# Predicting Metabolic Pathways from Metabolic Networks with Limited Biological Knowledge

S.Y. Leung
Department of Computer Science
University of Hong Kong
Pokfulam, Hong Kong
syleung@cs.hku.hk

Henry C.M. Leung
Department of Computer Science
University of Hong Kong
Pokfulam, Hong Kong
cmleung2@cs.hku.hk

Carlos L. Xiang
Department of Computer Science
University of Hong Kong
Pokfulam, Hong Kong
lxiang2@cs.hku.hk

S.M. Yiu
Department of Computer Science
University of Hong Kong
Pokfulam, Hong Kong
smyiu@cs.hku.hk

Francis Y.L. Chin
Department of Computer Science
University of Hong Kong
Pokfulam, Hong Kong
chin@cs.hku.hk

***Abstract*—Understanding the metabolism of new species (e.g. endophytic fungi that produce fuel) have tremendous impact on human lives. Based on predicted proteins and existing reaction databases, one can construct the metabolic network for the species. Next is to identify critical metabolic pathways from the network. Existing computational techniques identify conserved pathways based on multiple networks of related species, but have the following drawbacks. Some do not rely on additional information, so only locate short (of length at most 5), but not necessarily interesting, conserved paths. The others require extensive information (the complete pathway on one species). In reality, researchers usually know only partial information of a metabolic pathway and may not have a conserved pathway in a related species. The Conserved Metabolic Pathway (CMP) problem is to find conserved pathways from the networks with partial information on the initial substrates and final products of the target pathways. Experimental results show that our algorithm CMPFinder can predict useful metabolic pathways with acceptable accuracy.**

***Keywords: Metabolic Network, Conseved Metabolic Pathways, Building Block***

## I. INTRODUCTION

Metabolism refers to the set of cellular processes. These processes are not isolated events, but interrelated and can be modeled by a metabolic network in which each compound and each enzyme (reaction) is represented by a vertex in the network and an edge connects a compound and a reaction if the compound is involved in the reaction. A metabolic network captures the set of chemical reactions among substrates, compounds and enzymes that represent the metabolism within a cell. Conceptually, a metabolic network can be divided into functional pathways. Identifying different metabolic pathways of a species is an important topic in biological research. Any subtle shifts or malfunctions in metabolic pathway may result in diseases. For example, phenylketonuria (PKU) is a metabolic disorder caused by the lack of the enzyme, phenylalanine hydroxylase, which may cause mental retardation in a person. There may also be important metabolic activities that lead to the drug resistance property of pathogenic bacteria. This topic is particularly important for studying new species that have high impact, such as endophytic fungi that can produce fuel and pathogenic bacteria.

However, it is not an easy task to identify a metabolic pathway in laboratory. It involves many difficult subtasks such as metabolic flux analysis [1] and labeling techniques for dynamic metabolic profiling [2]. All these require advanced technologies which are expensive and time-consuming. Another direction is to make use of the comparative approach by comparing metabolic networks of related species. There is a lot of information available in databases, such as KEGG and EcoCyc, which contain information about individual reactions among substrates, enzymes and products. Even for new species, based on predicted genes/proteins, one can construct a metabolic network of the species from this information. The next step is to identify critical pathways from the network.

Since many metabolic activities are believed to be fundamental and conserved in living organisms, a traditional computational approach will try to extract conserved sub-networks (pathways) from multiple networks of related species so as to obtain useful information about the pathways. Many algorithms have been developed for finding conserved sub-networks from multiple networks. For example, algorithms in [3-5] find conserved dense subgraphs in PPI network for predicting protein complexes. However, these algorithms cannot be applied for finding metabolic pathways because metabolic pathways are usually sparse. Other algorithms [6-11] find conserved sub-networks by network alignment. As these algorithms are for general application, they cannot model the relationship between reactions, e.g. the order of a chain of reactions in a pathway.

Some algorithms [12-16] are developed specifically for finding conserved metabolic pathways. Some of them [12,13] find conserved reactions in multiple species and reconstruct the pathways using the conserved reactions. However, these

978-1-4244-8302-0/10/$26.00 ©2010 IEEE

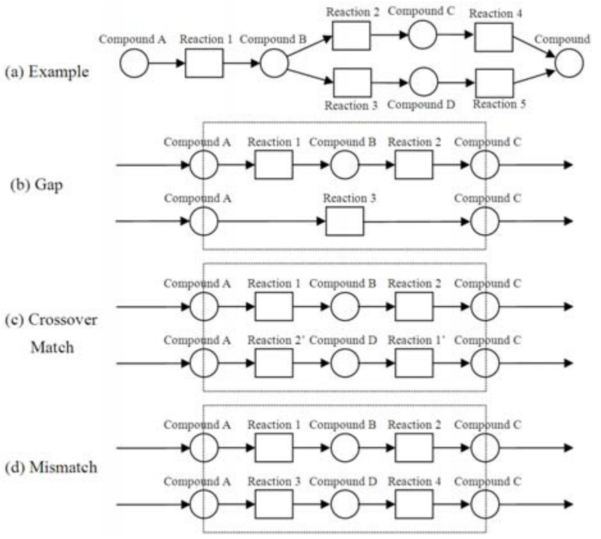

Figure 1. (a): An example of metabolic network. (b), (c), (d): All penalty blocks with unit penalty. *For crossover match, Reaction 1 (Reaction 2) and Reaction 1' (Reaction 2') are reactions using similar enzymes, i.e., the penalty block represents two chains of reactions for producing compound C from compound A with different order of reactions. For gap and mismatch, the penalty block represents two different chains of reactions for producing compound C from compound A.*

algorithms cannot capture the variations and mutations between the species and fail to find conserved pathways with slightly different in different species. Other algorithms [14-16] find conserved metabolic pathways directly by allowing gaps and differences. The conserved metabolic pathways found by these algorithms are useful not only for prediction of accurate pathways, but also for identifying and understanding the crucial processes that are present in multiple species.

However, there are still some drawbacks as these algorithms either assume there is no information of the pathway or require a known similar pathway as reference. Algorithms with no information of the pathway extract all conserved pathways from the metabolic networks. Besides computationally difficult, they can at best locate conserved paths of size at most 5 [14] and there can be thousands of such paths. It is difficult to interpret the results. Also, due to the constraint on the size of the reported paths (of at most 5), the results may not be very useful and cannot lead to a good understanding of the global picture of more complicated metabolic activities/processes.

At the other extreme, some algorithms assume the details (i.e., all compounds, enzymes, and how they interact) of a metabolic pathway of a species are known (e.g. [15,16]). They then locate a corresponding conserved pathway in another species by graph alignment. However, the assumption of knowing the details of the pathway of one species is not always realistic. Usually when a new metabolic activity is being investigated, only partial information about the pathway is available. For example, we may only know about what the initial substrates and some of the final products of a metabolic pathway are, without knowing any intermediate compounds or reactions in the pathway. Existing algorithms are not useful in solving this problem.

**Our contributions:** In this paper, we consider the CMP problem for predicting metabolic pathways with limited pathway information. Given a set of metabolic networks and a set of initial substrates and final products, the CMP problem aims at finding conserved pathways that can convert the initial substrates into the final products. We developed an algorithm CMPFinder to solve this problem with time complexity O($n^3$) where $n$ is the number of compounds and reactions in the input.

We have evaluated the performance of CMPFinder using three sets of real data. (a) a single E. coli network, (b) two networks of E. coli and S. cerevisiae; and (c) two networks of E. coli and H. sapiens. We assume that we have partial information (only initial substrates and final products) of pathways in E. coli and try to identify pathways (conserved pathways for two networks) with this minimal knowledge. The results show that we are able to identify useful pathway information and the accuracy is comparable to that of GraphMatch [23], which requires the whole pathway information of one species as input.

## II. CONSERVED METABOLIC PATHWAY

The *Conserved Metabolic Pathway* (CMP) problem is to predict the conserved metabolic pathway for producing a set of product compounds from a set of substrate compounds in multiple metabolic networks of related species.

The metabolic reactions that occur in a species can be represented by a graph with each reaction and its corresponding substrate and product compounds as vertices. There is a directed edge from a compound to a reaction if the compound is a substrate of the reaction and there is a directed edge from a reaction to a compound if the compound is a product of the reaction. If the reaction is reversible, i.e. two compounds A and B can be used to produce each other, we will represent it as two reactions using two vertices. In a metabolic network, the product of one reaction will become the substrate of another reaction; therefore, a metabolic network can be represented by a graph. An example of a metabolic network is shown in Figure 1.

Given a set of $k$ metabolic networks $G_1, G_2, \ldots, G_k$ which represent some known metabolic reactions of $k$ species, a conserved pathway is a set of reactions (can be represented as sub-network because there may be more than one initial substrate and one final product) which takes the same initial substrates to produce the same set of products in each network, in the sense that each reaction might not be identical but some of the intermediate compounds should be the same. A conserved pathway can be divided into short conserved sub-paths. Each sub-path have the same starting compound vertex $u$ and the same ending compound vertex $v$ in each network which can be aligned by forming *building blocks* similar to those formed by Li et al. [14]. A building block is made up of $k$ aligned sub-paths from $G_1, G_2, \ldots, G_k$ such that the first vertex in each sub-path refers to the same compound $u$ and the last vertex in each sub-path refers to the same compound $v$. A building block is an identical building block if the number of compound vertices (except for the

TABLE I. SIZES OF THE GRAPHS GENERATED FROM THE KEGG DATABASE

| Species | Vertices | Compound vertices | Reaction vertices | Edges |
|---|---|---|---|---|
| *Escherichia coli* | 2670 | 1103 | 1567 | 3439 |
| *Saccharomyces cerevisiae* | 2209 | 962 | 1237 | 2754 |
| *Homo sapiens* | 3701 | 1567 | 2134 | 4601 |

TABLE II. TOTAL NUMBER OF PATHWAYS FOUND BY CMPFINDER

| Accuracy $\sigma$ | Number of pathways with accuracy $\geq \sigma$ | | |
|---|---|---|---|
| | *E. coli* | *E. coli & S. cerevisiae* | *E. coli & H. sapiens* |
| 1 | 18 | 17 | 23 |
| $0.9 \leq \sigma < 1$ | 11 | 7 | 13 |
| $0.8 \leq \sigma < 0.9$ | 18 | 15 | 16 |
| $0.7 \leq \sigma < 0.8$ | 19 | 10 | 13 |
| $0.6 \leq \sigma < 0.7$ | 12 | 11 | 8 |
| $0.5 \leq \sigma < 0.6$ | 13 | 9 | 9 |
| Total | 91 (95%) | 69 (97%) | 82 (92%) |

first and last vertices) in each sub-path is 0, i.e. there is a reaction that turns $u$ into $v$ in all the $k$ networks. A building block is a penalty building block if the length of some sub-paths are larger than 2. The penalty of a building block is equal to the number of compound vertices (except for the first and last vertices) in the longest sub-paths. Identical blocks represent highly similar reactions while penalty blocks capture evolutionary diversity such as gaps [17], mismatches and crossover mismatches [18,19] between chains of reactions. Figure 1 shows all the unit penalty blocks for two species ($k = 2$).

Given $k$ paths $P_1, P_2, \ldots, P_k$ with the same starting and ending compound vertices in $G_1, G_2, \ldots, G_k$ respectively, these $k$ sub-paths are *conserved* if and only if we can divide each path $P_i$ into short sub-paths such that (1) these sub-paths can form building blocks in order with at most $g$ penalty blocks and (2) the penalty of each block is at most $l$. Given a set of metabolic networks $G_1, G_2, \ldots, G_k$ from similar species, a set of initial substrate compounds and a set of final product compounds, the CMP problem is finding a sub-graph from each metabolic networks $G_1, G_2, \ldots, G_k$ such that for each initial substrate compound $s$, we can find a chain of reactions from $s$ to any final product compounds $p$. Formally, the problem can be defined as follows:

**Conserved Metabolic Pathways (CMP) Problem**: Given $k$ directed graphs $G_1, G_2, \ldots, G_k$, a set of initial substrate compounds $s_1, s_2, \ldots, s_a$, a set of final product compounds $p_1, p_2, \ldots, p_b$, maximum number of penalty blocks $g$ and the maximum penalty $l$, we want to find $k$ acyclic subgraphs $S_1, S_2, \ldots, S_k$, one from each graph, such that (1) for each compound $s_1, s_2, \ldots, s_a$ and $p_1, p_2, \ldots, p_b$, there is are paths $P_1, P_2, \ldots, P_k$ in $S_1, S_2, \ldots, S_k$ respectively from $s_i$ to $p_j$, for some $i = 1, 2, \ldots, a$ and $j = 1, 2, \ldots, b$, such that $P_1, P_2, \ldots, P_k$ can be aligned with at most $g$ penalty blocks, each with at most penalty $l$, and (2) all compounds in $S_1, S_2, \ldots, S_k$, except $s_1, s_2, \ldots, s_a$, have positive in-degrees and all compounds in $S_1, S_2, \ldots, S_k$, except $p_1, p_2, \ldots, p_b$, have positive out-degrees.

## III. METHODOLOGY

We developed an algorithm CMPFinder for solving the CMP problem. CMPFinder first constructs a weighted directed graph G, where a vertex represents a common compound in the input graphs $G_1, G_2, \ldots, G_k$ and a directed edge $(u, v)$ in G represents a building block producing compound $v$ from compound $u$. The edge weight is 0 when the building block is an identical block and the weight is 1 when it is a penalty block. Hence, a path in G with total weight $g$ represents a conserved path, i.e. an alignment of a path from each input graph with $g$ penalty blocks, each of which has at most $l$ penalties. Then CMPFinder will discover all conserved path in G from each initial substrate compounds to final product compounds using Floyd-Warshall algorithm [22]. We will describe these two steps in details.

In order to construct graph G, CMPFinder first finds the list of common compounds in the input graphs $G_i$ which takes $O(kn)$ time. Then it determines whether there is a path of length at most $2l + 2$ from $u$ to $v$ in each input graph. It can be done by preprocessing all pairs of compounds for each input graph by multiplying the adjacency matrix of the input graph. For an input graph $G_i$ with $n$ vertices, the adjacency matrix $A_i$ is an $n \times n$ matrix where $A_i(u, v) = 1$ if and only if there is a path of length at most 1 from vertex $u$ to vertex $v$, i.e. there is an edge from vertex $u$ to vertex $v$ (we assume $A_i(u, u) = 1$); otherwise, $A_i(u, v) = 0$. Consider the $n \times n$ matrix $A_i^2$ which is the result of matrix $A_i$ multiplied by itself. $A^2(u, v) \geq 1$ if and only if there is a vertex $w$ with $A(u, w) = 1$ and $A(w, v) = 1$, i.e. there is a path of length at most 2 from $u$ to $v$. Note that a path of length 1 from $u$ to $v$ can also be discover as $A(u, u) = 1$ and $A(u, v) = 1$. As we want to know whether there exists a path of length $\leq 2l + 2$ from any two vertices in $G_i$, we consider the result of matrix $A_i$ multiplied by itself $2l + 2$ times. Similarly, $A_i^{2l+2}(u, v) \geq 1$ if and only if there is a path of length at most $2l + 2$ from vertex $u$ to vertex $v$. Since multiplying two $n \times n$ matrices takes $O(n^{2.4})$ [20], and using repeated squaring, $O(\lg l)$ matrix multiplications are needed. It takes $O(n^{2.4} \lg l)$ time to find all the pairs of vertices with distance at most $2l + 2$ and $O(kn^2)$ to construct all edges in G, thus the weighted directed graph G can be constructed in $O(kn + kn^2 + kn^{2.4} \lg l) = O(kn^{2.4} \lg l)$ time.

Based on the principle of least action [21] and the fact that essential metabolic pathway is usually small, a conserved pathway from initial substrate to final product with the least number of reactions is a candidate of real metabolic pathway. Given a set of initial substrates $s_1, s_2, \ldots, s_a$ and final products $p_1, p_2, \ldots, p_b$, CMPFinder finds all paths from $s_1, s_2, \ldots, s_a$ to $p_1, p_2, \ldots, p_b$ in graph G with less than $g$ penalty blocks. It can be done using the Floyd-Warshall algorithm [22] and the time needed is $O(n^3)$. Finally CMPFinder traces back all paths (at most $ab$ paths) of length at most $lg+n$ in the input graph(s). With the help of the common nodes in the conserved paths, the tracing process requires the information of the shortest path of length $O(l)$ from $u$ to $v$ in each input graph which can be got in $O(kln^2)$ times. Note that CMPFinder can detect those substrate

TABLE IV. PERFORMANCE OF CMPFINDER IN THE EXPERIMENTS, WITH STATISTICS BROKEN DOWN BY PATHWAY SIZES

| **Pathway size** | ***E. coli*** | | | | ***E. coli and S. cerevisiae*** | | | | ***E. coli and H. sapiens*** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | # | Sens | Spec | Acc | # | Sens | Spec | Acc | # | Sens | Spec | Acc |
| 3 to 10 | 15 | 0.921 | 0.957 | 0.937 | 17 | 0.912 | 0.984 | 0.940 | 29 | 0.821 | 0.912 | 0.859 |
| 11 to 20 | 23 | 0.791 | 0.924 | 0.848 | 23 | 0.772 | 0.888 | 0.820 | 24 | 0.829 | 0.875 | 0.844 |
| 21 to 30 | 11 | 0.712 | 0.889 | 0.785 | 10 | 0.707 | 0.869 | 0.769 | 10 | 0.797 | 0.817 | 0.800 |
| 31 to 50 | 25 | 0.653 | 0.811 | 0.712 | 13 | 0.555 | 0.827 | 0.667 | 17 | 0.658 | 0.739 | 0.687 |
| > 50 | 22 | 0.541 | 0.807 | 0.653 | 8 | 0.613 | 0.763 | 0.677 | 9 | 0.624 | 0.710 | 0.658 |
| Total | 96 | 0.709 | 0.869 | 0.775 | 71 | 0.739 | 0.883 | 0.798 | 89 | 0.769 | 0.838 | 0.795 |

TABLE III. TOTAL NUMBER OF PATHWAYS FOUND BY CMPFINDER

| **Accuracy σ** | **Number of pathways with accuracy ≥ σ** | | |
|---|---|---|---|
| | *E. coli* | *E. coli & S. cerevisiae* | *E. coli & H. sapiens* |
| 1 | 18 | 17 | 23 |
| 0.9 ≤ σ < 1 | 11 | 7 | 13 |
| 0.8 ≤ σ < 0.9 | 18 | 15 | 16 |
| 0.7 ≤ σ < 0.8 | 19 | 10 | 13 |
| 0.6 ≤ σ < 0.7 | 12 | 11 | 8 |
| 0.5 ≤ σ < 0.6 | 13 | 9 | 9 |
| Total | 91 (95%) | 69 (97%) | 82 (92%) |

compounds $s_i$ that cannot reach any product compounds $p_j$ or product compounds $p_j$ cannot be reached by any substrate compounds $s_i$, as the path penalty from $s_i$ to $p_j$ is larger than $g$. In this case, CMPFinder will report that there is no solution.

## IV. EXPERIMENTS

In this section, CMPFinder was tested on its performance to find metabolic pathways from a single metabolic network and from pairwise alignment of metabolic networks. We also compared CMPFinder with GraphMatch [23], a known query approach, on finding metabolic pathways.

We evaluated the performance of CMPFinder using the metabolic networks of E. coli, S. cerevisiae and H. sapiens constructed based on the reaction information from Release 50.2 of the species-specific KEGG databases [24]. Co-factors such as water, ATP or ADP were not included in the networks, as they are not major substrates and products in most reactions. We assume a one-to-one mapping between vertices of the two networks, under the assumption that metabolic networks are highly conserved. Only identical compounds which have the same compound ID are mapped together. The size of these three networks are shown in Table I.

For each known pathway defined in KEGG, we extracted the set of initial substrates and final products (instead of the whole pathway as required by other software such as GraphMatch) as input and evaluated CMPFinder by its ability to predict the known pathways given the initial substrates, final products and relevant metabolic networks. In our experiments, we tested CMPFinder on finding metabolic pathways in (1) a single E. coli network and (2) conserved metabolic pathways between E. coli and S. cerevisiae, and (3) between E. coli and H. sapiens. The results in using two metabolic networks were compared with those of GraphMatch.

The maximum number of penalty blocks $g$ = 3 and maximum penalty $l$ = 1 which support evolutionary like gaps, mismatches and crossover mismatches. We evaluate the performance of CMPFinder using accuracy defined as follows. For two species, only vertices common between the input metabolic networks were counted as known pathway vertices.

$$\text{sensitivity} = \frac{\text{number of correctly predicted vertices}}{\text{number of known pathway vertices}}$$

$$\text{specificity} = \frac{\text{number of correctly predicted vertices}}{\text{number of predicted vertices}}$$

$$\text{accuracy} = \sqrt{\text{sensitivity} \times \text{specificity}}$$

From the single E. coli network, CMPFinder found 91 out of 96 known pathways with accuracy ≥ 0.5. In experiments between E. coli and S. cerevisiae, 69 out of 71 known common pathways were found with accuracy ≥ 0.5. For E. coli and H. sapiens, 82 conserved pathways were found from the 89 known common pathways with accuracy ≥ 0.5 (Table II). The average accuracy of CMPFinder on finding metabolic pathways from a single E. coli network was 0.775, while the average specificity and sensitivity were 0.869 and 0.709. CMPFinder performed better when given two metabolic networks. The average accuracy for E. coli and S. cerevisiae was 0.798, while the average specificity and sensitivity were 0.883 and 0.739 respectively. For E. coli and H. sapiens, a similar result was obtained with accuracy of 0.795, specificity of 0.838, and sensitivity of 0.769 (Table III).

18 pathways found in E. coli, 17 pathways found between E. coli and S. cerevisiae and 23 pathways found between E. coli and H. sapiens exactly match with the corresponding known conserved pathways with an accuracy of 1 (Table II). Around half of the output graphs have accuracies higher than 0.8. For more than half of the pathways (37 of 71) conserved between E. coli and S. cerevisiae, CMPFinder performed better when given both metabolic networks rather than only the E. coli network. The accuracies for 12 pathways were the same for both cases.

In the majority of the cases, CMPFinder performed better with two input metabolic networks, as conservation between two species gave a better confidence to the pathways than basing solely on single species. Figure 2 shows the results of CMPFinder on the phenylalanine, tyrosine and tryptophan biosynthesis, based on the single E. coli network and the E. coli and S. cerevisiae networks. CMPFinder obtained a pathway (contains multiple paths) with the accuracy of 0.702

which given a single E. coli network and a pathway with the accuracy of 0.848 when given the E. coli and S. cerevisiae networks. As an alternative shortest path not in the known pathway was found in the single E. coli network, the specificity of the pathway found was lower, leading to its poorer performance. Lapses of accuracy occur when there are multiple reactions between two different compounds or loops in the metabolic pathway becasue our model is not able to capture multiple reactions and cycles.

We compared the performance of CMPFinder on two species alignment with GraphMatch [23]. GraphMatch is a graph matching algorithm which finds the optimal conserved graph given a metabolic network, a query network, and a mapping between the query vertices and the network vertices. As GraphMatch only accepts connected graphs as queries, a query was constructed from each isolated component from the known metabolic pathways in E. coli, i.e. if the known metabolic network has *y* isolated components, *y* queries were needed. The output for each isolated metabolic pathway was merged together as the final output, similar to our approach in running CMPFinder. Since GraphMatch uses exponential space depending on the query size, 29 queries failed to complete for the two experiments when using a machine with 8GB RAM. As a result, we could obtain the full results of conserved metabolic graphs from only 42 of the 71 common pathways between E. coli and S. cerevisiae (Table IV).

Considering only the pathways where GraphMatch could be completed for all the queries, the average accuracy of CMPFinder was 0.873 while the average accuracy of GraphMatch was 0.919. With the advantage of knowing the whole query graph, GraphMatch has a comparable performance to CMPFinder. CMPFinder outperformed GraphMatch in all the 29 pathways where a query failed to complete in GraphMatch. As a result, considering all 71 conserved pathways between E. coli and S. cerevisiae, the average accuracy of CMPFinder (0.798) was higher than GraphMatch (0.608) by 24%. Compared with GraphMatch, CMPFinder requires less information input and less memory, which makes it more practical to be used in a real biological setting, while yielding results of comparable or even better accuracy.

## V. Conclusion

In this paper, we considered a computational problem to predict metabolic pathways in multiple networks based only on the set of initial substrates and products. Algorithm CMPFinder is developed to solve this problem. According to the experimental results on real datasets, CMPFinder is effective and its performance is comparable with GraphMatch even using less information. We are in the process of modifying CMPFinder and so that additional information (such as known intermediate compounds/reactions or compounds that are known not to be in the pathways) about the pathways can be taken into account to make both algorithms more practical.

## Acknowledgment

This research is supported by the grant HKU 7116/08E

## References

[1] G.G. Harrigan and R. Godacre, “Metabolic Profiling: Its Role in Biomarker Discovery and Gene Function Analysis”, Kluwer Academic Publishers, 2003.

[2] W. Wiechert, M. Mollney, S. Petersen and A.A. Graaf, “A Universal Framework for 13C Metabolic Flux Analysis”, *Metabolic Engineering*, 2001, 3(3), pp 265-283.

[3] B. Adamcsek, G. Palla, I.J. Farkas, I. Derenyi and T. Vicsek, “CFinder: locating cliques and overlapping modules in biological networks”, *Bioinformatics*, 2006, 22(8), pp. 1021-1023.

[4] G.D. Bader and C.W.V. Hogue, “An automated method for finding molecular complexes in large protein interaction networks”, *BMC Bioinformatics*, 2003, 4(2).

[5] H. Leung, Q. Xiang, S.M. Yiu and F. Chin, “Predicting Protein Complexes from PPI Data: A Core-Attachment Approach”, *J of Computational Biology*, 2009, 16, 133-144.

[6] J. Flannick, A. Novak, B.S. Srinivasan, H.H. McAdams and S. Batzoglou, “Graemlin: general and robust alignment of multiple large interaction networks”, *Genome research*, 2006, 16, pp. 1169-1181.

[7] J. Flannick, A. Novak, C.B. Do, B.S. Srinivasan and S. Batzoglou, “Automatic Parameter Learning for Multiple Network Alignment”, *RECOMB*, 2008, 12, pp. 214-231.

[8] R. Sharan, S. Suthram, R.M. Kelley, T. Kuhn, S. McCuine, et al., “Conserved patterns of protein interaction in multiple species”, *PNAS*, 2005, 102(6), pp. 1974-1979.

[9] D. Croes, F. Couche, S.J. Wodak and J. Helden, “Metabolic PathFinding: Inferring Relevant Pathways in Biochemical Networks”, *Nucleic Acids Research*, 2005, 33, pp. W326-W330.

[10] R. Singh, J. Xu, B. Berger, “Pairwise global alignment of protein interaction networks by matching neighborhood topology”, *RECOMB*, 2007, 11, pp. 16-31.

[11] J. Berg and M. Lassig, “Cross-species analysis of biological networks by Bayesian alignment”, *PNAS*, 2006, 103(29), pp 10967-10972.

[12] J.C. Clemente, K. Satou and G. Valiente, “Finding conserved and non-conserved reactions using a metabolic pathway alignment algorithm”, *Genome informatics International Conference on Genome Informatics*, 2006, 17, pp 46-56.

[13] P.D. Karp, S.M. Paley, M. Krummenacker, M. Latendresse, J.M. Dale, *et al.*, “Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology”, *Brief Bioinform*, 2010, 11, pp 40-79.

[14] Y. Li, D. de Ridder, M.J.L. de Groot and M.J.T. Reinders, “Metabolic pathway alignment (M-Pal) reveals diversity and alternatives in conserved networks”, *Proc APBC*, 2008, 6, pp. 69-78.

[15] R.Y. Pinter, O. Rokhlenko, E. Yeger-Lotem and M. Ziv-Ukelson, “Alignment of metabolic pathways”, *Bioinformatics*, 2005, 21(16), pp. 3401-3408.

[16] Q. Yang and S.H. Sze, “Path matching and graph matching in biological networks”, *Journal of Computational Biology*, 2007, 14(1), pp 56-67.

[17] S. Roy, “Multifunctional enzymes and evolution of biosynthetic pathways: retro-evolution by jumps”, *Proteins*, 1999, 37, pp. 303-309.

[18] R.A. Jensen, “Enzyme recruitment in evolution of new function”, *Annu Rev Microbiol*, 1976, 30, pp. 409-425.

[19] S. Schmidt, S. Sunyaev, P. Bork and T. Dandekar, “Metabolites: a helping hand for pathway evolution?”, *Trends in Bioch Sci*, 2003, 28(6), pp. 336-341.

[20] D. Coppersmith and S. Winograd, “Matrix multiplication via arithmetic progressions”, *Journal of Symbolic Computation*, 1990, 9, pp. 251-280.

[21] R. Hoffmann, V.I. Minkin and B.K. Carpenter, “Ockham’s Razor and Chemistry”, *International Journal for Philosophy of Chemistry*, 1997, 3, pp. 3-28.

[22] R.W. Floyd, “Algorithm 97: Shortest path”, *Communications of the ACM*, 1962, 5(6), pp. 345.

[23] Q. Yang and S.H. Sze, “Path Matching and Graph Matching in Biological Networks”, *J of Computational Biology*, 2007, 14(1), pp. 56-67.

[24] M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, *et al*., “KEGG for linking genomes to life and the environment”, *Nucleic Acids Res*, 2008, 36, pp. 480-484.