

CAPACITATED TRANSSHIPMENT MODELS
FOR PREDICTING SIGNALING PATHWAYS

A Paper
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Ritika Sahni

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Department:
Computer Science
Option: Operation Research

December 2012

Fargo, North Dakota

North Dakota State University

Graduate School

Title

Capacitated Transshipment Models for Predicting Signaling Pathways

By

Ritika Sahni

The Supervisory Committee certifies that this *disquisition* complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Kendall E. Nygard

Chair

William Perrizo

Vasant Ubhaya

Eugene Berry

Approved by Department Chair:

12/7/2012

Date

Brian Slator

Signature

ABSTRACT

Signal transduction is a process of transmitting signals for controlling biological responses. The protein-protein interaction (PPI) data, containing signal transduction proteins, can be considered as a bi-directional, weighted network with the proteins as nodes, the interactions between them as edges, and the confidence score of the interaction as weights on edges. If the edges of this network are given a capacity of one, and if the starting and ending proteins are the supply and demand nodes, then this problem can be modeled as a capacitated transshipment model with pathways as the solutions. Our application concerns finding the signaling pathways for yeast's mitogen-activated protein-kinase (MAPK) pheromone response and filamentation growth using the model created in the SAS OPTMODEL. The results demonstrate that the proposed model is easier to understand and interpret, and is applicable to the PPI network to discover signaling pathways efficiently and accurately.

ACKNOWLEDGEMENTS

I am grateful to my adviser, Dr. Kendall E. Nygard, for his encouragement, support, and academic experience, which has been invaluable during my research. I would like to thank all the other committee members, Dr. William Perrizo, Dr. Vasant Ubhaya, and Dr. Eugene Berry, for their interest in my paper and for their valuable comments. I am also indebted to a friend for checking my paper for grammar and format.

TABLE OF CONTENTS

ABSTRACT	iii
ACKNOWLEDGEMENTS	iv
LIST OF TABLES	vii
LIST OF FIGURES	viii
CHAPTER 1. INTRODUCTION	1
1.1. Problem Statement	1
1.2. Organization of the Paper	3
CHAPTER 2. BACKGROUND	5
2.1. Signaling Pathways	5
2.1.1. Pheromone Response	6
2.1.2. Filamentation Growth	7
2.2. Depth-First Search	7
2.3. Capacitated Transshipment Problem	8
2.3.1. Definition of the Capacitated Transshipment Problem.....	9
2.3.2. Simplex Algorithm to Solve the Capacitated Transshipment Problem	10
2.4. SAS OPTMODEL and NETFLOW	18
2.4.1. OPTMODEL.....	19
2.4.2. NETFLOW	19
2.4.3. Example	19
CHAPTER 3. LITERATURE REVIEW	22
3.1. Color Coding.....	22
3.2. NetSearch.....	23

3.3. PathFinder	23
3.4. Integer Linear Programming	24
CHAPTER 4. APPROACH	27
4.1. Definition of the Problem	27
4.1.1. Steps Followed to Organize the Data as a Capacitated Transshipment Problem	27
4.2. Definition of the Model	30
CHAPTER 5. RESULTS AND EVALUATION	36
CHAPTER 6. CONCLUSION.....	46
REFERENCES	47

LIST OF TABLES

<u>Table</u>	<u>Page</u>
1. Comparison of methods for detecting the pheromone response signaling pathway.....	44
2. Comparison of methods for detecting the filamentation growth signaling pathway.....	44

LIST OF FIGURES

<u>Figure</u>	<u>Page</u>
1. KEGG's Yeast MAPK Signaling Pathways.....	8
2. Simplex Algorithm Example Network.	11
3. Initial Basis Tree in the Example.....	12
4. Initial Basis Node Potentials in the Example.....	12
5. Simplex Algorithm Cycle Formed in the Example.....	13
6. New Basis for the Example.....	14
7. Steps Followed to Achieve Optimal Solution in the Example.	15
8. Final Steps of Simplex Algorithm's Solution for the Example.....	16
9. Example Data Input in SAS.....	20
10. Solution Obtained Through OPTMODEL Procedure in SAS for the Example Data.....	21
11. Algebraic Representation of the Integer Linear Programming Model.	25
12. Pheromone Response Sub-Network.	28
13. Filamentation Growth Sub-Network.....	29
14. Input Data for Yeast's MAPK Pheromone Response Signaling Pathway.....	31
15. Input Data for Yeast's MAPK Filamentation Growth Signaling Pathway.....	32
16. Definition of the Capacitated Transshipment Model.....	34
17. An OPTMODEL Solution for Yeast's MAPK Pheromone Response Pathway.....	37
18. An OPTMODEL Solution for Yeast's MAPK Filamentation Growth Pathway.....	38
19. A NETFLOW Solution for Yeast's MAPK Pheromone Response Pathway.	38
20. A NETFLOW Solution for Yeast's MAPK Filamentation Growth Pathway.	39
21. Pheromone Response Pathways.....	40

22. Filamentation Growth Pathways..... 41

CHAPTER 1. INTRODUCTION

As defined by the National Institute of Health, "Bioinformatics is research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data" (Huerta, Haseltine, Liu, Downing, & Seto, 2000). The research work presented in this paper is in the field of bioinformatics.

With continuous, accelerated evolution in the field of biological science, the amount of such data is increasing at an exponential rate, significantly amplifying the need for sophisticated mathematical, statistical, and computational algorithms for storage, arrangement, retrieval, and mining of information to solve and analyze combinatorial problems. This paper focuses on the application of such a complex algorithm to solve a biological problem.

This chapter clearly states the problem which involves application of a mathematical algorithm for analyzing a biological problem. This chapter also provides information about the organization of the paper.

1.1. Problem Statement

Signal transduction is a significant process that controls different cell functions. It can be classified as a bi-directed, weighted network of protein interactions which can be used to predict signaling pathways (Vinayagam et al., 2011). These predicted pathways can be used as the starting point for the experiments that scientists need to conduct to find the actual signaling pathways. Discovering the signaling pathways plays an important role in understanding the interactions between the proteins and conducting further research in the field of drug/medicine discovery. The knowledge of mechanisms involved in signaling pathways can be used to find the

basis and causes of diseases which can be used to exploit these mechanisms to develop new therapeutic approaches (Wolkenhauer & Cho, 2003).

The procedure of finding the signaling pathways is like searching for optimal sub-networks based on a cost function. The problem can be regarded as an optimization network-flow problem.

The transshipment model is a minimum cost-flow network problem which includes the capacitated and un-capacitated transportation problems as well as the personnel assignment problem (Bradley, Brown, & Graves, 1977). It is used for a large number of diverse applications. It can be solved as linear programming problems with a constraint for each node and a variable for each arc (Bradley, Brown, & Graves, 1977).

The goal of the capacitated transshipment model presented in this paper is to determine how signals should be passed through the protein interaction arcs of a given network to minimize costs. We consider the protein-protein interaction network of baker's yeast as a bi-directed graph with N nodes (proteins) and E edges (interactions) (Yeang, Ideker, & Jaakkola, 2004). Each arc is an ordered pair of nodes (tail, head). Each arc has a "signal passing" cost per unit flow, and an upper bound on flow (also known as capacity) can be added based on the data obtained (Bradley, Brown, & Graves, 1977). Each node has either supply or demand where signal units enter or leave, or are transshipped. The problem is to minimize total costs with flows, to fulfill the associated lower bounds and capacities, and to conserve flow at each node using the capacitated transshipment model for predicting possible signaling pathways (Bradley, Brown, & Graves, 1977).

The main goal of this paper is applying a mathematical model, the capacitated transshipment model, to utilize the protein-protein interaction data of baker's yeast

(*Saccharomyces cerevisiae*) for predicting signaling pathways. The two main objectives to achieve this goal are described as follows:

- 1) Develop a capacitated transshipment model and classify the signaling pathways' prediction problem as a capacitated transshipment problem.
- 2) Carry out experiments in which the available biological data are used by the capacitated transshipment model to predict signaling pathways.

Through this paper, we attempt to prove that the capacitated transshipment model can be utilized effectively to predict signaling pathways. With continuous advancements in technology and high-throughput biological experiments, we intend to implement and improve the capacitated transshipment model presented as well as the cost function of the protein-protein interaction network.

1.2. Organization of the Paper

This paper is organized according to the format that is recommended by the NDSU Department of Computer Science. This paper concentrates on applying an optimization model, the capacitated transshipment model, to a biological problem of predicting signaling pathways.

The algorithm used, problem addressed, and tools used in this paper are sophisticated and complex, and they require a strong background. Chapter 2 focuses on providing the Background for the reader's understanding.

The idea presented in this paper is based on the previous work and accomplishments of other researchers. Chapter 3 addresses the Literature Review which contributes to and contrasts the research conducted.

Chapter 4 explains the presented Approach in detail. It demonstrates the methods and process used to accomplish the objectives.

Chapter 5 explains and evaluates the results of the research. The findings are compared to the approaches presented in Chapter 3 using a different statistical approach for the same problem. Chapter 5 also provides the significance of the obtained results and how to improve the results through future research.

Chapter 6 concludes the research. The chapter explains what has been achieved, suggesting future work and extensions.

CHAPTER 2. BACKGROUND

This chapter provides detailed information about Signaling Pathways, the Capacitated Transshipment Model, and the tools used. This chapter briefly describes a Depth-First Search for the unfamiliar reader's understanding.

2.1. Signaling Pathways

An organism's body is made up of billions of cells. A cell responds to different chemicals present in its environment. There are different types of proteins present in an organism's body. Certain proteins, such as hormones (signal initiator), act as chemical signals which provide instruction to the cell about how to react to changes in the conditions around the cell (Berg, Tymoczko, & Stryer, 2002).

Signal transduction can be defined as a process in which a signal from the cell's environment is received by a cellular component due to interaction with the signaling initiator and then converted, or transduced, into a different form of chemical signal which evokes a certain response (Albert et al., 2007). A signal transduction cascade behaves like a circuit which senses, processes, amplifies, and generates a response to stimulus (Chen & Yuan, 2006). Signal transduction is essential in many biological processes, such as metabolism, cell proliferation, and cell differentiation (Chen & Thorner, 2007). The biological processes involving signal transduction follow certain paths. Finding the signaling pathways is important to establish an understanding of these processes. This information can be used for drug discovery or understanding a disease like cancer (Berg, Tymoczko, & Stryer, 2002).

The biochemical processes that occur in a living cell for its proper functioning are extremely complex. Identifying every reaction and component involved with a simple process can take several years of experimental research (Feiglin, Moulton, Lee, Ofran, & Unger, 2012).

With recent developments and advancements in biotechnological techniques, different types of data have been accumulated. This huge amount of data contains information about various biological processes such as signal transduction. Therefore, the demand for computational methods to mine the details of these processes has increased.

Signal transduction involves many protein interactions, so protein-protein interaction data are a great source of information for pathways (Gitter, Klein-Seetharaman, Gupta, & Bar-Joseph, 2011). Therefore, the protein-protein interaction data can be analyzed to understand the mechanisms of the signaling processes.

In this paper, two processes, pheromone response and filamentation growth, of the mitogen-activated protein kinase (MAPK) signaling network and processes in yeast are studied (Zhao, Wang, Chen, & Aihara, 2008a). The MAPK signaling network is commonly found in eukaryotic cells such as baker's yeast which is known as *Saccharomyces cerevisiae* (Chen & Thorner, 2007). The MAPK signaling network involves five different cascades: pheromone response, cell-wall integrity, high osmolarity, filamentation growth, and spore-wall assembly (Liu & Zhao, 2004). The MAPK signaling pathways in yeast regulate each other. They also interact with other signaling networks to coordinate gene expression (Chen & Thorner, 2007).

2.1.1. Pheromone Response

Yeast cells can be either haploid or diploid. Like humans, yeast also has opposite genders (or mating types), and haploid cells from each of these mating types merge together to form a diploid (Gustin, Albertyn, Alexander, & Davenport, 1998). The release of small proteins (known as pheromones) for mating acts as a signal to prepare cells for mating by producing a cascade of protein interactions (Chen & Thorner, 2007). Responses generated by the stimulus provided by the pheromones are as follows: polarized growth in the direction of a mating partner; cell-cycle

arrest in the G₁ phase; and enhancement of the expression of proteins required for cell adhesion, cell fusion, and nuclear fusion (Gustin, Albertyn, Alexander, & Davenport, 1998).

2.1.2. Filamentation Growth

Diploid yeast cells form pseudohyphae under certain conditions of the culture (Chen & Thorner, 2007). The pseudohyphae are filaments made up of connected and extended cells. This response is generated if the cell is starving for nitrogen along with some other environmental factors which act as the stimulus (Chen & Yuan, 2006). This process affects cell elongation, growth, and selection of the budding site (Gustin, Albertyn, Alexander, & Davenport, 1998).

Figure 1 shows the baker's yeast MAPK signaling pathways for pheromone response and filamentation growth. Figure 1 has been obtained through KEGG's (Zhang & Wiemann, 2009) database.

2.2. Depth-First Search

The depth-first search algorithm is used to search a graph or network starting from a root, or starting, node and going down the tree to find the destination, or ending, node. An evolutionary tree can be considered as a graph with a starting node on top and the child nodes as leaves or branches (Korf, 1985). The depth-first search algorithm starts with the first child from the left, then down to the children on left until there is no child on left, and then, it moves to the second child of the root node on the right. The algorithm moves down from left to right until it finds the ending node or child. In this paper, the depth to which the algorithm searches to get to the ending node is constrained between three and nine (Korf, 1985).

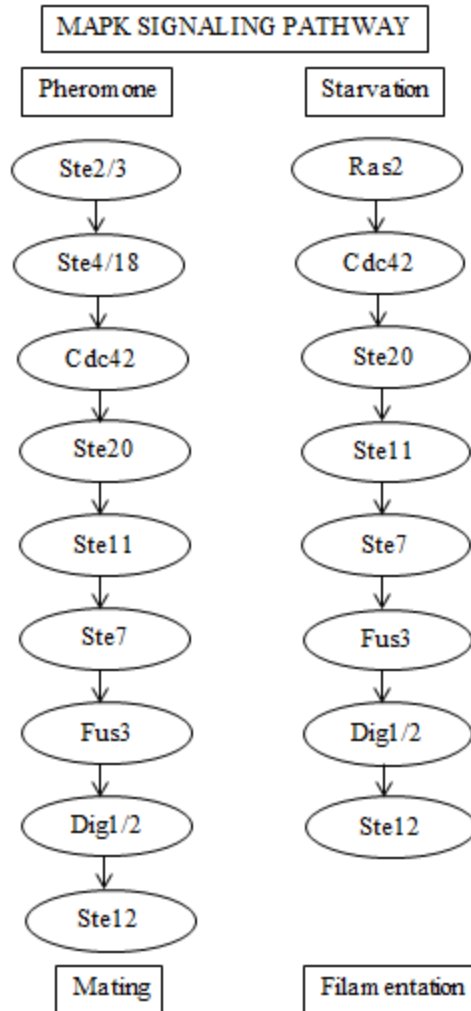


Figure 1. KEGG's Yeast MAPK Signaling Pathways.

2.3. Capacitated Transshipment Problem

A network that consists of the nodes and directed edges that connect them can be used to define network-flow optimization models. Network-flow optimization models are linear programming problems with a specialized structure (Pavlopoulos et al., 2011).

Network-flow optimization models have been applied to different types of data related to computer science; business; industry; and, recently, biology. There are several network-flow

optimization models available to solve large-scale optimization problems in a matter of seconds. The capacitated transshipment problem is a special type of network-flow optimization problem which is considered as a linear programming problem.

2.3.1. Definition of the Capacitated Transshipment Problem

The transportation problem aims to determine an optimal way to satisfy the demand of n nodes using the supplies of m nodes. It allows only shipments that go directly from supply nodes to demand nodes. In many situations, shipments are allowed between supply nodes or between demand nodes. This type of problem is known as a transshipment problem (Bradley, Brown, & Graves, 1977). This paper considers the capacitated transshipment problem as a linear programming problem solved with a primal, simplex computational algorithm.

The capacitated transshipment problem can be defined as a directed network with node set N and arc set A ; $i \in N$, $(i,j) \in A$ (Bradley, Brown, & Graves, 1977). The problem is to find the set of flows that minimize the total cost subject to constraint sets that requires flow balance at each node and a capacity restriction on each arc:

$$\text{Minimize } z = \sum_{(i,j) \in A} c_{ij} x_{ij}$$

Subject to

$$\sum_{i:(i,j) \in A} x_{ij} - \sum_{j:(i,j) \in A} x_{ji} = b_i \quad \text{and}$$

$$0 \leq x_{ij} \leq u_{ij} \quad ,$$

where c_{ij} = cost or unit of commodity flow on arc (i,j) , u_{ij} = capacity (upper bound) for commodity flow on arc (i,j) , b_i = supply of the commodity at node i (interpret negative b_i as a

demand of $-b_i$ units), x_{ij} = commodity flow on arc (i,j) , and z = objective function with an assumption that the total supply and demand are equal.

A primal simplex algorithm can be used to solve the capacitated transshipment problem by handling capacity constraints with upper-bounding techniques (Khurana, Verma, & Arora, 2012). Each arc with x_{ij} has an outflow at node i and an inflow at node j . Therefore, x_{ij} appears exactly in two equations. This matrix is called a node-arc incidence matrix (A). A basis is a collection of $m-1$ arcs where arcs are incident to every one of the m nodes at least once and where the arcs do not form any cycles (a spanning tree). A basis can be used as a starting point, and the primal simplex algorithm can be applied to obtain the optimal solution for the capacitated transshipment problem (Bradley, Brown, & Graves, 1977).

2.3.2. Simplex Algorithm to Solve the Capacitated Transshipment Problem

The capacitated transshipment problem can be considered as a minimum cost-flow problem, so it is a special case of a linear programming problem with some different properties that greatly improve the algorithm's performance. The simplex algorithm is one way to solve the capacitated transshipment problem (Bradley, Brown, & Graves, 1977). An example simplex algorithm is demonstrated below, followed by a step-by-step summary of the algorithm.

2.3.2.1. Example Problem

In the following example, the procedure to find the initial basis solution is not explained. It starts with a given initial-basis spanning tree. The example network is shown in Figure 2 (Nygard, 2009). In the figure, c_{ij} represents the cost or unit of commodity flow on arc (i,j) ; u_{ij} represents the capacity (upper bound) for commodity flow on arc (i,j) ; and positive b_i represents the supply of the commodity at node i while negative b_i represents the supply of the commodity at node i .

It is assumed that the initial, feasible basis tree is selected. Figure 3 (Nygard, 2009) shows the selected initial basis tree in which the number of arcs is one less than the number of the nodes present in the original network with flow that satisfies the condition of conservation of flow and is bounded by an upper and lower bound. The commodity flow on arc (i,j) is represented by x_{ij} in the Figure 3.

Node potentials are dual variables used to quantify the cost of sending flow (c_{ij}) from a node to the root. For this example, node 4 is arbitrarily chosen as the root. For this example, the potentials, π_i , are shown in Figure 4 (Nygard, 2009).

The reduced costs for the non-basic arcs are calculated as follows: $r_{ij} = c_{ij} - \pi_i + \pi_j$. For non-basic arc $(2, 3)$, r_{23} is -1 , and for arc $(3, 4)$, r_{34} is -1 . Both non-basic arcs are attractive to enter. Non-basic arc $(2, 3)$ is chosen randomly to enter. It forms a cycle. Figure 5 (Nygard, 2009) shows the cycle formed and the direction of the cycle. A unique cycle is always created when a non-basic arc is introduced in a basis tree.

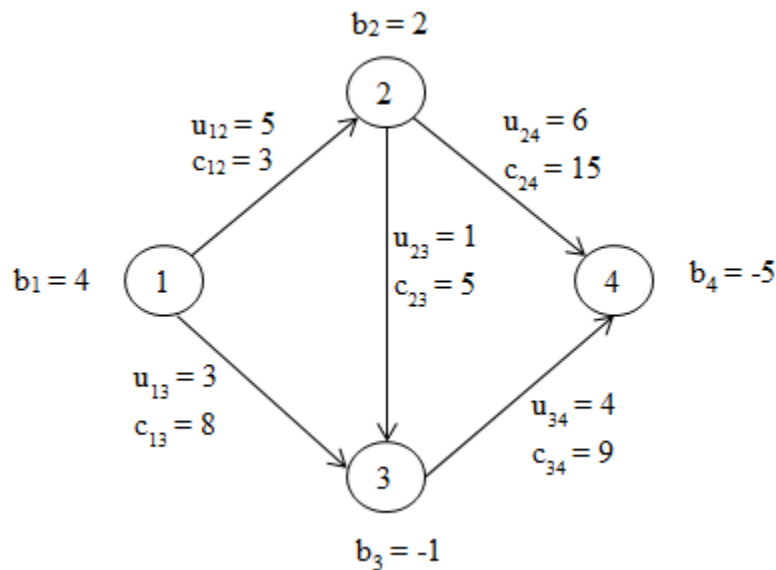


Figure 2. Simplex Algorithm Example Network.

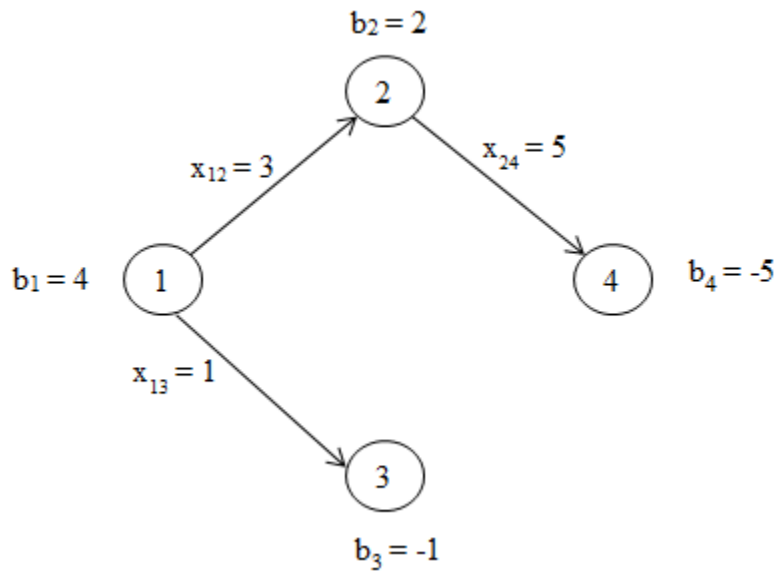


Figure 3. Initial Basis Tree in the Example.

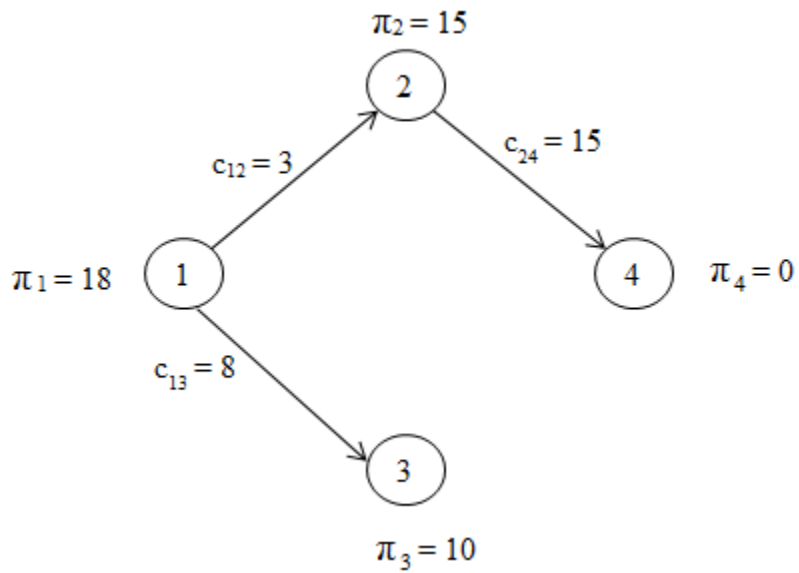


Figure 4. Initial Basis Node Potentials in the Example.

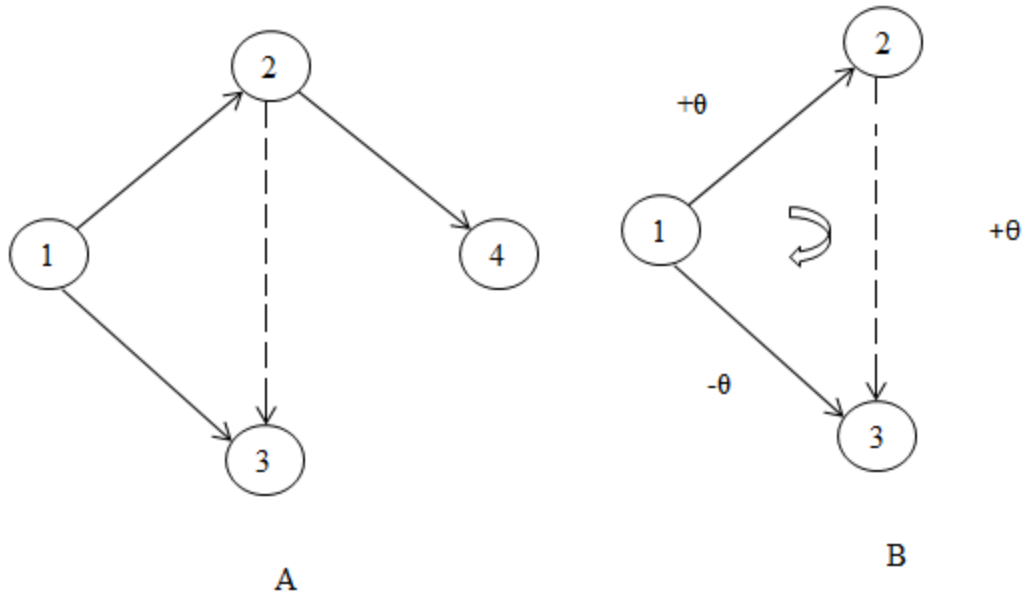


Figure 5. Simplex Algorithm Cycle Formed in the Example.

To find the leaving arc for the above cycle and to maintain the flow on the network, the simplex ratio test is performed in the following way:

$$\theta = \min \{ \text{capacity of } (i, j), \text{ flow: basic arc of the cycle with the opposite orientation of } (i, j), \min [\text{capacity} - \text{flow: basic arc of the cycle with the orientation of } (i, j)] \}$$

$$\theta = \min \{ \text{capacity of } (2, 3), \text{ flow: basic arc of the cycle with the opposite orientation of } (1, 3), \min [\text{capacity} - \text{flow: basic arc of the cycle with the orientation of } (1, 2)] \}$$

$$\theta = \min \{ 1, 1, 2 \}$$

The minimum is 1 for the ratio test. Let us say that x_{13} is leaving the basis and that x_{23} is entering the basis. The new basis tree is shown in Figure 6 (Nygard, 2009).

These above steps are repeated by computing new dual variables and reduced costs. The steps are demonstrated in Figure 7.

Figure 7(A) (Nygard, 2009) shows the new basis with the calculation of the node potentials. Figure 7(B) shows that the reduced cost for arc (3, 4) is -2 and for arc (1, 3) is 1. Arc

(3, 4) is chosen to enter the basis. Figure 7(C) shows the new cycle. In the simplex ratio test, arc (2, 3) leaves the basis with the flow shown in Figure 7(D). Figure 7(E) shows the new basis after the flow adjustments. When reduced costs for the non-basic arcs are calculated in Figure 7(F), arc (1, 3) is chosen to enter the basis because arc (2, 3) is non-basic and has a flow equal to the capacity of the arc. Figure 8(A) (Nygard, 2009) shows the cycle that formed. According to the simplex ratio test, arc (1, 3) leaves the basis and goes to its upper-bound without changing the basis as shown in Figure 8(B), where non-basic arcs with flows are illustrated as dotted lines. In this case, the non-basic arcs have flows equal to the upper-bound of the arcs.

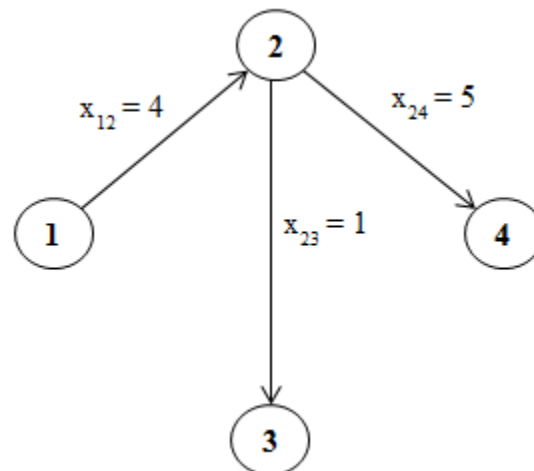
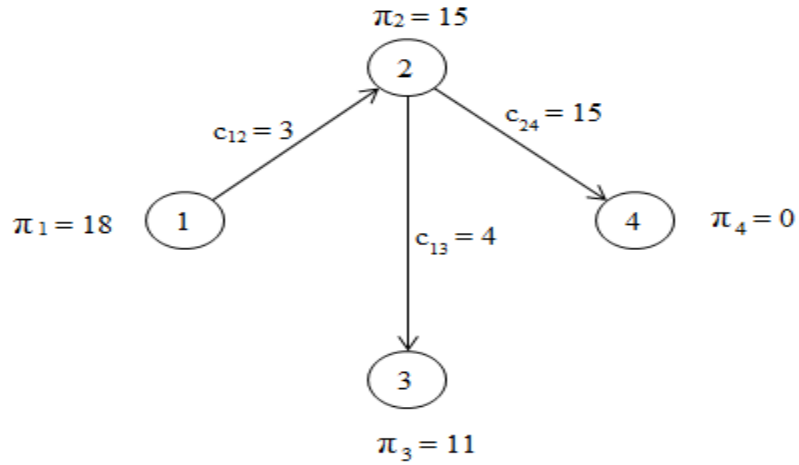


Figure 6. New Basis for the Example.

The current solution obtained, shown in Figure 8(B), is optimal when verified using the flow balance and simplex ratio test. The optimal solution equals 88 with arc (1, 2), arc (2, 4), and arc (3, 4) in the basis, and arcs (1, 3) and (2, 3) are not in the basis.



A

$$r_{13} = c_{13} - \pi_1 - \pi_3 = 8 - 18 + 11 = 1$$

$$r_{34} = c_{34} - \pi_3 - \pi_4 = 9 - 11 + 0 = -2$$

B

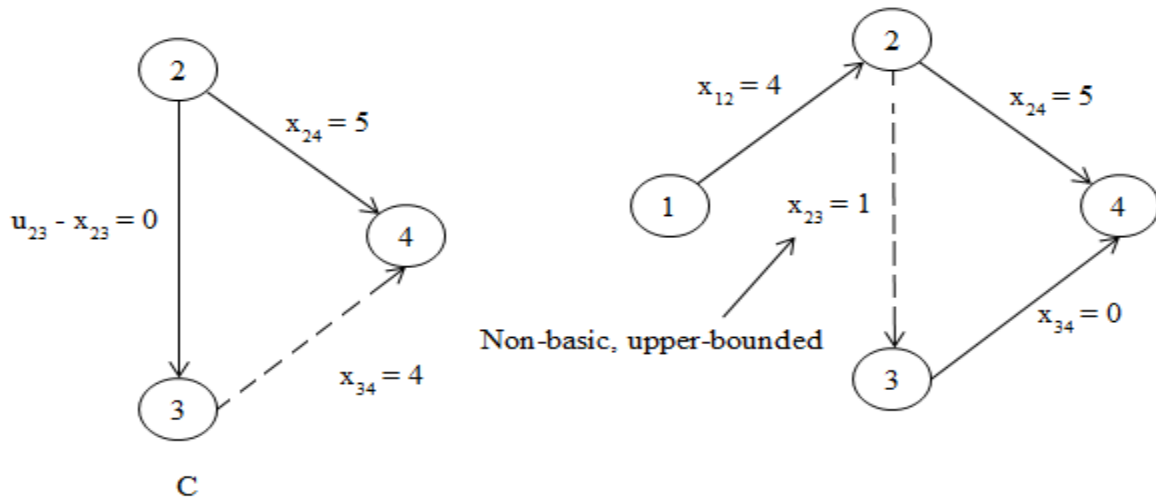


Figure 7. Steps Followed to Achieve Optimal Solution in the Example.

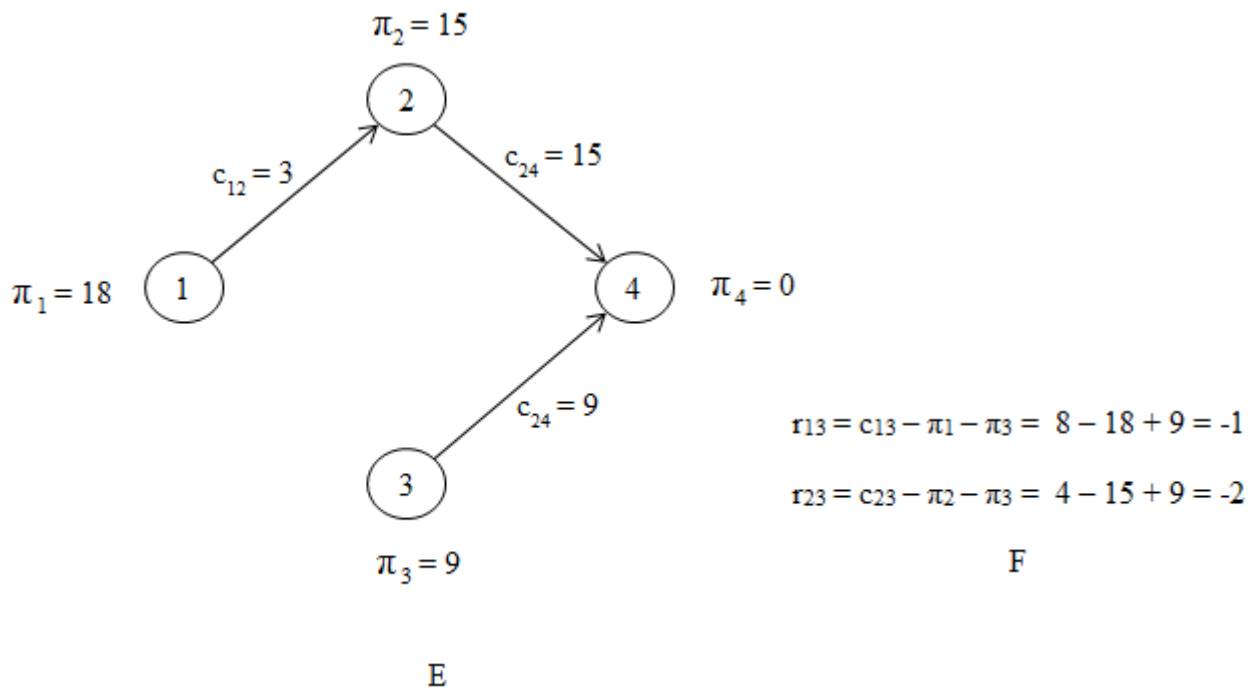


Figure 7 (continued). Steps Followed to Achieve Optimal Solution in the Example.

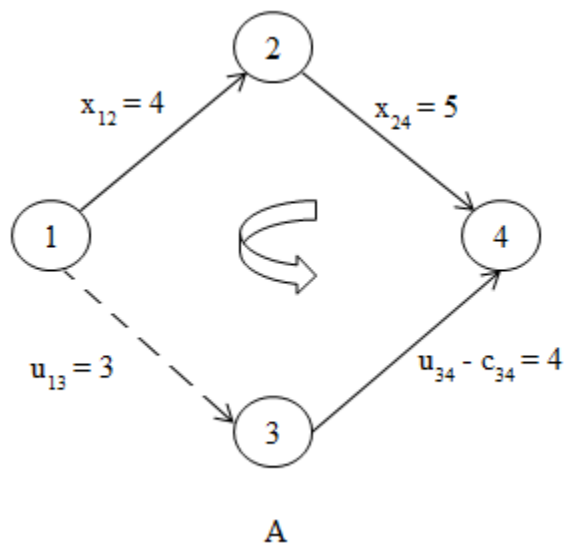


Figure 8. Final Steps of Simplex Algorithm's Solution for the Example.

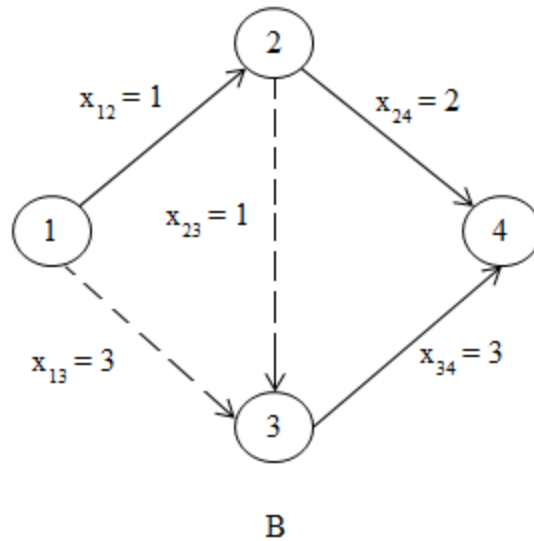


Figure 8 (continued). Final Steps of Simplex Algorithm's Solution for the Example.

The capacitated transshipment problem creates a framework of a general linear programming problem and specializes to the cost-flow simplex network problem. A primal simplex algorithm is used to solve the mathematical problem of finding the minimum total cost-flow objective function in a directed, weighted network bounded by upper-bound and lower-bound on the arcs.

2.3.2.2. Simplex Algorithm Steps

The following steps show the summary of the simplex algorithm applied in the above example. These steps explain how capacitated transshipment problem network can be solved.

2.3.2.2.1. Step 1: Initialization

For initialization of the primal simplex algorithm, a feasible solution is chosen, from the possible feasible solutions to the problem, as a basis tree. The arcs present in this basis tree are known as basic variables/arcs, and other arcs of the network are known as non-basic variables/arcs.

2.3.2.2.2. Step 2: Pricing

Dual variables, known as node potentials, are calculated for all the nodes. For all non-basic arcs, reduced costs are calculated to find the entering arc using the formula $r_{ij} = c_{ij} - \pi_i + \pi_j$. If r_{ij} is greater than or equal to zero for all lower-bounded arcs and if r_{ij} is less than or equal to zero for all the upper-bounded arcs, then the solution is considered optimal. Otherwise, the arc with the lowest reduced cost that is not at its upper-bound is considered as an entering arc.

2.3.2.2.3. Step 3: Find the Leaving Arc

For the cycle formed because of a non-basic arc's entry, the cycle direction is used to find the arc that is considered the leaving arc. The minimum value for function θ is calculated to change the flows of the new basis and to find the blocking, or leaving, arc.

2.3.2.2.4. Step 4: Update

The basis tree is changed, and the flows on the arcs are adjusted and updated. Then, the algorithm returns to Step 2 to check for optimality. This process continues until optimality is reached.

This process, followed with the simplex algorithm, gives an optimal solution for the capacitated transshipment problem. SAS has this model implemented in a procedure known as NETFLOW. On the other hand, OPTMODEL in SAS is used to create models for optimization. In this paper, both of these SAS models are used; they are explained in the next section.

2.4. SAS OPTMODEL and NETFLOW

SAS is business-analysis and business-intelligence software. It is used for information-technology management, human-resource management, financial management, customer-relationship management, and more (SAS Institute, Inc., 2008).

2.4.1. OPTMODEL

The OPTMODEL procedure in SAS constitutes a modeling language combined with alternative solvers for several types of mathematical programming problems. The programming language in the OPTMODEL procedure supports formulating optimization models using symbolic algebra (SAS Institute, Inc., 2008). The OPTMODEL procedure has higher efficiency because it has structured input and output and because the model's data reside in the memory and can be manipulated in a graph form very easily (SAS Institute, Inc., 2008). In this paper, the OPTMODEL procedure is used to develop a model for the simplex algorithm and to solve the capacitated transshipment problem using the linear programming solver.

2.4.2. NETFLOW

The NETFLOW procedure in SAS consists of an algorithm that exploits a specialized version of the simplex method to solve network problems (SAS Institute, Inc., 2008). This algorithm is used to find flow on each arc in a network such that the total cost of the flow is minimized and that the conservation of flow is satisfied (SAS Institute, Inc., 2008). Conservation of flow means that the supply at the node plus inflow through the arcs directed towards the node are equal to the demand at the node plus outflow through the arcs directed away from the node (Nygard, 2009). In this paper, the NETFLOW procedure is used to confirm the solution obtained through the optimization model built with the OPTMODEL procedure.

2.4.3. Example

Figure 9 shows the format for data input in SAS. Figure 9 represents the data of the example for the capacitated transshipment problem presented in this chapter. Data node0 represents the node-specific data required by the model. The first column represents the node name/number while the second column represents the supply or demand on that node. Data arc0

represents the arc-specific data. The first column represents the tail of the arc/end of the arc without the arrow head. On the other hand, the second column represents the head of the arc/end of the arc with the arrow head. The third column represents the cost per unit flow through that arc. The fourth column represents the upper-bound of the flow through that arc.

```

data nodes;
  format node $16. ;
  input node $ supdem;
  datalines;
1 4
2 2
3 -1
4 -5
;
data network;
  format from $16. to $16. cost 8. _capac_;
  input from $ to $ cost $ _capac_ ;
  datalines;
1 2 3 5
1 3 8 3
2 3 4 1
2 4 15 6
3 4 9 4
;

```

Figure 9. Example Data Input in SAS.

Figure 10 shows the results of the OPTMODEL procedure for the example data. The OPTMODEL solution gives a solution summary showing the optimal value of the objective function obtained. The objective value obtained going through one iteration of the dual simplex algorithm is 88. The flow on arcs is printed to compare the results obtained through the OPTMODEL procedure to the results shown in this chapter. The first column of the flow table represents the tail of the arcs; the second column of the flow table represents the head of the arcs; and the third column represents the flow on the arcs.

The OPTMODEL Procedure

Solution Summary	
Solver	Dual Simplex
Objective Function	obj
Solution Status	Optimal
Objective Value	88
Iterations	1
Primal Infeasibility	0
Dual Infeasibility	0
Bound Infeasibility	0

[1]	[2]	Flow
1	2	1
1	3	3
2	3	1
2	4	2
3	4	3

Figure 10. Solution Obtained Through OPTMODEL Procedure in SAS for the Example Data.

CHAPTER 3. LITERATURE REVIEW

Protein-protein interaction data contain information about signaling pathways and can be exploited to understand the mechanisms of signal transduction. Many algorithms have been developed and used to find signaling pathways. Some methods are described in this chapter.

3.1. Color Coding

The color-coding technique is used for finding simple paths or cycles of a certain length in a given network. It has been extended to solve biological problems (Scott, Ideker, Karp, & Sharan, 2006).

A protein-protein interaction network graph, G , with N nodes and E edges in which each node is a protein and each edge is the interaction between the proteins is considered to be a weighted interaction network such that each edge has a score that the interaction between the proteins exists. A weight can be assigned to the simple path of a given length in this network graph, G , which is equal to the product of scores assigned to the edges in that path (Hu, Yan, Huang, Han, & Zhou, 2005). The main objective of the color-coding method is to find the highest scoring paths. To adjust the color-coding technique to find signaling pathways, the weight on the edges is considered a negative logarithm of the original weight, so it can be used to calculate the sum of weights for the path. The new objective of this network is minimum weight paths (Scott, Ideker, Karp, & Sharan, 2006).

The main idea behind the color-coding technique is to assign each node a random color and to search for paths which have distinct colors rather than searching for distinct nodes (Scott, Ideker, Karp, & Sharan, 2006). This process reduces the complexity of the standard dynamic programming algorithm used to find simple paths. The color-coding algorithm needs to repeat these random trials. It is adjusted for application to a protein-protein interaction network by

restricting the proteins that can occur in a path and by restricting the order in which proteins occur on a path (Scott, Ideker, Karp, & Sharan, 2006).

3.2. NetSearch

NetSearch is an algorithm which is used to model signaling networks through protein-protein interaction data. The protein-protein interaction data are considered an un-weighted interaction network (Steffen, Petti, Aach, D'haeseleer, & Church, 2002). The NetSearch algorithm searches for paths of length that are eight or less and that begin at membrane proteins and end in transcription factors. These paths are usually found in millions, so they need to be scored; for this process, NetSearch uses a k-means algorithm to cluster genes based on their expression profiles under the sumprob scoring metric (Steffen, Petti, Aach, D'haeseleer, & Church, 2002).

The main idea behind NetSearch is to find a sub-network for a particular biological process (Steffen, Petti, Aach, D'haeseleer, & Church, 2002). Therefore, the sub-network is used to minimize the scope of the signaling pathways that needs to be established.

3.3. PathFinder

PathFinder is an algorithm designed for mining protein-protein interaction networks to extract signaling pathways (Bebek & Yang, 2007). The PathFinder algorithm collects functional annotations for known proteins to elucidate characteristics of known signaling pathways (Bebek & Yang, 2007). These characteristics are used to find unknown signaling pathways.

Association-rule mining is a process utilized to obtain the attributes which are used to determine patterns on these pathways (Bebek & Yang, 2007). The main idea behind this algorithm is to find unknown signaling pathways and to ensure that the proteins have a strong association/interaction with each other on these pathways (Bebek & Yang, 2007).

3.4. Integer Linear Programming

Integer linear programming has been used for solving many problems in a wide variety of fields. We discuss one of the integer linear-programming models designed to predict signaling pathways which is closest to the capacitated transshipment model presented in this paper.

A protein-protein interaction network graph, G , with N nodes and E edges in which each node is a protein and each edge is the interaction between the proteins is considered to be an undirected, weighted network where the weight represents the confidence score of the interaction between the proteins or the correlation coefficient obtained through gene-expression data (Allena, Fetrowb, Daniel, Thomas, & John, 2006). In this weighted network, using a starting and ending node, a linear pathway, with a weight equal to the sum of weights for the edges in the path, can be discovered (Baitaluk, Qian, Godbole, Raval, Ray, & Gupta, 2006). To discover this pathway, a sub-network, described as a signal transduction network, is deduced to limit the number of irrelevant interactions and proteins searched because many of them are not involved with the particular biological process in question; a depth-first search algorithm which prunes for lengths greater than nine is used (Zhao, Wang, Chen, & Aihara, 2008b).

Figure 11 (Zhao, Wang, Chen, & Aihara, 2008b) shows the algebraic representation of the integer linear-programming model. The integer linear-programming model utilized to discover the linear signaling pathways using this graph, G , is described in the following paragraph.

The objective function to be minimized is a negative summation of the product of the weight on the edge, w_{ij} , and the binary variable, y_{ij} , that denotes that the edge is the part of the pathway or not a part of the pathway which is added to the product of a penalty, λ , and summation of the binary variable, y_{ij} (Zhao, Wang, Chen, & Aihara, 2008b). The given objective

function is subjected to a number of constraints to define this issue for finding the signaling pathways as a type of integer linear programming problem.

$$\text{Minimize}_{\{x_i, y_{ij}\}} S = - \sum_{i=1}^{|V|} \sum_{j=1}^{|V|} w_{ij} y_{ij} + \sum_{i=1}^{|V|} \sum_{j=1}^{|V|} y_{ij}$$

Subject to:

$$y_{ij} \leq x_i$$

$$y_{ij} \leq x_j$$

$$\sum_{j=1}^{|V|} y_{ij} \geq 1 \quad \text{if } i \text{ is either a starting or ending protein}$$

$$\sum_{j=1}^{|V|} y_{ij} \geq 2x_i \quad \text{if } i \text{ is not a starting or ending protein}$$

$$x_i = 1$$

$$x_i \in \{0, 1\}, i = 1, 2, \dots, |V|$$

$$y_{ij} \in \{0, 1\}, i, j = 1, 2, \dots, |V|$$

Figure 11. Algebraic Representation of the Integer Linear Programming Model.

The model presented in Figure 11 is constrained. The constraints are as follows:

- 1) If protein i is a starting or ending protein, then the summation of binary variable y_{ij} is greater than or equal to 1, where $|V|$ denotes the total number of proteins involved in the

considered network. This constraint ensures that there is at least one edge connected to the ending and starting protein in the pathway (Zhao, Wang, Chen, & Aihara, 2008b).

- 2) If protein i is a starting or ending protein, then the summation of binary variable y_{ij} is greater than or equal to twice the value of binary variable x_i for protein i to denote whether protein i is in the pathway. This constraint ensures that binary variable x_i has at least two edges linked to it once it has been selected to be part of the pathway (Zhao, Wang, Chen, & Aihara, 2008b).
- 3) If and only if proteins i and j are selected to be part of the pathways, then the edge between them is considered to be part of the pathway (Zhao, Wang, Chen, & Aihara, 2008b).

This model helps to find the signaling pathway or pathways using protein-protein interaction data. This concept has been used as a basis to describe the protein-protein interaction network as a capacitated transshipment problem, a method which is explained in the next chapter.

CHAPTER 4. APPROACH

The objective of this paper is to apply the capacitated transshipment model to protein-protein interaction data of baker's yeast to predict signaling pathways. This approach is presented in two main steps. The first step is to classify signaling pathways' discovery in protein-protein interaction data as a capacitated transshipment problem; the second step is to use the SAS OPTMODEL to develop a capacitated transshipment model to solve the problem at hand.

4.1. Definition of the Problem

The protein-protein interaction network of baker's yeast is considered as a bi-directed graph with N nodes (proteins) and E edges (interactions). Each edge is an ordered pair of nodes (tail, head). These edges can be assigned a cost per unit flow, a lower bound on flow, and an upper-bound on flow (also known as capacity) (Kestler, Wawra, Kracher, & Kuhl, 2008). This network can be considered a special case of the capacitated transshipment problem. The starting and ending proteins involved in signal transduction are marked as supply and demand nodes while the other nodes have the supply/demand equal to zero where signal units enter or leave, or are transshipped. The problem is to minimize total cost with flows, where flows are in between the associated lower bounds and capacities, and to balance inflow and outflow at each node.

4.1.1. Steps Followed to Organize the Data as a Capacitated Transshipment Problem

First, the protein-protein interaction data for baker's yeast are obtained from the Database of Interaction Proteins (DIP) (Xenarios, Salwinski, Duan, Higney, Kim, & Eisenber, 2002). These protein-protein interaction data include 5,103 proteins and 24,247 interactions (Xenarios, Salwinski, Duan, Higney, Kim, & Eisenber, 2002). This idea of considering a undirected arc as a bi-directed arc implies that the main network formed using protein-protein interaction data

consists of 5,103 nodes and 48,494 arcs because the 24,247 interactions are considered bi-directional (two arcs in different directions for each interaction).

Then, for both the pheromone response pathway and the filamentation growth pathway, the starting and ending proteins are used to conduct a depth-first search to find pathways from the starting proteins to the ending proteins. Pathways with a length greater than nine and less than three are pruned to obtain a smaller network of proteins to reduce the scope of the problem.

Figure 12 shows a sub-network for the pheromone response in baker's yeast, and Figure 13 shows a sub-network for filamentation growth in baker's yeast after applying the customized depth-first search algorithm.

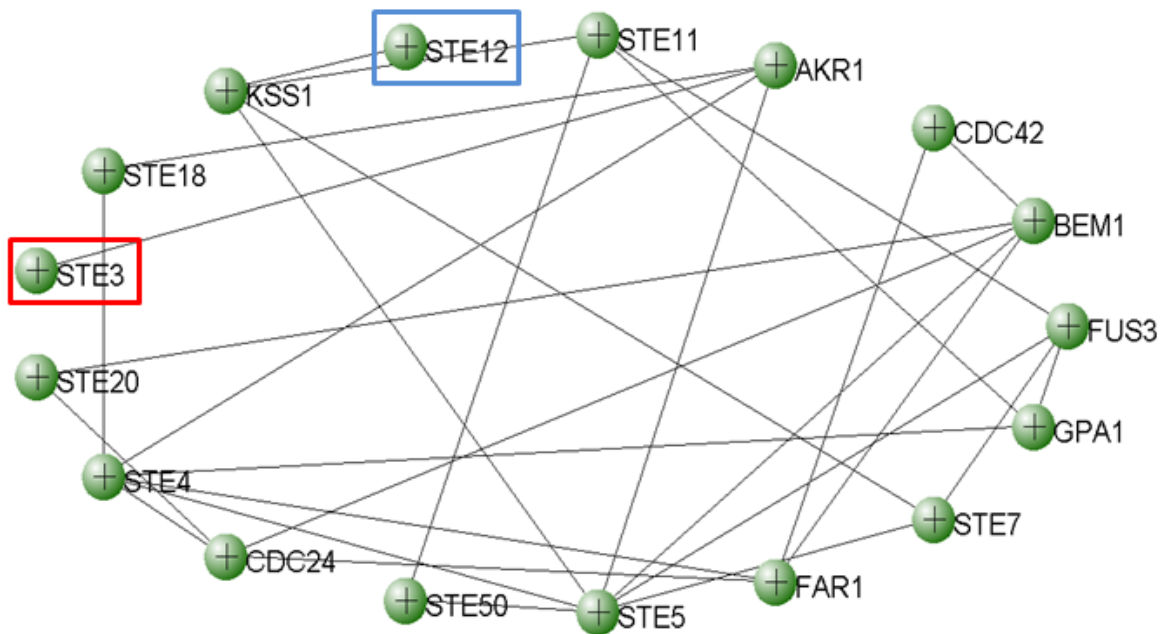


Figure 12. Pheromone Response Sub-Network.

The starting proteins are given a supply of 1, and the ending nodes are assigned a demand of -1. The starting proteins in the sub-networks are marked in red, and the ending proteins are

marked in blue. The sub-network obtained for these biological processes shows undirected edges; each of these edges can be represented as two directional edges from protein i to j and from protein j to i. Therefore, this sub-network is considered as a bi-directional network of protein-protein interactions.

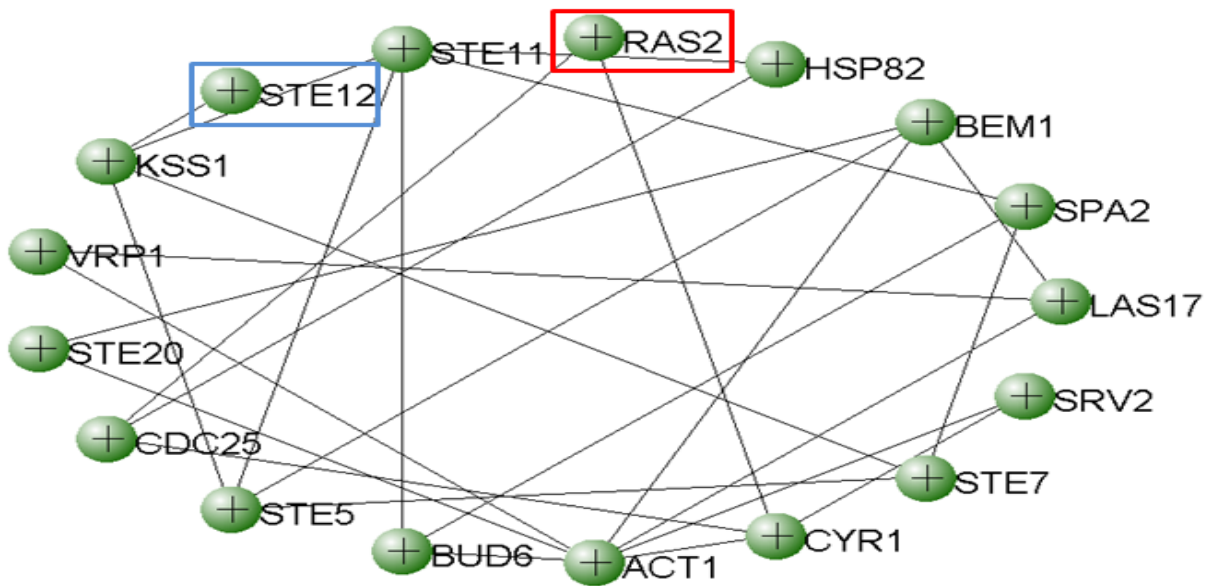


Figure 13. Filamentation Growth Sub-Network.

The sub-networks obtained are then used as input in the STRING (Szklarczyk et al., 2011) database to obtain the score of the edges involved in these sub-networks. The output data obtained from the STRING database are then downloaded in a tab-delimited text format and read using Microsoft Excel 2010. The edges are doubled by changing the head and tail components to make the graph bi-directional.



The cost of the edges is considered to be a function of the edges' score obtained from the STRING (Szklarczyk et al., 2011) database. The cost of the edges is defined as $1000 * (1 - (\text{score} - 0.4))$.

of the edges)) and is calculated in Microsoft Excel 2010. Figure 14 shows a screen with the data compiled in Microsoft Excel and used as input in SAS (refer to Section 4.2). The capacity of the edges is considered as 1 and the lower bound as 0 such that the flow on the edge is a binary variable which shows whether an edge is in the pathway or not. It is possible to have multiple alternative, optimal solutions, so adjustments should be made to the capacitated transshipment problem to obtain one or the other alternative optimal solution at a time.









In this paper, the data for both pheromone response and filamentation growth show that there are two alternative optimal solutions. These solutions are shown in Chapter 5.

4.2. Definition of the Model

Figure 14 shows the input data for the yeast's MAPK pheromone response signaling pathway. Figure 14(A) shows the node-specific data, and Figure 14(B) shows the arc-specific data. As in Figure 12, in Figure 14(A), the first column represents a protein involved in yeast's MAPK pheromone response while the second column represents the assigned supply or demand to that protein. As in Figure 12, in Figure 14(B), the first column represents the tail protein, and the second column represents the head protein of the interaction between two proteins, x and y , such that there are two arcs $x \rightarrow y$ and $y \rightarrow x$. The third column represents the cost per unit flow through that interaction such that, for both arcs $x \rightarrow y$ and $y \rightarrow x$, the cost is the same. The fourth column represents the upper-bound of the flow through these arcs.

	 _node_	 _supdem_
1	STE3	1
2	AKR1	0
3	STE18	0
4	STE4	0
5	FAR1	0
6	CDC24	0
7	STE20	0
8	BEM1	0
9	GPA1	0
10	STE5	0
11	STE50	0
12	STE11	0
13	KSS1	0
14	CDC42	0
15	FUS3	0
16	STE7	0
17	STE12	-1



A

	 _tail_	 _head_	 _cost_	 _capac_		 _tail_	 _head_	 _cost_	 _capac_
1	STE3	AKR1	177	1	31	STE18	AKR1	183	1
2	AKR1	STE18	183	1	32	STE4	AKR1	2	1
3	AKR1	STE4	2	1	33	STE5	AKR1	183	1
4	AKR1	STE5	183	1	34	STE4	STE18	2	1
5	STE18	STE4	2	1	35	FAR1	STE4	34	1
6	STE4	FAR1	34	1	36	CDC24	STE4	2	1
7	STE4	CDC24	2	1	37	STE5	STE4	20	1
8	STE4	STE5	20	1	38	GPA1	STE4	10	1
9	STE4	GPA1	10	1	39	CDC24	FAR1	2	1
10	FAR1	CDC24	2	1	40	BEM1	FAR1	2	1
11	FAR1	BEM1	2	1	41	CDC42	FAR1	35	1
12	FAR1	CDC42	35	1	42	BEM1	CDC24	1	1
13	CDC24	BEM1	1	1	43	STE20	CDC24	216	1
14	CDC24	STE20	216	1	44	BEM1	STE20	1	1
15	STE20	BEM1	1	1	45	CDC42	BEM1	1	1
16	BEM1	CDC42	1	1	46	FUS3	GPA1	2	1
17	GPA1	FUS3	2	1	47	STE11	GPA1	7	1
18	GPA1	STE11	7	1	48	STE7	STE5	2	1
19	STE5	STE7	2	1	49	STE50	STE5	6	1
20	STE5	STE50	6	1	50	FUS3	STE5	2	1
21	STE5	FUS3	2	1	51	STE11	STE50	2	1
22	STE50	STE11	2	1	52	KSS1	STE5	2	1
23	STE5	KSS1	2	1	53	KSS1	STE11	2	1
24	STE11	KSS1	2	1	54	STE12	KSS1	2	1
25	KSS1	STE12	2	1	55	STE5	BEM1	6	1
26	BEM1	STE5	6	1	56	FUS3	STE7	2	1
27	STE7	FUS3	2	1	57	STE11	FUS3	2	1
28	FUS3	STE11	2	1	58	KSS1	STE7	2	1
29	STE7	KSS1	2	1					
30	AKR1	STE3	177	1					

B

Figure 14. Input Data for Yeast's MAPK Phermone Response Signaling Pathway.

Figure 15 shows the input data for the yeast's MAPK filamentation growth signaling pathway. Figure 15(A) shows the node-specific data, and Figure 15(B) shows the arc-specific data. As in Figure 14(A), in Figure 15(A), the first column represents a protein involved in yeast's MAPK filamentation growth while the second column represents the assigned supply or demand to that protein. As in Figure 14(B), in Figure 15(B), the first column represents the tail protein, and the second column represents the head protein of the interaction between two proteins, x and y , such that there are two arcs, $x \rightarrow y$ and $y \rightarrow x$. The third column represents the cost per unit flow through that interaction such that, for both arcs $x \rightarrow y$ and $y \rightarrow x$, the cost is the same. The fourth column represents the upper-bound of the flow through these arcs.

	 _node_	 _supdem_
1	RAS2	1
2	CYR1	0
3	SRV2	0
4	ACT1	0
5	VRP1	0
6	LAS17	0
7	STE20	0
8	BUD6	0
9	CDC25	0
10	BEM1	0
11	SPA2	0
12	STE7	0
13	HSP82	0
14	STE11	0
15	STE5	0
16	KSS1	0
17	STE12	-1

A

Figure 15. Input Data for Yeast's MAPK Filamentation Growth Signaling Pathway.

	tail	_head_	_cost_	_capac_		_tail_	_head_	_cost_	_capac_
1	RAS2	CDC25	2	1	31	CDC25	CYR1	173	1
2	RAS2	CYR1	2	1	32	ACT1	CYR1	62	1
3	CYR1	SRV2	2	1	33	ACT1	SRV2	1	1
4	CYR1	CDC25	173	1	34	VRP1	ACT1	2	1
5	CYR1	ACT1	62	1	35	BUD6	ACT1	6	1
6	SRV2	ACT1	1	1	36	LAS17	ACT1	2	1
7	ACT1	VRP1	2	1	37	BEM1	ACT1	16	1
8	ACT1	BUD6	6	1	38	STE20	ACT1	81	1
9	ACT1	LAS17	2	1	39	LAS17	VRP1	1	1
10	ACT1	BEM1	16	1	40	BEM1	LAS17	4	1
11	ACT1	STE20	81	1	41	BEM1	STE20	1	1
12	VRP1	LAS17	1	1	42	SPA2	BUD6	183	1
13	LAS17	BEM1	4	1	43	HSP82	CDC25	173	1
14	STE20	BEM1	1	1	44	STE5	BEM1	6	1
15	BUD6	SPA2	183	1	45	STE7	SPA2	183	1
16	CDC25	HSP82	10	1	46	STE5	STE7	1	1
17	BEM1	STE5	6	1	47	KSS1	STE7	2	1
18	SPA2	STE7	183	1	48	STE11	BUD6	183	1
19	STE7	STE5	1	1	49	STE11	HSP82	1	1
20	STE7	KSS1	2	1	50	STE11	SPA2	47	1
21	BUD6	STE11	183	1	51	STE5	STE11	2	1
22	HSP82	STE11	1	1	52	KSS1	STE11	2	1
23	SPA2	STE11	47	1	53	KSS1	STE5	2	1
24	STE11	STE5	2	1	54	STE12	KSS1	2	1
25	STE11	KSS1	20	1					
26	STE5	KSS1	2	1					
27	KSS1	STE12	2	1					
28	CDC25	RAS2	2	1					
29	CYR1	RAS2	2	1					
30	SRV2	CYR1	2	1					

B

Figure 15 (continued). Input Data for Yeast’s MAPK Filamentation Growth Signaling Pathway.

The capacitated transshipment model created using OPTMODEL in SAS is described in Figure 16. Figure 16 represents the algebraic representation of the capacitated transshipment model. An algebraic model is a set of algebraic equations which explicitly or implicitly describes the solution to a problem. The algebraic equations in this model are represented according to the syntax provided by SAS OPTMODEL. In the algebraic equation, “con balance {i in NODES}: sum {<(i,j)> in ARCS} Flow[i,j] - sum {<j,(i)> in ARCS} Flow[j,i] = _supdem_[i]”, con represents constraint; balance {i in NODES} represents the variable balance at each node where i is the index of nodes; sum {<(i,j)> in ARCS} Flow[i,j] represents the summation of flow on arcs

(i, j) where i and j are indexes of tail and head of the arcs; $\sum \{ \langle j, (i) \rangle \text{ in ARCS} \} \text{Flow}[j, i]$ represents the summation of flow on arcs (j, i) where j and i are indexes of tail and head of the arcs; and $\text{supdem}[i]$ represents the supply or demand on node i.

```

proc optmodel;
  /* Set of nodes */
  set <str> NODES;
  /* Set of supply and demand at nodes where default value is 0 */
  num _supdem_ {NODES} init 0;
  /* Data for nodes and their supply and demand is obtained from data defined as node0 */
  read data node1 into NODES=[_node_]_supdem_;
  /* Set of arcs */
  set <str,str> ARCS;
  /* Set of capacity of arcs where default value is the maximum integer value */
  num _capac_ {ARCS} init .;
  /* Set of cost of unit flow on arcs */
  num _cost_ {ARCS};
  /* Data for arcs and their cost and capacity is obtained from data defined as arc0 */
  read data arc1 nomiss into ARCS=[_tail_]_head_]_capac_]_cost_];
  /* This shows that the arcs in the set ARCS involves only nodes in set NODES */
  NODES = NODES union (union {<i,j> in ARCS} {i,j});
  /* Flow on the arcs have lower bound = 0 */
  var Flow {<i,j> in ARCS} >= 0;
  /* Flow on the arcs have upper bound = capacity on the arcs */
  for {<i,j> in ARCS: _capac_[i,j] ne .} Flow[i,j].ub = _capac_[i,j];
  /* Objective function of cost-flow which is minimized */
  min obj = sum {<i,j> in ARCS} _cost_[i,j] * Flow[i,j];
  /* Conservation of flow where flow in and flow out at a node are equal*/
  con balance {i in NODES}: sum {<(i),j> in ARCS} Flow[i,j] - sum {<j,(i)> in ARCS}
Flow[j,i] = _supdem_[i];
  /*basic linear programming solver used to solve the given problem*/
  solve;

```

Figure 16. Definition of the Capacitated Transshipment Model.

With the SAS OPTMODEL procedure, a model can be defined with one or more declarations of variables, objective functions, constraints, and declarations in mathematical form to solve the linear programming statements. The following paragraph shows the model created to solve the capacitated transshipment model as an optimization or mathematical problem.

The model reads the data input for the node and arcs, considering that the arcs in the arc-specific (arc0/arc1) data consist of only the nodes represented in the node-specific data (node0/node1). The model represented above ensures that the arcs are formed of only nodes present in the node-specific data. The variable flows (Flow $\langle i,j \rangle$) on the arcs are assumed to have a lower bound of 0 and an upper-bound equal to the capacity ($_capac_ [i,j]$) of the arc from the arc-specific data. The objective function minimizes the summation of the product of cost per unit flow on the arcs and the actual flow through the arcs. In Figure 16, the objective function is represented by “obj”. The model is based on the capacitated transshipment model in Chapter 2, and the algebraic model represented here also follows the constraint of conserving flow balance; that is, the inflow and outflow to and from a node are equal to the supply or demand at the node. One of the basic solvers in SAS OPTMODEL is used to obtain the results which are presented in Chapter 5.

CHAPTER 5. RESULTS AND EVALUATION

The proposed method was tested using two experiments to detect signaling pathways in the sub-network of yeast's MAPK pheromone response and filamentation growth. Figures 17 and 18 show the solutions of the OPTMODEL procedure for yeast's MAPK pheromone response and filamentation growth, respectively. The OPTMODEL solution presented in Figures 17 and 18 shows the solution summary with the optimal values of the objective function obtained. The objective value of yeast's MAPK pheromone response obtained by going through 14 iterations of the dual simplex algorithm is 192 while the objective value of yeast's MAPK filamentation growth obtained by going through 16 iterations of the dual simplex algorithm is 19. The flow on arcs is printed to compare the results obtained through the OPTMODEL procedure to the results shown. The first column of the flow table represents the tail of the arcs; the second column of the flow table represents the head of the arcs; and the third column represents the flow on the arcs. The flow 0 represents the arc that is not included in the pathway obtained. On the other hand, flow 1 represents that the arc is included in the resultant signaling pathways.

Figures 19 and 20 show the solutions of the NETFLOW procedure for yeast's MAPK pheromone response and filamentation growth. The NETFLOW solution presented in Figures 19 and 20 show the solution summary with the optimal values of the objective function obtained. The objective value of yeast's MAPK pheromone response obtained by going through summation of the product of cost and flow for the arcs is 192 while the objective value of yeast's MAPK filamentation growth obtained by going through summation of the product of cost and flow for the arcs is 19. The flow on arcs is printed to compare the results obtained through the OPTMODEL procedure to the results shown. The first column of the flow table represents the tail of the arcs; the second column of the flow table represents the head of the arcs; the third

column represents the cost of per unit flow on the arcs; the fourth column represents the capacity of flow on the arcs; the fifth column represents the flow on the arcs; and the last column represents the product of the third column for arc cost and the fifth column for arc flow. The flow 0 represents the arc that is not included in the pathway obtained. On the other hand, flow 1 represents that the arc is included in the resultant signaling pathways.

Solution Summary		[1]	[2]	Flow	[1]	[2]	Flow
Solver	Dual Simplex	AKR1	STE18	0	KSS1	STE7	0
Objective Function	obj	AKR1	STE3	0	STE11	FUS3	0
Solution Status	Optimal	AKR1	STE4	1	STE11	GPA1	0
Objective Value	192	AKR1	STE5	0	STE11	KSS1	0
Iterations	14	BEM1	CDC24	0	STE11	STE50	0
Primal Infeasibility	0	BEM1	CDC42	0	STE12	KSS1	0
Dual Infeasibility	0	BEM1	FAR1	0	STE18	AKR1	0
Bound Infeasibility	0	BEM1	STE20	0	STE18	STE4	0
		BEM1	STE5	1	STE20	BEM1	0
		CDC24	BEM1	1	STE20	CDC24	0
		CDC24	FAR1	0	STE3	AKR1	1
		CDC24	STE20	0	STE4	AKR1	0
		CDC24	STE4	0	STE4	CDC24	1
		CDC42	BEM1	0	STE4	FAR1	0
		CDC42	FAR1	0	STE4	GPA1	0
		FAR1	BEM1	0	STE4	STE18	0
		FAR1	CDC24	0	STE4	STE5	0
		FAR1	CDC42	0	STE5	AKR1	0
		FAR1	STE4	0	STE5	BEM1	0
		FUS3	GPA1	0	STE5	FUS3	0
		FUS3	STE11	0	STE5	KSS1	1
		FUS3	STE5	0	STE5	STE4	0
		FUS3	STE7	0	STE5	STE50	0
		GPA1	FUS3	0	STE5	STE7	0
		GPA1	STE11	0	STE50	STE11	0
		GPA1	STE4	0	STE50	STE5	0
		KSS1	STE11	0	STE7	FUS3	0
		KSS1	STE12	1	STE7	KSS1	0
		KSS1	STE5	0	STE7	STE5	0

Figure 17. An OPTMODEL Solution for Yeast’s MAPK Pheromone Response Pathway.

Figures 21 and 22 show the results of three different models applied to the protein-protein interaction data for baker’s yeast to find the MAPK signaling pathways for pheromone response and filamentation growth, respectively. These results are graphical representations of the pathways obtained in Figures 17-20.

Solution Summary		[1]	[2]	Flow	[1]	[2]	Flow
Solver	Dual Simplex	ACT1	BEM1	0	LAS17	BEM1	0
Objective Function	obj	ACT1	BUD6	0	LAS17	VRP1	0
Solution Status	Optimal	ACT1	CYR1	0	RAS2	CDC25	1
Objective Value	19	ACT1	LAS17	0	RAS2	CYR1	0
Iterations	16	ACT1	SRV2	0	SPA2	BUD6	0
Primal Infeasibility	0	ACT1	STE20	0	SPA2	STE11	0
Dual Infeasibility	0	ACT1	VRP1	0	SPA2	STE7	0
Bound Infeasibility	0	BEM1	ACT1	0	SRV2	ACT1	0
		BEM1	LAS17	0	SRV2	CYR1	0
		BEM1	STE20	0	STE11	BUD6	0
		BEM1	STE5	0	STE11	HSP82	0
		BUD6	ACT1	0	STE11	KSS1	0
		BUD6	SPA2	0	STE11	SPA2	0
		BUD6	STE11	0	STE11	STE5	1
		CDC25	CYR1	0	STE12	KSS1	0
		CDC25	HSP82	1	STE20	ACT1	0
		CDC25	RAS2	0	STE20	BEM1	0
		CYR1	ACT1	0	STE5	BEM1	0
		CYR1	CDC25	0	STE5	KSS1	1
		CYR1	RAS2	0	STE5	STE11	0
		CYR1	SRV2	0	STE5	STE7	0
		HSP82	CDC25	0	STE7	KSS1	0
		HSP82	STE11	1	STE7	SPA2	0
		KSS1	STE11	0	STE7	STE5	0
		KSS1	STE12	1	VRP1	ACT1	0
		KSS1	STE5	0	VRP1	LAS17	0
		KSS1	STE7	0			
		LAS17	ACT1	0			

Figure 18. An OPTMODEL Solution for Yeast's MAPK Filamentation Growth Pathway.

Obs	from	to	cost	capac	FLOW	FCOST	Obs	from	to	cost	capac	FLOW	FCOST
1	STE3	AKR1	177	1	1	177	31	GPA1	STE11	7	1	0	0
2	STE18	AKR1	183	1	0	0	32	STE50	STE11	2	1	0	0
3	STE4	AKR1	2	1	0	0	33	FUS3	STE11	2	1	0	0
4	STE5	AKR1	183	1	0	0	34	KSS1	STE11	2	1	0	0
5	FAR1	BEM1	2	1	0	0	35	KSS1	STE12	2	1	1	2
6	CDC24	BEM1	1	1	1	1	36	AKR1	STE18	183	1	0	0
7	STE20	BEM1	1	1	0	0	37	STE4	STE18	2	1	0	0
8	CDC42	BEM1	1	1	0	0	38	CDC24	STE20	216	1	0	0
9	STE5	BEM1	6	1	0	0	39	BEM1	STE20	1	1	0	0
10	STE4	CDC24	2	1	1	2	40	AKR1	STE3	177	1	0	0
11	FAR1	CDC24	2	1	0	0	41	AKR1	STE4	2	1	1	2
12	BEM1	CDC24	1	1	0	0	42	STE18	STE4	2	1	0	0
13	STE20	CDC24	216	1	0	0	43	FAR1	STE4	34	1	0	0
14	FAR1	CDC42	35	1	0	0	44	CDC24	STE4	2	1	0	0
15	BEM1	CDC42	1	1	0	0	45	STE5	STE4	20	1	0	0
16	STE4	FAR1	34	1	0	0	46	GPA1	STE4	10	1	0	0
17	CDC24	FAR1	2	1	0	0	47	AKR1	STE5	183	1	0	0
18	BEM1	FAR1	2	1	0	0	48	STE4	STE5	20	1	0	0
19	CDC42	FAR1	35	1	0	0	49	BEM1	STE5	6	1	1	6
20	GPA1	FUS3	2	1	0	0	50	STE7	STE5	2	1	0	0
21	STE5	FUS3	2	1	0	0	51	STE50	STE5	6	1	0	0
22	STE7	FUS3	2	1	0	0	52	FUS3	STE5	2	1	0	0
23	STE11	FUS3	2	1	0	0	53	KSS1	STE5	2	1	0	0
24	STE4	GPA1	10	1	0	0	54	STE5	STE50	6	1	0	0
25	FUS3	GPA1	2	1	0	0	55	STE11	STE50	2	1	0	0
26	STE11	GPA1	7	1	0	0	56	STE5	STE7	2	1	0	0
27	STE5	KSS1	2	1	1	2	57	FUS3	STE7	2	1	0	0
28	STE11	KSS1	2	1	0	0	58	KSS1	STE7	2	1	0	0
29	STE7	KSS1	2	1	0	0							
30	STE12	KSS1	2	1	0	0							192

Figure 19. A NETFLOW Solution for Yeast's MAPK Pheromone Response Pathway.

Obs	from	to	cost	_capac	_FLOW	_FCOST	Obs	from	to	cost	_capac	_FLOW	_FCOST
1	CYR1	ACT1	62	1	0	0	28	ACT1	LAS17	2	1	0	0
2	SRV2	ACT1	1	1	0	0	29	VRP1	LAS17	1	1	0	0
3	VRP1	ACT1	2	1	0	0	30	BEM1	LAS17	4	1	0	0
4	BUD6	ACT1	6	1	0	0	31	CDC25	RAS2	2	1	0	0
5	LAS17	ACT1	2	1	0	0	32	CYR1	RAS2	2	1	0	0
6	BEM1	ACT1	16	1	0	0	33	BUD6	SPA2	183	1	0	0
7	STE20	ACT1	81	1	0	0	34	STE7	SPA2	183	1	0	0
8	ACT1	BEM1	16	1	0	0	35	STE11	SPA2	47	1	0	0
9	LAS17	BEM1	4	1	0	0	36	CYR1	SRV2	2	1	0	0
10	STE20	BEM1	1	1	0	0	37	ACT1	SRV2	1	1	0	0
11	STE5	BEM1	6	1	0	0	38	BUD6	STE11	183	1	0	0
12	ACT1	BUD6	6	1	0	0	39	HSP82	STE11	1	1	1	1
13	SPA2	BUD6	183	1	0	0	40	SPA2	STE11	47	1	0	0
14	STE11	BUD6	183	1	0	0	41	STE5	STE11	2	1	0	0
15	RAS2	CDC25	2	1	1	2	42	KSS1	STE11	2	1	0	0
16	CYR1	CDC25	173	1	0	0	43	KSS1	STE12	2	1	1	2
17	HSP82	CDC25	173	1	0	0	44	ACT1	STE20	81	1	0	0
18	RAS2	CYR1	2	1	0	0	45	BEM1	STE20	1	1	0	0
19	SRV2	CYR1	2	1	0	0	46	BEM1	STE5	6	1	0	0
20	CDC25	CYR1	173	1	0	0	47	STE7	STE5	1	1	0	0
21	ACT1	CYR1	62	1	0	0	48	STE11	STE5	2	1	1	2
22	CDC25	HSP82	10	1	1	10	49	KSS1	STE5	2	1	0	0
23	STE11	HSP82	1	1	0	0	50	SPA2	STE7	183	1	0	0
24	STE7	KSS1	2	1	0	0	51	STE5	STE7	1	1	0	0
25	STE11	KSS1	20	1	0	0	52	KSS1	STE7	2	1	0	0
26	STE5	KSS1	2	1	1	2	53	ACT1	VRP1	2	1	0	0
27	STE12	KSS1	2	1	0	0	54	LAS17	VRP1	1	1	0	0
													19

Figure 20. A NETFLOW Solution for Yeast’s MAPK Filamentation Growth Pathway.

The first two pathways shown in Figure 21 represent two alternative solutions obtained using the capacitated transshipment model for the baker’s yeast MAPK pheromone response. If these two pathways are combined, it can provide a signal transduction network for the pheromone response. One of these solutions is identical to the solution obtained through the color coding presented in Chapter 3. The signal transduction network formed by combining the two alternative solutions is part of the signal transduction network obtained with the integer linear-programming model presented in Chapter 3.

The first two pathways shown in Figure 22 represent two alternative solutions obtained using the capacitated transshipment model for the baker’s yeast MAPK filamentation growth. If these two pathways are combined, it can provide a signal transduction network for filamentation growth. One of these solutions is identical to the solution obtained through the color coding

presented in Chapter 3. The signal transduction network formed by combining the two alternative solutions is part of the signal transduction network obtained with the integer linear-programming model presented in Chapter 3.

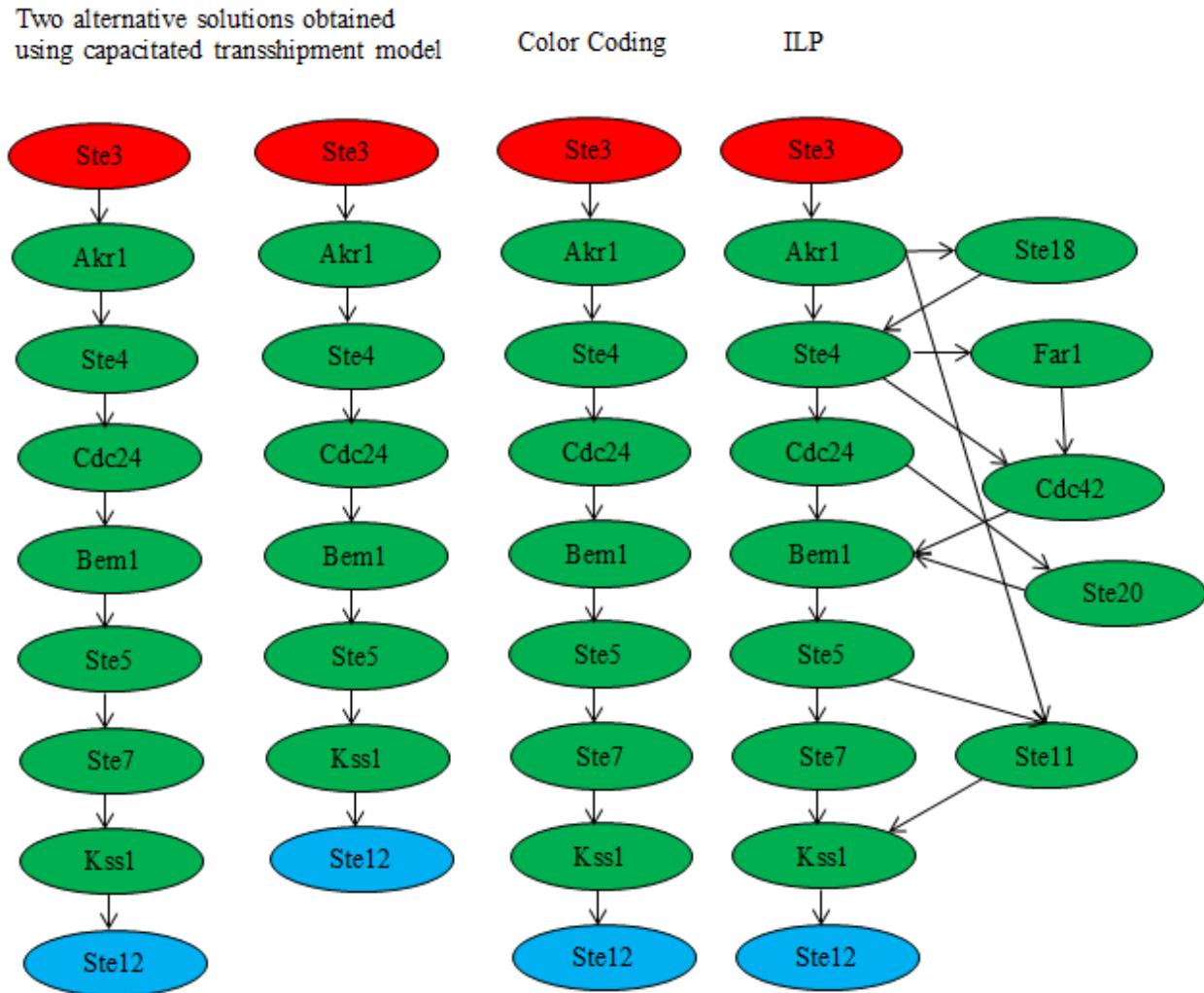


Figure 21. Pheromone Response Pathways.

The results obtained from the capacitated transshipment model created in OPTMODEL give identical findings if the same problem is solved using the existing network cost-flow model,

NETFLOW, in SAS which confirmed the accuracy of the model as shown in Figures 17-20.

When the results of the capacitated transshipment model are compared to the results of the color-coding and integer linear programming methods presented in Chapter 3, it is found that the results given by the method proposed in this paper are close to the results of the color-coding method.

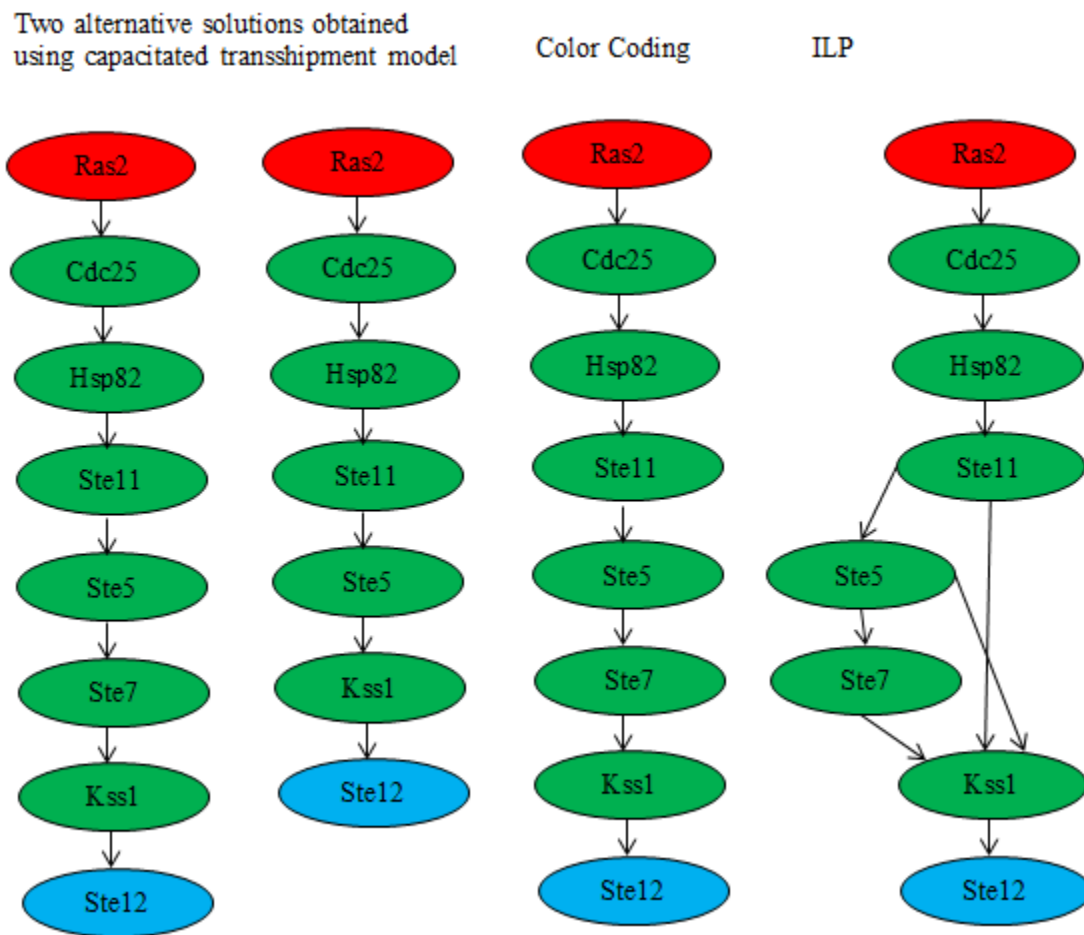


Figure 22. Filamentation Growth Pathways.

When the presented capacitated transshipment model is compared to other methods in terms of memory and time consumption, the presented method is more efficient. The following paragraph shows the comparison for the time and data of the presented method with the other methods explained in Chapter 3.

The color-coding method assigns random color to the nodes and then finds pathways by finding distinct colors. Finding distinct colors requires a number of random trials are conducted to find an accurate pathway because colors are assigned randomly to the nodes. On the other hand, the capacitated transshipment model has a more efficient algorithm because it can provide an accurate solution as color coding in one run.

NetSearch first employs a search algorithm to find pathways and then uses k-means clustering and a scoring method to find the pathway with the highest geometric probability, so it needs a lot of time to go through this process. On the other hand, the presented model employs a customized depth-first search and integer linear programming. Therefore, NetSearch involves more steps to follow as compared to the presented mode.

PathFinder requires much more data and more time to analyze those data while the presented method only uses protein-protein interaction data. Therefore, the presented method takes less storage space and time to analyze the smaller amount of data.

Integer linear programming first uses a depth-first search, and then, the integer linear-programming model is applied, a time-efficient process. On the other hand, the presented model also uses a depth-first search but with more pruning, and then, capacitated transshipment model with binary integer linear programming. Therefore, the presented model is faster because the depth-first search prunes more branches and uses less time in orders of magnitude as compared

to the integer linear-programming model because it bounds the flow on the arcs of the network shown in Chapter 3.

When the capacitated transshipment model is applied to find the pheromone response pathway on the given sub-network, Ste3 is considered the starting protein, and Ste12 is considered the ending protein. The linear pathways we found are as follows: Ste3 to Akr1 to Ste4 to Cdc24 to Bem1 to Ste11 to Ste5 to Kss1 to Ste12 and Ste3 to Akr1 to Ste4 to Cdc24 to Bem1 to Ste11 to Ste5 to Ste7 to Kss1 to Ste12. The latter pathway is identical to that of the color-coding method. The combination of both possible pathways can be considered as part of the sub-network obtained with the integer linear-programming model. This comparison shows that the capacitated transshipment model is more specific. Therefore, the presented model produces fewer interaction pathways which are biologically reasonable because the signals are conducted through less energy-consuming pathways.

When the presented model is applied to find the filamentation growth pathway on the given sub-network, Ras2 is considered the starting protein, and Ste12 is considered the ending protein. The linear pathways we found are as follows: Ras2 to Cdc25 to Hsp82 to Ste11 to Ste5 to Kss1 to Ste12 and Ras2 to Cdc25 to Hsp82 to Ste11 to Ste5 to Ste7 to Kss1 to Ste12. The latter pathway, in this case, too, is identical to that of the color-coding method. The combination of both possible pathways in this case is comparable to the sub-network obtained with the integer linear-programming model. This comparison shows that the capacitated transshipment model is very close to the integer linear-programming model.

Tables 1 and 2 show the comparison of the algorithms described in Chapter 3 for predicting the two signaling pathways used in the experiment. The tables compare precision, defined as the percentage of relevant proteins in the current method's pathway proteins, and

recall, defined as the percentage of the current method's pathway proteins in the actual pathway proteins. The tables show that the presented method has about 83% precision and 75% recall for the pheromone response while there is about 29% precision and 74% recall for the filamentation growth.

Table 1. Comparison of methods for detecting the pheromone response signaling pathway.

Method	Precision (%)	Recall (%)
CTP	83	75
ILP ($\lambda = 0.50$)	47	80
Color coding	83	75
Pathfinder	88	75
NetSearch	74	70

Table 2. Comparison of methods for detecting the filamentation growth signaling pathway.

Method	Precision (%)	Recall (%)
CTP	28	74
ILP ($\lambda = 0.50$)	29	73
Color coding	28	74
Pathfinder	28	82
NetSearch	33	64

When compared to the simple pathways obtained from KEGG's database shown in Figure 1, the proposed method for the capacitated transshipment model has recall and precision that are closer to the recall and precision of the color-coding method. The resultant pathway is missing some proteins involved in the actual process, but the resultant pathway is dependent on

the sub-network obtained through a depth-first search. Therefore, there is a need for a more efficient and accurate way of finding a sub-network or a need for widening the scope of the capacitated transshipment model by taking complete protein interaction data as input and by adding extra constraints, such as the length of the pathway. Another way to improve the presented method is to use a different attribute of the protein-protein interaction, that is, the energy consumed or released during the protein-protein interaction process. As known, most of the reactions occurring in nature and in an organism's body are based on energy, so considering energy consumption as the cost function would be an appropriate step towards improvement. When using energy consumption as a cost function, there is a need for high-throughput experiments to calculate energy consumption when two proteins interact with each other.

The main objective of creating these models to predict signaling pathways is to provide scientists a direction in which experimental research can be conducted to confirm the existing processes and to find new, unknown processes. The discovery of new information can be used to understand the functioning of an organism. On the other hand, utilizing these models to predict known signaling pathways can help scientists conduct research on the new proteins found through these predictions to establish their significance and to confirm their interactions.

If the scoring method of the sub-network and the depth-first search method to find the sub-network are enhanced, the precision and recall for the presented method can be improved to a great extent. Hence, the scope and accuracy of the proposed model can be greatly increased.

CHAPTER 6. CONCLUSION

In this paper, a novel model based on the capacitated transshipment problem is formulated to detect signaling networks in a protein-protein interaction network. As compared to the other methods, excluding the integer linear-programming method, the method presented in this paper is not a heuristic method, but a simpler method, because it detects signaling data directly from protein-protein interaction data.

We were able to formulate the protein-protein interaction data as a capacitated transshipment model and to configure a solver using the SAS OPTMODEL to obtain computation results. This approach was evaluated by comparing it with other approaches and another network flow optimization model built-in SAS, known as NETFLOW.

The results obtained with the proposed method for yeast MAPK signaling pathways demonstrate that protein-protein interaction data can be considered as a capacitated transshipment model and, to a great extent, can help uncover the signaling pathways. These results of known signaling pathways help scientists to design research for finding/confirming the involvement of new proteins and/or interactions in these pathways. If this approach is applied to other networks, new signaling pathways can be discovered, helping to understand the proper functioning of an organism.

The presented model performed comparably well to other methods shown in this paper for detecting the signaling pathways, so it can be concluded that the new model is a good addition to existing methods. If this model is exploited in a more advanced manner, it can be made more efficient by improving the scoring/cost method, by improving the model, and by improving the process of finding the sub-network.

REFERENCES

- Albert, R., DasGupta, B., Dondi, R., Kachalo, S., Sontag, E., Zelikovsky, A., et al. (2007). A novel method for signal transduction network inference from indirect experimental evidence. *Journal of Computational Biology*, 14(7):927-949.
- Allena, E. E., Fetrowb, J. S., Daniel, L. W., Thomas, S. J., & John, D. J. (2006). Algebraic dependency models of protein signal transduction networks from time-series data. *Journal of Theoretical Biology*, 238(2):317-330.
- Baitaluk, M., Qian, X., Godbole, S., Raval, A., Ray, A., & Gupta, A. (2006). PathSys: Integrating molecular interaction graphs for systems biology. *BMC Bioinformatics*, 7:55.
- Bebek, G., & Yang, J. (2007). PathFinder: Mining signal transduction pathway segments from protein-protein interaction networks. *BMC Bioinformatics*, 8:335.
- Berg, J., Tymoczko, J., & Stryer, L. (2002). Chapter 15, Signal-Transduction Pathways: An Introduction to Information Metabolism. In T. J. Berg JM, *Biochemistry*. New York: W. H. Freeman and Company. <http://www.ncbi.nlm.nih.gov/books/NBK21205/>.
- Bradley, G. H., Brown, G. G., & Graves, G. W. (1977). Design and implementation of large scale primal transshipment algorithms. *Management Science*, 24(1):1-34.
- Chen, J. C., & Yuan, B. (2006). Detecting functional modules in the yeast protein-protein interaction network. *Bioinformatics*, 22(18):2283-2290.
- Chen, R. E., & Thorner, J. (2007). Function and regulation in MAPK signaling pathways: Lessons learned from the yeast *Saccharomyces cerevisiae*. *Biochimica et Biophysica Acta*, 1773(8):1311-1340.
- Feiglin, A., Moulton, J., Lee, B., Ofran, Y., & Unger, R. (2012). Neighbor overlap is enriched in the yeast interaction network: Analysis and implications. *PLoS One*, 7(6): e39662.

- Gitter, A., Klein-Seetharaman, J., Gupta, A., & Bar-Joseph, Z. (2011). Discovering pathways by orienting edges in protein interaction networks. *Nucleic Acids Research*, 39(4):e22.
- Gustin, M. C., Albertyn, J., Alexander, M., & Davenport, K. (1998). MAP kinase pathways in the yeast *Saccharomyces cerevisiae*. *Microbiology and Molecular Biology Reviews*, 62(4):1264-1300.
- Hu, H., Yan, X., Huang, Y., Han, J., & Zhou, X. J. (2005). Mining coherent dense subgraphs across massive biological networks for functional discovery. *Bioinformatics*, 21(1):213-221.
- Huerta, M., Haseltine, F., Liu, Y., Downing, G., & Seto, B. (2000, July 17). *Biomedical Information Science and Technology Initiative*. Retrieved November 15, 2012, from National Institute of Health: <http://www.bisti.nih.gov/docs/CompuBioDef.pdf>
- Kestler, H. A., Wawra, C., Kracher, B., & Kuhl, M. (2008). Network modeling of signal transduction: Establishing the global view. *BioEssays*, 30:1110-1125.
- Khurana, A., Verma, T., & Arora, S. R. (2012). An algorithm for solving time minimizing capacitated transshipment problem. *International Journal of Management Science and Engineering Management*, 7(3):192-199.
- Korf, R. E. (1985). Depth-first iterative-deepening: An optimal admissible tree search. *Elsevier Science Publishers B.V. (North-Holland)*, 97-109.
- Liu, Y., & Zhao, H. (2004). A computational approach for ordering signal transduction pathway components from genomics and proteomics data. *BMC Bioinformatics*, 5:158.
- Nygaard, K. (2009). Classroom notes. *An Example of Capacitated Transshipment Problem*. Fargo, North Dakota: Department of Computer Science, NDSU.

- Pavlopoulos, G. A., Secrier, M., Moschopoulos, C. N., Soldatos, T. G., Kossida, S., Aerts, J., et al. (2011). Using graph theory to analyze biological networks. *BioData Mining*, 4:10.
- SAS Institute, Inc. (2008). SAS/OR® 9.2 user's guide: Mathematical programming. Cary, NC.
- Scott, J., Ideker, T., Karp, R. M., & Sharan, R. (2006). Efficient algorithms for detecting signaling pathways in protein interaction networks. *Journal of Computational Biology*, 13(2):133-144.
- Steffen, M., Petti, A., Aach, J., D'haeseleer, P., & Church, G. (2002). Automated modelling of signal transduction networks. *BMC Bioinformatics*, 3:34.
- Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguéz, P., et al. (2011). The STRING database in 2011: Functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Research*, Database issue:D561–D568.
- Vinayagam, A., Stelzl, U., Foulle, R., Plassmann, S., Zenkner, M., Timm, J., et al. (2011). A directed protein interaction network for investigating intracellular signal transduction. *Science Signaling*, 4(189):rs8.
- Wolkenhauer, O., & Cho, K.-H. (2003). Analysis and modelling of signal transduction pathways in systems biology. *Biochemical Society Transactions*, 31(Pt 6):1503-1509.
- Xenarios, I., Salwinski, L., Duan, X. J., Higney, P., Kim, S.-M., & Eisenber, D. (2002). DIP, the database of interacting proteins: A research tool for studying cellular networks of protein interactions. *Nucleic Acids Research*, 30(1):303-305.
- Yeang, C.-H., Ideker, T., & Jaakkola, T. (2004). Physical network models. *Journal of Computational Biology*, 11(2-3):243-262.
- Zhang, J., & Wiemann, S. (2009). KEGGgraph: A graph approach to KEGG Pathway in R and bioconductor. *Bioinformatics*, 25(11):1470-1471.

- Zhao, X.-M., Wang, R.-S., Chen, L., & Aihara, K. (2008a). Automatic modeling of signal pathways from protein-protein interaction networks. *Proceedings of the 6th Asia Pacific Bioinformatics Conference, Vol. 6 of Series on Advances in Bioinformatics and Computational Biology* (pp. 287-296). Singapore: Imperial College Press.
- Zhao, X.-M., Wang, R.-S., Chen, L., & Aihara, K. (2008b). Uncovering signal transduction networks from high-throughput data by integer linear programming. *Nucleic Acids Research*, 36:e48.