

# Pathway Mapping With Operon Information: An Integer-Programming Method

Fenglou Mao(*Fenglou@csbl.bmb.uga.edu*)  
Victor Olman(*olman@csbl.bmb.uga.edu*)  
Ying Xu(*xyn@bmb.uga.edu*)\*

Zhengchang Su(*zhx@csbl.bmb.uga.edu*)  
David Chuang(*wchung@csbl.bmb.uga.edu*)

*Computational Systems Biology Lab, Biochemistry & Molecular Biology Department, University of Georgia and Computational Biology Institute, Oak Ridge National Laboratory, USA*

## Abstract

*Biological pathway mapping is an important problem in the post-genomic era. We now present a new algorithm for pathway mapping in microbes. The algorithm considers not only sequence similarity among the template and target genes, but also the operon structures in the target genome. We formulated the mapping problem as a graph finding problem, and solved it by an Integer-Programming (IP) method. The goal is to minimize a linear object function subject to six constraints, such that maximal sequence similarity among the template and target genes are achieved, and at the same time, a minimal number of operons are covered in the target genome. Compared to our previous Minimal Spanning Tree (MST) algorithm, the IP method has the following advantages: i) It is much faster and thus can map larger pathway involving a much large set of genes. ii) The IP method looks into the details of genes in the operons, and consequently avoids the many-to-one mapping mistakes that sometimes occur in the MST algorithm. We have compiled a large pathway training set to optimize the parameters of the program, and tested it by mapping 16 complex pathways from BioCyc onto E.coli K12 genome and the results are very promising.*

## 1. Introduction

Comparative genomics analysis is a powerful approach for gene function predictions for newly sequenced genomes, and this is mainly done by identification of their orthologous genes in other well studied genomes. Current efforts for orthologues identification are mainly based on sequence similarity. While these methods including the popular bi-directional-best-hit (BDBH)[1] scheme, Clusters of Orthologous Groups (COGs)[2], multiple sequence alignment based method [3] and phylogeny tree based method [4] may work well for closed related genomes, they may fail on remotely related genomes. Therefore, more information is needed to achieve more robust orthologue predictions.

We have recently proposed a new framework for mapping the orthologous genes between genomes at pathway level by considering not only sequence similarity but also other gene relationships such as operon

structures and transcription binding site information, since genes in the same operon and regulon are often involved in a same biological process. Our previous solution to this problem was a Minimal Spanning Tree (MST) algorithm which is not capable of solving problems involving a large set of operons. Here, we re-formulated the problem as combinatory optimization problem and solved it by Integer Programming (IP). This new algorithm overcomes the shortcomings in the previous method and can work on large pathways.

## 2. Problem Formulation

The pathway mapping problem between a template and a target genome is illustrated in figure 1. There are  $m$  genes in the template pathway, and  $n$  candidate genes in target set. Each open circle is a gene and each shadowed eclipse is an operon.  $x_{ij}$ ,  $y_{kl}$  are variables. If gene  $i$  in the template is mapped to gene  $j$  in the target genome, then  $x_{ij} = 1$ , otherwise, 0. If genes  $k$  and  $l$  in the target genome are mapped sequentially, then  $y_{kl} = 1$ , otherwise, 0.  $h_{ij}$  reflects the sequence similarity of gene  $i$  in template pathway and gene  $j$  in the target genome.  $s_{kl}$  has a large value if gene  $k$  and  $l$  are in different operons, otherwise a small value. We formulate the pathway mapping problem as an IP problem as follows:

$$\text{Objective function: } \sum_{i=1}^m \sum_{j=1}^n h_{ij} x_{ij} + \sum_{l=1}^n \sum_{k=l+1}^n s_{kl} y_{kl}$$

Constraints:

$$1. \sum_{j=1}^n x_{ij} = 1, \text{ for } i=1 \dots m, \text{ to guarantee each template gene}$$

is exactly mapped to one target gene.

$$2. 0 \leq \sum_{i=1}^m x_{ij} \leq 1, \text{ for } j=1 \dots n, \text{ to guarantee each target gene}$$

is only mapped by one template gene, or not mapped by any gene.

$$3. 0.5 \times \sum_{k=1, k \neq l, l=j}^n y_{kl} \leq \sum_{i=1}^m x_{ij} \leq \sum_{k=1, k \neq l, l=j}^n y_{kl}, \text{ for } j=1 \dots n, \text{ to}$$

guarantee if a target gene is mapped, it must have at least one edge (up to 2 edges) connected to other target gene(s), if it is not mapped, it will not be connected to any target gene.

$$4. \sum_{k>l} y_{kl} \leq 1,$$

$$5. \sum_{k < l} y_{kl} \leq 1,$$

$$6. \sum_{k=1}^n \sum_{k > l} y_{kl} = m - 1$$

The Integer Programming formulation was proved mathematically a solution to the pathway mapping problem: map one template pathway to a target pathway with maximal sequence similarity and minimal number of operons covered in the target genome, the solution is global optimal.

### 3. Results

#### 3.1 Speed improvement

The pathway mapping problem is a NP hard problem. The MST method is an enumerate algorithm which takes the computational complexity  $o(2^n)$ . As shown in Table 1, it could not solve problems involving a large number of genes. In contrast, the IP method resolves the problem by optimizing a linear objective function subject to some constraints and integer bindings for some variables. It takes "Branch, Cut and Price" algorithm, and solved the practical mapping problems involving up to 200 genes in a reasonable time scale (Table 1). COIN-OR[5] is employed to program the problem.

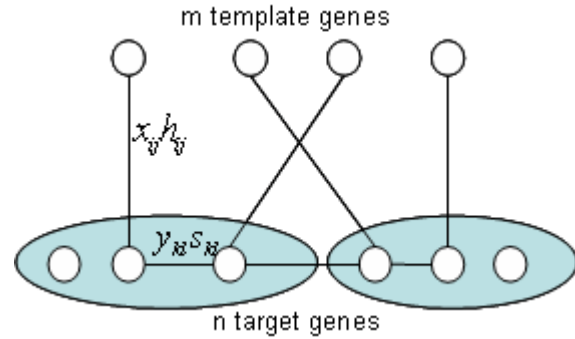


Figure 1

#### 3.2 Experimental Results

We have tested the algorithm by mapping three experimentally verified pathways (phenylalanine, tyrosine and tryptophan biosynthesis superpathway; peptidoglycan biosynthesis pathway; KDO2-lipid A biosynthesis pathway) of 15 microbes stored in BioCyc database [6] to E.Coli K12 genome. Among the 172 genes in the these pathways, our program correctly predicted 163 orthologous in E.Coli K12, with only 9 mistake matches, yielding a total accuracy rate of 94.8%.

Table 1: Speed comparison between IP and MST<sup>3</sup> methods

	Template Gene Number	Target Operon Number	Edge Number For Each Gene	Time For IP	Time For MST
1 <sup>1</sup>	8	3	3	0.02s	0.01s
2	20	10	5	0.03s	0.01s
3	40	20	3	0.20s	0.30s
4	50	30	6	1.46s	20.71s
5	60	40	6	5.09s	8511.17s
6	100	50	5	6.96s	- <sup>2</sup>
7	200	100	7	69.27s	- <sup>2</sup>

1. The problems are generated by computer; 2. The problem can not be resolved in a reasonable time scale.
3. MST algorithm takes computational complexity of  $o(2^n)$ , please refer to another paper in this proceeding.

#### Acknowledgement:

1. DOE office of Biological/Environmental Research, Genome to Life Project, "Carbon Sequestration in Synechococcus sp.: From Molecular Machines to Hierarchical Modeling."
2. NSF award number: DBI-0354771, title: A Computational Capability for Fast and Reliable Characterization of Protein Complexes.
3. NSF award number IIS-0407204, title: ITR Collaborative Research: Combinatorial Algorithms for Biological Data Clustering.
4. GA cancer coalition award
5. GA research alliance award

#### Reference:

1. Mushegian A. R. and Koonin E. V., "A minimal gene set for cellular life derived by comparison of complete bacterial genomes", Proc Natl Acad Sci U S A, 93:10268-10273, 1996.

2. Tatusov R. L., Koonin E. V. and Lipman D. J., "A genomic perspective on protein families", Science, 278:631-637, 1997.
3. Wall D. P., Fraser H. B. and Hirsh A. E., "Detecting putative orthologs", Bioinformatics, 19:1710-1711, 2003.
4. Arvestad L., Berglund A. C., Lagergren J. and Sennblad B., "Bayesian gene/species tree reconciliation and orthology analysis using MCMC", Bioinformatics, 19 Suppl 1:117-115, 2003.
5. Robin Lougee-Heimer, "The Common Optimization INterface for Operations Research: Promoting open-source software in the operations research community", IBM Journal of Research and Development, Vol. 47, No. 1, 2003, pp. 57-66.
6. Karp P.D., Arnaud M., Collado-Vides J., Ingraham J., Paulsen I.T., Saier M.H. Jr., "The E.Coli EcoCyc Database: No Longer Just a Metabolic Pathway Database". ASM News 70(1): 25-30, 2004.