Pacific Symposium on Biocomputing 12:496-507(2007)

COMPARATIVE PATHWAY ANNOTATION WITH PROTEIN-DNA INTERACTION AND OPERON INFORMATION VIA GRAPH TREE DECOMPOSITION

JIZHEN ZHAO, DONGSHENG CHE AND LIMING CAI

Department of Computer Science, University of Georgia, Athens, GA 30602, USA. Email: {jizhen, che, cai}@cs.uga.edu, Fax: (706) 542-2966.

Template-based comparative analysis is a viable approach to the prediction and annotation of pathways in genomes. Methods based solely on sequence similarity may not be effective enough; functional and structural information such as protein-DNA interactions and operons can prove useful in improving the prediction accuracy. In this paper, we present a novel approach to predicting pathways by seeking high overall sequence similarity, functional and structural consistency between the predicted pathways and their templates. In particular, the prediction problem is formulated into finding the maximum independent set (MIS) in the graph constructed based on operon or interaction structures as well as homologous relationships of the involved genes. On such graphs, the MIS problem is solved efficiently via non-trivial tree decomposition of the graphs. The developed algorithm is evaluated based on the annotation of 40 pathways in *Escherichia coli* (*E. coli*) K12 using those in *Bacillus subtilis* (*B. subtilis*) 168 as templates. It demonstrates overall accuracy that outperforms those of the methods based solely on sequence similarity or using structural information of the genome with integer programming.

Keywords: pathway annotation, pathway prediction, protein-DNA interaction, operon, tree decomposition, independent set

1. Introduction

A challenge in the post-genomic era is the understanding at different levels of the genomes that have been sequenced [12]. Many efforts have been made in gene finding and assigning predicted or determined functionalities to found genes. However, higher order functional analysis of organisms from their genomic information remains in demand [14]. Assigning a biological pathway to a set of genes, known as pathway annotation, is one such analysis, which is essential to understanding cellular processes and organism behaviors in a larger context [14]. Biological pathways could be determined experimentally but this is usually expensive and laborious. As

Pacific Symposium on Biocomputing 12:496-507(2007)

more and more genomes are sequenced and annotated, it is feasible to employ comparative genomic analysis in pathway prediction and annotation at the genome scale. Based on an annotated pathway from one genome as template, a pathway for a target genome can be predicted by identifying a set of orthologues based on sequence similarity to the genes in the template pathway. A naive approach for orthology assigning is choosing the best BLAST hit for each gene (BH). A more often used technique is by reciprocal BLAST search, called bidirectional best-hit (BBH) [8], where gene pairs are regarded as orthologues if they are the best hits in both directions of the search. However, these and other sequence similarity based approaches share the same limitation [7]: the best hits may not necessarily be the optimal orthologues, thus compromising the prediction accuracy.

It is observed that homology relationships exist not only at the sequence level, but also at functional and structural levels [15], e.g., those of operon structures and protein-DNA interactions such as transcriptional regulation patterns of some genes by transcriptional factors (TFs). Recently, substantial operon and transcriptional regulation information have been curated from the scientific literatures for a number of genomes [6,11]. Computational methods [10,11,15] have also been developed to predict operon structures and co-regulated genes. The structural information about transcriptional regulation patterns that are needed may be gathered in a number of ways, although they may not necessarily be complete or extremely accurate. By considering such high level information among genes along with the sequence similarity, it becomes possible to improve the pathway prediction accuracy. However, the optimal prediction of pathways at the genome scale becomes difficult combinatorial optimization problems if sophisticated structural information is incorporated. PMAP [7] is an existing method that overcomes the difficulty by incorporating partial structural information (it i.e. structural information of the target genome only) with integer programing. In this paper, a novel approach is introduced based on integrating data in sequence similarity, experimentally confirmed or predicted operons, transcriptional regulations, as well as available functional information of related genes, in both template pathway and target genome. The new approach has led to an efficient graph-theoretic algorithm called TdPATH for pathway prediction.

Algorithm TdPATH predicts a pathway in a target genome based on a template pathway by identifying an orthologous gene in the target genome for each gene in the template pathway, such that the overall sequence and structural similarity between the template and the predicted path-

Pacific Symposium on Biocomputing 12:496-507(2007)

ways achieves the highest. In particular, homologes for each gene in the template pathway are first identified by the BLAST search [1]. Functional information is then used to filter out genes unlikely to be orthologues. The structural information are used to further constrain the orthology assignment. One of the homologes is eventually chosen to be the ortholog for the gene. The pathway prediction is formulated into the maximum independent set (MIS) problem and the maximum CLIQUE problem by taking protein-DNA interaction constraints and operon constraints respectively. Because both problems are computationally intractable, we solve them efficiently with non-trivial techniques based on tree decompositions of the graphs constructed from the structural constraints.

Our algorithm TdPATH has been implemented and its effectiveness is evaluated against BH, BBH and PMAP in the annotation of 40 pathways of *E. coli* K12 using the corresponding pathways of *B. subtilis* 168 as templates. The results showed that overall, in terms of the accuracy of the prediction, TdPATH outperforms BH and BBH that based solely on sequence similarity, as well as PMAP that uses partial structural information. In term of average running time to predict a pathway, it outperforms PMAP. Algorithm TdPATH is dynamic programing based on tree decomposition techniques. The running time of the algorithm is dominated by function $2^t n$, where t is the tree width of the underlying graphs of n vertices constructed from the structural constraints. In particular, the statistics on the tree width of these graphs shows that about 87% of the graphs have tree width at most 5, while 94% have tree width at most 8. Therefore, the tree decomposition based algorithm for pathway prediction is both theoretically and practically efficient than the integer programming based algorithm PMAP..

2. Methods and Algorithm

2.1. *Problem formulation*

A *pathway* is defined as a set of molecules (genes, RNAs, proteins, or small molecules) connected by links that represent physical or functional interactions. It can be reduced to a set of genes that code related functional proteins. An *operon* is a set of genes transcribed under the control of an operator gene. Genes that encode transcriptional factors are called *tf* genes. In the work described in this paper, a known pathway in one genome is used as a template to predict a similar pathway in a target genome. That is, for every gene in the template pathway, we identify some gene in target

Pacific Symposium on Biocomputing 12:496-507(2007)

genome as its ortholog if there is one, under the constraints of protein-DNA interaction (*i.e.*, transcriptional regulation) and operon information. The problem of predicting pathways is defined as:

INPUT: a template pathway model $P = \langle A_P, R_P, O_P \rangle$ and a target genome T, where A_P is a set of genes in P, R_P is a set of relationships between *tf* genes and genes regulated by corresponding *tf* gene products, and O_P is a set of operons;

OUTPUT: a pathway $Q = \langle A_Q, R_Q, O_Q \rangle$ for T and an orthology mapping $\pi : A_Q \rightarrow A_P$ such that the overall sequence similarity between all genes in pathway Q and their corresponding orthologues in the template P, as well as the consistency of the operon and regulation structures between pathways P and Q are as high as possible.

2.2. *The methods*

Our approach consists of the following steps:

(1) For every gene in the template pathway P, find a set of homologes in the target genome T with BLAST;
(2) Remove from the homologes genes unlikely to be orthologues to the corresponding gene in the template P. This is done based on functional information, e.g., Cluster of Orthologous Groups (COG)[16], which is available. In particular, genes that are unlikely orthologous would have different COG numbers.
(3) Obtain protein-DNA interactions and operon structures for the homologous genes in the template pathway and target genome from related databases [6,11], literatures or computational tools [10,15].
(4) Exactly one of the homologous genes is eventually assigned as the ortholog for the corresponding gene in the template P. This is done based on the constraints by the protein-DNA interaction and operon information (for any gene that is not covered by the structural information due to the incomplete data or other reasons, we simply assign the best BLAST hit as the ortholog). Such an orthology mapping or assignment essentially should yield a predicted pathway that has overall high sequence similarity and structural consistency with the template pathway.

By incorporating sophisticated structural information, the pathway prediction problem may become computationally intractable. We describe in the following in detail how an efficient algorithm can be obtained to find

Pacific Symposium on Biocomputing 12:496-507(2007)

the orthology mapping between the template pathway and the one to be predicted. We consider in two separate steps structural constraints with protein-DNA interactions and those with operons.

2.2.1. *Constraints with protein-DNA interactions*

We use available protein-DNA interaction information, *i.e.* the transcriptional regulation information, to constrain the orthology assignment. This is to identify orthologs with consistent regulation structures to the corresponding genes in the template pathway. Think genes as vertices and relations among the genes as edges, the template pathway and the corresponding homologs in target genome can be naturally formed into two graphs. Thus the problem can be converted to finding the optimal common subgraph of these two graphs. It is in turn to be formulated into the maximum independent set (MIS) problem. Details are given below. For convenience, we call a *regulon* in this paper to be a gene encoding a transcription factor and all the genes regulated by the factor.

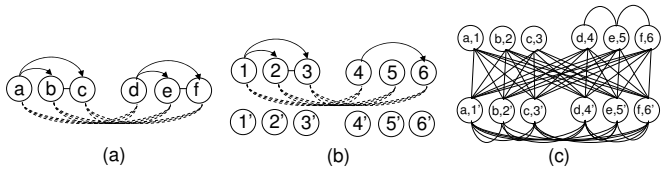

Figure 1. Constraints with transcriptional regulations. (a) Regulation graph G_1 for template pathway. A directed edge points from a *tf* gene to a gene regulated by the corresponding TF, a solid edge connects two genes regulated by a same TF, a dashed edge connects two genes belonging to different regulons. (b) Regulation graph G_2 for the homologous genes in the target genome, constructed in similar way to (a). (c) Merged graph G from G_1 and G_2. Each node is a pair of homologous genes.

(1) A regulation graph $G_1 = (V_1, E_1)$ is built for the template pathway P, where vertex set V_1 represents all genes in template pathway P, and edge set E_1 contains three types of edges: an edge of type-1 connects a *tf* gene and every gene regulated by the corresponding product; an edge of type-2 connects two genes regulated by the same *tf* gene product; and edges of type-3 connect two genes from different regulons if they are not yet connected (Figure 1(a)).
(2) A regulation graph $G_2 = (V_2, E_2)$ is built for the target genome in

Pacific Symposium on Biocomputing 12:496-507(2007)

the similar way, where V_2 represents homologous genes in the target genomes (Figure 1(b)).

(3) Graphs G_1 and G_2 are merged into a single graph $G = (V, E)$ such that V contains vertex $[i, j]$ if and only if $i \in V_1$ and $j \in V_2$ are two homologous genes. A weight is assigned to vertex $[i, j]$ according to the BLAST score between genes i and j. Add an edge $([i, j], [i', j'])$ if either (a) $i = i'$ or $j = j'$ but not both, or (b) edges $(i, i') \in E_1$ and $(j, j') \in E_2$ are not of the same type (Figure 1(c)).

(4) Then the independent set in the graph G with the maximum weight should correspond to the desired orthology mapping that achieves the maximum sequence similarity and regulation consistency. This assigns one unique orthologous gene in this template pathway to each gene in the pathway to be predicted, as long as they are covered by the known protein-DNA interaction structures.

2.2.2. *Constraints with operon structures*

We now describe how to use confirmed or predicted operon information to further constrain the orthology assignment. This step applies to the genes that have not been covered by protein-DNA interaction structures.

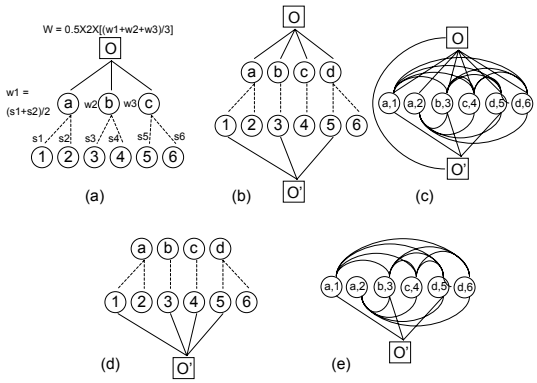

Figure 2. Constraints with operon information. See description for details. A dashed line connects two homologes. a) Setting weight for an operon. b) A pair of partially conserved operons in template pathway and target genome. (c) A mapping graph formed according to (b). (d) An operon only appears in target genome. (e) The mapping graph formed according to (d).

Pacific Symposium on Biocomputing 12:496-507(2007)

We first assign to each gene i with a weight w_i. w_i is set according to the average of its BLAST scores with its top m (say, 5) homologes. The weight of an operon o is set as $0.5(n-1)\sum_{i\in o} w_i/n$, where n is the number of genes in the operon (Figure 2(a)). The factor 0.5 allows an operon in one genome to only contribute 50% and a conserved operon in the other genome to contribute the other 50%. We use term $n-1$ in the formula since we want to exclude the operons that have only one gene from consideration, since they do not introduce structural information.

We then sort the operons according the non-decreasing of their sizes and then use the following greedy iterative process to constrain the orthology mapping as long as there is an operon unexamined. Repeat the following 4 steps:

(1) Select the largest unexamined operon and consider the related homologes in another genome as well as the available operon structures in them;
(2) Build a mapping graph $G_m = (V_m, E_m)$ (Figure 2(b)-(e)), where V_m contains the following two types of vertices: an operon vertex presents each of the involved operons and a mapping vertex $[i, j]$ presents each pair of homologous genes i and j. Edge set E_m also contains three types of edges: an edge connects every pair of mapping vertices $([i, j], [k, l])$ if $i \neq k$ and $j \neq l$, an edge connects an operon node and a mapping node if one of the two genes in the mapping node belongs to the operon, and an edge connects every pair of involved operons between the target genome and the template pathway;
(3) Find the maximum clique C on G_m;
(4) Remove the template genes appeared in the mapping nodes of C and their homologes. Remove an operon if all genes in it have been removed. If only a subset of the genes in an operon have been removed, leave the remaining genes as a reduced operon. Resort the remaining operons.

By this formulation, an edge in graph G_m denotes a consistent relationship between two nodes connected by it. A maximum clique denotes a set of consistent operon and mapping nodes that have the maximum total weight and thus can infer a optimal mapping. Note that an operon in one genome could have zero or more, complete or partial conserved operons in another genome [10]. If it has one or more (Figure 2(b)), the constraint can be obtained from both of the genomes and thus is called a *two side con-*

Pacific Symposium on Biocomputing 12:496-507(2007)

straint. The procedure can find the orthology mapping that maximizes the sequence similarity and the operon structural consistency. Otherwise, it is called called an *one side constraint* (Figure 2(b)). The procedure can find the orthology mapping that minimizes the number of involved operons.

2.3. *Tree decomposition based algorithm*

Based on section 2.2, constraining the orthology mapping with protein-DNA interactions and with operon structures can be reduced to the problems of maximum independent set (MIS) and maximum clique (CLIQUE) on graphs formulated from the structural constraints. Both problems are in general computationally intractable; any naive optimization algorithm would be very inefficient considering the pathway prediction is at the genome scale.

Our algorithm techniques are based on graph tree decomposition. A tree decomposition [13] of a graph provides a topological view on the graph and the tree width measures how much the graph is tree-like. Informally, in a tree decomposition, vertices from the original graph are grouped into a number of possibly intersecting bags; the bags topologically form a tree relationship. Shared vertices among intersecting bags form graph separators; efficient dynamic programming traversal over the graph is possible when all the bags are (*i.e.*, the tree width is) of small size [3].

In general, the graphs formulated from protein-DNA interactions and operon structures have small tree width . We employ the standard tree decomposition-based dynamic programming algorithm [3] to solve MIS and CLIQUE problems on graphs of small tree width. On graphs with larger tree width, especially on dense graphs, our approach applies the tree decomposition algorithm on the complement of the graph instead. The running time of the algorithms is $O(2^t n)$, where t and n are respectively the tree width and the number of vertices in the graph. Such a running time is scalable to larger pathways. Due to the space limitation, we omit the formal definition of tree decomposition and the dynamic programming algorithm. Instead, we refer the reader to [3] for details.

We need to point out that finding the optimal tree decomposition (*i.e.*, the one with the smallest tree width) is NP-hard [2]. We use a simple, fast approximation algorithm greedy fill-in [4] to produce a tree decomposition for the given graph. The approximated tree width t may affect the running time of the pathway prediction but not its accuracy.

Pacific Symposium on Biocomputing 12:496-507(2007)

3. Evaluation Results

We evaluated TdPATH against BH, BBH and PMAP by using 40 known pathways in *B. subtilis* 168 from KEGG pathway database [5] as templates (Table 1) to infer corresponding pathways in *E. coli* K12. For TdPATH, the operon structures are predicted according to the method used in [10] and experimentally confirmed transcriptional regulation information is taken from [6] for *B. subtilis* 168 and from [11] for *E. coli* K12. For PMAP, predicted operon and regulon information is obtained according to the method used in [7]. Both TdPATH and PMAP include the COG filtering.

Table 1. Template pathways of *B. subtilis* 168, taken from KEGG pathway database.

bsu00040	bsu00100	bsu00130	bsu00190	bsu00193	bsu00401	bsu00430
bsu00471	bsu00480	bsu00511	bsu00530	bsu00531	bsu00602	bsu00604
bsu00660	bsu00720	bsu00730	bsu00750	bsu00760	bsu00900	bsu00903
bsu00930	bsu00950	bsu01031	bsu01032	bsu02040	bsu03020	bsu03030
bsu03060	bsu00220	bsu00450	bsu00770	bsu00780	bsu01053	bsu02030
bsu00520	bsu00920	bsu03010	bsu00240	bsu00400		

We evaluated the accuracy of the algorithms. The accuracy was measured as the arithmetic mean of sensitivity and specificity. Let K be the real target pathway, H be the homologous genes searched by BLAST according to the corresponding template pathway. Let R be the size of $K \cap H$, *i.e.* the number of genes common in both the real target pathway and the candidate orthologues. We use this number as the number of real genes to calculate sensitivity and specificity because that is the maximum number of genes a sequence based method can predict correctly. Since BH (or BBH) can be considered a subroutine of PMAP and TdPATH, we only evaluated efficiency for PMAP and TdPATH. Running times from reading inputs to output the predicted pathway were collected. For TdPATH, we also collected the data on tree width of the tree decompositions on the constructed graphs or their complement graphs. For all of the algorithms, program NCBI *blastp* [1] was used for BLAST search and the E-value threshold was set to 10^{-6}. The experiments ran on a PC with 2.8 GHz Intel(R) Pentium 4 processor and 1-GB RAM, running RedHat Enterprise Linux version 4 AS. Running times were measured using the "time" function. The testing results are summarized in Table 2.

On average, TdPATH has accuracy of 0.88, which is better than those of other algorithms. We give two examples here to show the improvement is good for small as well as large pathways. One is the nicotinate and nicotinamide metabolism, which has 13 genes in *B. subtilis* 168 while 16

Pacific Symposium on Biocomputing 12:496-507(2007)

genes in *E. coli* K12. The prediction accuracy of TdPATH is 0.9, better than 0.79, 0.83 and 0.79 of BH, BBH and PMAP respectively. Another is the pyrimidine metabolism pathway, which has 53 genes in *B. subtilis* 168 and 58 in *E. coli* K12. TdPATH has prediction accuracy of 0.82, better than 0.79, 0.80, 0.79 of BH, BBH and PMAP respectively. PMAP has second highest accuracy, which means prediction accuracy could be improved even by incorporating structural information partially.

Table 2. Evaluation results. T: time (in seconds), A: accuracy ((sensitivity+specificity)/2).

	BH	BBH	PMAP		TdPATH	
	A	A	A	T	A	T
min	0.33	0.45	0.33	12.8	0.50	1.2
max	1.00	1.00	1.00	27.3	1.00	33.3
ave	0.84	0.85	0.86	16.4	0.88	11.5

For efficiency, TdPATH has average of 11.5 seconds for predicting a pathway, which is slightly better than 16.4 seconds of PMAP. The tree width distribution is shown in Figure 3. On average, tree width of the tree decompositions on the constructed graphs or their complement graphs is 3. 87% of them have tree width at most 5 while 94% at most 8. Since theoretically the running time to find the maximum independent set by the tree decomposition based method is $O(2^t n)$ (where t is the tree width), we can conclude that most of the time our algorithm is efficient based on the statistics of the tree width.

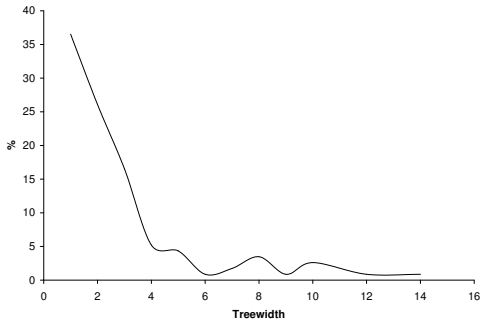

Figure 3. Distribution of the tree width of the tree decompositions on the constructed graphs or their complement graphs.

Pacific Symposium on Biocomputing 12:496-507(2007)

4. Discussion and Conclusion

We have shown our work in utilizing functional information and structural information including protein-DNA interactions and operon structures in comparative analysis based pathway prediction and annotation. The structural information used to constrain the orthology assignment between the template pathway and the one to be predicted appears to be critical for prediction accuracy improvement. It was to seek the sequence similarity and the structural consistency between the template and the predicted pathways as high as possible. Technically, the problem was formulated as finding the maximum independent set problem on the graphs constructed based on the structure constraints. Our algorithm, based on the non-trivial tree decomposition, coped with the computational intractability issue well and ran very efficiently. Evaluations on real pathway prediction for *E coli* also showed the effectiveness of this approach. It could also utilize incomplete data and tolerate some noise in the data.

Tree decomposition based algorithm is sophisticated yet practically efficient. Simpler algorithms are possible if only functional information and sequence similarity are considered. However, computationally incorporating structure information such as protein-DNA interactions and operons in optimal pathway prediction appears to be inherently difficult. Naive optimization algorithms may not be scalable to larger pathway at the genome scale. In addition to the computational efficiency, our graph-theoretic approach also makes it possible to incorporate more information such as gene fusion and protein-protein interactions [12] to further improve the accuracy simply because such information may be represented as graphs as well.

On the other hand, when a template pathway is not well conserved in the target genome, the method may fail to predict the pathway correctly. Multiple templates could be used to rescue this problem since the conserved information could be compensated with each other. We are trying to build profiles from multiple template pathways and use them to do the pathway prediction.

References

1. S. F. Altschul, T. L. Madden, A. A. Schffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", *Nucleic Acids Res*, 25, 3389-3402, 1997.
2. H. L. Bodlaender, "Classes of graphs with bounded tree-width", *Tech. Rep. RUU-CS-86-22, Dept. of Computer Science, Utrecht University, the Netherlands*, 1986.

Pacific Symposium on Biocomputing 12:496-507(2007)

3. H. L. Bodlaender, "Dynamic programming algorithms on graphs with bounded tree-width", In *Proceedings of the 15th International Colloquium on Automata, Languages and Programming*, *Lecture Notes in Computer Science*, 317, 105-119, Springer Verlag, 1987.
4. I. V. Hicks, A. M. C. A. Koster, E. Kolotoglu, "Branch and tree decomposition techniques for discrete optimization", In *Tutorials in Operations Research: INFORMS – New Orleans*, 2005.
5. M. Kanehisa, S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, M. Hirakawa, "From genomics to chemical genomics: new developments in KEGG", *Nucleic Acids Res.* 34, D354-357, 2006.
6. Y. Makita, M. Nakao, N. Ogasawara, K. Nakai, "DBTBS: database of transcriptional regulation in Bacillus subtilis and its contribution to comparative genomics", *Nucleic Acids Res.*, 32, D75-77, 2004
7. F. Mao, Z. Su, V. Olman, P. Dam, Z. Liu, Y. Xu, "Mapping of orthologous genes in the context of biological pathways: An application of integer programming", *PNAS*, 108 (1), 129-134, 2006.
8. D. W. Mount, *Bioinformatics: sequence and genome analysis*, Cold Spring Harbor Lab Press, 516-517, 2000.
9. R. Nielsen, "Comparative genomics: Difference of expression", *Nature*, 440, 161-161, 2006.
10. M. N. Price, K. H. Huang, E. J. Alm, A. P. Arkin, "A novel method for accurate operon predictions in all sequenced prokaryotes", *Nucleic Acids Res.*, 33, 880-892, 2005.
11. H. Salgado, S. Gama-Castro, M. Peralta-Gil, etc., "RegulonDB (version 5.0): Escherichia coli K-12 transcriptional regulatory network, operon organization, and growth conditions", *Nucleic Acids Res.*, 34, D394-D397, 2006.
12. J. L. Reed, I. Famili, I. Thiele, B. O. Palsson, "Towards multidimensional genome annotation.", *Nature Reviews Genetics*, 7, 130-141, 2006.
13. N. Robertson and P. D. Seymour, "Graph minors ii. algorithmic aspects of tree width", *J. Algorithms*, 7, 309-322, 1986.
14. P. Romero, J. Wagg, M. L. Green, D. Kaiser, M. Krummenacker, P. D. Karp, "Computational prediction of human metabolic pathways from the complete human genome", *Genome Biology*, 6, R2, 2004.
15. Z. Su, P. Dam, X. Chen, V. Olman, T. Jiang, B. Palenik, Y. Xu, "Computational Inference of Regulatory Pathways in Microbes: an Application to Phosphorus Assimilation Pathways in Synechococcus sp. WH8102", *Genome Informatics*, 14, 3-13, 2003.
16. R. L. Tatusov, E. V. Koonin, D. J. Lipman, "A Genomic Perspective on Protein Families", *Science*, 278 (5338), 631-637, 1997.