KADIR HAS UNIVERSITY
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

GLOBAL ALIGNMENT OF METABOLIC PATHWAYS AND PROTEIN-PROTEIN INTERACTION NETWORKS

GRADUATE THESIS

GAMZE ABAKA

September, 2014

Gamze Abaka	M.S. Thesis	2014

GLOBAL ALIGNMENT OF METABOLIC PATHWAYS AND PROTEIN-PROTEIN INTERACTION NETWORKS

by

Gamze Abaka

Bachelor's degree, Computer Engineering, Kadir Has University, 2012

Submitted to the Graduate School of

Science and Engineering in partial fulfillment of the requirements for the degree of

Computer Engineering

Master of Science

Kadir Has University

2014

GLOBAL ALIGNMENT OF METABOLIC PATHWAYS AND PROTEIN-PROTEIN INTERACTION NETWORKS

APPROVED BY:

Assoc. Prof. Dr. Cesim Erten ..... 
(Thesis Supervisor)

Asst. Prof. Dr. Öznur Yaşar Diner .....

Asst. Prof. Dr. Şebnem Eşsiz Gökhan .....

DATE OF APPROVAL: .....

ABSTRACT

GLOBAL ALIGNMENT OF METABOLIC PATHWAYS AND PROTEIN-PROTEIN INTERACTION NETWORKS

Metabolic pathways and protein interaction networks are essential at almost every function for living organisms. Simply, while reactions produce life energy within cells, protein interaction networks provide biological functions. Also, abnormal reactions or interactions cause various disorders. Thus, in bioinformatics, most of the studies are based on these networks in order to find hopeful results for these disorders and biological challenges. Solving alignment problem is one of these studies such that it tries to find similar reactions, proteins or functions. In this thesis, we focus on that problem within both metabolic pathways and protein interaction networks. Firstly, we propose a constrained alignment algorithm, CAMPways, for one-to-many alignment of metabolic pathways and we extend the framework, CAPPI, for one-to-one protein interaction network alignment with necessary changes. Afterwards, we provide the computational intractability of the problem and finally we compare our algorithm with different algorithms on actual metabolic pathways and protein interaction networks.

ÖZET

METABOLİK YOLAKLARIN VE PROTEİN ETKİLEŞİM AĞLARININ HİZALANMASI

Metabolik yolaklar ve protein etkileşim ağları yaşayan canlıların neredeyse tüm fonksiyonlarında hayati önem taşmaktadır. En basit haliyle, reaksiyonlar hücre içinde yaşam enerjisi üretirken, protein etkileşim ağları biyolojik fonksiyonların gerçekleşmesini sağlamaktadır. Ayrca, normal olmayan reaksiyonlar ya da etkileşimler çeşitli hastalıklara neden olmaktadır. Bu nedenle, biyoinformatik alanındaki bir çok çalışma, bu hastalıklara ve biyolojide çözülmesi gereken sorunlara umut verici sonuçlar alabilmek amacıyla, bu ağlara dayanmaktadır. Hizalama probleminin çözülmesi, bu çalşmalardan biridir ve bu problem, benzer reaksiyonları, proteinleri ya da fonksiyonları bulmaya çalşır. Bu tez kapsamında, hem metabolik yolaklar hem de protein etkileşim ağları için hizalama problemi ele alınmaktadır. Öncelikle metabolik yolakların bire-çok hizalanması için kısıtlandırılmış bir algoritma (CAMPways) sunulmakta, daha sonra bu algoritma protein etkileşim ağlarının bire-bir hizalanması için geliştirilmekte (CAPPI) ve gerekli değişiklikler uygulanmaktadır. Problemin işlemsel karmaşıklığı verilip, gerçek veriler üzerinde diğer algoritmalar ile karşlaştırmaları yapılmaktadır.

ACKNOWLEDGEMENTS

It is my great pleasure to thank many people who helped me for my studies and inspired me for my life. Firstly, I would like to thank Assoc. Prof. Cesim ERTEN, my thesis advisor, for his valuable and constructive suggestions. Also, I would like to offer my special thanks to him for endearing bioinformatics area to me.

I am particularly grateful for the assistance given by research assistances, my friends. Especially, many thanks to Aykut Çayır for his helps about programming knowledge and Serkan Altuntaş for his valuable ideas and inspirations about bioinformatics and biological processes. I would also thank to all computer engineering professors for their worthful courses and my high school teacher Şenay Öztürk for her inspiration.

I would like to express my very great appreciation to my family and my close friends. I thank my parents Nermin & Doğan Abaka for their endless confidence and support. Finally, I would like to thank my friends Barış Karataş, Çiğdem Öztürk, Kadir Özbek and Seda Çam and all my other friends who inspire me.

To everyone who endeared world to her,

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

The main purpose of biology is to understand the cell system such that the fundamental questions can be answered: How the cellular functions happen and which interactions are established, in particular, how the proteins interact with each other to perform proper functions and how the reactions occur to maintain essential processes in the cell.

In this introductory chapter, there exist three main sections such that in the first section, the definition of metabolic pathways is given in order to indicate the significance of the subject. Afterwards, the notions that are used for the implementation of these pathways in the alignment problems are given. In the second section, protein-protein interaction networks and their notions are defined as well, and finally in the third main section, network alignment problem is described as a summary for a better understanding of the subject.

1.1. Metabolic Pathways

In this section, we give definitions of metabolic pathways, reactions and metabolic networks with the significance of them in the bioinformatics and we define the notions for a better understanding of the next problem sections.

1.1.1. Metabolism and Metabolic Reactions

In order to show the big picture of the metabolism by providing the important aspects of the subject, we first need to begin from the smallest piece of the metabolism: A *chemical reaction* is an occurrence of the interaction of two or more chemical substances that have accurate characteristic properties, and a transformation of them to others. Generally, the transformation separates the reactions as *catabolic* and *anabolic*. Whereas catabolic reactions

provide *Adenosine triphosphate (ATP)* that refers to chemical energy is usually used for the occurrence of new reactions such as anabolic reactions within cells and metabolism of living organisms by breaking down the complex organic molecules into smaller ones such as breaking down the sugar to obtain energy, anabolic reactions use the provided energy by catabolic reactions and gather the small molecules together to obtain complex organic molecules such as attaining protein by gathering the amino acids together. Both catabolic and anabolic reactions can happen at any moment but when the living organism is young, the number of anabolic reactions is greater than the number of catabolic reactions for growth and the living organism gets older, firstly the number of catabolic reactions balances with the number of anabolic reactions and afterwards, exceeds anabolic reactions number.

Normally, the chemical reaction happens due to some important components such as physiological pH and temperature. The temperature is the vital component for many living organisms and it helps to maintain life and carry out some medical procedures. For instance, even if many animals have stable body temperature, some of them are affected from the cold temperature that decreases the metabolism significantly and causes the hibernation. Also, for the important surgeries such as heart and brain, the temperature of the operating room is reduced to slow up the metabolism.

The biochemical reaction is usually catalyzed by an enzyme that is a protein or RNA. Of course, each enzyme does not perform same tasks. Whereas several mechanisms affect the catalytic action of the enzyme such as the interaction or the shape of the molecule, there are two major tasks of enzymes such as increasing the rate of the reaction and helping to produce product by providing high activity [3]. Also, when the enzyme catalyzes the reaction, it uses the substrates. A *substrate* refers to a required molecule for the cell and is used as a input by the reaction. Also substrates transform to *products* when the reaction is completed. Both a substrate and a product is defined as a molecule that can be in one of three major categories: Carbohydrates, fats and proteins. But in some cases, the nucleic acids such as

ribose and deoxyribose can be the substrate or the product. Hereby, the important point is the products of the reaction can be the substrates of the functionally-related reactions and the series of reactions are created. Because these series are essential to grow, repair, reproduce and respond to environmental conditions, they are one of the most interesting and popular topics in the bioinformatics.

The *metabolism* that is sometimes called *intermediate metabolism* is the set of all these chemical reactions and biological processes on them and refers to a connection between phenotype and genotype of the species. In general, the metabolism is divided two categories such as catabolism and anabolism. It is possible to see that while catabolism is the set of catabolic reactions, anabolism consists of anabolic reactions.

1.1.2. Significance of Metabolism

As it is mentioned before, the reactions within the metabolism keep cells and organisms alive by producing energy for growth and maintaining the life. They are used in the essential processes such as growing and repairing damages. Besides, the determination of the substances as nutritiuos or poisonous is made by the metabolic system of the living organism and this determination helps to life-sustaining reactions. For example, whereas hydrogen sulfide is nutritiuous for prokaryotes, it is poisonous for animals [4].

The knowledge of all reactions and their interactions within the metabolism is also important for the medicine and pharmacy such that if the genetic enzyme-catalyzed reactions are known, proper diagnosis and treatments are developed for the most important and common metabolic diseases such as gout and diabetes. According to Danaei [5], *"The number of people with diabetes increased from 153 (127-182) million in 1980, to 347 (314-382) million in 2008."* The increasing number shows that understanding, analyzing and developing methods and models for metabolisms are essential to the human life. The set of metabolic

reactions helps not only to the metabolic diseases but also the diseases that are not affected especially by the metabolic abnormal such as autoimmune and neurological diseases [6]. At this point, the metabolism leads to drug targets such that antimicrobial drugs which are given for the diseases usually affect the vital enzymes of the reactions. The modifications such as elimination and rejection of the affects of drugs and also activation of enzymes and drugs are made by the metabolism. In this manner, understanding the metabolism processes is crucial, as well.

1.1.3. Metabolic Pathways and Their Functions

First of all, there are thousands of reactions and limited number of metabolic resources in a cell. So, the usage of these resources must be efficient for regular working. As it is mentioned in the previous subsection, organizing and analyzing these metabolic reactions and resources are important in order to get reasonable and useful results. Thus, the mathematical models have been searched to maximize efficiency of the resources and complex reaction networks have been created to predict utilization and remaining rate of products and substrates. Afterwards, the limited number of biochemical reactions has been considered and sets of these reactions have been organized such that the organization of these chemical reactions within in a cell is defined as *metabolic pathways*. Thus, metabolic pathways can be defined as the subnetworks of the complex reaction networks [1] as shown in Figure 1.1. These subnetworks are actually step by step processes: Initially, the substrate is used as the input by the enzyme to catalyze the reaction. When the reaction is completed, the substrate turns into the product. The next reaction uses that product as the substrate and the interconnected reactions continue until the exact products and the processes that are required by the cell are obtained.

Several types of metabolic pathways exist and the types can show differences due to organisms: Some of them consists of both catabolic and anabolic reactions such as citric

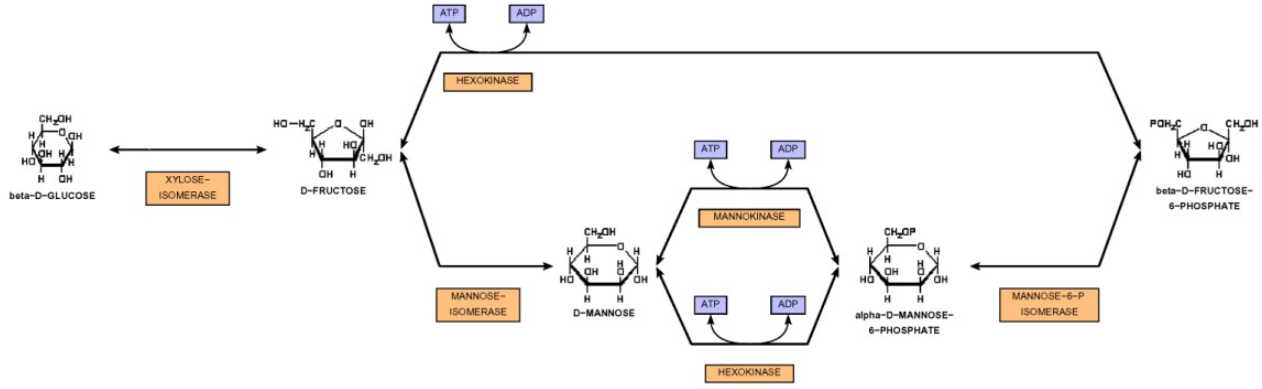

Figure 1.1. A metabolic pathway (taken from [1])

acid cycle that is the last step of chemical processes and are called *amphibolic*. On the other side, whereas a species can have a specific metabolic pathway, the central pathway that is called *glycolysis* that is the degradation of the glucose exists for all living organisms. Nevertheless, previous studies have argued that there exists remarkable variations even in that pathway [7]. According to these variations and the differences, the comparative analysis of metabolic pathways have been become the central subject in the bioinformatics. The comparative analyses are helpful for major challenges of biology. Firstly, the evolutionary relationships can be found between organisms and classification of organisms can be made more meaningful. In the second place, as it is mentioned in the previous sections, drug targets can be determined purposefully based on the species. Furthermore, the unknown parts of the metabolic pathway can be determined by comparing it with well-known pathways.

In order to organize and analyze the metabolic pathways, some databases have been developed such as KEGG [8], MetaCyc [9] and Reactome [10]. Whereas KEGG consists of the data-oriented and organism-specific resources with analysis tools, MetaCyc provides experimentally elucidated pathways with applications that help to make prediction. Reactome is different from KEGG and MetaCyc such that it is prefered mostly for visualization.

The representation of the metabolic pathway differs depending on the databases, algorithms and methods. Generally, the representation of the metabolic pathway is made by using graph formulations. First modeling is to use a directed hyper-graph such that the vertices are substrates or products and the hyper-edges represent the enzymes or reactions. The directed hyper-edge is added between the vertices depending on the producer and consumer relationships and the direction of the metabolic processes. However, the directed hyper-edges are not preferred when the challenges depend on the visualization and simulation of the metabolic pathways [1, 11]. The bipartite graphs are used where the vertices corresponding to the reactions and the edges corresponding to binary relations in KEGG for these purposes [1]. But, because it is hard to implement, use and adapt these models to metabolic pathways, the simple directed graph representations have been suggested such that the vertices represent the enzyme or the reaction and the directed edges refer to the direction of the metabolic processes, as well [11]. The simple directed graphs are useful when the product or the substrate details are negligible for the challenges such as finding alignment results. So, the directed edge addition depends on knowing which reaction is producer and which one is consumer: The directed edge is added from the node corresponding to the producer reaction or enzyme to the node corresponding to the consumer reaction or enzyme. If the nodes represent the enzymes in the metabolic pathway, the new challenge occurs such that because the enzyme may come up more than once in the metabolic pathway, a few different nodes may refer to the same enzyme in the graph corresponding to that pathway. In such a case, the nodes that represent the same enzyme may be merged. But, for more simplicity, the definition of nodes is changed and the nodes indicate the reactions.

Because the biological networks are mostly defined as the graphs, theoretical methods are needed to solve graph problems and analyze these biological networks. Solving these problems is important to obtain significant information and improve the aspects for biological networks such as finding common patterns, functional relationships, structural and sequential similarities and evolutionary classifications between organisms. Additionally, ac-

cording to Koyutürk [11], *"two key problems on graphs are aligning multiple graphs, and finding frequently occurring subgraphs in a collection of graphs".* Even if these are the major problem definitions, it is possible to extend and classify them as global alignment, local alignment, subgraph isomorphism, motif, graph and subgraph matching, graph clustering and also graph mining [12]. Even if all of them provides remarkable frameworks to detect functional relationships in general, whereas the alignment, matching and clustering algorithms help to find common patterns and functionally similar groups [13, 14], the results of graph mining algorithms provide subgraph similarities in detail [11].

1.2. Protein-protein Interaction Networks

Proteins are vital macromolecules in the cell of living organisms and most of the studies in the bioinformatics area are based on their interactions. Whereas some proteins perform their functions independently, almost all proteins interact with other proteins to perform proper biological processes. Protein-protein interactions represent purposeful physical contacts between two or more proteins depending on biochemical and physiologic events and often occur in order to carry out their biological functions. Figure 1.2 represents a protein interaction network with directed interactions and proteins. They are essential at almost every function of living organisms, for instance, in the signal transduction across the biological membranes, the movement of substances in and out of cells and RNA/DNA synthesis such as replication, transcription and translation. In addition to essential functions, previous studies have shown that abnormal interactions between proteins cause various disorders such as Alzheimer's disease and cancer [15]. Hence, identifying these interactions is crucial to understand and control the cellular functions at molecular level. Towards this goal, in the recent years, various high-throughput experimental techniques have been presented to identify, characterize and discover protein interactions such as yeast two-hybrid [16] and co-immunoprecipitation [17]. These techniques have provided promise to predict new interactions and have been supplement for new discovery methods. Following on the discovery and

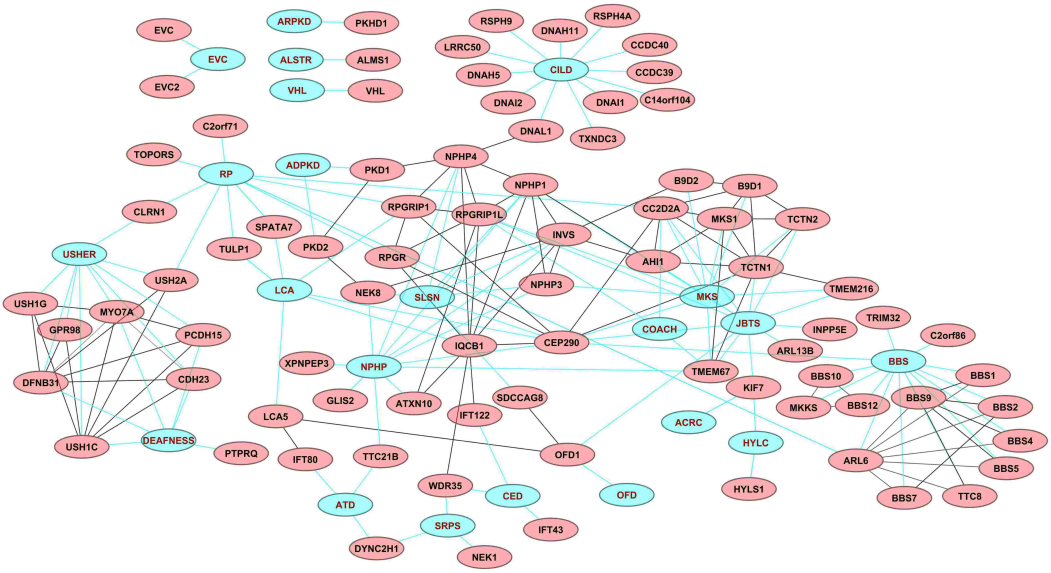

Figure 1.2. A protein interaction network (taken from [2])

prediction methods, amount of available data on protein-protein interactions has increased rapidly for different species such as human, worm, fly and yeast.

Studying the protein-protein interaction data has become a crucial problem because of the high noise levels in the data such that possibly helpful methods, models and computational approaches are required to enhance its' understanding. The available protein-protein interaction data has been represented as a network for comprehensible and reasonable studies such that each protein corresponds to a node and each direct physical interaction between two proteins corresponds to an edge in the network. Besides, as the amount of available protein-protein interaction network data increases, computational methods have been developed to make comparative protein-protein interaction network analysis and attain new predictions involving high accuracy across species.

The results of protein-protein interaction network comparisons provide crucial aspects of similarities and differences between species at the biological level and lead to find functional ortholog proteins [18]. At this point, we need to give the definition of functional ortholog.

In general, the term *ortholog* refers to genes of different species such that they come from a common ancestor and evolve after speciation and also, it is assumed that orthologs perform same functions. Thus, to determine functional ortholog proteins is a crucial step in both network alignment problems and in bioinformatics in terms of evolutionary aspect. On the other side, similarities between proteins indicate the evolutionary conservation across species and this helps to predict the biological function of individual proteins. The first measurement of protein similarities is to compare sequence similarity which means to find similarity between amino acid sequences of proteins. The similarity information between sequences reveals an idea in molecular biology such that similar protein sequences carry out similar functions. Because of the importance of this information, various homology-based algorithms, tools and powerful methods with high probability and low computationally cost have developed to compare and search protein sequences such as BLAST [19] and Hidden Markov Model (HMM)-based search methods [20]. In bioinformatics, Basic Local Alignment Tool (BLAST) is the commonly used similarity tool such that it makes sequence-based comparisons for DNA and protein sequences faster. Furtermore, the results of BLAST identify structural, functional or evolutionary relationships between sequences. Conveniently, it is supplement for matching algorithms that require approximate solutions.

The idea of sequence similarity corresponds to functional similarity has been accepted as the main concept in molecular biology for a long time. With the increase in the number of comparative sequence alignment tools, stating functional orthologs and functions of proteins have got difficult because a protein have had sequence similarity to many proteins [21]. Because only sequence similarity is not sufficient for determining true orthologs, new topology-based similarity approaches have been improved such that these methods elaborate on conserved pathways across multiple species [22]. Consequently, the measurement of conserved protein networks includes both protein sequence similarity and interaction topology.

The next section describes the types of network alignment problem that are mostly used in the previous works.

1.3. Network Alignment Problem

Network alignment problem is interested in predicting interactions and functions, finding conserved functional modules, verifying existing biological networks such as metabolic pathways and protein-protein interaction networks and discovering unknown parts of metabolic pathways and protein complexes within k different networks belonging to different organisms, spanning different challenges such as local alignment, global alignment, network querying and multiple network alignment. In general case, many formulations have been found to solve the network alignment problem, but all of them have proven that this problem is NP-Hard which means there is no polynomial time algorithm for the solution [23]. Therefore, different heuristic algorithms have been presented to align k networks for different major goals such as finding conserved regions [24] and identifying conserved functional modules of arbitrary networks [25].

Generally in the network alignment problem concept, whereas a metabolic pathway is modeled by an undirected simple graph, a protein-protein interaction network is represented as a simple directed graph. For the simple directed or undirected graph $G = \{V, E\}$, $V=$ $\{$ V_1, V_2,V_3 ... ,V_n $\}$ is a finite set of vertices corresponding to N proteins or reactions and $E \subset V \times V$ is the set of edges corresponding to interactions between proteins or the relationships between metabolic pathways such that $(u, v) \in E$ represents an interaction between proteins or a relationship between reactions where $u \in V$ and $v \in V$.

1.3.1. Global and Local Alignment Problem

The global alignment provides an end-to-end alignment of the sequences or the nodes of the graphs corresponding to the biological networks which is the best match in their entirety, even though there are suboptimal regions in the alignment. It is most helpful when the query sequences or the graphs are similar enough and their total sizes are nearly close. Furthermore, it is often used to understand variations of species by comparing genomic sequences or interactomes that are the interaction networks with details and may help to detect functional orthologs and predict functions of the biological components [26].

In general, the global alignment of the sequences is based on a dynamic programming algorithm which is called Needleman-Wunsch algorithm and the algorithm consists of two steps: Finding highest possible score by using dynamic programming and determining one or more alignment with that score. But, on the other side, when the problem consists of whole interaction-based or relation-based networks, there exists many different algorithms and studies [26, 27, 28] to align globally for both metabolic and protein interaction networks in the bioinformatics area.

On the other side, the local sequence alignment provides the best *subsequence* alignment between sequences. In general, the local alignment is due to Smith & Waterman algorithm [29] that use dynamic programming to find best local alignment using a score function and substitution matrix. There are many subjects that the alignment can be useful such as comparing both protein sequences which have common conserved patterns or domains and genomic DNA sequences against protein sequences. Besides, it is more sensitive for especially comparing highly diverged sequences.

Afterwards, the idea is extended to work on biological networks such that the local network alignment provides the subgraph(s) that has the best local alignment score between

k graphs corresponding to different biological networks. The resulting subgraphs usually show the conserved structures of the networks. But, finding the local sequence alignment or local network alignment has a challenge: Initially, the beginning and ending positions of the resulting subsequences or subgraphs are unknown. According to this challenge, obviously, finding an optimal local alignment is more complex than finding an optimal global alignment. Nevertheless, most of the previous works [23, 25, 30, 31] depend on the local alignments.

1.3.2. Pairwise and Multiple Alignment Problem

In general, the pairwise alignment provides useful information for detecting similar regions such that these regions may denote possibly functional, structural and evolutionary relationships between two biological sequences through comparisons. It has an important place in molecular biology and bioinformatics such that the vast majority of sequence analysis tools depend on pairwise alignment. These tools provide valuable insight into phylogenetic analysis, structure prediction and similarity searches within the classifications and databases.

The pairwise alignment is important not only for two sequences, but also for the biological networks in general. Various efficient computational methods have been proposed for aligning two networks and identifying their conserved pathways based on the sequence and function similarity [30, 27].

In the second place, the multiple alignment is a kind of alignment methods such that three or more biological sequences are aligned. In many cases, an evolutionary relationship such as having common ancestor is assumed between the input sequences. This alignment method is often used to conduct phylogenetic trees and dedicate both sequence homology and conservation between these sequences for evolutionary analysis such as showing historical relationships between organisms or genes and evolution of molecules and phenotypes. As it is mentioned in the previous subsections, the sequence-based idea is extended to network

alignments, as well.

Needles to say, the multiple alignment is more computationally complex than pairwise alignment not only for the sequences, but also for networks as a whole. Correspondingly, more heuristic algorithms [14, 26] are proposed rather than optimization algorithms which are computationally expensive for multiple network alignments.

1.4. The Scope and Contribution of the Thesis

With this study, we proposed a constrained one-to-many alignment algorithm that was inspired by the model suggested in SubMap algorithm [27] for metabolic pathways such that it was accepted by *Bioinformatics* and published in 2013 [32]. Furthermore, we adapted that algorithm for global one-to-one pairwise protein interaction network alignment by making the necessary changes and additions in order to get reasonable and useful results. This algorithm was implemented in C++ programming language using LEDA library [33].

First of all, we focused on global one-to-many network alignment problem and we provided the formal description for this problem. Next, we provided a novel constrained alignment framework appropriate for both one-to-one and one-to-many alignments model. Secondly, we proposed an algorithm which implements this framework efficiently. We showed the computational intractability of the constrained alignment problem and we improved the constrained alignment framework for protein-protein interaction network challenges. Finally, we presented experimental evaluations that are performed on actual metabolic pathways and protein interaction networks and also demonstrated that our algorithm gives better results in terms of biological meaning.

2. METHODS AND ALGORITHMS

In this chapter, firstly, we define the problem of global one-to-many alignment of pairwise metabolic pathways. Afterwards, we indicate the constrained alignment framework and the algorithm that is appropriate for this framework for metabolic pathways and also, we extend the algorithm with necessary changes for one-to-one pairwise alignment of protein-protein interaction networks. The algorithms are called CAMPways and CAPPI for metabolic pathways and protein interaction networks, respectively.

2.1. Problem Definition for Metabolic Pathway Alignment

Initially, we preffer to use *reaction-based* representations that are employed in SubMap [27] for metabolic pathways. Let $\mathcal{P}$ be a metabolic pathway, we use a directed unweighted graph $G_p(V_p, E_p)$ for its representation. As each node $u_{r_i} \in V_p$ is representing the reaction $r_i \in \mathcal{P}$, a directed edge (u_{r_i}, u_{r_j}) is added between the nodes u_{r_i} and u_{r_j} if the output compound (product) of r_i is the input compound (substrate) of r_j in the pathway. The extension is made due to reversibility of the reactions such that if the input compound of r_i is the output compound of r_j, then the edge existence condition is considered, as well. Similarly, the same case is considered for r_j. So, if both reactions are reversible, the existence of the edge is handled in four cases.

Hereby, we need to give a definition for the legal alignment and allowed types of mappings due to one-to-many alignment restriction. Let G_p, G'_p be the graph representations of the metabolic pathways $\mathcal{P}, \mathcal{P}'$ and R_x be a subset of V_p such that the nodes in R_x indicate an induced subgraph that is connected in its underlying graph. Let R_k indicate the set of such subsets such that the size of each subset is greater than zero and less than or equal to k and R'_k represent the similar set for G'_p. The mapping sets (R_x, R'_x) for $R_x \in R_k, R'_x \in R'_k$

indicate a *legal alignment* $\mathcal{A}$ between G_p, G'_p such that the following are satisfied:

- For $(R_x, R'_x) \in \mathcal{A}$, $|R_x|$ or $|R'_x|$ is 1.
- For $(R_x, R'_x) \in \mathcal{A}$ and For $(R_y, R'_y) \in \mathcal{A}$, $R_x \cap R_y = \emptyset$ and $R'_x \cap R'_y = \emptyset$.

As the first condition ensures that there must be only one reaction in one side of the mapping to obtain one-to-many alignment, the second one indicates the uniqueness such that two mappings cannot contain same reactions. For instance, if the reaction $r_x \in \mathcal{P}$ aligns with the reactions $r'_x, r'_y, r'_z \in \mathcal{P}'$ for k equals to three, then the aligned reactions r_x, r'_x, r'_y, r'_z cannot be in other mapping of a legal alignment $\mathcal{A}$.

In the second place, we need to define the quality of the alignment problem for metabolic pathways. Generally, the definition of the alignment is the similarity measure that includes both homological and topological similarities. The homological similarity of the alignment is defined as a sum of all sequence-based similarity scores of all mappings. When the problem is about proteins, only amino acid sequence similarities are considered, but when the subject is the metabolic pathways, the computation of the homological similarity becomes more complex due to compounds and enzymes. Thus, for the mapping (R_x, R'_x), the homological similarity is computed due to input compounds, output compounds and enzymes of R_x and R'_x. In this study, we use the homological similarty scores that are produced by SubMap [27]. First of all, the enzyme sets E_x, E'_x are produced by unifying the enzymes of the reactions that are in the reaction subsets R_x and R'_x, respectively. The computation of the enzymatic homology score between the enzyme sets E_x, E'_x is calculated by creating a bipartite graph such that the first partition of the graph corresponds to the enzymes in the enzyme set E_x and the other partition corresponds to the enzymes in the enzyme set E'_x. An edge is added between every enzyme that belong to different enzyme sets and a similarity score is assigned to that edge as the weight. Afterwards, total homological score is obtained for E_x, E'_x by making the maximum weight bipartite matching on the bipartite

graph. Similar computations can be made for the unions of the input compounds I_x, I'_x and the unions of the output compounds O_x, O'_x corresponding to R_x, R'_x respectively. Totally, the homological similarity score of (R_x, R'_x) is a convex combination of the scores that are computed independently for input compounds, output compounds and enzymes. On the other side, the topological similarity of the alignment is defined as a sum of all conservation-based similarity scores of all mappings. For the mappings $(R_x, R'_x) \in \mathcal{A}$ and $(R_y, R'_y) \in \mathcal{A}$, the score is computed based on the conserved edge numbers. If there exists an edge from a reaction in R_x to a reaction in R_y and an edge from a reaction in R'_x to a reaction in R'_y, or vice versa, then it is accepted that there is a conserved edge between the mappings (R_x, R'_x) and (R_y, R'_y). Totally, the topological similarity is defined as a score that is proportional to total conserved edge number. When both homological and topological similarity scores are obtained, the network alignment problem becomes a problem that maximizes the convex combination of these scores.

2.2. Constrained Alignment Framework

In this subsection, we give a formal definition for our constrained alignment framework within one-to-many metabolic pathway alignment. We propose a constrained alignment framework that aims to maximize only topological similarity while satisfying some constraints on homological similarity, rather than maximizing the convex combination of homological and topological similarities.

For a metabolic pathway representation $G_p = (V_p, E_p)$, the k^{th} extension of G_p is denoted by G_p^k that is the directed edge-weighted graph and each node u_{R_x} in G_p^k corresponds to a reaction subset $R_x \in R_k$. If there is an edge from u_{r_i} to u_{r_j} in G_p, a directed edge (u_{R_x}, u_{R_y}) is added in G_p^k, where $r_i \in R_x$ and $r_j \in R_y$. At this point, the weight $w(u_{R_x}, u_{R_y})$ is assigned as the total number of such edges. Surely, the same definition can be used for ${G'_p}^k$. Let $Cons(u_{R_x})$ which is the subset of possible nodes that the node u_{R_x} can mapped to,

denote the constraints set of u_{R_x} in G_p^k. Similarly, this definition can be used for the nodes of ${G'_p}^k$. Hereby, there is a symmetry such that $u_{R'_y} \in Cons(u_{R_x})$, if and only if $u_{R_x} \in Cons(u_{R'_y})$ depending on $|Cons(u_{R_x})| \leq k_1$ and $|Cons(u_{R'_y})| \leq k_2$ for any nodes $u_{R_x} \in G_p^k$ and $u_{R'_y} \in {G'_p}^k$ and fixed constants k_1 and k_2, respectively. A bipartite *similarity* graph may be used to represent all constraints such that while the first partition of the graph consists of G_p^k nodes and the second partition consists of the nodes of ${G'_p}^k$, the edges correspond to the constraints of the nodes. At this point, the constraint alignment problem turns into a problem that aims to find the subset of the constraints. When the bipartite similarity graph is considered, the problem corresponds to find the subset of edges in that graph such that the subset provides a legal alignment and also maximum number of conserved edges in the result alignment. It is important to emphasize that the constrained alignment definition has been given in the previous study [34] for the global one-to-one alignment of protein-protein interaction (PPI) networks. In this sense, our constrained alignment framework is more general consisting of the previous model completely and can be used for the alignment of undirected PPI networks while the previous model may not be used for some instances. For a given two nodes u_{R_x}, u_{R_y}, if $Cons(u_{R_x}) \cap Cons(u_{R_y}) \neq \emptyset$, then the previous model applies $Cons(u_{R_x}) = Cons(u_{R_y})$. There is a restriction in the case where $Cons$ reflects high homological similarity such that some pairs that are homological similar are missed or long homologically similar chains of nodes are created incorrectly.

We first need to clarify that for a very restricted case, the constrained alignment problem is computationally intractable.

Proposition 2.2.1. *The constrained alignment problem where $k = k_1 = 1$ and $k_2 = 3$ is NP-Complete.*

Proof. In order to provide integrity, the proof is given in Chapter 3.

□

Secondly, we clarify the point that the computationally intractable starts dissolving for better understanding of the constrained alignment framework.

Proposition 2.2.2. *The constrained alignment problem where $k = k_1 = 1$ and k_2 any positive integer constant, is polynomially solvable if one of the directed graphs G_p or G'_p is acyclic.*

Proof. In order to provide integrity, the proof is given in Chapter 3.

□

2.3. CAMPways Algorithm

Even though Proposition 2.2.2 provides an affirmative perspective, there is a restriction in the usage. Even if our constrained alignment algorithm provides high quality alignments, it may not give optimum results in some cases. Our algorithm consists of three major steps assuming $G_p^k, G_p'^k$, the constants k_1, k_2 and the homological similarity score of $(u_{R_x}, u_{R'_y})$ is given where u_{R_x} and $u_{R'_y}$ are any nodes in $G_p^k, G_p'^k$, respectively. These major steps are shown in Figure 2.1 on a metabolic pathway pair. The details are explained in the next subsections.

2.3.1. Constructing Bipartite Similarity Graph

In the first step, $Cons(u_{R_x})$ and $Cons(u_{R'_y})$ are created for every node u_{R_x} in G_p^k and $u_{R'_y}$ in $G_p'^k$ where $|Cons(u_{R_x})| \leq k_1$ and $|Cons(u_{R'_y})| \leq k_2$. Let we have an edge-weighted bipartite graph where the first partition corresponds to the nodes of G_p^k, the other partition corresponds to the nodes of $G_p'^k$ and also, an edge between two nodes includes the homological score of these nodes. At this point, finding a subset of edges that provides the

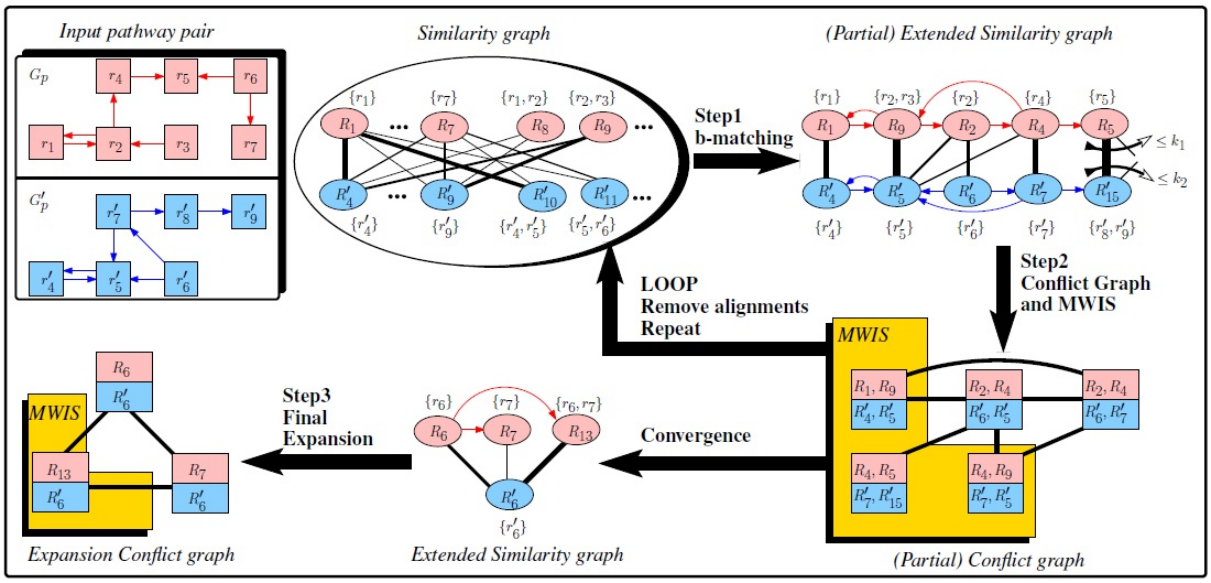

Figure 2.1. CAMPways algorithm

degree constraints k_1 and k_2 and also, maximizes the total weight of the edges in that subset is the major goal of the algorithm. In Figure 2.1, the thickness of the edges represents the weight of the edges such that the thickest edge corresponds to the highest score.

The major goal of the algorithm turns the problem into *b-matching* or *degree constrained subgraph problem* that have been studied in the previous works [35] such that polynomial time solutions, network-flow algorithms and also, belief propagation methods have been suggested [36, 37]. Nevertheless, instead of using them, we prefer to use a simple greedy algorithm to provide the efficiency. The greedy algorithm selects the heaviest edge in the bipartite graph considering the degree constraints k_1 and k_2 for both end points and the output edge set. When there exists no edges that are appropriate for selecting due to edge weight and degree constraints k_1 and k_2, the algorithm stops and we have a bipartite graph that consists of the selecting edges and nodes that are connected by these edges. Afterwards, we called the obtained bipartite graph as the similarity graph, S.

2.3.2. Conflict Graph Generation and Conflict Resolution

Let we assume that the bipartite similarity graph S is extended by the directed edges of G_p^k and $G_p'^k$ due to a restriction such that if there exists an edge (u_{R_x}, u_{R_y}) in the graph G_p^k, then the edge (u_{R_x}, u_{R_y}) is added to the similarity graph. Of course, same restriction is valid for the $G_p'^k$, as well. After the extension of the similarity graph, an undirected node-weighted conflict graph is created where the nodes of that graph corresponds to a set of four nodes that provides conserved edges in the similarity graph S. Actually, a node that corresponds to a $4-tuple \prec u_{R_x}, u_{R_y}, u_{R'_x}, u_{R'_y} \succ$ is added to the conflict graph if and only if the following are satisfied:

1. $R_x \cap R_y = \emptyset$ and $R'_x \cap R'_y = \emptyset$.
2. Either $(u_{R_x}, u_{R_y}), (u_{R'_x}, u_{R'_y})$ are in $G_p^k, G_p'^k$ respectively,or $(u_{R_y}, u_{R_x}),(u_{R'_y}, u_{R'_x})$ are in $G_p^k, G_p'^k$ respectively.
3. $\{u_{R_x}, u_{R'_x}\}, \{u_{R_y}, u_{R'_y}\}$ are undirected edges in S.

For each node that corresponds to a 4-tuple in conflict graph is called as c_4 and a score is assigned as a weight to every c_4 such that while the score is 1 if only the first part of the second condition is satisfied, 2 is assigned as the score if all parts of the second condition is provided. At this point, it is possible to see that every node c_4 of the conflict graph corresponds to a pair of reaction subset mappings and provides at least one conserved edge in the output alignment set. Furthermore, the weight of a c_4 represents the conserved edge number that is provided by that node. In figure 2.1, the exact conflict graph that is obtained from the partial similarity bipartite graph is showed. Whereas the $4-tuple \prec u_{R_9}, u_{R_2}, u_{R'_5}, u_{R'_6} \succ$ may denote a c_4, it does not happen due to condition *1* such that the reaction subsets of R_9 and R_2 share the common reaction r_2 in the partially extended similarity graph as shown in figure 2.1. Also, when examining the weight of the c_4s, it is possible to understand that the weight of the c_4 corresponding to the $4-tuple \prec u_{R_1}, u_{R_9}, u_{R'_4}, u_{R'_5} \succ$ is two while other c_4s

weight are one according to the same figure.

Let the conflict nodes C_1, C_2 correspond to the $4-tuples \prec u_{R_x}, u_{R_y}, u_{R'_x}, u_{R'_y} \succ$ and $\prec u_{R_w}, u_{R_z}, u_{R'_w}, u_{R'_z} \succ$, respectively. Moreover, let S_1, S_2 be the elements of $\{R_x, R_y\}$, $\{R_w, R_z\}$ and S'_1, S'_2 be the elements of $\{R'_x, R'_y\}$, $\{R'_w, R'_z\}$, respectively. For a c_4 node C_i, let $M_{C_i}(u)$ denotes the neighbour of u in C_i from the opposite network. In this case, an edge is added between two c_4 nodes if and only if at least one of the following satisfied:

1. $\exists S_1, S_2$ such that $S_1 \neq S_2$ and $S_1 \cap S_2 \neq \emptyset$.
2. $\exists S'_1, S'_2$ such that $S'_1 \neq S'_2$ and $S'_1 \cap S'_2 \neq \emptyset$.
3. $\exists S_1, S_2$ such that $S_1 = S_2$ and $M_{C_1}(S_1) \neq M_{C_2}(S_2)$.
4. $\exists S'_1, S'_2$ such that $S'_1 = S'_2$ and $M_{C_1}(S'_1) \neq M_{C_2}(S'_2)$.

Totally, these conditions indicate that there exists an edge between two c_4s in the conflict graph such that the conserved edges corresponding to these c_4 nodes cannot coexist in a legal alignment set. For instance, the edge is added between the c_4 nodes corresponding to the $4-tuples \prec u_{R_1}, u_{R_9}, u_{R'_4}, u_{R'_5} \succ$ and $\prec u_{R_2}, u_{R_4}, u_{R'_6}, u_{R'_7} \succ$ in the conflict graph due to condition *1* such that the reaction subsets R_9 and R_2 share a common reaction. Accordingly, there is no legal alignment that consists of both these c_4s due to shared reactions. On the other side, the edge is added between the c_4 nodes corresponding to the $4-tuples \prec u_{R_4}, u_{R_5}, u_{R'_7}, u_{R'_1 5} \succ$ and $\prec u_{R_2}, u_{R_4}, u_{R'_6}, u_{R'_5} \succ$ due to condition *3* such that whereas the reaction subset R_4 matches with the reaction subset R'_7 in one c_4, in the other one, it matches with the reaction R'_5 and matching between different reaction subsets is not allowed to be in the legal alignment set at the same time. In addition, these conditions and the construction of the conflict graph supports the following proposition:

Proposition 2.3.1. *The maximum weight independent set (MWIS) of C provides an optimum solution to the constrained alignment problem.*

Before the maximum weight independent set solution, we need some modifications to make our conflict graph model more useful in the framework. Firstly, in order to increase the quality of the alignment, we propose two weighting formulas for the conflict graph nodes. Let $w_s(e)$ be the weight of the edge e in the similarity graph S such that this weight indicates the homological score between the reaction subsets corresponding to the end points of the edge e. The first weighting scheme is denoted by W_1 that equals to α x $H(C_I)+(1-\alpha)$ x $I(C_I)$ where C_I corresponds to the conflict node that represents the $4-tuple \prec u_{R_x}, u_{R_y}, u_{R'_x}, u_{R'_y} \succ$ and $H(C_I), I(C_I)$ correspond to the following:

$$H(C_1) = \frac{1}{2} \times (w_S(u_{R_x}, u_{R'_x}) + w_S(u_{R_y}, u_{R'_y}))$$

$$I(C_1) = \frac{1}{2(k^2+1)} \times \sum_{\substack{i,j \in \{u_{R_x}, u_{R_y}\}, i \neq j \\ i',j' \in \{u_{R'_x}, u_{R'_y}\}, i' \neq j'}} w(i,j) + w(i',j')$$

In order to calculate $I(C_1)$, the total number of directed edges that are between R_x, R_y and between R'_x, R'_y is normalized with the maximum number of possible directed edges in any conflict node c_4. The parameter α is a balance parameter such that it balances the relationship between homological similarity score and topological similarity score. On the other side, our second weighting scheme that is denoted by W_2 does not check the conserved edge number between the reaction subsets due to knowledge of providing at least one conserved edge by each c_4. Furthermore, depending on the evolutionary distances between the organisms that provide input pathways for our algorithm, differentiating between one-to-many

alignments and one-to-few alignments is more meaningful. We use additional parameters $a_1, a_2..., a_k$ in second weighting scheme W_2 in order to make a such differentiation such that $a_1 + a_2 + ... + a_k = 1$ and each a_i corresponds to importance of one-to-i mappings in the total alignment. Thus, for the node $C_1 = \prec u_{R_x}, u_{R_y}, u_{R'_x}, u_{R'_y} \succ$, W_2 is calculated as $a_{|R_x|}$ x $|R_x| + a_{|R_y|}$ x $|R_y|$ where $|R_x| \geq |R'_x|$ and $|R_y| \geq |R'_y|$.

After the construction of the conflict graph, second important issue is solving that conflict graph which means solving the maximum weight independent set (MWIS) problem on the conflict graph and obtaining the maximum number of conserved edges. In general, the maximum weight independent set problem is in NP-Complete problem set [38]. In order to solve MWIS problem, several greedy heuristic algorithms have been proposed [39]. We implement and test the performance of all greedy heuristic algorithms and decide on GWMIN2 algorithm that gives best results for our algorithm. GWMIN2 algorithm, firstly, selects a node u in the conflict graph C such that the node u maximizes the score of $\mathcal{W}(u)/\sum_{v \in N_{\mathcal{C}}^+(u)} \mathcal{W}(v)$ where $N_c^+(u)$ denotes the node u and all neighbors of it. This process goes on until there is no node in the conflict graph. Besides, the algorithm provides a theoritical guarantee such that the weight of the output independent set is at least $\sum_{u \in V_{\mathcal{C}}} [\mathcal{W}(u)^2 / \sum_{v \in N_{\mathcal{C}}^+(u)} \mathcal{W}(v)]$ where V_c denotes the vertex set of the conflict graph C. Depending on the results of our performance tests and the theoreticall guarantee of GWMIN2 algorithm, we prefer to use that algorithm to solve conflict graph.

Consequently, it is possible to see that we find a mapping set that consists of the edges in the bipartite similarity graph S based on the process of the Step 1 and also depending on the constraints k_1, k_2, our mapping set is limited. Obviously, extending the mapping set increases the meaningful results. In order to extend the alignment set, firstly we restore all homological edges and we remove the mapped nodes from G_p^k,$G_p'^k$ after the steps 1 and 2 are over and afterwards, we repeat the steps 1 and 2. The loops go on until the conflict graph C produce empty set. For the sample input pathway pair in Figure 2.1, the loop iterates

only once such that after the step 1 and 2 works once, remaining extended similarity graph consists of the nodes R_6, R_7, R_13, R'_6 and no conflict graph is produced by these nodes.

2.3.3. Final Alignment Expansion

Step 1 and Step 2 produce mappings based on the maximization of the conserved edge number and depending on the loops, it is possible to see that after the loops are over, the algorithm cannot produce more conserved edges anymore. But, still there may exist potential matchings that have high homological scores and these may be added the output alignment set. In order to provide such an extension, we restore all homological similarity edges and remove all matched nodes from the graphs G_p^k,$G_p'^k$. At this point, we create a new conflict graph that is conceptually different from the conflict graph which is produced in Step 2, based on the remaining bipartite similarity graph S. The conflict graph is called *expansion conflict graph* and each node in that graph corresponds to a $2-tuple \prec u_{R_x}, u_{R'_x} \succ$ where $\{u_{R_x}, u_{R'_x}\}$ is an edge in the remaining bipartite similarity graph S. An edge is added between two nodes in the expansion conflict graph if and only if the intersection of the reaction subsets which belong to the same pathway is not empty. The expansion conflict graph construction is shown in Figure 2.1. After the construction of the expansion conflict graph, GWMIN2 algorithm is used to solve conflicts on that graph as same as in Step 2 and finally, the output matching of GWMIN2 algorithm are added in the output alignment set.

2.4. Extension of Constrained Alignment Framework and CAMPways Algorithm

In this section, we extend the constrained alignment framework and CAMPways algorithm for one-to-one pairwise protein-protein interaction network alignment by making the necessary changes and additions in order to get reasonable and useful results. We give the problem definition for this problem and define major steps of CAPPI algorithm.

2.4.1. Problem Definition for PPI Network Alignment

Let simple undirected graphs $G_1(V_1,E_1)$ and $G_2(V_2,E_2)$ be the input PPI networks where V_1,V_2 denote the sets of nodes corresponding to the proteins and E_1,E_2 denote the sets of edges corresponding to the interactions, respectively. Moreover, let undirected edge-weighted bipartite graph S be the similarity graph where the partitions of S are V_1,V_2 and each edge (u, u') in S has a positive real weight $w(u, u')$. In many studies, the weight is a sequence similarity score $w(u, u')$ that is usually obtained by using BLAST between sequences of u and u', where $u \in V_1$ and $u' \in V_2$. BLAST bit score is the most preferred score that is a log-scaled score and indicates biological relevance of a finding. But, when you compare the sequences of different species by using BLAST, you may not obtain all pairwise scores such that some pairwise scores are found as zero. So, the number of scored sequences which are taken as input may not be sufficient in some cases in order to get remarkable results. According to Aladağ and Erten [28], *"most of global network alignment algorithms can be viewed to proceed in two phases. For each pair $u_i \in V_1$,$v_j \in V_2$, an estimate confidence score is sought at an initial coarse-grained phase. The score represents the level of confidence that the match (u_i, v_j) is in the optimum alignment maximizing the global score. This is usually followed by a fine-grained phase that consists of refining an initial global alignment based on the estimate scores attained in the previous phase"*. Correspondingly, we prefer to use *estimate confidence scores* instead of BLAST bit scores and in this case, we obtain some advantages such as increase in the number of scored sequences and decrease in the running time. Thus, formally, the weight $w(u, u')$ of each edge (u, u') is the *estimate confidence score* that is produced in SPINAL coarse-grained phase in our study.

Hereby, we need to give a definition for the legal alignment such that the definition is simpler than the problem definition of metabolic pathways within one-to-many alignment perspective. Because we focus on only one-to-one mappings, the connected subsets that are employed in the metabolic pathway alignment problem are not considered. Thus, in a simple

way, a legal alignment $\mathcal{A}$ occurs between G_1 and G_2 if for any matched pairs $(u, u') \in \mathcal{A}$ and $(v, v') \in \mathcal{A}$, $u \neq v$ and $u' \neq v'$. This condition implies the uniqueness of the output alignment such that each node in G_1 can match with only one node in G_2 and vice versa. Afterwards, the important point is the quality of the alignment. As it is mentioned in the previous sections, the quality of the alignment corresponds to the similarity measure in terms of both homological and topological similarities. Because the subject is PPI networks, while we use estimate confidence score that is mentioned before as homological similarity score, we give the definition of the topological similarity score as in the problem definition of metabolic pathways in terms of conserved edge number. There exists a conserved edge for any matched pairs $(u, u'), (v, v')$ where u, v in V_1 and u', v' in V_2, if there is an undirected edge (u, v) in G_1 and an undirected edge (u', v') in G_2. Consequently, in a similar way, the major goal of the PPI network alignment is maximization of homological and topological scores.

2.4.2. CAPPI Algorithm

As it is mentioned before, because the major goal consists of both homological and topological similarities, we propose an algorithm that balances these scores with a parameter. While high-valued parameters handle the problem within conserved edge maximization, low-valued parameters give alignment results based on better biological meaning. As both versions are explained in same sections, in general, CAPPI algorithm consists of four main steps assuming G_1, G_2, S, the constants $k_1, k_2, \alpha, f, b, i$ and the homological similarity score (estimate confidence score) $w(u, u')$ is given where u and u' are any nodes in G_1 and G_2, respectively. The details are given in the next sections.

2.4.3. Finding Maximum Weight Bipartite Matching

Because the general framework is based on the conserved edge maximization, while obtaining the conserved edges, some maximum homologically weighted pairs may be missed

if they don't provide any conserved edge or they may not be selected due to conflict status. In order to handle this case, we employ a maximum weight bipartite matching (MWBM) on the bipartite similarity graph S. Let b be a parameter that is used to define the number of matchings that are taken from the alignment set of maximum weight bipartite matching such that the first b maximum weighted pairs are taken from the alignment set of MWBM and added to the actual alignment output set of CAPPI algorithm. Afterwards, the nodes and edges that are in the selected pairs are removed from G_1, G_2 and S. Next steps of $CAPPI$ algorithm go on the remaining graphs. These processes are based on the value of the constant f such that if f equals to one, then finding maximum weight bipartite matching step is employed but when the value of f equals to zero, this step is not performed and the original graphs G_1, G_2 and S are used in the next sections.

2.4.4. Constructing Reduced Bipartite Similarity Graph

Initially, we assume that we have a bipartite similarity graph such that the first partition nodes correspond to the nodes of G_1 and second one includes the nodes correspond to the nodes of G_2. The edges that are between two partitions have estimate confidence scores. But we change these scores according to the goal of the algorithm. When the goal is finding more conserved edges, high-valued α parameter is used. However when the goal focuses on biological meaning, low-valued α parameter is preferred. Thus, the score is based on both homological and topological score and the score equals to $\alpha \times min(|E_u|, |E'_u|) + (1-\alpha) \times w(u, u')$ for any node u in G_1 and any node u' in G_2. In this formula, while $w(u, u')$ denotes the estimate confidence score between the nodes u, u', $|E_u|$ and $|E'_u|$ denotes the number of edges of u and u' in the original graphs G_1 and G_2, respectively. We take the minimum number of edges and it is possible to see that the minimum number of edges indicates the possible conserved edge number for a node pair and if all edges are legal in the conflict graph, then the pair gives maximum $min(|E_u|, |E'_u|)$ conserved edges. Also, it is possible to understand that α is a balance parameter between the homological and topological scores.

In order to explain this step, we need to give the constrained definition for this problem depending on the constrained alignment framework. In a similar way, $Cons(u)$ denotes the constraints set of u in G_1 and includes the possible nodes that the node u can mapped to. Of course, the same definition can be used for the nodes of G_2. The same symmetry that is mentioned in the constrained framework can be used for this problem as well such that $u' \in Cons(u)$ if and only if $u \in Cons(u')$ depending on $|Cons(u)| \leq k_1$ and $|Cons(u')| \leq k_2$ for any nodes u in G_1 and u' in G_2 and fixed constants k_1, k_2.

This step reduces the original bipartite similarity graph based on these constraints such that all constraints $Cons(u), Cons(u')$ are created for every node u of G_1 and u' of G_2 where $|Cons(u)| \leq k_1$ and $|Cons(u')| \leq k_2$. In the fact, the problem is to find an edge subset that maximizes the sum of edge weights by providing the constraints k_1 and k_2. In order to solve this problem, we use the same greedy algorithm that is used for metabolic pathways alignment in CAMPways algorithm and obtain reduced bipartite similarity graph.

2.4.5. Conflict Graph Generation and Conflict Resolution

In general concept, conflict graph generation and conflict resolution is same as CAMPways algorithm. But, in order to get better results we make some changes in this step. Let the reduced bipartite similarity graph be extended with the edges of G_1 and G_2 and afterwards, an undirected node-weighted conflict graph is created such that each node in the conflict graph corresponds a 4-tuple $\prec u, u', v, v' \succ$ and is denoted as c_4, as well. In detail, the node that corresponds to 4-tuple $\prec u, u', v, v' \succ$ is added to the conflict graph if and only if the following are satisfied:

1. $u \neq v$ and $u' \neq v'$.
2. The undirected edge (u, v) is in G_1 and the undirected edge (u', v') is in G_2.
3. $\{u, u'\}$ and $\{v, v'\}$ are undirected edges in S.

In the conflict graph, a weight is assigned to each node such that the weight of c_4 that corresponds to the 4-tuple $\prec u, u', v, v' \succ$ equals to the following:

$$W(c_4) = (\frac{1}{2} \times (w(u, u') + w(v, v')))^{|e|}$$

where $|e|$ denotes the number of possible conserved edges such that it is possible to see that each c_4 denotes one conserved edge and it is important to check that if that conflict node is selected in resolution phase, what is the contribution of that node to the conserved edge number in the output alignment. Thus, the number of possible conserved edges $|e_p|$ that are contribution of the conflict node to the output alignment is added to one and $|e| = 1 + |e_p|$. It is clear that whereas in the first loop, that score is only one, but in the next loops the score is changed due to output alignment set that is provided by conflict resolution.

In this step, the second important issue is to add edges between the conflict nodes. Let C_1, C_2 be two conflict nodes corresponding to 4-tuples $\prec u, u', v, v' \succ$ and $\prec w, w', z, z' \succ$, respectively. Let S_1, S_2 and S_1', S_2' be the unions of the nodes $\{u, v\}$,$\{w, z\}$ and $\{u', v'\}$,$\{w', z'\}$, respectively. Furthermore, for a c_4 node C_i, let $M_{C_i}(u)$ denotes the neighbor of u in C_i from the opposite network. In this case, the condition of adding an edge is same as in CAMPways algorithm. Thus, the conditions are not given again in order to prevent tautology.

After the construction of the conflict graph, we need to solve the conflicts in an optimum way. Similarly, we use GWMIN2 algorithm that is used in CAMPways algorithm within same definitions. However, because this algorithm is heuristic, we extend it with some modifications. Without loss of generality, we use the swap idea in the algorithm such that the impact of that idea is negligible on the running time and it helps to increase the size of the alignment set that is provided by GWMIN2. The swap idea have been used in both the alignment problems and bioinformatics studies in order to get better results [40, 41]. At this point, we use a simple swap process such that after GWMIN2 is completed, we try to swap the nodes that are in the alignment set with the nodes in the conflict graph that are legal for being in that set. The swap iteration starts from the first node in the alignment set, removes this node from the set and finds the legal nodes that are not conflict with the nodes in the remaining alignment set. Afterwards it compares the score of the node in the alignment set with the total score of legal nodes. If the total score is greater than the score of the node in the alignment set, then it swaps these nodes. The iteration goes on until all nodes are checked in the alignment set.

Obviously, the alignment set that is provided by GWMIN2 includes the node pairs based on the reduced bipartite similarity graph. Still, for the original bipartite similarity graph, there may exist some matching that are created conflict graphs. Thus, in order to extend the alignment set and obtain possible matching based on the conflict graphs, we restore the bipartite similarity graph and remove the nodes that are in the alignment set of GWMIN2 from the similarity graph. Afterwards, we repeat step 2 and 3 until the bipartite similarity graph does not produce any conflict graph. The constant i defines the number of such iterations. When the goal is maximization of the conserved edges, then the value of constant i is higher and in that time, we observe that the homological score decreases depending on the natural concept of the framework such that when the iterations are employed, while the conserved edge number increases, the biological meaning decreases. Thus when the better results in terms of biological meaning are aimed, then second and

third steps are employed only once.

2.4.6. Final Alignment Expansion

Final Alignment Expansion parts of CAMPways and CAPPI are completely different. While CAMPways try to find conflicts due to 2-tuples, CAPPI uses maximum weight bipartite matching such as in step 1 of CAPPI algorithm. However, depending on the constant f, maximum weight bipartite matching algorithm uses different values. As it is mentioned before, when the constant f equals to zero, the algorithm aims to find more conserved edges. Thus, when the iterations are over, the edge weights of remaining bipartite similarity graph are changed depending on the aim. For each pair of the remaining bipartite similarity graph, the conserved edge contribution number is calculated such as in conflict graph generation step. The possible conserved edge number that is provided by the pair if it is selected for being in the alignment graph is assigned as a weight to the considered edge. Afterwards, the maximum weight bipartite matching algorithm is used on the remaining bipartite similarity graph within these scores. The alignment set that is produced by that algorithm is added to the actual alignment set. However, when the constant f equals to one, the goal is maximization of the biological meaning. Thus, the edge weights of remaining bipartite graph are selected as estimate confidence scores and similarly, the alignment set that is produced by that algorithm is added to the actual alignment set.

3. COMPLEXITY ANALYSIS

3.1. NP-Hardness Proof of Constrained Alignment Problem

Proposition 3.1.1. The constrained alignment problem where $k = k_1 = 1$ and $k_2 = 3$ is NP-Complete.

Proof. As it is defined previously, the problem refers to one-to-one alignment between the nodes of G_p and G'_p in case k equals to one. In addition, the constraints that express $k_1 = 1$ and $k_2 = 3$ mean that each node of G_p can be aligned with one node of G'_p and on the other side, each node of G'_p can be aligned with one of at most 3 nodes of G_p.

Because of a problem x in NP-Complete is also in both NP and NP-Hard, we need to handle the proof from both directions. So, under these considerations, according to general proof strategy, we first need to show that the problem is in NP by giving an efficient certification. Hereby, the set of mappings between the nodes of G_p and G'_p gives the certification and shows that the problem is in NP which means the problem is a decision problem within yes or no answers that *yes* answers can be proved in polynomial time. For this problem, *yes* answers correspond to checking whether the provided alignment is legal or not within all these considerations and whether is it giving at least a fixed number f of conserved edges or not. In order to show NP-Hardness of the problem, we use reduction from *Monotone 1in3SAT* that is a restricted version of 3SAT such that while every clause has exactly three literals and exactly one of them is true, no negations in the clauses are allowed. Whereas the reduction is based on the undirected graphs, it can be adapted to directed graphs as well. In order to generalize the reduction for directed graphs, we make each edge of G_p and G'_p bidirectional.

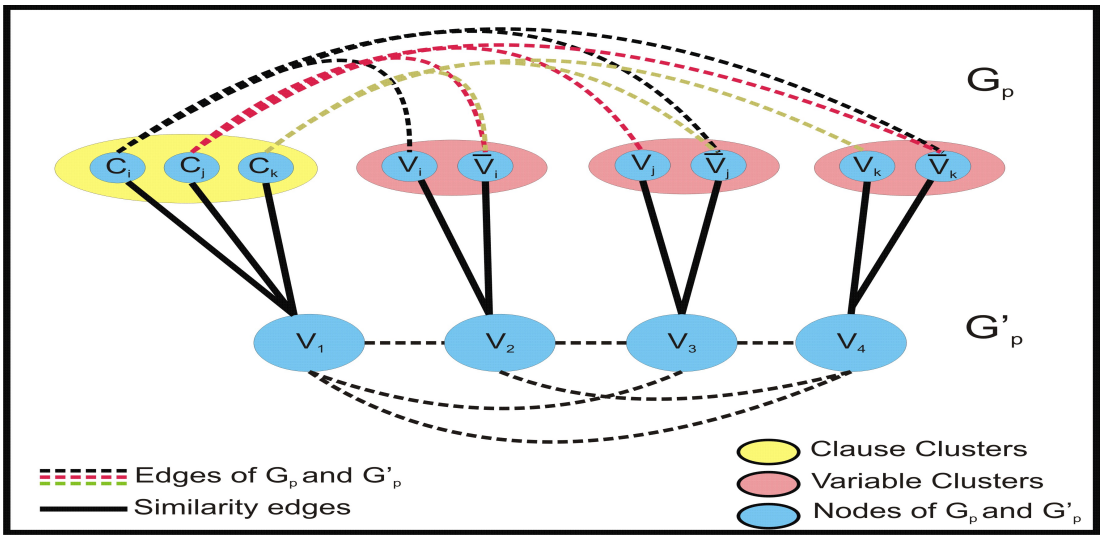

Figure 3.1. NP-Hardness Proof Graph

According to the reduction idea, firstly, we need to create graphs that represent the variables and clauses. Thus, we start by creating G_p. A *clause cluster* is created for each clause $(x_i \vee x_j \vee x_k)$ in a given Monotone 1in3SAT instance ϕ where the nodes c_i, c_j, c_k of the cluster correspond to x_i, x_j, x_k in the clause. Furthermore, a *variable cluster* is created for each variable x_t in ϕ where the nodes $v_t, \bar{v_t}$ correspond to $x_t, \bar{x}_t$. Each node c_i in a clause cluster is connected to three nodes $v_i, \bar{v_j}, \bar{v_k}$ in variable clusters. Thus, G_p becomes a bipartite graph where one partition consists of clause clusters and the other partition consists of variable clusters. Creating G'_p is simpler than G_p such that the nodes are created corresponding to clause and variable clusters of G_p. The edges are added between all possible node pairs and a complete graph is obtained. Eventually, in order to represent the similarity edges, we add an edge between a node of G'_p and its corresponding clusters in G_p. The figure 3.1 illustrates the graph definitions.

Hereby, our claim is that there exists a valid satisfying Monotone 1in3SAT assignment of variables in ϕ if and only if the global alignment score is at least $f = 3|C|$ such that $|C|$ refers to the number of clauses in ϕ. According to graph definitions and Figure 3.1, it is possible to see G'_p refers to the maximization of the number of conserved edges in

the alignment. In this manner, the problem becomes selecting exactly one node from each cluster in G_p such that the problem supports Monotone 1in3SAT restrictions and makes maximum the number of edges in the induced subgraph of G_p. After this point, G'_p becomes negligible and we focus on a new problem that is defined on G_p: Assume that there is a valid satisfying assignment A_ϕ for variables of ϕ instance. We select v_t if x_t is assigned as true in A_ϕ but otherwise $\bar{v_t}$ is selected for the variable cluster corresponding to x_t variable of G_p. For the clause cluster corresponding to $(x_i \vee x_j \vee x_k)$ in G_p, we select only the one that corresponds to true literal in A_ϕ. Let x_i be the only true literal in $(x_i \vee x_j \vee x_k)$. As it is mentioned before, we focus on v_i, $\bar{v_j}$ and $\bar{v_k}$ for x_i. Since the nodes v_i, $\bar{v_j}$, $\bar{v_k}$ are exactly selected from their respective variable clusters, for each clause cluster, exactly three edges are in the induced subgraph on all selected nodes and totally, $3|C|$ edges are obtained as shown in Figure 3.1. For the reverse direction, we can handle this case by selecting the nodes from each cluster such that the induced subgraph on all selected nodes contains at least $3|C|$ edges. Thus, we give a new definition for a valid satisfying assignment A_ϕ: x_t is assigned to true if v_t is selected from the variable cluster corresponding to x_t ; otherwise false. Let only c_i be selected from the clause cluster corresponding to clause $(x_i \vee x_j \vee x_k)$. At this point, we need to show that v_i, $\bar{v_j}$ and $\bar{v_k}$ must be selected nodes from the variable clusters corresponding to x_i,x_j and x_k, respectively. As shown in Figure 3.1, since every node in a clause cluster has three edges to variable cluster nodes and only one node is selected from each cluster, we can say that there must be $3|C|$ edges in the induced subgraph on all selected nodes. Consequently, if the node c_i is selected from the clause cluster, then the nodes v_i, $\bar{v_j}$, $\bar{v_k}$ are selected from variable clusters corresponding to x_i,x_j and x_k and a valid satisfying assignment A_ϕ is provided such that there exists exactly one true literal in the each clause of the instance ϕ. It is possible to see that, this proof is valid for one-to-one alignment of protein interaction network within undirected graphs and the value of k in detail such that both algorithms involve this proof.

□

3.2. Polynomial Time Solution of the Alignment Problem

In this section, we prove that the simple pathway alignment problem within the constrained alignment framework is solvable in the polynomial time.

Proposition 3.2.1. The constrained alignment problem where $k = k_1 = 1$ and k_2 any positive integer constant, is polynomially solvable if one of the directed graphs G_p or G'_p is acyclic.

Proof. Let us begin the proof by creating a conservation graph such that each node C'_x in that graph corresponds to $u_{R'_x} \cup Cons(u_{R'_x})$ where $u_{R'_x} \in G'_p$. Let C'_x and C'_y be two nodes of the conservation graph. There exist a directed edge (C'_x, C'_y) if the edge from C'_x to C'_y is induced such that there is a directed edge $(u_{R'_x}, u_{R'_y})$ in the graph G'_p and there is a directed edge (u_{R_w}, u_{R_z}) in the graph G_p, as well where $u_{R_w} \in Cons(u_{R'_x})$ and $u_{R_x} \in Cons(u_{R'_y})$. At this point, the size of vertex set is $|V'_p|$ and the size of edge set is $O(|E'_p|)$ of the conservation graph, respectively. Thus, this size is polynomial due to the problem size. It is possible to verify that if the conservation graph includes a directed edge, both G_p and G'_p certainly contains a directed edge, as well. A cycle $C_{x'_1}, ..., C_{x'_t}, C_{x'_1}$ is possible in such as case: If there is a cycle $u_{R'_{x_1}}, ..., u_{R'_{x_t}}, u_{R'_{x_1}}$ in G'_p and at the same time if there is a cycle $u_{R_{x_1}}, ..., u_{R_{x_t}}, u_{R_{x_1}}$ in G_p where $u_{R_{x_1}} \in Cons(u_{R'_{x_1}})..., u_{R_{x_t}} \in Cons(u_{R'_{x_t}})$. When we assume that at least one of graphs G_p, G'_p is acyclic, the conservation graph must be acyclic, as well. Let T be the topological ordering of the conservation graph. A dynamic programming approach that traverses the nodes depending on T and calculates the score of each k_2 possible mappings that represents the conserved edge number based on the neighbors scores is used and finally, when it reaches the node whose out-degree 0 in T, it obtains the optimum matching for the last node. For the remaining nodes, the optimum matching is obtained by backtracking and traversing the nodes in the opposite direction of T. During both traversals, because the time that is spent for each node is polynomial, the total algorithm is completed in polynomial time. □

4. DISCUSSION OF RESULTS

In this section, first of all, we give our comparative experimental results on actual metabolic pathways that are taken from KEGG database such as SubMap [27]. The comparisons are made between CAMPways and SubMap because of the same problem definitions. In other words, both CAMPways and SubMap algorithms try to find one-to-many mappings for a pair of metabolic pathways. It must be known that a new version of SubMap is proposed based on the original version whose running time is decreased [42]. But, it is not possible to make comparison between CAMPways and new compressed version of SubMap because of the lack of publicly available implementation of that algorithm. According to Ay et al [42] reported results, the improvement on the running time causes %50 diminution on the accuracy. At this point, the accuracy is measured in terms of Pearson's correlation coefficient between the original version and compressed version of SubMap. On the other side, the experimental results show that our algorithm provides both running time efficiency and more accuracy without any loss on the time. Thus, the alignment results of CAMPways provide more accuracy than the results of original version of SubMap. Afterwards, we compare CAPPI algorithm with SPINAL, IsoRank [43] and MI-GRALL [44] algorithms in terms of conserved edge numbers and biological meaning. The experimental results on actual PPI networks show that CAPPI algorithm gives better results than those of other algorithms in general and finally, we give memory requirements and running time analysis of our algorithm.

4.1. Discussion of Results for CAMPways

Although KEGG database provides detailed metabolism categories such as *Glycerolipid metabolism* and *Tryptophan metabolism*, these pathways are not suitable for directly usage on the algorithm. The most important reason is the lack of the gold standard that is the

base of objective comparisons. On the other side, there is another issue such that the sizes of pathways are quite small and in such a case, it is hard to predict the behavior of the algorithm in order to obtain realistic results. Therefore, the mechanism that merges all pathways in the detailed metabolism categories is used to handle these problems. Also, these detailed metabolism categories are gathered under the more general categories. Depending on the first 11 high-level categories, we merge all pathways that are gathered under the same high-level categories and thus, we obtain more extended metabolic pathways. In this way, we have 11 metabolic networks such that each one corresponds to following, respectively: 1.1 Carbohydrate metabolism, 1.2 Energy metabolism, 1.3 Lipid metabolism, 1.4 Nucleotide metabolism, 1.5 Amino acid metabolism, 1.6 Metabolism of other amino acids, 1.7 Glycan biosynthesis and metabolism, 1.8 Metabolism of cofactors and vitamins, 1.9 Metabolism of terpenoids and polyketides, 1.10 Biosynthesis of other secondary metabolites, 1.11 Xenobiotics biodegradation and metabolism. The number of metabolic pathways changes between 2 and 15 in these extended metabolic networks. The experimental comparisons that are mentioned in this section are performed on these extended metabolic networks.

Next 2 subsections include the experimental comparisons between the output alignments of CAMPways and original version of SubMap based on their accuracy. For this purpose, we perform two accuracy experiments. While first one is based on reverse engineering successes of the output alignments, second one includes experiments on the functional group conversion categorization that are provided by KEGG database. Afterwards, the experiments consist of the comparisons on the running times of these algorithms.

4.1.1. Reverse Engineering Metabolic Pathways

The natural accuracy measure is the capacity of reverse engineering of output alignments. The matched reactions in the output alignment that belong to the same KEGG pathway provide higher quality. Thus, our gold standard is the pathways that are provided

by detailed metabolism categories in KEGG. In this sense, it is important to remember that we employ the algorithms on the general pathways that are created by merging the detailed metabolism categories. But hereby, we assume that the metabolic pathways are noise-free which means the pathways are completely valid and there is no missing data or poorly designed pathway in that database. Let X and X' be two organisms and G_x, G'_x be their metabolic networks, respectively such that these networks correspond to the metabolism 1.m that are defined above. Moreover, let $\prec u_{R_x}, u_{R'_x} \succ$ be the mapping in the alignment of G_x, G'_x. Hereby, while R_x corresponds to the reaction subset of X, R'_x corresponds to the reaction subset of X', similarly and let $R_x = \{r_x\}$ such that it is the subset which contains only one reaction in the one-to-many alignment. Additionally, let $P_1, ..., P_x$ represent the pathways that includes the reaction r_x. These pathways belong to the metabolism 1.m of the species X. Then, a mapping is called correct if all reactions in R'_x is included in at least one of the pathways $P'_1, ..., P'_x$ such that a pathway P'_i is a pathway that is in the metabolism 1.m of the species X' and corresponds to the pathway P_i of the species X. Within this knowledge, we make two different experiments such that the experiments are performed between same-domain species and between across-domain species. We select *H.Sapiens (hsa)* and *M.Musculus (mmu)* as the representative species for eukaryote domain and *A.tumefaciens(atc)* and *E.coli (eco)* as the representative species for bacteria. Also, k value is fixed as 3 which means only one reaction of the one network match with at most three reactions of the other network. Furthermore, for CAMPways algorithm $k_1 = k_2 = 3$ is selected.

4.1.1.1. Same-domain Alignments. The experimental results of output alignments of hsa-mmu and atc-eco pairs based on 11 metabolic networks that are mentioned before are given in Table 4.1.1. In the table, while Total Reactions column represents the number of total reactions of the network pair, Coverage column indicates the number of total reactions that are covered by the matching in the alignment. The column which indicates the correct mapping number in the alignment is called Correct Mappings and Ratio column gives the

ratio of the correct mapping number that is produced by the alignment to the total mapping number in that alignment. Also, each subcolumn indicates the algorithm names such that whereas the subcolumn which is represented as S corresponds to the alignment scores of SubMap algorithm, the subcolumn C_1 indicates the weighting scheme W_1 of CAMPways algorithm with $\alpha = 0.3$. When we try different α values, we obtain similar results to the results of $\alpha = 0.3$. So, we only give the results of CAMPways algorithm when $\alpha = 0.3$ in the weighting scheme W_1. Similarly, the subcolumn C_2 represents the weighting scheme W_2 of CAMPways algorithm with $\alpha_1 = 0.4, \alpha_2 = 0.5, \alpha_3 = 0.1$.

When we compare the alignment results of SubMap and CAMPways, both algorithms provide similar coverage values in general. In some cases, SubMap produces more coverage but in the others, both versions of CAMPways produce better results for the coverage. When we look the results of correct mappings, the results of CAMPways are overwhelmingly superior than the results of SubMAP. Even if SubMap provides more coverage than CAMPways for the alignment of atc-eco pair within 1.11 Xenobiotics biodegredation metabolic network (153 versus 134), the correct mapping number of CAMPways algorithm is still better than the SubMap results (60 versus 53). In this case, it is possible to see that even if SubMap provides more coverage which means more matched reactions in the alignment, the reactions in that alignment are not share same pathways and these matches are meaningless. In 5 instances of 22 results, SubMap does not give results due to excessive memory consumption and these results' entries are empty in Table 4.1.1. Whereas for 16 instances, CAMPways presents more correct mappings, in only one instance, both algorithms give same number of correct mappings. Additionally, results of ratio shows that CAMPways gives better results than SubMap algorithm. Hereby, we need to remember that the ratio does not normalize the correct mapping number with the coverage number, conversely, it normalize the correct mapping number with the total reaction number in the output alignment. Thus, the ratio value gives us the correct mapping ratio in that alignment.

4.1.1.2. Across-domain Alignments. We perform the same tests for the across-domain pairs within the same metabolism networks and two remarkable observations are obtained from the alignment results. The first one is the difference between the correct mapping numbers and between the correctness ratios. When we focus on these results, we observe that the results are decreased by comparison with Table 4.1.1. The prime reason of the decrease is the evolutionary distance such that when the evolutionary distance between the species increases, the reactions which are in the different pathways are matched by the algorithms. Secondly, when we compare the alignment qualities, the trend is same as in the same-domain experiments. In almost all cases, CAMPways provides more and better results in terms of correctness ratio than SubMap. In 4 instances of 20 results, SubMap does not produce any alignment due to excessive memory consumption as in the same-domain experiments. In 7 instances, both algorithms produce same results in terms of correct mapping number and in 16 instances, CAMPways algorithm gives more correct mappings than SubMap. On the other hand, in only 1 instance, the results of SubMap are better than CAMPways. These results can be shown in Table 4.1.1.1 and all column names are same as in Table 4.1.1.The results in that table belong to the metabolisms 1.2, 1.6, 1.7, 1.9, 1.10 and 1.11, respectively. It is important to emphasize that CAMPways produces all alignment within 11 metabolism networks. In order to provide compactness, we do not give all results in the table, but we define our results that are taken from the other metabolism network alignments. The average correctness ratios of the output alignments within 1.1, 1.3, 1.4, 1.5 and 1.8 are 0.7, 0.88, 0.97, 0.64 and 0.77, respectively. According to these results, it is possible to observe that totally CAMPways works better than SubMap but there are some exceptional cases such as the alignment results of 1.7 Glycan biosynthesis and metabolism and 1.10 Biosynthesis of other metabolites. In these exceptional cases, the sizes of correct alignments of both algorithms are quite small. Therefore, drawing a conclusion from these results is hard and it shows that these results are negligible. Even if we consider these results, in all output alignment results of CAMPways and SubMap, the correctness ratio of CAMPways is %5.3 better than the correctness ratio of SubMap on average.

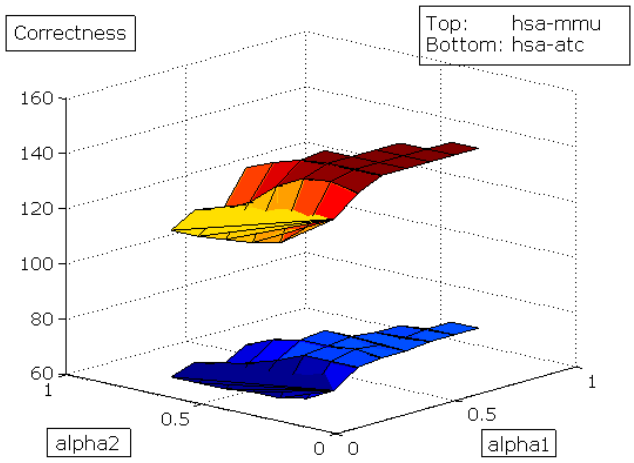

Figure 4.1. Top: Same-domain (hsa-mmu). Bottom: Across-domains (hsa-atc)

At this point, we need to explain that we make experiments to see the results of weighting scheme W_2 for various $\alpha_1, \alpha_2, \alpha_3$ values in terms of correctness values of the output alignments and the change in the number of 1-to-i mappings.

4.1.1.3. Correctness and Sizes of Mappings. We perform many tests in order to obtain the output alignments and experimental results of these depending on the various a_1, a_2, a_3 values of the weighting scheme W_2 of CAMPways. In Figure 4.1, while top 3D surface represents the correct mapping number of the same-domains (hsa-mmu) pair, bottom 3D surface indicates the correct mapping number of the across-domains (atc-eco) pair. In this figure, the plots are drawn due to changing values of a_1, a_2, a_3 parameters. The values are limited between 0.1 and 0.8 and in each experiment the values increase 0.1. The value of a_3 is not clearly defined because of the parameters relationship such that $a_3 = 1 - a_1 - a_2$. The correct mapping number is given in z-axis. This number is an average number of the correct mapping number of the alignments within 1.1 and 1.11 metabolism networks. The best average number of correct mappings is obtained when $a_1 = 0.4, a_2 = 0.3$ and $a_3 = 0.3$. In this sense, we need to remember that each a_i represents the importance of 1-to-i mappings. According to this representation, we also perform some tests in order to observe the changing in the number of

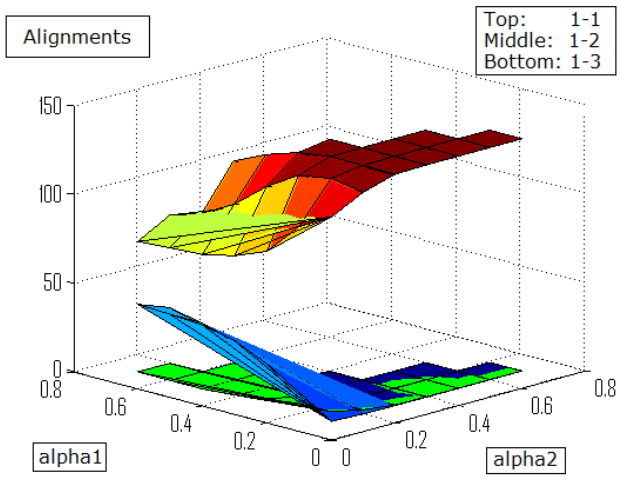

Figure 4.2. Same-domain (hsa-mmu) results. Top: 1-to-1 mappings. Middle: 1-to-2 mappings. Bottom: 1-to-3 mappings

1-to-i mappings in the alignment. These results are shown in Figure 4.2. Both same-domain results and across-domain results are similar, therefore we give only same-domain (hsa-mmu) results in that figure. As in the other figure, the plot shows the average number of 1-to-i mappings within 1.1 and 1.11 metabolism networks. For same-domain tests, the maximum number of 1-to-3 mappings is around 10 and this number is obtained when $a_1 = 0.1$ and $a_2 = 0.1$. Similarly, the maximum number of 1-to-2 mappings is around 40 and as shown in figure 4.2, while the value of a_1 increases, the number of 1-to-2 mappings decreases. The remarkable difference between same-domain and across-domain tests, 1-to-2 and 1-to-3 mapping number does not decrease under 0. This implies that when the evolutionary distance increases between two species, the reaction in one network match with two or three reactions in other network, mainly.

4.1.2. Biochemical Significance of the Alignments

In order to compare the alignment qualities of both algorithms, we use functional group conversion (FGC) hierarchy. This hierarchy is provided by RCLASS database in KEGG database [45]. The reactions in that database are organized in the hierarchical func-

tional group categories. Same functional group indicates same or similar chemical reactions independently of the molecule size [46]. Thus, the inter-species alignment of a pair is called biochemically valid if the matched reaction subsets are in same FGC category. There are five levels in KEGG database such that the root level includes eight high-level FGC categories: Carbon-related, hydrogen-related, isomerization-related, nitrogen-related, oxygen-related, phosphorus-related, oxygen-related and halogen-related. The accuracy measure is same as in the previous sections such that for fixed level i, a mapping is called valid if all reactions in the mapping are in at least one of i. level categories. Assume that root level is defined as $i = 1$ and starting with the root level, we performed some tests and evaluate the accuracy degrees of CAMPways and SubMap algorithm alignments for first five levels.

We perform two type of experiments such as in the previous subsections: experiments on same-domains and across-domains. These results are given in Table 4.1.2. The metabolism network pairs, rows and subrows are same as in Table 4.1.1. While the subcolumns that are presented as S indicates the alignment results of SubMap, the subcolumns that are marked as C shows the results of weighting scheme W_1 of CAMPways algorithm. The weighting scheme W_2 gives similar results to W_1. Thus, we don't give the results of weighting scheme W_2 in that table. The main column title indicates the first five levels in FGC hierarchy and it helps to understand results easily level by level. Each table record corresponds to the correct mapping number of the alignment. It is possible to understand that the results of CAMPways are superior than SubMap in that table. For same-domain pairs, the number of correct mappings decreases from the abstract categories of root level to less abstract categories. Also, it is important to define that for 1.7 Glycan biosynthesis and metabolism, while the size of mapping is around 80 for hsa-mmu pair, both algorithms provide too few correct mapping. The ratio of correct mappings to mapping size is only %6. So, this case is inconsistent with the results that are given in Table 4.1.1 that provides %90 accuracy ratio. The main reason is the lack of reaction classification in FGC categories. In fact, this lack provides a useful opportunity such that if there exists a reaction

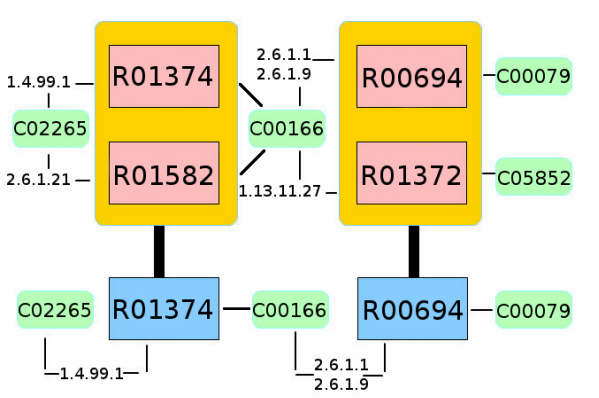

Figure 4.3. Sample mapping from the CAMPways alignment

in the mapping whose FGC classification is unknown, then FGC category of this reaction can be predicted or defined considering the reactions whose FGC classification is known are in the same mapping. When we look across-domain experiment results that are given in Table 4.1.2, same results are obtained such that CAMPways algorithm provides better results than Submap for almost all instances in all hierarchical levels. The only exception is 1.10 metabolism alignment for has-atc species. In this instance, both algorithms don't give remarkable results.

The experiments that are based on RCLASS data are extended by RPAIR data. For the alignment of atc-eco pair on Amino acid metabolism network, an instance is analyzed in detail for both algorithms. In this sense, a reactant pair is determined as a pair of a substrate and a product such that the reactant pair is used as a chemical substructure in the enzymatic reactions. In fact, RCLASS database classification provides information about the reactant pairs but there is a difference between RCLASS and RPAIR databases. Whereas RCLASS classifications are made due to molecular methods or computational methods with chemical structure information, RPAIR classifications are made by molecular alignments and manually compiled reactant pairs with biochemical information. The sample that is analyzed of CAMPways algorithm alignment is shown in Figure 4.3. In this matching, atc reactions R01374 (D-phenylalanine: acceptor oxidoreductase (deaminating)) and R01582 (D-Phenylalanine:

2-oxoglutarate aminotransferase) are together matched with eco reaction R01374. Additionally, the reactions R00694 (L-Phenylalaline: 2-oxoglutarate aminotransferase) and R01372 (Phenylpyruvate: oxygen oxidoreductase (hydroxylating, decarboxylating)) of atc species are together matched with the reaction R00694 of eco species. The output compound of reactions R01374 and R01374 is C00166 (Phenylpyruvate) and also, the output compound of these reactions is input compound of R00694 and R01372 reactions. As a consequence, in atc pathway, there is a directed edge from the node that corresponds to the reaction subset of R01374 and R01582 reactions to the node that correspond to the reaction subset of R00694 and R01372 reactions. Similarly, in eco pathway, there is a directed edge from the node that corresponds to the reaction R01374 to the node that corresponds to the reaction R00694. Thus, it is possible to see that this sample provides a conserved edge. Moreover, when we look the classifications, the first five levels in FGC categorization of the reactions R01374 and R01582 are same and this implies that the alignment is biologically valid depending on RCLASS classification. Also, both reactions are under the same RCLASS entry RC00006. For more accuracy, when we look RPATH data that includes manually compiled reactant pairs and provides more realism, both reactions belong to the same reactant pair RP00289. On the other side, in the alignment of SubMap, the reactions R01582 and R01373 (Prephenate hydro-lyase (decarboxylating phenylpyruvate - forming)) of atc species are matched with the single reaction R01373 of etc species. The reactions R01373 and R01582 shows differences from the second level in FGC categorization and also these reactions belong to the different RCLASS entries. Furthermore, there is not a remarkable relationship between these reactions considering RPAIR database. Consequently, CAMPways algorithm provides more meaningful in terms of biological significance alignment results.

4.1.3. Execution Speed and Memory Requirements

When we assume that degree of each nodes in G_p and G'_p is limited with a fixed number, the total running time of CAMPways is $O(|V_p|^2 log^2 |V_p|)$. In this sense, it is assumed that $|V_p|$

is greater than $|V'_p|$. The detailed running time analysis is given in the next section. If any comparison is needed between SubMap and CAMPways algorithms, there is no running time analysis for SubMap. The experimental results that are given in this section are provided by performing the tests on the system that consists of Intel(R) Xeon(R) CPU, 2.67 GHz and 24 GB memory. All required CPU times are given in Table 4.1.5. Whereas the first three rows indicate the experimental results of same-domain pairs, the other rows represent the results of across-species pairs. For each instances, the total reaction number is denoted by TR column in that table. The algorithm names' abbreviations are same as in previous tables. There is a limitation for SubMap algorithm such that some experiments are not completed due to excessive memory consumption. For the alignment of hsa-atc pair within 1.1 Carbonhydrate-metabolism, whereas CAMPways algorithm gives the alignment around 3 minutes, SubMap algorithm does not complete after 2 hours execution. In 15 instances of 17 results, the experiments that are performed on same-domain pairs show that CAMPways' execution time is better than the execution time of SubMap. In this sense, the important point is the differences between the execution times of CAMPways and SubMap is large when CAMPways execution is faster but when SubMap algorithm is completed in a small time, the difference between the execution times is quite small. When we look in terms of computational efficiency of both algorithms, the difference between same-domain alignments and across-domain alignment is interesting. In fact the difference between computationally efficiency corresponds to the difference between the algorithms. The metabolic networks that belong to the same-domain species are close to each other in terms of evolutionary distance. Therefore, the aligned networks include more conserved edges. In fact, there are many instances that make the network alignment sensible such that both homological similarity and topological similarity are optimized. Most of the reactions are aligned in the main loop of CAMPways algorithm and conflict graph sizes are large due to higher edge conservation. But when the species are not close in terms of evolutionary distance, naturally, the number of conserved edge decreases and in this sense, both algorithm preferred to give alignments that includes higher homologically similar matched reactions.

4.1.4. Running Time Analysis

We assume that the degree of each node in G_p and G'_p is limited with a constant $\triangle$. This limitation is sensible when the metabolic pathways are considered. In addition to this limitation, we assume that $|V'_p| = O(|V_p|)$. Each node in G_p can be denoted in at most $(1-\triangle^k)/(1-\triangle)$ such that the node subsets in G_p that consist of the connected nodes give rise to the nodes of G_p^k. This implies that the number of nodes in the extended graph G_p^k is at most $|V_p|$ x $(1-\triangle^k)/(1-\triangle)$. Because we assume that both $\triangle$ and k are constant, the size of node set of G_p^k is limited with $O(|V_p|)$. Similar arguments are employed for $G_p'^k$, as well. Also, the degree of each node in the extended graphs is limited with $kx\triangle$. According to these limitations, degree values are constant. The running time of step 1 is limited with the time of sorting the edges in the complete bipartite graph according to the edge weights. Therefore, the running time of Step 1 is $O(|V_p|^2 log|V_p|)$. Each node in the extended graph G_p^k is shown in at most k_1^2 x k x $\triangle$ node in the conflict graph. This implies that the number of nodes in the conflict graph is at most $|V_p|$ x k_1^2 x k x $\triangle$. Because the values of k_1, k and $\triangle$ are constant, the number of nodes in the conflict graph is $O(|V_p|)$. So, in simple terms, the construction of the conflict graph requires $O(|V_p|^2)$. Hereby, there is an important point that the degree of each node in the conflict graph is also limited with a constant. Two nodes in the conflict graph share an edge if these nodes share a common node in their original pathways G_p, G'_p. Therefore each original pathway node is represented with at most k_1^2 x k x $\triangle$ x $(1-\triangle^k)$ / $(1-\triangle)$ conflict node such that this value is also a constant. When we focus on maximum weight independent set solution, GWMIN2 heuristic needs the weight of the neighbors of the node to make a calculation for this node. Therefore, each node can be calculated in a constant time and this part of Step 2 needs $O(|V_p|log|V_p|)$. It is important to remember that both Step 1 and Step 2 remain until the loops are over which means until the convergence is provided. Because the degree of each node in the conflict graph is limited with a constant, in each iteration, after the aligned nodes are removed from the conflict node, a constant number of conflict nodes remains. Because any node of extended

graph can be represented in constant number of conflict graph, ongoing iteration remains with the constant number of extended graph nodes. This implies that the iteration number is $O(log|V_p|)$. As a consequence, the total running time that consists of both Step 1 and Step 2 is $O(|V_p|^2 log^2|V_p|)$. This is also an upper bound for the final expansion step and provides an upper bound on the algorithm. At this point, it is important to remember that the convergence is obtained in a few iterations on the mentioned metabolic networks.

4.1.5. Discussion of Results for CAPPI

In this section, we compare the results of CAPPI with the results of SPINAL, IsoRank and MI-GRAAL algorithms. Firstly, we compare biological significance of these results. When the subject is protein interaction network, the biological significance is evaluated in terms of gene ontology (GO) consistency scores. In order to make such a comparison, we use the formula that is given in [28]. The results are given in table 4.6. Because, mostly SPINAL gives more superior results, IsoRank and MI-GRAAL results are not given in that table. In order to obtain these results, we run SPINAL algorithm with various parameters and produce similarity files within these processes. Afterwards, we run CAPPI algorithm with related similarity files. For CAPPI algorithm, α equals to 0, the constant b equals to %55 and the constants k_1, k_2 are 3. Also, the iteration number is given as one. For 18 instances of 25 results, CAPPI algorithm gives better results in terms of go consistency scores. Furthermore, similarly, for 18 instances of 25 results, CAPPI algorithm produces more conserved edges even if the priority is finding better biological results. For the instances where IsoRank gives better results than SPINAL, even if CAPPI algorithm does not pass IsoRank, it increases the score of SPINAL. On the other hand, when we look conserved edges, interestingly, CAPPI algorithm passes SPINAL where SPINAL gives low results. The results are given in 4.7. MI-GRAAL results are not given in that table for integrity because MI-GRAAL algorithm does not use any constant value. For ce-dm species and $\alpha = 0.7$, while MI-GRAAL algorithm produces 2390 conserved edges, CAPPI algorithm gives 2374 conserved edges and SPINAL

produces 2258 conserved edges. It implies that both biological meaning evaluations and conserved edge evaluations, CAPPI algorithm may be an alternative or a supplement for SPINAL algorithm.

Table 4.1. Same-domains reverse engineering experiment.

Total Reactions	Coverage			Correct Mappings			Ratio		
	S	C1	C2	S	C1	C2	S	C1	C2
437	-	435	435	-	211	213	-	0.99	0.98
458	-	416	416	-	166	171	-	0.82	0.83
62	62	62	62	29	31	31	0.96	1	1
116	105	110	110	45	51	51	0.93	0.94	0.94
745	-	726	726	-	361	361	-	0.99	0.99
264	244	254	254	96	105	103	0.82	0.82	0.83
320	-	320	320	-	159	159	-	0.99	0.99
296	280	262	262	110	128	128	0.90	0.98	0.98
496	491	481	481	221	239	239	0.96	0.99	0.99
369	352	340	339	122	143	143	0.79	0.86	0.86
134	128	130	130	59	64	64	0.96	0.98	0.98
108	102	97	97	37	39	39	0.78	0.82	0.82
168	148	168	168	73	76	76	1	0.90	0.90
73	69	64	64	31	31	31	0.96	0.96	0.96
307	-	306	307	-	150	151	-	0.98	0.98
334	325	324	326	129	143	144	0.87	0.89	0.90
31	28	28	28	12	14	14	1	1	1
51	43	43	44	15	17	17	0.78	0.80	0.77
35	34	34	34	16	17	17	1	1	1
23	21	20	20	8	9	9	0.8	0.9	0.9
207	201	200	200	87	100	100	0.92	1	1
175	153	134	134	53	60	60	0.81	0.89	0.89

Table 4.2. Across-domains experiment.

Total Reactions	Coverage			Correct Mappings			Ratio		
	S	C1	C2	S	C1	C2	S	C1	C2
93	71	61	63	23	26	27	0.76	0.86	0.87
85	74	61	61	19	25	25	0.63	0.83	0.83
85	74	61	61	19	25	25	0.63	0.83	0.83
93	71	61	63	23	26	27	0.76	0.86	0.87
128	122	117	117	41	46	46	0.74	0.80	0.79
114	108	100	100	37	41	41	0.77	0.85	0.85
118	110	101	101	38	41	41	0.79	0.83	0.83
124	118	117	115	38	44	41	0.73	0.77	0.74
125	79	78	80	7	7	7	0.19	0.18	0.17
116	61	63	63	6	6	6	0.22	0.19	0.19
116	61	63	63	6	6	6	0.22	0.19	0.19
125	79	78	80	7	7	7	0.19	0.18	0.17
39	37	34	34	8	12	12	0.53	0.70	0.70
43	34	27	27	9	12	12	0.69	0.92	0.92
46	40	33	33	12	15	15	0.75	0.93	0.93
36	36	28	28	7	11	11	0.50	0.78	0.78
30	24	26	26	4	4	4	0.40	0.36	0.36
28	21	21	19	2	2	1	0.25	0.22	0.12
27	21	18	18	2	1	1	0.25	0.14	0.14
31	24	26	26	4	4	4	0.40	0.36	0.36
174	156	135	135	43	46	46	0.67	0.68	0.68
208	198	198	198	35	40	40	0.39	0.41	0.41
215	208	214	214	42	46	46	0.45	0.44	0.44
167	156	136	136	36	40	40	0.56	0.59	0.59

Table 4.3. Same-domains biochemical significance experiments.

Level 1		Level 2		Level 3		Level 4		Level 5	
S	C	S	C	S	C	S	C	S	C
-	193	-	193	-	193	-	192	-	192
-	154	-	154	-	151	-	144	-	138
23	23	22	23	22	23	21	23	21	22
32	41	32	41	32	39	32	39	32	39
323	343	323	343	323	343	318	340	316	338
97	105	97	105	97	104	93	103	92	102
-	103	-	103	-	101	-	101	-	101
66	84	66	84	64	80	64	80	63	80
209	229	209	229	208	229	205	227	205	227
117	143	110	139	104	132	97	130	93	127
53	57	53	57	52	57	52	57	52	56
37	35	37	35	34	33	33	33	33	32
5	6	5	6	5	6	5	6	5	6
20	21	20	21	20	21	20	21	19	21
-	123	-	123	-	123	-	123	-	123
96	115	94	114	93	111	93	110	90	109
9	13	9	13	9	13	9	13	9	13
16	17	16	16	16	16	15	15	14	15
14	16	14	16	13	16	13	16	13	16
7	9	7	9	7	9	6	8	6	8
79	97	78	97	76	97	76	97	76	97
44	59	44	58	42	55	42	55	42	54

Table 4.4. Across-domains biochemical significance experiments.

Level 1		Level 2		Level 3		Level 4		Level 5	
S	C	S	C	S	C	S	C	S	C
17	19	16	18	15	18	13	18	12	16
13	19	12	18	10	17	9	17	8	15
13	19	12	18	10	17	9	17	8	15
17	19	16	18	15	18	13	18	12	16
37	40	37	40	35	40	34	39	31	36
33	35	33	35	32	35	30	34	26	30
35	37	35	37	32	36	31	35	26	31
32	38	32	38	30	36	30	36	27	36
2	3	2	3	2	3	1	2	1	2
2	3	2	3	2	3	1	2	1	2
2	3	2	3	2	3	1	2	1	2
2	3	2	3	2	3	1	2	1	2
8	12	7	10	7	10	5	8	4	8
9	11	6	11	6	11	6	11	5	10
9	13	7	11	6	10	6	10	6	10
9	11	7	9	6	9	5	8	4	8
5	5	5	5	4	5	4	5	4	5
1	2	1	2	1	2	1	1	1	1
1	0	1	0	1	0	1	0	1	0
5	5	5	5	4	5	4	5	4	5
43	55	36	50	32	48	30	44	29	44
51	66	37	53	34	46	30	40	28	40
56	69	43	53	39	48	35	43	33	41
39	50	32	44	28	43	26	40	25	39

Table 4.5. Running Time Analysis Table

TR	S	C	TR	S	C	TR	S	C	TR	S	C	TR	S	C	TR	S	C
62	3.04	0.30	116	62.81	2.26	264	454.21	13.39	296	1620	15.73	496	975.31	39.87	369	121.43	25.23
134	48.09	1.42	108	17.99	0.94	168	0.32	2.94	73	0.50	0.28	334	1788.84	25.17	31	0.06	0.04
51	0.15	0.09	35	0.09	0.04	23	0.04	0.02	207	3.25	1.00	175	0.67	5.39			
93	33.16	2.79	85	6.64	0.82	85	6.51	0.72	93	34.68	2.72	128	40.46	1.67	114	21.52	1.17
118	20.7	1.13	124	42.0	1.45	125	0.44	10.25	116	0.3	6.64	116	0.38	6.08	125	0.41	10.19
39	0.07	0.09	43	0.09	0.05	46	0.10	0.11	36	0.08	0.07	30	0.04	0.03	28	0.05	0.02
27	0.06	0.03	31	0.05	0.03	174	1.26	10.95	208	1.85	20.03	215	1.77	13.24	167	1.27	9.56

Table 4.6. GOC evaluations

		GOC scores					Conserved Interactions				
Dataset	Algorithm	α=0.3	α=0.4	α=0.5	α=0.6	α=0.7	α=0.3	α=0.4	α=0.5	α=0.6	α=0.7
ce-dm	SPINAL$_I$	235.28	234.90	231.87	230.80	225.99	575	585	611	624	655
	CAPPI	235.51	233.28	231.11	230.87	224.80	624	624	650	659	676
ce-hs	SPINAL$_I$	100.83	100.31	100.31	99.43	99.45	518	537	535	562	605
	CAPPI	101.75	99.73	100.57	99.47	99.60	547	565	566	592	624
ce-sc	SPINAL$_I$	148.53	150.59	149.51	148.93	148.75	810	815	815	814	809
	CAPPI	149.01	149.76	150.70	149.52	149.22	797	805	810	807	803
dm-sc	SPINAL$_I$	392.41	390.64	389.28	388.99	385.42	1645	1653	1647	1646	1681
	CAPPI	392.51	391.13	390.61	391.23	387.22	1661	1660	1642	1664	1678
hs-sc	SPINAL$_I$	341.15	342.38	342.07	342.56	340.08	2209	2234	2226	2254	2262
	CAPPI	342.02	343.48	343.03	341.18	339.96	2223	2251	2247	2282	2280

Table 4.7. Conserved edge evaluations

Conserved Edges						
DataSet	Algorithm	0.3	0.4	0.5	0.6	0.7
ce-dm	$SPINAL_{II}$	2343	2320	2300	2337	2258
	CAPPI	2343	2372	2335	2361	2374
	IsoRank	335	329	325	327	328
ce-hs	$SPINAL_{II}$	2370	2446	2437	2487	2512
	CAPPI	2420	2387	2411	2403	2405
	IsoRank	299	287	290	300	293
ce-sc	$SPINAL_{II}$	2326	2384	2323	2361	2398
	CAPPI	2312	2309	2295	2322	2306
	IsoRank	410	385	385	360	339
dm-sc	$SPINAL_{II}$	5203	5150	5311	5283	5360
	CAPPI	4802	4827	4854	4754	4828
	IsoRank	840	856	837	781	763
hs-sc	$SPINAL_{II}$	5703	5593	5651	5706	5798
	CAPPI	5264	5275	5246	5262	5184
	IsoRank	786	824	817	763	761

REFERENCES

1. U. Brandes, T. Dwyer, and F. Schreiber. Visual understanding of metabolic pathways across organisms using layout in two and a half dimensions. *JOURNAL OF INTEGRATIVE BIOINFORMATICS*, 1(1):2004, 2004.

2. Jeroen van Reeuwijk, Heleen H. Arts, and Ronald Roepman. Scrutinizing ciliopathies by unraveling ciliary interaction networks. *Human Molecular Genetics*, 20(R2):R149–R157, 2011.

3. G. M. Cooper. *The Cell - A Molecular Approach 2nd Edition*. Sunderland (MA): Sinauer Associates, 2000.

4. Cornelius G. Friedrich. Physiology and genetics of sulfur-oxidizing bacteria. volume 39 of *Advances in Microbial Physiology*, pages 235 – 289. Academic Press, 1997.

5. Y Lu GM Singh MJ Cowan CJ Paciorek JK Lin F Farzadfar Y-H Khang GA Stevens G Danaei, MM Finucane. National, regional, and global trends in fasting plasma glucose and diabetes prevalence since 1980: systematic analysis of health examination surveys and epidemiological studies with 370 country-years and 27 million participants, 2011.

6. M. P. Heyes, K. Saito, J. S. Crowley, L. E. Davis, M. A. Demitrack, M. Der, L. A. Dilling, J. Elia, M. J. P. Kruesi, A. Lackner, S. A. Larsen, K. Lee, H. L. Leonard, S. P. Markey, A. Martin, S. Milstein, M. M. Morradian, M. R. Pranzatelli, B. J. Quearry, A. Salazar, M. Smith, S. E. Strauss, T. Sunderland, S. W. Swedo, and W. W. Tourtellotte. Quinolinic acid and kynurenine pathway metabolism in inflammatory and non-inflammatory neurological disease. *Brain*, 115(5):1249–1273, 1992.

7. T. Dandekar, S. Schuster, B. Snel, M. Huynen, and P. Bork. Pathway alignment: application to the comparative analysis of glycolytic enzymes. *Biochem. J.*, 343(1):115–124, 1999.

8. M. Kanehisa and S. Goto. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 28:27–30, 2000.

9. Dreher K Fulcher CA Subhraveti P Keseler IM Kothari A Krummenacker M Latendresse M Mueller LA Ong Q Paley S Pujar A Shearer AG Travers M Weerasinghe D Zhang P Karp PD Caspi R, Altman T. The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic Acids Res*, 40(D1):D742–D753, 2012.

10. David Croft, Gavin OKelly, Guanming Wu, Robin Haw, Marc Gillespie, Lisa Matthews, Michael Caudy, Phani Garapati, Gopal Gopinath, Bijay Jassal, Steven Jupe, Irina Kalatskaya, Shahana Mahajan, Bruce May, Nelson Ndegwa, Esther Schmidt, Veronica Shamovsky, Christina Yung, Ewan Birney, Henning Hermjakob, Peter DEustachio, and Lincoln Stein. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Research*, 39(suppl 1):D691–D697, 2011.

11. Mehmet Koyutrk, Ananth Grama, and Wojciech Szpankowski. An efficient algorithm for detecting frequent subgraphs in biological networks. *Bioinformatics*, 20(suppl 1):i200–i207, 2004.

12. Biopathways and protein interaction databases. a lecture in bioinformatics tools for comparative genomics: A short course.

13. Alexander W. Rives and Timothy Galitski. Modular organization of cellular networks. *Proceedings of the National Academy of Sciences*, 100(3):1128–1133, 2003.

14. Yukako Tohsato, Hideo Matsuda, and Akihiro Hashimoto. A multiple alignment algorithm for metabolic pathway analysis using enzyme hierarchy. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, pages 376–383. AAAI Press, 2000.

15. Alexander P. Ducruet, Andreas Vogt, Peter Wipf, and John S. Lazo. Dual specificity

protein phosphatases: Therapeutic targets for cancer and alzheimer's disease. *Annual Review of Pharmacology and Toxicology*, 45(1):725–750, 2005. PMID: 15822194.

16. R L Finley and R Brent. Interaction mating reveals binary and ternary connections between drosophila cell cycle regulators. *Proceedings of the National Academy of Sciences*, 91(26):12980–12984, 1994.

17. R Aebersold and M Mann. Mass spectrometry-based proteomics. *Nature*, (422):198–207, 2003.

18. Sourav Bandyopadhyay, Roded Sharan, and Trey Ideker. Systematic identification of functional orthologs based on protein network comparison. *Genome Research*, 16(3):428–435, 2006.

19. Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, Oct 1990.

20. Xiaoning Qian, Sing-Hoi Sze, and Byung-Jun Yoon. Querying pathways in protein interaction networks based on hidden markov models. *Journal of Computational Biology*, 16(2):145–157, 2009.

21. Kimmen Sjlander. Phylogenomic inference of protein molecular function: advances and challenges. *Bioinformatics*, 20(2):170–179, 2004.

22. Brian P. Kelley, Roded Sharan, Richard M. Karp, Taylor Sittler, David E. Root, Brent R. Stockwell, and Trey Ideker. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proceedings of the National Academy of Sciences*, 100(20):11394–11399, 2003.

23. Johannes Berg and Michael Lssig. Local graph alignment and motif search in biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(41):14689–14694, 2004.

24. Mehmet Koyutrk, Yohan Kim, Shankar Subramaniam, Wojciech Szpankowski, and Ananth Grama. Detecting conserved interaction patterns in biological networks. *Journal of Computational Biology*, 13(7):1299–1322, 2006.

25. Jason Flannick, Antal Novak, Balaji S. Srinivasan, Harley H. McAdams, and Serafim Batzoglou. Grmlin: General and robust alignment of multiple large interaction networks. *Genome Research*, 16(9):1169–1181, 2006.

26. Rohit Singh, Jinbo Xu, and Bonnie Berger. Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proceedings of the National Academy of Sciences*, 105(35):12763–12768, 2008.

27. Ferhat Ay and Tamer Kahveci. Submap: Aligning metabolic pathways with subnetwork mappings. In *Proceedings of the 14th Annual International Conference on Research in Computational Molecular Biology*, RECOMB'10, pages 15–30, Berlin, Heidelberg, 2010. Springer-Verlag.

28. A. E. Aladag and C. Erten. SPINAL: Scalable Protein Interaction Network Alignment. *Bioinformatics*, 29(7):917–924, April 2013.

29. T.F. Smith and M.S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195 – 197, 1981.

30. B. P. Kelley, B. Yuan, F. Lewitter, R. Sharan, B. R. Stockwell, and T. Ideker. PathBLAST: A Tool for Alignment of Protein Interaction Networks. *Nucleic Acids Research*, 32(Web Server issue), July 2004.

31. Zhi Liang, Meng Xu, Maikun Teng, and Liwen Niu. Netalign: a web-based tool for comparison of protein interaction networks. *Bioinformatics*, 22(17):2175–2177, 2006.

32. G. Abaka, T. Biyikoglu, and C. Erten. Campways: constrained alignment framework for the comparative analysis of a pair of metabolic pathways. *Bioinformatics*, 29(13):i145–i153, 2013.

33. K. Mehlhorn and S. Näher. *LEDA: A Platform for Combinatorial and Geometric Computing* . Cambridge University Press, November 1999.

34. Mikhail Zaslavskiy, Francis Bach, and Jean-Philippe Vert. Global alignment of protein-protein interaction networks by graph matching methods. *Bioinformatics*, 25(12):i259–1267, 2009.

35. Jack Edmonds. Maximum matching and a polyhedron with 0, 1-vertices. *Journal of Research of the National Bureau of Standards B*, 69:125–130, 1965.

36. Harold N. Gabow. Scaling algorithms for network problems. *J. Comput. Syst. Sci.*, 31(2):148–168, September 1985.

37. Mohsen Bayati, Christian Borgs, Jennifer T. Chayes, and Riccardo Zecchina. Belief-propagation for weighted b-matchings on arbitrary graphs and its relation to linear programs with integer solutions. *CoRR*, abs/0709.1190, 2007.

38. Michael R. Garey and David S. Johnson. *Computers and Intractability; A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., New York, NY, USA, 1990.

39. Shuichi Sakai, Mitsunori Togasaki, and Koichi Yamazaki. A note on greedy algorithms for the maximum weighted independent set problem. *Discrete Applied Mathematics*, 126(23):313 – 322, 2003.

40. Leonid Chindelevitch, Cheng-Yu Ma, Chung-Shou Liao, and Bonnie Berger. Optimizing a global alignment of protein interaction networks. *Bioinformatics*, 29(21):2765–2773, 2013.

41. Frederic Rousseau, Joost Schymkowitz, and LauraS. Itzhaki. Implications of 3d domain swapping for protein folding, misfolding and function. In JacquelineM. Matthews, editor, *Protein Dimerization and Oligomerization in Biology*, volume 747 of *Advances in Experimental Medicine and Biology*, pages 137–152. Springer New York, 2012.

42. Ferhat Ay, Michael Dang, and Tamer Kahveci. Metabolic network alignment in large scale by network compression. *BMC Bioinformatics*, 13(Suppl 3):S2, 2012.

43. Rohit Singh, Jinbo Xu, and Bonnie Berger. Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proceedings of the National Academy of Sciences*, 105(35):12763–12768, 2008.

44. Oleksii Kuchaiev and Nataa Prulj. Integrative network alignment reveals large regions of global network similarity in yeast and human. *Bioinformatics*, 2011.

45. Minoru Kanehisa, Susumu Goto, Yoko Sato, Miho Furumichi, and Mao Tanabe. Kegg for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, 40(D1):D109–D114, 2012.

46. J. March. *Advanced Organic Chemistry: Reactions, Mechanisms, and Structure (3rd ed.)*. Wiley, New York, 1985.

Curriculum Vitae

Gamze Abaka was born in 1991 at Bakırköy, İstanbul. She graduated from Kırımlı İsmail Rüştü Olcay Anatolian High School in 2008. Then, she enrolled into Kadir Has University in 2008 and graduated in 2012. She has a bachelor degree of Computer Engineering. Afterwards, she continued on her education in Kadir Has University at the graduate program of Computer Engineering. She also worked as a research assistant in Kadir Has University. Her main research interests are programming languages, bioinformatics and designing.