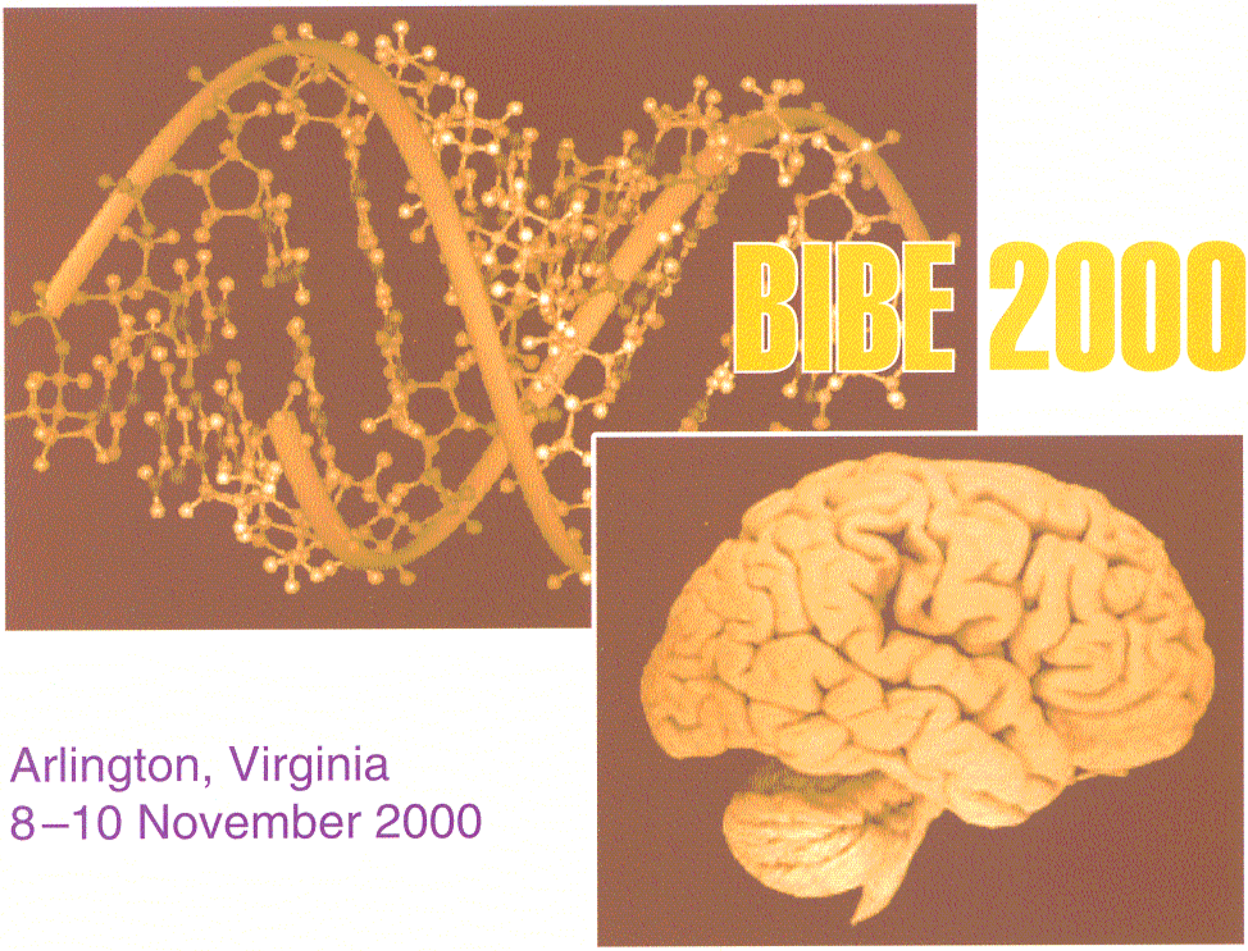
Proceedings

IEEE International Symposium on

Bio-Informatics and Biomedical Engineering



BIBE 2000

Arlington, Virginia
8–10 November 2000

Sponsors:
IEEE Computer Society
IEEE Engineering in Medicine and Biology Society
IEEE Computer Society Task Force on Virtual Intelligence (TFVI)

IEEE
COMPUTER
SOCIETY

Proceedings of the IEEE International Symposium on Bio-informatics and Biomedical Engineering
Arlington, VA, 8-10 November 2000, ISBN 0-7695-0862-6, Library of Congress Number 00-107812

A Framework of Automated Reconstruction of Microbial Metabolic Pathways

Arvind K. Bansal
Department of Mathematics and Computer Science
Kent State University, Kent, OH 44242, USA
arvind@mcs.kent.edu

and

Intellibio Software and Consultancy Corporation
3109 Killingworth Lane, Twinsburg, OH 44087, USA
arvind@intellibiosoft.com

Abstract

This paper describes a framework of automated reconstruction of metabolic pathways using the information about orthologous and homologous gene-groups derived by the automated comparison of the whole genomes archived in the GenBank. The method integrates automatically derived orthologs, orthologous, and homologous gene-groups (http://www.mcs.kent.edu/ ~arvind/orthos.html), the biochemical pathway template available at Kegg database (http://www.genome.ad.jp), and enzyme information derived from SwissProt enzyme database (http://expasy.hcuge.ch/) and Ligand database (http://www.genome.ad.jp). The technique is useful to identify refined metabolic pathway based on operons, and to derive the non-enzymatic genes within a group of enzymes. The technique has been illustrated by the comparison of E. coli with B. subtilis genome.

Keywords: bacteria, drug-discovery, gene-groups, metabolic pathway, microbes, operons, orthologs

1. Introduction

Understanding the microbial machinery has important advantages in bio-remediation, development of more effective antibiotics, and anti-bacterial agents based upon the regulation of metabolic pathways. At this point 31 microbial genomes (including many pathogens) have been sequenced, and more than 50 are underway. An important aspect of understanding the microbial machinery is to derive the metabolic pathway — the control flow diagram of network of enzymatic reactions, and the regulation of these pathways. By blocking a part of a metabolic pathway or by regulating the gene-expression, the production of a specific biochemical product can be regulated. This regulation of pathways will help us to understand the dynamic behavior of the pathway, and to study the rate of change of biochemical product. The study of dynamic behavior of pathway will facilitate the development of more effective anti-bacterial drugs with little side effects [16]. However, in order to understand the metabolic mechanism metabolic pathway should be mapped accurately.

Previous techniques to derive metabolic pathways [9, 14, 17] are based upon deriving the best homologs, from genome comparison and identifying the corresponding enzymes using Ligand or SwissProt database, and using the knowledge of biochemical pathways and enzyme reaction available from the literature. These techniques are sound and have provided excellent base to understand metabolic pathways of newly sequenced genomes. However, these techniques are limited as follows:

The current techniques do not take complete advantage of orthologous (having the same functionality) and homologous (having similar functionality) gene-groups. 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 functionality within the same metabolic pathway, and co-regulated genes possibly sharing the same control region.

In this paper, we describe a framework of an automated technique to reconstruct metabolic pathways of newly sequenced genomes based upon first mapping the automatically derived orthologous and homologous gene-groups on a standard pathway template as given in Kegg database (see http://www.genome.ad.jp/kegg-bin/ mk_point_html), refining the template using the information of gene-groups, and extending the gene-groups using the adjacent genes information.

0-7695-0862-6/00 $10.00 © 2000 IEEE

The result presented in this paper shows that orthologous gene-groups and a large class of homologous gene-groups obtained from pair-wise genome comparison map on local sections of the metabolic pathway. This is a significant finding since the local sections of a metabolic pathway become the seed points to trace the metabolic pathway for newly sequenced genomes.

The first attempt to use orthologs and gene-groups to reconstruct metabolic pathway manually was made by Mushgesian et. al [17]. However, their reasoning was limited to comparing evolutionary close microbes such as *E. coli* and *H. influenzae*. Recent studies by the author [4] have shown that all genomes share many gene-groups, the number of shared gene-groups is also a function of the size of the genomes, and different genomes may share different gene-groups. This finding has significant impact since a newly sequenced genome can be compared with a large number of genomes to identify almost all the gene-groups. This work significantly extends the work in [12] by using all the genomes for the automated comparison to identify a comprehensive set of gene-groups, and present a comprehensive technique for automation.

The major contributions of this paper are:

(i) 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.

(ii) Current techniques to reconstruct metabolic pathway have been refined and extended by using a comprehensive set of genomes to identify gene-groups using pair-wise genome comparison.

(iii) The current metabolic pathways derived by ortholog analysis and biochemical reactions can be curated. This is possible since an enzyme with slightly different functionality within a gene-group will substitute for the enzyme found using ortholog analysis.

(iv) Non-enzymatic genes embedded inside a group of enzymes have been identified. These genes may be involved in the regulation of the gene-group, or they may themselves be affected in some way by the reactions in the gene-group.

(v) It establishes that a large percentage of the multi gene groups involved are sensor or regulatory proteins

The database results from this technique will also help the wet lab experimenters to identify the remaining enzymes in the genomes, by looking for missing enzymes in a pathway trace. Another advantage is that the association of gene-group with metabolic pathway modules will improve the efficiency of the metabolic pathway deduction.

The result in this paper is limited to *E. coli* vs. *B. subtilis* comparison. The rationale for choosing *E. coli* and *B. subtilis* is to verify the validity of this approach since a large percentage of the enzymes in *E. coli* and *B. subtilis* have been established using wet lab techniques.

The paper is organized as follows: Section 2 describes the background, orthologs, homologs [4, 7], homologous gene-groups [3, 4], and enzymes and metabolic pathways [9, 15]. Section 3 describes briefly the techniques to derive orthologs and homologous gene-groups [3, 4]. Section 4 describes a technique to reconstruct the metabolic pathways integrating homologous gene-groups, chemical reactions of enzymes (as given in Ligand and SwissProt database), and the standard template of the metabolic pathways available at Kegg site (see http://www.genome.ad.jp). Section 5 describes the results to substantiate the hypothesis that the orthologous and homologous gene-groups mark a large subset of the local sections in a metabolic pathway. Section 6 describes the future work and 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 pathway, and provide some new definitions needed to explain the concepts.

2.1. Modeling genomes

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

2.2. Orthologs

An *ortholog* is a functional counterpart of a gene in another genome that has arisen from speciation [7]. Due to the inherent uncertainty in phylogeny due to the lateral transfer of genes [4,11], gene insertions and deletions, gene fusion and splitting [3], and difference in the evolutionary trees based upon different criterion, this paper uses a definition based upon sequence similarity to define putative orthologs. A *putative ortholog* is defined as a gene γ_{2J} in a genome Γ_2 such that it has the best similarity score (above a threshold) with another gene γ_{1I} in a genome Γ_1 during the pair-wise comparison of genomes Γ_1 and Γ_2.

2.3. Classes of gene groups

A *gene group* Σ is a cluster of neighboring genes $<\gamma_I, \gamma_J, \gamma_K\ldots>$ with at least two distinct genes which have a natural pressure to occur in close proximity. *Close proximity* of two gene positions indexed by I and J is defined as $0 < I - c < J < I + c <$ genome size and $0 < J - c < I < J + c <$ genome size where c is a small constant experimentally limited by 12.

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

$M < N < P$ when $I < J < K$, or (1)
$M > N > P$ when $I > J > K$ (2)

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

($M > N$ when $I < J$) or ($N > P$ when $J < K$) (3)
(M < N when I > J) or (N < P *when* J > K) (4)
(M ≠ N *and* I = J) or (N ≠ P *and* J = K) (5)

Orthologous gene-groups comprise only orthologous genes. The study of orthologous gene groups is important to annotate the function of gene groups in genomes.

A *multigene group* Σ_{1I} in a genome Γ_1 satisfies the following two conditions:

(i) Σ_{1I} has a corresponding gene group Σ_{2M} in the genome Γ_2, and

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

A fused gene group has a gene γ_{1I} in the genome Γ_1 such that two (or more) non-overlapping [3] protein subsequences $\delta_{1M}, \delta_{1N} \subseteq \pi_{1I}$. $\delta_{1M} \not\subset \delta_{1N}$ and $|\delta_{1M} - \delta_{1N}| \leq \in$) align with the protein subsequences $\delta_{2U} \subseteq \pi_{2J}$ and $\delta_{2V} \subseteq \pi_{2K}$ ($J \neq K$, and J and K are in close proximity).

2.4. Enzymes and metabolic 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 directed graph (including cyclic subgraphs) with biochemical products as the nodes and the enzymes as the edges.

The complete metabolic pathway has been partitioned into multiple subgraphs based upon the functionality of individual partitions. Each such partition will be referred as *metabolic pathway module* in this paper. Some of the examples of the metabolic pathways modules are amino acid Biosynthesis, glycolysis, purine and pyramidine metabolism, folate metabolism, fatty acid biosynthesis, pentose and gluconate conversions, biosynthesis of lipids, citrate cycle, CO_2 fixation cycle, urea cycle, CoA biosynthesis [9, 15]. A detailed description of the various metabolic pathway modules is available at Kegg site at http://www.genome.ad.jp).

A *pathway trace* is a set of strongly connected subgraphs within a metabolic pathway module such that each subgraph contains two or more edges. For example, the 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 Kegg database) "Valine, leucine, and isoleucine biosynthesis" form a pathway trace.

2.5. Nomenclature

This paper uses Genbank nomenclature for gene-names and abbreviates *B. subtilis by Bs*, *E.coli* by *Ec*, *H. influenzae* by *Hi*. For gene names, when there is no ambiguity, *E. coli* name (as present in Genbank) has been used. 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*. An unannotated gene (gene with no functionality) has been annotated as *orf.seq* where '*seq*' is the position of the coding region in the genome.

3. Deriving orthologs and gene-groups

This section briefly describes the algorithm used to identify orthologs, homologous and orthologous gene groups [3, 4].

3.1. Identifying orthologs

The pair-wise comparison of two genomes was modeled as a weighted bipartite graph matching problem [3]. The weights of the edges were 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 (π_{1I}, π_{2J}) of the highest weight, all the edges involving the nodes π_{1I} and π_{2J} are deleted. The process is continued until there are no more edges. The edges starting or ending in genes inside a gene group are biased positively as genes within a gene group are better candidates for preserving a common functionality. Two edges with close weights, if two weights are above a threshold, suggest multiple orthologs or gene-fusion, and need further analysis to identify the fused genes. A detailed algorithm is given in [3].

3.2. Identifying gene groups

A neighboring group S_0 for a gene in Γ_1 is marked. A *neighboring group* is a group of genes in the close proximity. Then a set S_1 (in Γ_2) of homologs for S_0 is marked. Then the set S_2 — a union of all the sets of homologs of S_1 — in Γ_1 is marked. A non-empty intersection of the sets S_0 and $S_{2,}$ with more than one element in the intersection, marks the presence of the start of a homologous gene group. After marking the start of homologous gene groups, the genome Γ_1 is traversed one node at a time, checking for the presence of an edge in the close proximity of the last homologous gene in the genome Γ_2. The method identifies gene groups of any variable size. A detailed discussion of the algorithm is available in [3].

4. Reconstruction using gene-groups

The reconstruction was done in many steps. First a standard template of metabolic pathway was taken from Kegg database (see http://www.genome.ad.jp/kegg-bin/mk_point_html). Genomes were extracted from the Genbank using .gbk file format. The gene-groups obtained from pair-wise comparison of two genomes Γ_1-Γ_2 was obtained using Goldie 2.0 — an automated tool to identify orthologs and homologous gene-groups from Genbank database. The process is repeated by identifying gene-groups of the genome Γ_1 by varying the genome Γ_2. For example, *E. coli* was compared successively against remaining genomes in the Genbank starting with *B. subtilis*. *B. subtilis* was chosen as the first one as it is experimentally well explored bacteria. A union of the set of gene-groups was taken. Overlapping gene-groups were collapsed to form bigger gene-groups. The set of collapsed gene-groups was used to identify the local sections of metabolic pathway.

The gene-groups were classified in various groups: orthologous gene-groups, homologous ordered gene-groups, homologous unordered gene-groups, duplicated domain gene-groups, and multigene-groups. The enzymes name and gene-name correspondence (as present in Genbank) in the gene-groups was extracted using Goldie 2.0 (http://www.mcs.kent.edu/~arvind/intellibio/orthos. html)

The standard template of the metabolic pathway was procured from the Kegg database (http://www. genome.ad.jp/kegg-bin/mk_point_html). The gene-groups containing the enzymes were mapped using the EC numbers of the genes in the gene-groups, and the locations of the enzymes in the gene-groups were marked in the pathways using a subset of the enzymes in the gene-group. The end point of the pathway trace was detected. Any gene without an enzyme number in the gene-group was matched against the enzymes in the pathway trace. The functionality of the gene was identified from the Genbank annotation, SwissProt enzyme annotation, and Ligand database annotation using the gene-name match if given. If the functionality 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 exactly match then the functions of the generic class of the enzymes in the pathway trace were identified and matched with the function associated with the gene-name. If the generic function description matched then the enzyme in the pathway trace was replaced by the generic description of the enzyme class.

The gene-groups were mapped in the order orthologous groups first, followed by ordered homologous gene-groups, followed by gene-groups having multiple duplicated domains, followed by multi-gene groups. The mapping of the fused genes in all the enzyme databases (SwissProt and Ligand) is missing as sets of fused have recently been identified by the author [4], and are yet to be studied. After mapping the gene-groups, on the template, the gene-group on the genomes were extended by identifying the genes in the neighborhood of the genomes next to the gene-groups, and the pathway trace was extended by identifying the enzymes in the neighborhood of the gene-groups. The extended gene-group and extended pathway-trace in the standard pathway template were matched until the EC number of the genes in the neighborhood of the gene-group and the EC number in the extended pathway trace did not match. This extended pathway trace was recorded. For the orthologous gene-groups and single homologous gene-groups (ordered or unordered) and duplicated domain gene-groups, the EC number of genes (or function of the gene) corresponding to the gene-groups were preferred over the EC number of an enzyme at the corresponding position in Kegg database (see http://www.

genome.ad.jp/kegg-bin/mk_point_html). In the case of multi-gene groups, the decision was based upon the reaction of the proteins in the gene-group.

5. Results and discussion

All the gene-groups obtained from *E. coli* and *B. subtilis* complete genome comparison (http://www.mcs.kent.edu /~arvind/intellibio/orthos.html) were mapped on the standard template obtained from Kegg standard template (see http://www.genome.ad.jp). The results are summarized in Tables 1 and 2.

Table 1 describes a limited representative subset of gene-groups acting as seeds for pathway traces. The complete set is available on the site (http://www.mcs.kent.edu/~arvind/intellibio/orthos.html# pathway).

Table 2 shows the extended gene-group in *E. coli* genome acting as the pathway trace derived from the gene-groups in Table 1. Only extended gene-groups larger than the homologous or orthologous gene-group (as shown in Table 1) are shown. The genes with the enzyme numbers are well defined in Kegg database and Genbank. The genes without enzyme numbers show that there is no corresponding EC number in Kegg database. EC number without gene name shows that the enzyme is present in Kegg database. However, the ortholog does not occur within the same gene-group.

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 they are genes with similar functionality and are controlled 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 the similar functionality — both are dehydratases: *EC 4.2.1.9* is involved in "Valine, Leucine, and Isoleucine metabolic pathway", and *EC 4.2.1.16* is involved in serine. 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-groups may belong to two or more modules (many times similar). Sometimes gene-group may be involved in a reaction which spans adjacent metabolic pathway modules despite the pathway trace involving adjacent nodes. This is due to artificial splitting of complete metabolic pathway in different modules. The order of the gene-groups in the large pathway traces is not significant.

Some of the genes such as *upp:2.4.2.9*, *udhA:1.-.-.-* etc. were missing from SwissProt and Kegg database, and could not be tested despite being in the same gene-group. Few genes such as *ppc:4.1.3.1* belong to a different pathway despite being a gene-group belonging to a different metabolic pathway. A finer analysis based of the control regions in future will remove this discrepancy. Some enzymes such as *thtR:2.8.1.1* were absent from Kegg database, and could not be verified.

Table 1. A set of gene-groups as pathway seeds

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*)
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*), (*cysC:2.7.1.25*, cysH:2.8.2.-), (*prsA:2.7.6.1*, ychB), (*purD:6.3.4.13*, purH:2.1.2.3),
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*)
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: (*succC:6.2.1.5*, *succC:6.2.1.5*)
Phopholipid degradation: (*glpQ:3.1.4.46*, *glpT*)
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.-*)
Pentose phosphate cycle: (*prsA:2.7.6.1*, *ychB*), (*deoC:4.1.2.4*, *deoA:2.4.2.4*)
Glycine, serine, and threonine metabolism: (*gcvP:1.4.4.2, gcvT: 2.1.2.10*), (*betB:1.2.1.8, orf.308*), (*ycaJ:2.7.7.7*, *serS:6.1.1.11*), (*thrA:2.7.2.4, thrB:2.7.1.39, thrC:4.2.99.2*)
Glycolysis: (*ascF:2.7.1.69, ascB:3.2.1.86*)
Pyrimidine metabolism: (*carA:6.3.5.5, carB:6.3.5.5*), (rhlB, *trxA:1.6.4.5*), (deoC:4.1.2.4, deoA:2.4.2.4), (*tmk:2.7.4.9*, *holB:2.7.7.7*)
Glyoxylate and dicarboxylate metabolism: (*def:3.5.1.27*, *fmt:2.1.2.9, fmu*), (*yddg, fdnG:1.2.1.2*)
Glycerolipid metabolism: (*glpQ:3.1.4.46*, *glpT*), (*glpK:2.7.1.30*, glpF)
Riboflavin metabolism: (*ribD*, *ribH:2.5.1.9*)
Glutamate metabolism: (*gltA:1.4.1.3, gltB:1.4.1.3*)
Porphyrin and cholorophyll metabolism: (*proB:2.7.2.11, proA: 1.2.1.41*)
Arginine and proline metabolism: (*argH:4.3.2.1, oxyR*)
Pantothenate and CoA biosynthesis: (*panD:4.1.1.11*, *panC:4.3.2.1, panB:2.1.2.11*)

Table 2. Extended gene-groups 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, succC: 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*)

The gene-groups in multi gene-groups can be classified in four different categories:

(i) Gene-groups which map on a pathway trace in the same module,

(ii) A subgroup of a larger gene-group occurring at a different location in the genome. Obviously, the role of multiple subgroups of the same group of enzymes at different location in the genome is not clear, and may be related to enhanced reaction and production of a specific biochemical product

(iii) The regulatory and sensor proteins have multiple occurrences. For example, EC 2.7.3.- (sensor protein) along with the transcriptional regulator has multiple occurrences.

(iv) Random duplication may bring enzymes from two metabolic pathway may come together. However, the such random duplication are rare, and their role in metabolic pathway is not clear.

6. Conclusion

In this paper, we have described a scheme to automate the reconstruction of metabolic pathways of newly sequenced genomes using the automatically derived orthologous and homologous gene-groups. While benefiting from the previous results and databases [2, 8, 9], this scheme improves upon the previously derived semi-automatic databases based upon homologs and orthologs by marking the local sections in metabolic pathways, and extending the local sections. In addition, the gene-group analysis also identifies non-enzymatic 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. An interesting finding is that gene-groups containing regulatory proteins and sensor protein are duplicated quite often.

The scheme is still limited by the lack of the available functionality of remaining genes in *E. coli* and *B. subtilis* genomes from the wet laboratories.

Due to the differences in the annotations of Genbank, Kegg, and Swissprot, some of the annotations were different, and could not be tested.

Currently, the author is developing a fully automated software based upon the above scheme to refine the metabolic pathways.

References

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

[2] A. Bairoch, "The ENZYME database in 2000," *Nucleic Acids Research*, (2000) 304-305

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

[4] A. K. Bansal, "An Automated Comparative Analysis of 17 Complete Microbial Genomes," *Bioinformatics*, **15:11** (1999), 900 - 908.

[5] 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.*, **283** (1998) 707-725.

[6] S. J. Cordwell, "Microbial genomes and missing enzymes: redefining biochemical pathways," *Arch Microbiol*ogy, **172:5** (1999) 269-79.

[7] W. M. Fitch, "Distinguishing Homologous from Analogous Proteins," *Systematic Zoology*, **19** (1970) 99 - 113.

[8] S. Goto, T. Nishioka, M. Kanehisa, "LIGAND: chemical database for enzyme reactions," *Bioinformatics.* **14:7** (1998) 591-599.

[9] H. Bono, H. Ogata, S. Goto, M. Kanehisa, "Reconstruction of amino acid biosynthesis pathways from the complete genome sequence," *Genome Res.* **8:3** (1998) 203-210.

[10] T. M. Gruber and D. A. Bryant, "Molecular Systematic Studies of Eubacteria, Using σ^{70} — Type Sigma factors of Group1 and Group 2," *J. of Bacteriology*, **179** (1997) 1734 - 1747

[11] M. A. Huynen and P. Bork, "Measuring Genome Evolution," *Proc. Natl. Acad. Sci.*, USA, **95** (1998) 5849 – 5856

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

[13] C. H. Papadimitrou and K. Steiglitz, "Combinatorial Optimization: Algorithm and Complexity," *Prentice Hall*, 1982

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

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

[16] 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,* **17:2** (1999) 53-60.

[17] 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*, **6** (1996) 279 – 291

[18] Waterman,M.S., "Introduction to Computational Biology: Maps, Sequence, and Genomes," *Chapman & Hall*, London, UK, 1995