the KEGG database. (http://www.genome.ad.jp), the enzyme information derived from SwissProt enzyme database (http://expasy.hcuge.ch/) and Ligand database (http://www.genome.ad.jp). The technique refines existing pathways (based upon network of reactions of enzymes) by associating corresponding non-enzymatic, regulatory, and co-transcribed proteins to enzymes. The technique has been illustrated by deriving a major pathway of *M. tuberculosis* by comparison with seven microbial genomes including *E. coli* and *B. subtilis* ? two microbes well explored in wet laboratories.

Keywords: Automation, bacteria, co-regulation, co-transcription, drug-discovery, enzymes, gene-groups, homologs, metabolic pathway, microbes, operons, orthologs, pathogenicity

1. Introduction

Understanding microbial machinery has important advantages in bio-remediation, management of environmental waste, efficient utilization and generation of energy, and development of more effective antibiotics based upon the regulation of metabolic pathways. At this point 55 microbial genomes (including many pathogens) have been sequenced and archived, and more than 182 are 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 inter-connected 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.

The study of the dynamic cellular behaviors dictated by metabolic pathways will facilitate the development of effective anti-bacterial drugs with reduced side effects [19]. However, in order to understand the metabolic mechanisms, biochemical pathways and their variations across different microbes need to be mapped accurately.

Previous techniques to derive metabolic pathways are incomplete and restricted. The schemes either did not take grouping of genes in account [7], or compared very closely related genomes [20], or used strict ordering of gene-groups of orthologs [20] ? strictly functionally equivalent genes, or used only co-transcribed gene-groups without any pair-wise genome comparison to derive groups [16]. The schemes based simply on co-transcribed gene-groups are not sufficient since multiple Co-transcribed gene-groups (containing even non-enzymatic genes) may be in the same pathway. In addition most of these schemes were semi-automated [16, 17, 20]. These schemes did not integrate all the information available.

It has now become clear that

1. many times functionally similar genes can substitute for the needed functionality despite them not being orthologs,
2. larger microbial genomes have a large number of functionally similar gene-groups irrespective of their evolutionary classification [4], and
3. neighboring genes involved in the same gene-group (co-transcribed or identified by pair-

0-7695-1423-5/01 $10.00 © 2001 IEEE

wise genome comparison) tend to be in closely related pathways [5, 6] even though the genes in a group may not always preserve order [5, 6]

In our previous automated pathway reconstruction technique [5, 6], we integrated automatically derived orthologs and homologous groups of neighboring genes derived by comparison with multiple genomes (including evolutionary distant genomes). This scheme also benefited from the use of homologous fused genes [3, 4] and genes having a similar domain which can substitute for the functionality of a given enzyme.

The previous technique was based upon 1) identifying orthologous (functionally equivalent) and homologous (functionally similar) gene-groups and marking them as seed points for the identification of pathways, 2) merging two adjacent seed points heuristically and taking union of sets of genes in orthologous gene-groups to identify a pathway trace, and 3) merging pathway traces based upon common enzymatic reaction.

The scheme had two drawbacks:

1. The scheme did not take into account the rich information, which can be derived by co-transcribed gene-groups.
2. The merging of adjacent gene-groups was based upon the user defined proximity between two adjacent seeds.

This paper describes an alternate scheme to remove the first problem, and reduce the second problem. The extended scheme is as follows:

1. Automatically derived co-transcribed gene-groups ? groups of genes which are expressed together and have a common control region — with the pair-wise gene-group information have been integrated with other information in the previous scheme, and
2. Gene-group information from multiple pair-wise genome comparisons has been merged to fill in the missing information from individual pair-wise genome comparison. Multiple genomes are needed for pair-wise genome comparisons due to the genome rearrangement and insertion and deletion of genes from pathways. This extension makes our scheme more robust, and less dependent upon heuristics to merge two pathway seeds. In addition, the comparison of one genome against multiple genomes significantly reduces the possibilities of missing genes in individual homologous gene-groups. This integration improves upon the schemes based solely on co-transcribed gene-groups since groups derived from pair-wise genome comparison also contain genes from two adjacent co-transcribed gene-groups.

The paper is organized as follows: Section 2 describes the background. Section 3 describes briefly the techniques to derive orthologs, homologous gene-groups [3, 4], and co-transcribed gene-groups. Section 4 briefly describes the merger techniques to reconstruct the metabolic pathways using the enzymes in the pathways derived from wet laboratories (such as pathways in *E. coli* and *B. subtilis*) available at the KEGG site (http://www.genome.ad.jp). Section 5 describes a detailed example of citrate cycle pathway derivation for *M. tuberculosis* using seven microbial genomes (including *E. coli and B. subtilis*) present in GenBank (ftp://ftp.ncbi.nlm.nih.gov/GenBank/genomes/Bacteria/). Section 6 discusses the results. Section 7 concludes the paper.

2. Background and Definitions

This section describes background concepts related to genome comparison, orthologs, gene-groups, enzymes, metabolic pathways, and some new definitions to explain the technique to automatically derive pathways.

2.1 Modeling Genomes

A genome Γ is modeled as an ordered set of genes $<g_1, g_2, \ldots, g_N>$ where N is the number of genes in a genome. Each gene γ_I is a pair of the form (y_I, h_I) where y_I is a control region and h_I is the corresponding protein coding region. The set of protein sequences corresponding to the protein coding regions in a genome Γ_1 is modeled as $<p_1, , p_2, \ldots, p_N>$ where p_I is the amino-acid sequence corresponding to the nucleotide sequence $h_{I.}$. There may be more than one different subsequence in a protein those are homologous to sequences corresponding to different genes in another genome. These subsequences include one or more protein domains.

2.2. Orthologs and Gene -groups

A *putative ortholog* is defined as a gene g_{2J} in a genome Γ_2 such that it has either a unique matching or the best similarity score (above a threshold) with another gene g_{1I} in a genome Γ_1 during the pair-wise comparison of genomes Γ_1 and Γ_2. A *paralog* is defined as a duplicated gene caused by gene duplication. Paralogous genes correspond to functionally similar yet not functionally equivalent genes.

A *homologous gene group* is a cluster of neighboring genes $<g_I, g_J, g_K\ldots>$ with at least two distinct genes those have a natural pressure to occur in close proximity. *Close proximity* of two gene positions indexed by I and J is defined as $0 < I - c < J < I + c <$ genome size and $0 < J - c < I < J + c <$ genome size where c is a small constant [3].

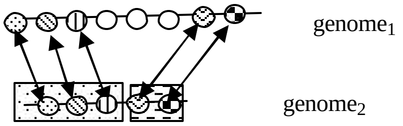


Figure1. Homologous gene-groups and genome rearrangement

Orthologous gene-groups comprise orthologous genes, and are important to identify operons ? smallest functional unit putatively in the same pathway.

A *co-transcribed* gene-group has two or more genes, which are expressed together; the control region of the first gene in the group is involved in the expression of all other genes, and the control region of other genes is either missing or very small to be involved individually in transcription.

2.3 Enzymes and Pathways

An *enzyme* is a protein catalyzing a reaction. A metabolic pathway is the complete network of reactions catalyzed by enzymes to transform (break up or synthesize) proteins and biochemical products.

A *metabolic pathway* is modeled as a network of reaction graphs. 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 detailed description of the various metabolic pathway modules is available at the KEGG site (http://www.genome.ad.jp).

A *pathway seed* is a gene-group (identified using pair-wise genome comparison and the identification co-transcribed gene-group) consisting of at least two genes and at least one enzyme in a pathway. A pathway seed will be modeled as a set of genes.

C*orresponding pathway seeds* in one genome are the set of pathway seeds identified by pair-wise comparisons of the genome (in question) with other genomes such that all the pathway seeds share at least one gene.

A *pathway trace* is a set of proteins in a pathway module such that the corresponding enzyme reaction-graph is a subgraph within a metabolic pathway module. A pathway trace contains at least one pathway seed.

2.4 Nomenclature

This paper uses GenBank nomenclature for gene-names and abbreviates gene names. The enzymes names have identified as EC (enzyme classification) number, and the correspondence of an enzyme with a gene has been given as a pair of the form *gene name: EC number*. A co-transcribed group is placed within square brackets, and a gene-group is placed within curly brackets. The number of nucleotides in a control sequence Y_I are denoted by $|Y_I|$.

3. Deriving Orthologs and Gene-groups

This section briefly describes algorithms used to identify orthologs, homologous and orthologous gene groups [3], and co-transcribed gene-groups.

Genomes were extracted from the GenBank using '.gbk' file format. The gene-groups obtained from pair-wise comparison of two-genomes?Γ_1 and Γ_2 were obtained using Goldie 2.0 [3] (see http://www.cs.kent.edu/~arvind/orthos.html) — a software library for automated comparison of genomes — to identify orthologs, homologous gene-groups, and gene-fusion.

3.1. Identifying Orthologs

The pair-wise comparison of two genomes is modeled as a weighted bipartite graph matching problem [3, 4]. The weights of the edges are identified using the Smith-Waterman algorithm. In order to improve the execution efficiency, a number of gene-pairs are pruned based on BLAST similarity techniques [1].

After identifying the weights of the edges, the edges are sorted in the descending order. The set of nodes corresponding to the highest weighted edges are collected as putative orthologs. After finding an edge (p_{1I}, p_{2J}) of the highest weight, all the edges involving the nodes p_{1I} and p_{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 biased positively as genes within a gene group being better candidates for preserving a common function. A detailed algorithm is given in [3].

3.2. Gene-groups using Genome Comparison

A neighboring group S_0 for a gene (in Γ_1), that has a corresponding homolog in Γ_2, is marked. A *neighboring group* is a group of genes in the close proximity where proximity is user defined. 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 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. A detailed discussion of the algorithm is available in [3].

3.3 Co-transcribed Gene-groups

Genes in a *co-transcribed gene-group* are expressed together. A co-transcribed gene-group forms a single smallest functional unit in a pathway.

Co-transcribed gene-groups are identified by the size of the control region of each gene. A gene (y_I, h_I) has a separate regulatory region if the size $|y_I|$ is greater than a threshold value. The default value chosen was 65 nucleotides. A gene with a separate regulatory region starts a new set of genes. All the following genes (y_I, h_I) y_J ($J > I$) that have the size of the control region is less than the threshold are collected in the same set. The old set is closed, and a new set is started. Each set with two or more genes forms Co-transcribed group of genes.

0-7695-1423-5/01 $10.00 © 2001 IEEE

4. Identifying Pathway Traces

In this section, we describe a technique to identify pathway traces. We will illustrate this technique with respect to the derivation of citrate cycle pathway in *M. tuberculosis* genome. The identification of pathways traces has following steps:

<u>Step 1:</u> First the orthologs are identified using pair-wise genome comparisons. Functional/enzyme annotations present in other well annotated genomes and SwissProt database are also used to annotate the genes in *M. tuberculosis*. The annotations from SwissProt database are preferred.

<u>Step 2:</u> After annotating the *M. tuberculosis* genes with enzyme numbers, the genes corresponding to the enzymes in the pathways (to be derived) of *E. coli* and *B. subtilis* are identified since the pathways in two genomes have been experimentally investigated in wet laboratories. KEGG database was used to collect the corresponding enzymes of *E. coli* and *B. subtilis.* The union of this set of genes formed the starting point. Let this set be S_1.

<u>Step 3:</u> The orthologous genes in *M. tuberculosis* corresponding to the enzymes in S_1 are collected. In the absence of orthologs all the paralogs with highest similarity are selected. The rationale is that one of the paralogous genes could substitute the function of the corresponding enzyme. Let this set be S_2.

<u>Step 4:</u> Other remaining genomes are compared against *E. coli* and *B. subtilis* first using NIH-BLAST [1] (using a cutoff of high-score value of 50 and chance score value less than 1.0 X 10^{-4}) and then the Smith-Waterman alignment [21]. After the Smith-Waterman alignment, All the homologs of S_1 with greater than 28% identity and at least 30 % of the size of the larger gene in a homolog pair are collected for other genomes, and the gene-groups containing these homologs are collected. Let the set of gene-groups be S_3.

<u>Step 5:</u> The gene-groups in *M. tuberculosis* those are homologous to gene-groups in the set S_3 are identified. These gene-groups become the pathway seeds. The *M. tuberculosis* genes in these gene groups are annotated using the information in Step 1. Let this annotated set of gene-groups be S_4.

<u>Step 6:</u> The set of pathway seeds in S_4 is compressed into initial set of pathway traces by merging two pathway seeds sharing a common gene. The merged pathway seed contains union of the genes in the two pathway seeds sharing a common gene. After the merger, two pathway seeds are replaced by the resulting merged pathway seed. The process is repeated until no element in the set shares any common gene with other elements. After the merger, if the group $\{\gamma_I, \ldots, \gamma_J\}$ missed a gene γ_K such that $I < K < J$, then γ_K was included in the merged group. Let this set of pathway seeds be S_5.

<u>Step 7:</u> The co-transcribed gene-groups were identified for *M. tuberculosis* using the technique described in Section 3.3. Let this set be S_6.

<u>Step 8:</u> For each gene-group in the set S_5, the corresponding set of co-transcribed gene-group was identified. The union of these set of co-transcribed gene-groups identified an initial pathway trace. This process was repeated for every gene-group in S_5 to identify other pathway traces. At the end of the process, the individual pathway traces were collected. These pathway traces contain extra information about non-enzymatic genes and regulatory genes in addition to some additional enzymes. Let this set be S_7.

<u>Step 9:</u> The pathway traces in set S_7 containing the orthologs in the set S_2 were identified. These traces were filtered as the final pathway traces. Other pathway traces were discarded.

<u>Step 10:</u> Our computational experiment has shown that the pathway traces making up a pathway may be scattered on different parts of the genomes. In order to build a complete pathway, the various traces have to be connected if a product in one pathway trace is consumed in the second pathway trace.

Multiple pathway traces were merged (as described in the following Subsection) using reaction-graph using orthologous enzymes if they shared a common substrate or one of the products from one pathway trace is consumed by another pathway trace.

To merge the pathway traces, KEGG database (see http://www.genome.ad.jp/kegg-bin/mk_point_html), LIGAND database and SwissProt database were used to identify the reactions and their classification.

5. An Example – Deriving Citrate Cycle

In this section, we apply our technique to identify *citrate cycle* pathway of *Mycobacterium tuberculosis*.

Mycobacterium tuberculosis is the cause of a deadly 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.

We compared *M. tuberculosis* with seven genomes: *Bacillus halodurans, Bacillus subtilis, Escherichia coli strain K12, Halobacterium sp., Helicobacter pylori* strain 26695, *Lactococcus lactis,* and *Vibrio cholera* to identify homologous gene-groups of citrate cycle. *E. coli* and *B. subtilis* are well explored in wet laboratories, *H. pylori*, *L. lactis*, and *V. cholera* are pathogens, and *Halobacterium sp*. is an archaea. After the Step 1, 14 enzymes (see Table I) were identified in citrate cycle for *E. coli* and *B. subtilis*.

Only the homologs (of enzymes) those were at least 30 % of the size of other genome and had an identity match at least 27% were selected. The homologs were matched in descending order of similarity. After the step 3, the corresponding orthologs of the enzyme was identified for *M. tuberculosis*. The data from the comparisons of seven genomes are summarized in Table I. The last column of Table IB contains the orthologs. For convenience, we have abbreviated *E. coli* by *EC*, B. subtilis by *BS*, *B. halodurans* by *BH*, *Halobacterium sp.* by *HB*, *H. pylori* by HP, *L. lactis* by *LL*, and *V. cholera* by *VC*.

Enzyme	***EC***	***BS***	***BH***	***HB***
4.1.1.9	*pckA*	*pckA*	*pckA*	*none*
6.4.1.1	*accC* *carB*	*pycA*	*pycA* *BH1132..34* *accC* *carB* *pyrAB*	*acc* *carB*
1.1.1.37	*mdh*	*lctE* *citH*	*citH* *lctE*	*mdhA*
4.2.1.3	*acnB* *leuC*	*leuC*	*leuC*	*can*
4.1.3.7	*gltA* *orf.325*	*citZ* *citA* *mmgD*	*citZ* *mmgD*	*citZ*
4.2.1.2	*fumC* *fumA*	*citG* *ansB* *argH* *purB*	*citG* *ansB* *argH* *purB* *BH0894*	*fumC* *purB* *argH*
1.3.99.1	*sdhA,sdhB* *frdA,frdB* *nadB*	*sdhA,sdhB* *nadB*	*sdhA.sdhB* *nadB*	*sdhA,sdhB* *nadB*
4.1.3.34	*citE*	*none*	*none*	*citE*
4.1.3.6	*citF*	*none*	*none*	*none*
6.2.1.5	*sucC/D* *orf.312* *orf.509*	*sucC/D*	*sucC/D*	*sucC/D*
2.3.1.61	*sucB* *aceF*	*odhB* *bfmBB* *pdhC* *acoC*	*BH2205* *acoC* *pdhC* *BH0778* *BH0215*	*dsa*
1.2.4.2	*sucA*	*odhA*	*BH2206*	*none*
1.1.1.42	*icdA* *leuB*	*citC* *leuB* *ycsA*	*citC* *leuB* *BH1070*	*icd*
1.8.1.4	*lpdA* *udhA* *gor* *nirB* *ykgC*	*pdhD* *ycgT* *acoL* *yqiV*	*pdhD* *BH0216* *BH0779* *acoL* *BH2764* *BH3776*	*lpdA* *merA*

Table IA: Best homologous genes to enzymes

Enzyme	***HP***	***LL***	***VC***	***MT orthologs***
4.1.1.9	*none*	*none*	*VC2738*	*none*
6.4.1.1	*HP0370* *HP0919*	*pycA* *accC* *carB*	*VC0295* *VC0550* *VC2389* *VC2413*	*pca.2961*
1.1.1.37	*none*	*ldhB* *ldhX* *ldh*	*VC0432*	*mdh.1238*
4.2.1.3	*HP0779*	*none*	*VC0604*	*acn.1471*
4.1.3.7	*HP0026*	*gltA*	*VC1337* *VC2092*	*gltA2.894*
4.2.1.2	*HP1325* *HP0649* *HP1112*	*argH* *purB*	*VC1573* *VC2698* *VC1304* *VC2641*	*fum.1096*
1.3.99.1	*HP0192*	*frdC+ycF*	*VC2088/9* *VC2656/7* *VC2469*	*sdhA.3312,* *sdhB3311* *frdA.1548,* *frdB.1549*
4.1.3.34	*none*	*citE*	*VC0798*	*citE.2493*
4.1.3.6	*none*	*citF*	*VC0799*	*none*
6.2.1.5	*none*	*none*	*VC2085/4*	*sucC,sucD*
2.3.1.61	*none*	*pdhC*	*VC2086* *VC2413*	*sucB.2209*
EC 1.2.4.2	*none*	*none*	*VC2087*	*sucA.1246*
1.1.1.42	*HP0027*	*icd* *leuB*	*VC2491*	*leuB.2981*
1.8.1.4	*none*	*phdD* *yvdG* *gshR* *ahpF* *noxD*	*VC2412* *VC0151* *VC2638*	*RV0462.461*

Table IB: Homologs of enzymes with high similarity

After the step 4, the corresponding gene-groups containing the homologs were identified in other genomes: *B. halodurans, Halobacterium sp., H. pylori strain* 26695, *L. lactis,* and *V. cholera.* After the step 5, the corresponding homologous groups were identified in *M. tuberculosis*. These pathway seeds were compressed to give S_5 — a set of pathway seeds. After the step 8, we identified the set of pathway traces from homologs. After the step 9, only the set of traces, containing the orthologs (or best paralogs) of the corresponding enzymes were kept, all other traces were discarded. This set is given in Table II. The orthologs (or best paralogs) are marked in boldface.

These set of pathway traces were merged in the step 10 to give the corresponding pathway for *M. tuberculosis*. The list of pathways involving these enzymes is given in Table III, and the corresponding pathway connectivity is given in Figure 2.

Enzymes	Pathway traces
4.1.1.9	none
6.4.1.1	[***pca:6.4.1.1*** (fused gene), *Rv2968c, Rv2969c*]
1.1.1.37	{[***mdh:1.1.1.37***, *Rv1241*], *Rv1242, PE_PGRS, lpqZ,* [*Rv1245c, Rv1246c, Rv1247c*], *sucA:1.2.4.2, Rv1249c, Rv1250*}
4.2.1.3	([*ctpD:3.6.1.-, trxA:thioredoxin, trxB1.6.4.5, echA12:4.2.1.17, Rv1473*], [*Rv1474c,* ***acn:4.2.1.3***])
4.1.3.7	{[*Rv0888, citA:4.1.3.7*], [*Rv0890, Rv0891*], [*Rv0892, Rv0893*], *Rv0894, Rv0895,* [***gltA2:4.1.3.7***, *Rv0897, Rv0898*] }
4.2.1.2	[*Rv1096, Rv1097,* ***fum:4.2.1.2***, *Rv1099*]
1.3.99.1	{[*sdhC:1.3.99.1, sdhD:1.3.99.1*], [***sdhA:1.3.99.1, sdhB:1.3.99.1*]}**
4.1.3.34	[***citE:4.1.3.34***, *Rv2499c, fadE19, accA1:6.3.4.14, accD1:6.4.1.2:, scoB:2.8.3.5, scoA:2.8.3.5*]
4.1.3.6	none
6.2.1.5	{[*Rv0941c, Rv0942, Rv0943c, Rv0944, Rv0945:1.-.-.-, pgi:5.3.1.9], Rv0947c, Rv0948c, uvrD:3.6.1._,* [*Rv0950,* ***sucC:6.2.1.5***, ***sucD:6.2.1.5***, *Rv0953*]}
2.3.1.61	***sucB:2.3.1.61***
1.2.4.2	{[*gcvT:2.1.2.10, Rv2212, pepB:3.4.11.1, ephD:1.-.-.-*], [***sucA:1.2.4.2***, *Rv2216, lipB: 6.-.-.-, lipA:3.1.1.3, Rv2219*]}
1.1.1.42	{[*leuD:4.2.1.33, leuC:4.2.1.33*], [*Rv2989, Rv2990c*], *Rv2991*, [*gltS:6.1.1.17, Rv2993c*], *Rv2994,* ***leuB:1.1.1.85***}
1.8.1.4	[*Rv0458, Rv0459, Rv0460, Rv0461,* ***Rv0462:1.8.1.4***, *Rv0463, Rv0464c, Rv0465c*]

Table II: Pathway traces after merging the pathway seeds and filtering the trace containing the orthologs

6. Discussions

Table I shows that there are variations in the pathway. Many times the genes are fused or split, and the genes may be deleted/inserted in a pathway. Since we started with well-explored pathways, we could not investigate the insertion of new enzymes other than those derived by homologous gene-groups. However, the absence of any homolog of enzymes from many genomes demonstrates the variations in metabolic pathway. For example, homologs for enzymes *EC 4.1.3.34* and *EC 4.1.3.6* are missing both from *B. subtilis* and *Halodurans sp..* Similarly, multiple enzymes such as *EC 4.1.1.9*, *EC 4.1.3.34*, *EC 4.1.3.6*, *EC 6.2.1.5*, *EC 2.3.1.61*, *EC 1.2.4.2*, *EC 1.8.1.4* are missing from the citrate cycle of *H. pylori*.
Similarly, homologues for the enzymes *EC 4.1.1.49* and *EC 4.1.3.6* are missing from *M. tuberculosis*.

KEGG pathways	Enzymes/co-transcribed Enzymes in Table II
Glycolysis	*1.8.1.4, 5.3.1.9*
Citrate cycle	*1.1.1.37, 1.2.4.2, 1.3.99.1, 1.8.1.4, 2.3.1.61, 4.1.3.6, 4.1.3.7, 4.1.3.34, 4.2.1.2, 4.2.1.3, 6.2.1.5, 6.4.1.1*
Pentose phosphate cycle	*5.3.1.9*
Fatty acid biosynthesis I	*6.3.4.14, 6.4.1.2*
Fatty acid biosynthesis II	*4.2.1.17*
Fatty acid metabolism	*4.2.1.17*
synthesis and degradation of ketone bodies	*2.8.3.5*
Ubiquinone biosynthesis	*6.-.-.-*
Oxidative phosphorylation	*1.3.99.1*
Pyrimidine metabolism	*1.6.4.5, thioredoxin*
Glutamate metabolism	*6.1.1.17*
Alanaine and aspartate metabolism	*6.4.1.1*
Tetracycline biosynthesis	*6.4.1.2*
Glycine, serine, and threonine metabolism	*1.8.1.4, 2.1.2.10*
Valine, leucine, isoleucine degradation	*2.8.3.5, 4.2.1.17*
Valine, leucine, isoleucine biosynthesis	*1.1.1.85, 4.2.1.33*
Lysine degradation	*1.2.4.2, 2.3.1.61, 4.2.1.17*
Tryptophan metabolism	*1.2.4.2, 4.2.1.17*
β-alanine metabolism	*4.1.1.9, 4.2.1.17*
Starch and sucrose metabolism	*3.6.1.-, 5.3.1.9*
Glycerolipid metabolism	*2.3.1.62, 3.1.1.3*
Pyruvate metabolism	*1.1.1.37, 1.8.1.4, 6.4.1.1, 6.4.1.2*
Glyoxylate & Dicarboxylate metabolism	*1.1.1.37, 4.1.3.7, 4.2.1.3*
Propanoate metabolism	*4.1.1.9, 4.2.1.17, 6.2.1.5, 6.4.1.2*
Butanoate metabolism	*1.3.99.1, 2.8.3.5, 4.2.1.17*
C5 branched dibasic acid metabolism	*6.2.1.5*
Carbon fixation	*1.1.1.37*
Reductive carboxylate cycle	*1.1.1.37, 1.3.99.1, 4.1.3.6, 4.2.1.2,*
Folate biosynthesis	*3.6.1.-, 6.-.-.-*
Porphyrin and chloraphyll	*1.-.-.-, 6.1.1.17*
aminoacyl-tRNA biosynthesis	*6.1.1.17*
Nitrogen metabolism	*2.1.2.10*

ble III: Associated pathways in KEGG database

0-7695-1423-5/01 $10.00 © 2001 IEEE

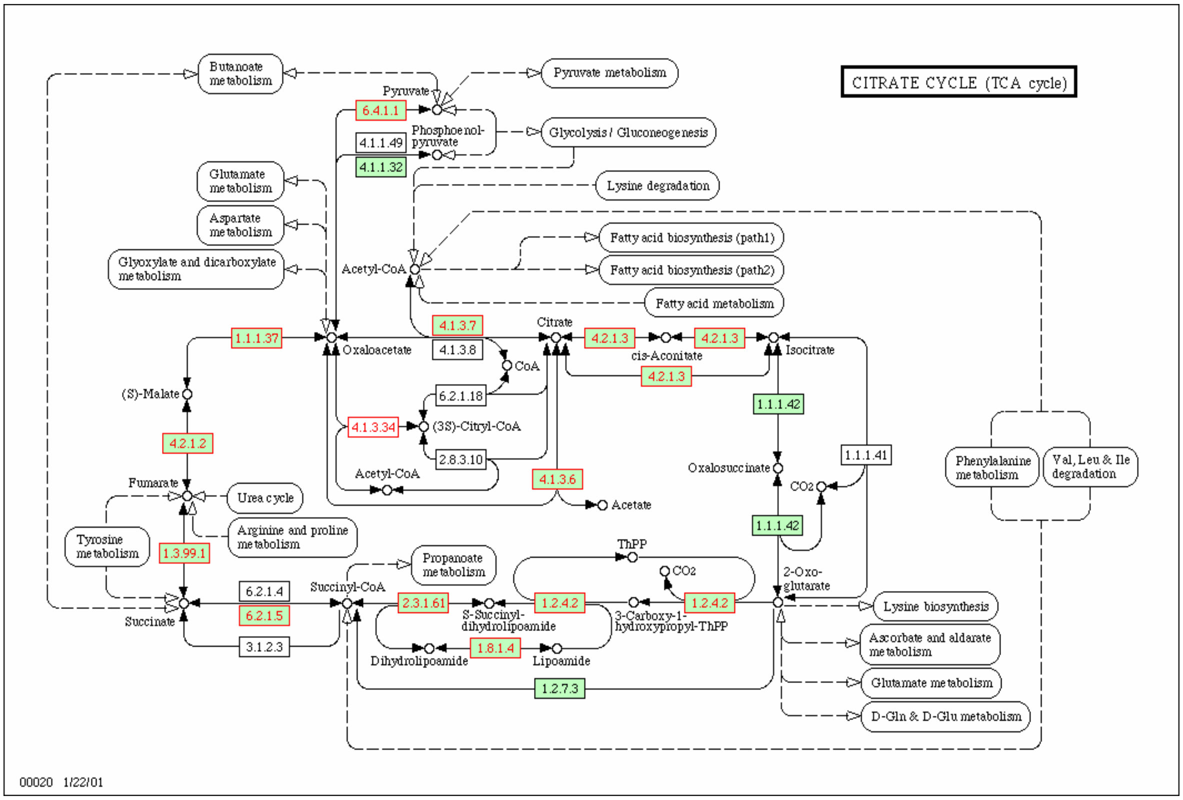


Figure 2. Citrate cycle of *M. tuberculosis* in KEGG pathway database

After merging the pathway seeds, the citrate cycle for *M. tuberculosis* was reconstructed. Two enzymes *EC 4.1.1.9* and *EC 4.1.3.6* were missing. Using BLAST sequence comparison, no similarity was found for these two enzymes. The enzyme *EC 4.1.3.6* is present only in proteobacteria in the set of seven genomes. Such missing enzymes may be used for phenotype characterization of the group of bacteria. However, the enzyme *EC 4.1.1.9* is missing from pathogens *M. tuberculosis, L. lactis, and H. pylori* as well as *Halobacterium sp., which* does not give any association with the classification of bacteria. It would be interesting to see if such missing enzymes may be correlated with the life style. Such missing enzymes may point out either to an abundance of supply of the product catalyzed by the enzyme obviating the need for the enzyme, or a need for the abundance of the product catalyzed by missing enzymes to trigger the pathway.

Table II also shows many co-transcribed genes those are absent from KEGG pathway. For example, enzymes *EC 4.2.1.17, EC 4.2.1.13, thioredoxin, EC 1.6.4.5, EC 6.3.4.14, EC 6.4.1.2, EC 2.8.3.5, EC 1.-.-.-, EC 5.3.1.9, EC 3.6.1.-, EC 4.2.1.33, EC 6.1.1.17, EC 1.1.1.85* are either co-transcribed with the enzymes or they are in the same pathway trace after the step 10 of the technique. A closer analysis of the pathways involving these enzymes shows that most of these enzymes are in the neighboring pathways of citrate cycle with majority in the adjacent pathways.

For example, butanoate metabolism, oxidative phosphorylation, pyruvate metabolism, glyclolysis/gluconeogenesis, lysine degradation, fatty acid biosynthesis I, fatty acid biosynthesis II, fatty acid metabolism, propanoate metbolism, Valine/leucine/ isoleucine degradation, glutamate metabolism, aspartate metabolism, glyoxylate and dicarboxylate metabolism, tryptophan metabolism (not shown in KEGG diagram in Figure 2), synthesis and dergradation of ketone bodies (not shown in Figure 2), and tetracycle bio-synthesis (not shown in Figure 2) are all adjacent (connected by the product acetyl-coA) to citrate cycle. Citrate cycle is also closely linked to reductive carboxylate cycle. All the enzymes shared in the adjacent pathways are at the periphery connecting to citrate cycle. Few pathways such as pentose phosphate pathway, tetracycline biosynthesis, starch and sucrose metabolism, folate biosynthesis, chloraphyll and porphyrin metabolism, valine/leucine/isoleucine biosynthesis and aminoacyl-tRNA biosynthesis do not seem to be adjacent to citrate cycle. However, a closer analysis shows pentose phosphate cycle shares a common enzyme *EC 5.3.1.9* with glycolysis pathway which is at periphery of citrate cycle and glycolysis pathway; valine/leucine/isoleucine biosynthesis is adjacent to valine/leucine/isoleucine degradation; carbon fixation is adjacent to reductive carboxylate cycle; β-alanine metabolism is adjacent to fatty acid biosynthesis, and is connected to acetyl-coA (in citrate cycle) through enzymes *EC 6.4.1.2* and *EC 6.3.4.14*; porphyrin and cholorophyll metabolism is adjacent to glutamate metabolism; nitrogen metabolism and C-5 branched carbon cycle are adjacent glutamate metabolism. Thus it can be concluded from enzymes in co-transcribed groups and pathway trace that pathway separation is artificial, and multiple adjacent pathways are related to each other through co-transcribed enzymes or enzymes in gene-groups identified by pair-wise genome comparisons.

8. Conclusion

In this paper, a novel technique has been described to automate and refine the reconstruction of metabolic pathways of newly sequenced microbial genomes using the pathways derived in wet laboratories. The technique extends the previous technique by integrating the co-transcribed gene-groups with homologous gene-groups with high similarity derived by multiple pair-wise genome comparisons, and orthologs. The technique reduces the need for heuristics due to the integration of multi genome comparison and co-transcribed gene-groups.

The analysis of homologs for the enzymes for citrate cycle shows that pathways vary a lot. This variation may be used for phenotype characterization of family of genomes, and may be indicative of their life style. As shown by *H. pylori*, multiple missing enzymes present in other closely related non-pathogenic genome may show the need for the end-product of the set of missing enzymes for the related pathways to remain active. In order to refine the pathways, there is a need to integrate more biochemical information. This study leads to yet another comparative study to understand the variations in similar pathways. Such an understanding will facilitate our understanding of diseases caused by metabolic pathway variations and imbalances.

References

[1] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, “Basic Alignment Search Tools,” *J. Mol. Biol.*, Vol. 215, 403 – 410, 1990

[2] A. Bairoch, “The ENZYME database in 2000,” *Nucleic Acids Research*, pp. 304-305, 2000

[3] A. K. Bansal, P. Bork. P.J. Stuckey, “Automated Pair-wise Comparisons of Microbial Genomes,” *Mathematical Modeling and Scientific Computing*, 9:1, pp. 1 – 23, 1998

[4] A. K. Bansal, “An Automated Comparative Analysis of 17 Complete Microbial Genomes,” *Bioinformatics*, 15:11, pp. 900 – 908, 1999

[5] A. K. Bansal, "A Framework of Automated Reconstruction of Microbial Metabolic Pathways," *Proceedings of IEEE International Conference of Bioinformatics and Biomedical Engineering*, pp. 184-190, 2000, Washington D. C., USA

[6] A. K. Bansal and C. J. Woolverton, Applying Automatically Derived Gene-groups to Automatically Predict and Refine Microbial Pathways,” Special issue on Bioinformatics, IEEE Transactions of Knowledge and Data Engineering, to appear, in press

[7] H. Bono, H. Ogata, S. Goto, M. Kanehisa, “Reconstruction of amino acid biosynthesis pathways from the complete genome sequence,” *Genome Res.*, Vol. 8, Issue 3, pp. 203-210, 1998

[8] P. Bork,, T. Dandekar, Y. Diaz-Lazcoz, F. Eisenhaber, M. A. Huynen, and Y. Yuan, “Predicting Function: From Gene to Genomes and Back,” *J. Mol. Biol.*, Vol. 283, pp. 707-725, 1999

[9] M. W. Covert, C. H. Schilling, I. Famili, J. S. Edwards, I. I. Goryannin, E. Selkov, and B. O. Palsson, "Metabolic Modeling of Microbial Strains in Silico," Trends in Biochemical Sciences, Vol. 26, No. 3, March 2001

[10] R. Brosch, S.V. Gordon, A. Pym, et al., "Comparative genomics of the Mycobacteria," J. Med. Microbiol. 290:2 (2000) 143-152

[11] S. J. Cordwell, “Microbial genomes and missing enzymes: redefining biochemical pathways,” *Arch Microbiol*ogy, 172:5 pp. 269-79, 1999

[12] W. M. Fitch, “Distinguishing Homologous from Analogous Proteins,” *Systematic Zoology*, Vol. 19, pp. 99 – 113, 1970

[13] S. Goto, T. Nishioka, M. Kanehisa, “LIGAND: chemical database for enzyme reactions,” *Bioinformatics.* 14:7, pp. 591-599, 1998

[14] P. D. Karp, M. Krummenacker, S. Paley, J. Wagg, “Integrated pathway-genome databases and their role in drug discovery,” *Trends Biotechnol*ogy, Vol. 17, Issue 7 (1999) 275-281.

[15] W. C. Lathe, B. Snel, and P. Bork, "Gene Context Conservation of a Higher Order than Operons," Trends Biochem Sci 2000, 25(10): 474-9.

[16] R. Overbeek et. al., "WIT: Integrated System for High Throughput Genome Sequence Analysis and Metabolic Reconstruction, Nucleic Acids Research, Vol. 28, pp. 123-125.

[17] E. Selkov, N. Maltsev, G. J. Olsen, R. Overbeek, W. B. Whitman, “A Reconstruction of the Metabolism of *Methanococcus jannaschi* from Sequence Data,” *Gene*, Vol. 197, Issues (1-2): GC (1997) 11-26

[18] E. Selkov Jr, Y. Grechkin, N. Mikhailova, E. Selkov, “MPW: the Metabolic Pathways Database,” *Nucleic Acids Research*, Vol. 26, Issue 1, (1998) 43-45.

[19] S. Schuster, T. Dandekar, D. A. Fell, “Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering,” *Trends Biotechnology, Vol.* 17, Issue 2 (1999) 53-60.

[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, (1996) 279 – 291

[21] M.S. Waterman, “Introduction to Computational Biology: Maps, Sequence, and Genomes,” *Chapman & Hall*, (1995)

IEEE International Symposium on Bio-informatics and Biomedical Engineering (BIBE'01)
0-7695-1423-5/01 $10.00 © 2001 IEEE