Spring 5-6-2012

Mathematical Methods for Network Analysis, Proteomics and Disease Prevention

Kun Zhao
Georgia State University

MATHEMATICAL METHODS FOR NETWORK ANALYSIS, PROTEOMICS AND DISEASE PREVENTION

by

KUN ZHAO

Under the Direction of Dr. Igor Belykh, Dr. Guantao Chen and Dr. Jenny J. Yang

ABSTRACT

This dissertation aims at analyzing complex problems arising in the context of dynamical networks, proteomics, and disease prevention. First, a new graph-based method for proving global stability of synchronization in directed dynamical networks is developed. This method utilizes stability and graph theories to clarify the interplay between individual oscillator dynamics and network topology. Secondly, a graph-theoretical algorithm is proposed to predict Ca^{2+}-binding site in proteins. The new algorithm enables us to identify previously-unknown Ca^{2+}-binding sites, and deepens our understanding towards disease-related Ca^{2+}-binding proteins at a molecular level. Finally, an optimization model and algorithm to solve a disease prevention problem are described at the population level. The new resource allocation model is designed to assist clinical managers to make decisions on identifying at-risk population groups, as well as selecting a screening and treatment strategy for chlamydia and gonorrhea patients under a fixed budget. The resource allocation model and algorithm can have a significant impact on real treatment strategy issues.

INDEX WORDS: Dynamical system, Complex network, Synchronization, Graph Theory, Proteomics, Combinatorial Optimization, Disease control

MATHEMATICAL METHODS FOR NETWORK ANALYSIS, PROTEOMICS AND DISEASE PREVENTION

by

KUN ZHAO

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

in the College of Arts and Sciences

Georgia State University

2012

MATHEMATICAL METHODS FOR NETWORK ANALYSIS, PROTEOMICS AND DISEASE PREVENTION

by

KUN ZHAO

Committee Chair:	Dr. Igor Belykh
Committee:	Dr. Guantao Chen
	Dr. Jenny J. Yang
	Dr. Vladimir Bondarenko

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

May 2012

DEDICATION

To my family

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my advisors, Professor Igor Belykh and Professor Guantao Chen, for their consistent support, guidance, and encouragements for the past few years. Their advice always inspires me to think broader and deeper. I feel fortune to have worked with them over these years. I am also deeply indebted to Professor Jenny J. Yang for sharing her expertise in biochemistry. The philosophy of science of Professor Belykh, Professor Chen, and Professor Yang, is an indispensable force that drives me further on the road of scientific research. My sincere gratitude also goes to Professor Vladimir Bondarenko, for serving on my Ph.D committee and for his time in reviewing this work. I would like to thank Professor Andrey Shilnikov for introducing me to the fantastic world of advanced dynamic systems.

In addition, I am obliged to Dr. Robert Wohlhueter and Dr. Michael Kirberger for their help on reviewing my manuscripts regarding a binding-site prediction project. I would like to express my appreciation to Dr. Hing-Cheung Wong, Dr. Xue Wang, Dr. Jiawei Liu, and the members in Dr. Yang's lab with whom I have worked, for their inspiring discussions for the project. I would like also thank Mr. Xin Wei and Dr. Fasheng Qiu for working with me on an optimization project. I thank Mr. Abdoul Sylla for the K^{th} shortest path project. I am grateful to Ms. Xia Hu, Ms. Sajiya Jalil and Mr. Jeremy Wojcik for their fruitful discussions on the augmented graph project.

I would like to thank my family for supporting me to pursue a Ph.D. in the United States. Special thanks to my beautiful wife, Lin Miao, for her support, encouragement, and companionship. For all that, and for being everything I am not, she has my everlasting love.

Finally, I am thankful for the financial support from the NSF (grant: DMS-1009744) and the GSU MBD fellowship program.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

This dissertation aims at analyzing complex problems arising in the context of dynamical networks, proteomics, and disease prevention. In Chapter 1, building on observations that synchronization has been observed in many complex networks (i.e. firing synchronization in neural networks is relevant for neurological disorders, for example, Parkinson's disease [1]), we extended the Connection Graph method [2] for proving synchronization in directed networks. Our approach, called the Augmented Graph Stability method, is based on the transformation of the directed graph into an undirected graph. This is done by replacing each direct link between node i node j with an undirected edge whose coupling strength depends on the mean node unbalance between the two nodes. In addition, we augment the graph by adding an extra edge, connecting node i and node j if there is no directed link between them and their mean node unbalance is negative. Different weights are also associated with each path between any two nodes of the augmented undirected network, according to the mean node unbalance. Upper bounds on the coupling strength sufficient for synchronization in this augmented symmetrized network also guarantee global stability of synchronization in the original directed network. We show that the new Augmented Graph Stability method is more effective than the connection graph method in sparse networks. In Chapter 2, we propose a graph theory algorithm to predict the Ca^{2+}-binding site in proteins at a molecular level. Predicting the Ca^{2+}-binding site is important as Ca^{2+} and Ca^{2+}-binding proteins (CaBP) are relevant to many diseases (i.e. Alzheimer's disease [3], heart disease [4], diabetes [4] , leukemia [5, 6], and cancers [7–10]). In order to understand the mechanism of diseases related to CaBP, it was first necessary to discover where the proteins bind to Ca^{2+}. We hypothesize that the second, hydrophobic shell of carbon atoms enclosing a Ca^{2+}-binding site could sufficiently determine the site's location in either X-ray or NMR structures. Then we validate the hypothesis with the new algorithm on various structural datasets. Chapter 3 addresses a real clinical issue and seeks to find a way to help publicly-funded programs that have only limited resources regarding

screening and treating Chlamydia trachomatis (CT) and Neisseria gonorrhoeae (GC). In this chapter, we develop a combinatorial optimization (a.k.a. resource allocation) model and algorithm for health care management to distribute its funds efficiently at a population level. The solutions generated by the new model can be used to assist clinical managers to make decisions on identifying at-risk population groups, as well as selecting a screening and treatment strategy for CT and GC patients under a fixed budget. We then propose a two-step branch-and-bound algorithm tailor-made for solving the model. The solutions calculated by the new algorithm have been compared to those calculated by commercial software application. The main contributions of this dissertation are summarized in the last section, "Major Findings and Significance".

CHAPTER 1

MATHEMATICS IN NETWORK ANALYSIS

1.1 Introduction

The phenomenon of synchronization in large complex networks of coupled dynamical systems has attracted a great deal of attention over the past decade. Research on this topic spans various scientific disciplines such as mathematics, physics, engineering, and other fields of science. The examples include coupled synchronized lasers [11, 12], networks of computer clocks [13], synchronized neuronal firing and calcium signals [14–17]. The utilization of mathematical methods in studying synchronization not only deepens our understanding towards the formation of this phenomenon in general, but also can have some practical implications. For example, the presence of synchronization in the human brain has been suggested as particularly relevant for neurological disorders, e.g. Parkinson's disease [1] and Alzheimer's disease [18]. The information regarding how the firing dynamics are synchronized in the neural network with a specific topology, can assist neurologists to discover the causes of incurable diseases such as Parkinson's and Alzheimer's diseases and to create better treatment. Motivated by mathematics and its applications, this Chapter will mainly focus on methodologies for studying network synchronization [19].

The strongest form of synchrony in oscillator networks is complete synchronization (when all oscillators do the same thing at the same time) [20–22]. The most important question in the synchronization studies is: What are the conditions for the stability of the synchronized state, especially with respect to coupling strengths and coupling configurations of the network? This problem was intensively studied for networks of limit-cycle oscillators [23–27] and chaotic dynamical systems [28–40].

Complete synchronization in networks of continuous time identical oscillators typically becomes stable when the coupling strength between the oscillators exceeds a critical value.

In light of this, an important problem is to identify the bounds on the coupling strengths so that the stability of synchronization is guaranteed. Many methods for determining stability for synchronized chaotic systems have been developed. Most of them are based on the calculation of two quantities: (i) the eigenvalues of the coupling matrix for different network topologies and (ii) a term that depends on the dynamics of the individual oscillators [28, 31, 34, 36–40].

One example of the methods mentioned above is the Master Stability function. Developed by Pecora and Carroll [34], it is a general approach to the local synchronization of chaotic systems for any linear coupling scheme. This approach is based on the calculation of the maximum Lyapunov exponent for the least stable transversal mode of the synchronous manifold, in conjunction with the eigenvalues of the connectivity matrix. An analog of the Master Stability function for global synchronization of chaotic systems was also proposed [36, 37]. However, the eigenvalues of the coupling matrix can often be calculated only for simple regular topologies such as local, star-like, and all-to-all networks. In more complex networks, the calculation of the eigenvalues becomes extremely difficult such that is often impossible to obtain analytical bounds for the synchronization thresholds. Moreover, for networks with a time-varying coupling, the application of the eigenvalue-based methods is difficult and often impossible.

As an alternative approach to calculate the synchronization condition, Belykh et al. [2, 41] proposed the Connection Graph method, which does not depend on explicit knowledge of the spectrum of the connectivity matrix. To guarantee complete synchronization with respect to arbitrary initial conditions, this method utilizes the Lyapunov function approach together with graph theoretical reasoning. It is also applicable to time-dependent networks. This method was originally developed for undirected graphs [2], and was later applied to asymmetrically directed networks [41].

In this Chapter, we present a modification of the Generalized Connection Graph method that gives tighter bounds on the coupling strength required for the onset of stable synchronization in sparse directed networks. We demonstrate how the directed network can be

turned into an augmented undirected network with weighted connections. As a result, the stability conditions for synchronization in this augmented directed network also ensure stable synchronization in the original directed network.

The layout of this study is the following. First, in Sec. 1.2, we state the problem in the study. Then, in Sec. 1.3, we present the derivations of the graph-based criterion for global synchronization in undirected networks. In Sec. 1.4, we introduce the new method and compare it to the existing Connection Graph method, using specific network examples. We also discuss computational algorithms for solving Short Path (SP) problems of how to choose a short path between two nodes of the network; this notion is heavily used in our graph-based Method. We show that the new Augmented Graph method is more effective than the original Connection Graph method, for proving synchronization in sparse directed networks.

1.2 Problem Statement

1.2.1 Complex network model

We consider a network of n interacting nonlinear d-dimensional dynamical systems (oscillators). We assume that the individual oscillators are all identical, even though our results can be generalized to slightly non-identical systems. The composed dynamical system is described by the $n \times d$ ordinary differential equations [2]

$$\dot{x}_i = F(x_i) + \sum_{j=1}^{n} \varepsilon_{ij}(t) P(x_j - x_i), \quad i = 1, ..., n, \tag{1.1}$$

where $x_i = (x_i^1, ..., x_i^d)$ is the d-vector containing the coordinates of the i-th oscillator. The non-zero elements of the $d \times d$ matrix P determine by which variables the oscillators are coupled. Without loss of generality, we shall consider a vector version of the coupling with the diagonal matrix $P = \mathrm{diag}(p_1, p_2, ..., p_d)$, where $p_h = 1$, $h = 1, 2, ..., s$ and $p_h = 0$ for $h = s + 1, ..., d$.

$G = (\varepsilon_{ij}(t))$ is an *asymmetric* $n \times n$ zero-row sum matrix with nonnegative off-diagonal elements such that $\varepsilon_{ij} \geq 0$ for $i \neq j$, and $\varepsilon_{ii} = -\sum_{j=1;\ j\neq i}^{n} \varepsilon_{ij}, \quad i = 1, ..., n$. This matrix represents an arbitrary directed network of asymmetrically connected oscillators. The zero-row sum condition is a necessary condition for the existence of the synchronous solution.

The connectivity matrix G corresponds to a directed graph with n vertices and m edges. The number of directed edges m is defined by the number of non-zero non-diagonal elements of the matrix G. The individual oscillators correspond to the vertices of the connection graph. To ensure synchronization of all oscillators, there must be at least one oscillator that directly or indirectly influences all the others. This amounts to the existence of a directed tree that involves all the vertices (oscillators).

1.2.2 Definition of global complete synchronization

The main goal of this study is to obtain stability conditions of complete synchronization in the system (1.1). Global complete synchronization in the system (1.1) amounts to global stability of the linear invariant manifold $M = \{x_1 = x_2 = ... = x_n\}$. The manifold M has the dimension of a single oscillator, and is called the synchronization manifold. This manifold contains completely synchronous solutions of all types (multi-stable, periodic, and chaotic oscillations).

Definition 1.1. *Complete synchronization occurs in the network (1.1), if*

$$\lim_{t\to\infty} ||x_i(t) - x_j(t)|| = 0 \ \text{ for } \ \forall i, j. \tag{1.2}$$

We want to determine upper bounds for the coupling strength sufficient for complete synchronization, and to identify the dependence of the threshold values on the network topology and the properties of the individual oscillator.

1.3 Connection Graph Method for Undirected Network Synchronization: Review

In this section, we follow the steps of the previous study by Belykh et al. [2, 41] to review the derivation of the Connection Graph method for undirected networks [2]. We assume that the connectivity matrix G in (1.1) is *symmetric*, and therefore the network is undirected.

1.3.1 Stability system for the difference variables

To prove the stability of complete synchronization, we have to show that the differences between the oscillators' corresponding variables become zero. Therefore, we introduce the notation for the differences

$$X_{ij} = x_j - x_i, \;\; i, j = 1, ..., n, \tag{1.3}$$

and derive the stability system for the difference variables [2]

$$\dot{X}_{ij} = F(x_j) - F(x_i) + \sum_{k=1}^{n} \{\varepsilon_{jk} P X_{jk} - \varepsilon_{ik} P X_{ik}\}, \;\; i, j = 1, ..., n. \tag{1.4}$$

We use the vector analog of the Mean Value Theorem for the function difference to re-write the difference $F(x_j) - F(x_i)$ as follows

$$F(x_j) - F(x_i) = \int_0^1 \frac{d}{d\beta} F(\beta x_j + (1-\beta) x_i) d\beta = \left[\int_0^1 DF(\beta x_j + (1-\beta) x_i) d\beta \right] X_{ij},$$

where DF is a $d \times d$ Jacobi matrix of F.

Consequently, the stability system becomes

$$\dot{X}_{ij} = \left[\int_0^1 DF(\beta x_j + (1-\beta) x_i) d\beta \right] X_{ij} + \sum_{k=1}^{n} \{\varepsilon_{jk} P X_{jk} - \varepsilon_{ik} P X_{ik}\}, \tag{1.5}$$

where $i, j = 1, ..., n$. It is worth noticing that one can calculate the Jacobian DF explicitly via the parameters of the individual oscillator.

Notice that the stability system (1.5) has n^2 equations, and $n(n-1)$ of them define the stability of synchronization in the corresponding pair of oscillators. Technically, only $n(n-1)/2$ difference variables are required to describe synchronization in the network, and the stability system (1.5) is redundant. The use of the redundant stability systems is the key ingredient of the Connection Graph method [2].

Let us study the redundant stability system (1.5).

We add and subtract an additional term AX_{ij} from the system (1.5) to obtain the following system

$$\begin{aligned} \dot{X}_{ij} &= \left[\int_0^1 DF(\beta x_j + (1-\beta)x_i)d\beta - A\right] X_{ij} + AX_{ij} + \\ &+ \sum_{k=1}^{n} \{\varepsilon_{jk} P X_{jk} - \varepsilon_{ik} P X_{ik}\}, \end{aligned} \tag{1.6}$$

where $i, j = 1, ..., n$ and the matrix $A = aP$, where P is the projection matrix from the system (1.1) and a is a constant.

The trivial equilibrium of the stability system (1.6) corresponds to the synchronization manifold of the system (1.1). In the following, we shall obtain conditions under which the trivial equilibrium is globally stable and therefore prove global asymptotical stability of complete synchronization.

The addition of the matrix $-A$ helps to damp instabilities caused by the Jacobian DF. On the other hand, the addition of the matrix $+A$ causes the instability that can be in turn damped by the coupling terms.

We shall study the stability of system (1.6) in two steps. First, we introduce the auxiliary system

$$\dot{X}_{ij} = \left[\int_0^1 DF(\beta x_j + (1-\beta)x_i)\, d\beta - A\right] X_{ij}, \quad i, j = 1, ..., n. \tag{1.7}$$

This system is identical to the stability system (1.6) where the coupling terms are removed.

The first step is to prove that this auxiliary system can be made stable by increasing parameter a. To do so, we assume that there exist Lyapunov functions

$$W_{ij} = \frac{1}{2}\, X_{ij}^T \cdot H \cdot X_{ij},\ i,j = 1,...,n, \tag{1.8}$$

where $H = \mathrm{diag}(h_1, h_2, ..., h_s, H_1)$, $h_1 > 0, ..., h_s > 0$, and the $(d-s) \times (d-s)$ matrix H_1 is positive definite.

We require their derivatives with respect to the system (1.7) to be negative

$$\dot{W}_{ij} = X_{ij}^T H \left[\int_0^1 DF(\beta x_j + (1-\beta)x_i)d\beta - A \right] X_{ij} < 0, \quad X_{ij} \neq 0. \tag{1.9}$$

This amounts to requiring global stability of the auxiliary system. This is a crucial component of the Connection Graph method. As a result, we require that that all oscillators of the system (1.1) can be synchronized when the coupling among the oscillators is sufficiently large. It is important to stress that this property is not always true as some networks such as x-coupled Rössler systems cannot be globally synchronized even if the coupling is made infinitely strong [30, 42], and the requirement (1.9) cannot be fulfilled.

The conditions that guarantee the requirement (1.9) are based upon the individual node's dynamics and the way the oscillators are coupled (matrix P). Therefore, the requirement (1.9) has to be proven for each specific network as this condition depends on the intrinsic dynamics of the individual oscillators and the projection matrix P. The condition was proved for various limit-cycle and chaotic oscillators, including Lorenz systems [43], double-scrolls [28, 40], Hodgkin-Huxley-type models and different P matrices [2, 41].

The proof of requirement (1.9) for the coupled chaotic Lorenz oscillators is given in [2, 43]; however for an illustrative purpose, we present the sketch of the proof and calculation of parameter a in Appendix A. One can prove the requirement (1.9) for other coupled chaotic oscillators as long as these oscillators can be synchronized.

To prove the global stability of the synchronization manifold, we also need to make an additional assumption on the eventual dissipativeness of the coupled system (1.1).

We need to assume that the individual oscillator $\dot{x}_i = F(x_i)$ is eventually dissipative, i.e. there exists a topological ball B which attracts all trajectories from the outside. This implies that there are no trajectories which escape to infinity. This is a natural assumption for most known chaotic oscillators.

To prove global stability of the synchronization manifold, we construct the Lyapunov function for the stability system (1.6)

$$V = \frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{n} X_{ij}^T \cdot H \cdot X_{ij}, . \tag{1.10}$$

The corresponding time derivative along the trajectories of (1.6) is

$$\begin{aligned} \dot{V} = \; & \tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \dot{W}_{ij} + \tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} X_{ij}^T A X_{ij} - \\ & -\tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\{\varepsilon_{jk} X_{ji}^T H P X_{jk} + \varepsilon_{ik} X_{ik}^T H P X_{ij}\}. \end{aligned} \tag{1.11}$$

We have to show the negative definiteness of the quadratic form $\dot{V}$. The first sum S_1 is negative definite due to the requirement (1.9). Hence, it is sufficient to analyze the last two sums S_2 and S_3. Recall that the coupling matrix G is assumed to be symmetric, we can calculate the sum S_2 as follows

$$S_2 = \sum_{i=1}^{n-1}\sum_{j>i}^{n} A X_{ij}^2 \, . \tag{1.12}$$

The contribution of this sum, which is always positive, must be compensated by the third sum

$$S_3 = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\{\varepsilon_{jk} X_{ji}^T H P X_{jk} + \varepsilon_{ik} X_{ik}^T H P X_{ij}\}. \tag{1.13}$$

Switching the summation index i and index j in the second term, we get

$$S_3 = -\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\varepsilon_{jk}X_{ji}^T HPX_{jk}\,. \tag{1.14}$$

As $X_{jj} = 0$, this formula transforms into

$$S_3 = -\sum_{i=1}^{n}\sum_{k=1}^{n-1}\sum_{j>k}^{n}\varepsilon_{jk}X_{ji}^T HPX_{jk} - \sum_{i=1}^{n}\sum_{k=1}^{n-1}\sum_{j<k}^{n}\varepsilon_{jk}X_{ji}^T HPX_{jk}\,. \tag{1.15}$$

We use the fact that $\varepsilon_{ij} = \varepsilon_{ji}$ to obtain the following

$$\begin{aligned} S_3 = \; & -\sum_{i=1}^{n}\sum_{k=1}^{n-1}\sum_{j>k}^{n}\varepsilon_{jk}X_{ji}^T HPX_{jk} - \sum_{i=1}^{n}\sum_{j=1}^{n-1}\sum_{k<j}^{n}\varepsilon_{jk}X_{ki}^T HPX_{kj} = \\ & = -\sum_{i=1}^{n}\sum_{k=1}^{n-1}\sum_{j>k}^{n}\varepsilon_{jk}(X_{ji}^T + X_{ik}^T)HPX_{jk}\,. \end{aligned} \tag{1.16}$$

The form S_3 can be further simplified using $X_{ji}^T + X_{ik}^T = \left[x_i^T - x_j^T + x_k^T - x_i^T\right] = X_{jk}^T$

$$S_3 = -\sum_{i=1}^{n}\sum_{k=1}^{n-1}\sum_{j>k}^{n}\varepsilon_{jk}X_{jk}^T HPX_{jk} = -\sum_{k=1}^{n-1}\sum_{j>k}^{n} n\varepsilon_{jk}X_{jk}^T HPX_{jk}\,. \tag{1.17}$$

Finally, we can make the claim that the time derivative $\dot{V}$ of the Lyapunov function V is negative if

$$S_2 + S_3 = \sum_{i=1}^{n-1}\sum_{j>i}^{n} X_{ij}^T H[A - n\varepsilon_{ij}P]X_{ij} \tag{1.18}$$

is negative definite. Therefore, $\dot{V} < 0$ if

$$\sum_{i=1}^{n-1}\sum_{j>i}^{n}\varepsilon_{ij}X_{ij}^T HPX_{ij} > \frac{1}{n}\sum_{i=1}^{n-1}\sum_{j>i}^{n} X_{ij}^T HAX_{ij} \tag{1.19}$$

This statement can summarized in the following theorem.

Theorem 1 [2]. *Under the above assumptions, synchronization in the network (1.1) with*

a symmetric connectivity matrix G is globally asymptotically stable if the following holds

$$\sum_{k=1}^{m} \varepsilon_{i_k j_k} X_{i_k j_k}^2 > \frac{a}{n} \sum_{i=1}^{n-1} \sum_{j>i}^{n} X_{ij}^2 \,, \tag{1.20}$$

where $X_{i_k j_k}$ $k = 1, ..., m$ are defined by m existing links. Note that m is the number of non-zero above diagonal elements in matrix G. Variables $X_{i_k j_k}$ correspond to the scalars $X_{i_k j_k}^{(l)}$, $l = 1, ..., s$.

To derive the bounds for the synchronization thresholds, we have to get rid of the difference variables in (1.20). This constitutes the second step of the Connection Graph method. In the simplest case of a complete graph, this calculation is straightforward. To illustrate this, let us assume that the graph is complete such that $\varepsilon_{i_k j_k}(t) \geq \varepsilon > 0$, $k = 1, ..., n(n-1)/2$ for all $i_k, j_k \in \{1, ..., n\}$. Therefore, by Theorem 1, the synchronization threshold becomes

$$\varepsilon(t) > \varepsilon^* = \frac{a}{n} \,.$$

Eliminating the variables X_{ij} and $X_{i_k j_k}$ in the inequality (1.20) requires re-calculating X_{ij} via the variables $X_{i_k j_k}$ that correspond to the edges on the connection graph.

1.3.2 Eliminating the difference variables using the connection graph

Our goal is to find the condition on the coupling strength ε that satisfies inequality (1.20)

$$\sum_{k=1}^{m} \varepsilon_k(t) \tilde{X}_k^2 > \frac{a}{n} \sum_{i=1}^{n-1} \sum_{j>i}^{n} X_{ij}^2, \tag{1.21}$$

where we have relabeled the variables as follows $\tilde{X}_k = X_{i_k j_k}$ and $\varepsilon_k = \varepsilon_{i_k j_k}$, $m \geq n - 1$.

We should recalculate all difference variables X_{ij}, $i, j = 1, ..., n$ through the difference variables $\tilde{X}_k$, $k = 1, ..., m$ corresponding to edges of the connection graph. This will allow one to eliminate the difference variables X_{ij} and $\tilde{X}_k$ in the inequality (1.21), and therefore derive the bound on the coupling strength.

To do so, for any pair of vertices (i, j), we choose a path P_{ij} from node i to node j. If edge k belongs to the path P_{ij}, we denote it by $k \in P_{ij}$. The path length P_{ij} is denoted by $z(P_{ij})$, representing the number of edges comprising P_{ij}. If the path P_{ij} passes through vertices $i, m_1, m_2, ..., m_\nu, j$ then $X_{ij} = X_{i,m_1} + X_{m_1,m_2} + ... + X_{m_{\nu},j}$. As a result, we get

$$X_{ij}^2 = \left(\sum_{k \in P_{ij}} \pm \tilde{X}_k \right)^2 \leq z(P_{ij}) \sum_{k \in P_{ij}} \tilde{X}_k^2 \,, \tag{1.22}$$

where we have applied the Cauchy-Schwarz inequality.

Therefore, the RHS of (1.21) can be bounded as follows

$$\sum_{i=1}^{n-1} \sum_{j>i}^{n} X_{ij}^2 \leq \sum_{k=1}^{m} \left(\sum_{i=1}^{n-1} \sum_{j>i;\, k \in P_{ij}}^{n} z(P_{ij}) \right) \tilde{X}_k^2 \,. \tag{1.23}$$

Plugging the bound (1.23) into the inequality (1.21) and canceling out the difference variables, we get

$$\varepsilon_k(t) > \frac{a}{n} \cdot \sum_{j>i;\, k \in P_{ij}}^{n} z(P_{ij}) \quad \text{for} \;\; k = 1, ..., m. \tag{1.24}$$

This criterion constitutes the Connection Graph method for synchronization in directed networks [2] which is formulated in the following theorem.

Theorem 2 [2]. *Under the assumption (1.9), complete synchronization of system (1.1) with a symmetrical connectivity matrix G is globally asymptotically stable if the following holds*

$$\varepsilon_k(t) > \frac{a}{n}\, b_k(n, m) \quad \text{for} \;\; k = 1, ..., m \;\; \text{and for all} \;\; t, \tag{1.25}$$

where $b_k(n, m) = \sum_{j>i;\, k \in P_{ij}}^{n} z(P_{ij})$ represents the sum of the lengths of all chosen paths P_{ij} which pass through a given edge k on the connection graph.

More details on the derivation of the Connection Graph method and its application to specific undirected networks can be found in [2].

1.4 Graph-based Stability Method for Directed Networks with Examples

In this section, we extend the Connection Graph method to directed graphs and derive an effective approach to proving synchronization in directed networks. In three subsections, we will use specific network topologies to illustrate how the methods work in these cases. In the first subsection, we will start from an introduction of the previously developed Generalized Connection Graph method [41]. Then, we will calculate a lower bound by using our new method. To show that our new method is more effective for sparse directed networks, we will compare the two methods in three more network configurations in the second subsection. In the last subsection, we will gave an example of utilizing the new method for a 30-node network, demonstrating that the computation task of the method could be laborious and the pseudo-code given in this section can be a solution for calculating the synchronization bound.

1.4.1 Five-node undirected networks

Let's consider a simple asymmetric directed graph (Fig. 1.1A). Let d denote the coupling strength in general. Specifically, d_{ij} denotes the coupling strength from node i to node j. D_i^c denotes the node unbalance at the node i, which is the difference between the sum of the coupling coefficients of all edges directed outward from node i and the sum of the coupling coefficients of all the edges directed to node i. D_{ij} denotes the mean value of the node unbalance between node i and j. e_{ij} denotes an edge from node i to node j in a directed graph and between node i to node j in a symmetrized graph. Therefore, we can calculate the following quantities.

The node balance D_i^c for each node of the graph:

$$D_1^c = d - 2d = -d \qquad D_2^c = d - d = 0 \qquad D_3^c = d - d = 0$$
$$D_4^c = d - d = 0 \qquad D_5^c = 2d - d = d.$$

The mean node unbalance D_{ij}, which is equal to $\frac{D_i^c+D_j^c}{2}$ for each nodes i and j :

$$D_{12}:\tfrac{D_1^c+D_2^c}{2}=-\tfrac{d}{2} \qquad D_{13}:\tfrac{D_1^c+D_3^c}{2}=-\tfrac{d}{2} \qquad D_{14}:\tfrac{D_1^c+D_4^c}{2}=-\tfrac{d}{2} \qquad D_{15}:\tfrac{D_1^c+D_5^c}{2}=0$$
$$D_{23}:\tfrac{D_2^c+D_3^c}{2}=0 \qquad D_{24}:\tfrac{D_2^c+D_4^c}{2}=0 \qquad D_{25}:\tfrac{D_2^c+D_5^c}{2}=\tfrac{d}{2}$$
$$D_{34}:\tfrac{D_3^c+D_4^c}{2}=0 \qquad D_{35}:\tfrac{D_3^c+D_5^c}{2}=\tfrac{d}{2}$$
$$D_{45}:\tfrac{D_4^c+D_5^c}{2}=\tfrac{d}{2}.$$

1.4.2 Existing Generalized Connection Graph method

To find an upper bound for the synchronization threshold in concrete networks, one can use the previously published Generalized Connection Graph method [41] and follow its steps.

Step 1. Symmetrize the graph by replacing each directed edge by an undirected edge with half the coupling strength: $d_{ij}=\frac{d}{2}$ (see Fig. 1.1B). The coupling strength is adjusted, based on the mean node unbalance D_{ij}. If $D_{ij}<0$ and there is an edge in the symmetrized graph linking directly i and j, then we calculate the quantity $\left|\frac{D_{ij}}{5}\right|$ and add this additional coupling strength to d_{ij}. For example in Fig. 1.1B, we only added a weight $\frac{d}{10}$ to edge e_{12} because $D_{12}=-\frac{d}{2}<0$ and there is an edge between node i and j.

Step 2. Choose a path P_{ij} between any pair of nodes i,j of the symmetrized graph (Fig. 1.1B). For convenience, we choose the shortest path. Note that the choice of the shortest path does not always lead to the lowest synchronization threshold [44].

Our choice of paths is

$$P_{12}:e_{12} \qquad P_{13}:e_{12},e_{23} \qquad P_{14}:e_{15},e_{45} \qquad P_{15}:e_{15}$$
$$P_{23}:e_{23} \qquad P_{24}:e_{23},e_{34} \qquad P_{25}:e_{12},e_{15}$$
$$P_{34}:e_{34} \qquad P_{35}:e_{34},e_{45}$$
$$P_{45}:e_{45}.$$

Step 3. For each edge of the graph we determine the following inequality.

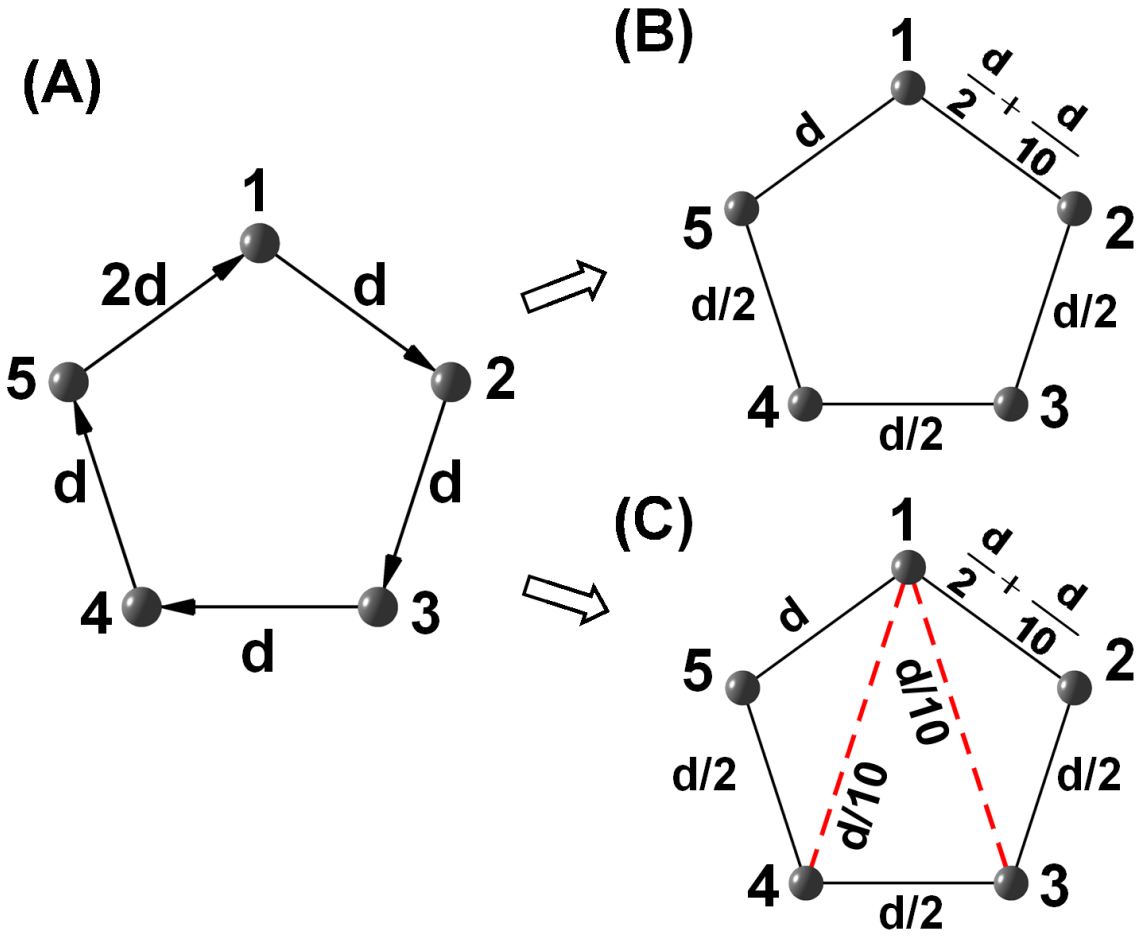


Figure 1.1. A five-node network.
(A) Original directed graph with different weights. (B) Symmetrized graph obtained using the previously developed method. The synchronization threshold is e_{45}: $d > 5a$. (C) Symmetrized graph obtained via the Augmented Graph method. The synchronization threshold is e_{34}: $d > \frac{10a}{3}$.

For e_{12} (link between nodes 1 and 2):

$$d_1 + D_1 = \frac{d}{2} + \frac{d}{10} > \frac{a}{5}\, b_k, \;\; \text{where} \;\; b_k = \sum_{j>i;\; k \in P_{ij}}^{n} L(P_{ij}).$$

The chosen paths that pass through the e_{12} are $P_{12},\ P_{13},\ P_{25}$. Their weighted lengths $L(P_{ij})$ are:

$L(P_{12}) = |P_{12}| = 1$ since $D_1^c + D_2^c < 0$; and there is an edge between nodes 1 and 2
$L(P_{13}) = |P_{13}|\chi(1 + \frac{D_{13}}{a}) = |P_{13}|(1 + \frac{0}{a}) = 2$
$L(P_{25}) = |P_{25}|\chi(1 + \frac{D_{25}}{a}) = |P_{25}|(1 + \frac{d}{2a}) = 2(1 + \frac{d}{2a})$

Summing up all the lengths, we get

$$\frac{d}{2} + \frac{d}{10} > \frac{a}{5}\ \left[1 + 2 + 2\left(1 + \frac{d}{2a}\right)\right].$$

Therefore, the synchronization condition for the e_{12} is $d > \frac{5a}{2}$.

Exactly as for the e_{12}, we can calculate the synchronization bounds for the other edges. These bounds are

$$\mathrm{e}_{12} : d > \tfrac{5a}{2} \quad \mathrm{e}_{15} : d > \tfrac{5a}{4} \quad \mathrm{e}_{23} : d > 2a$$
$$\mathrm{e}_{34} : d > \tfrac{10a}{3} \quad \mathrm{e}_{45} : d > 5a.$$

Hence, according to the Generalized Connection Graph method [41], the synchronization bottleneck for the entire network is the edge e_{45} where the maximum coupling strength is required to synchronize all oscillators of the network.

1.4.3 New Augmented Graph Stability method

In this subsection, we extend the Generalized Connection Graph method for proving synchronization in directed networks. Our approach, which we called the Augmented Graph Stability method, is based on the transformation of the directed graph into an undirected graph. This is done by replacing each direct link between node i node j with an undirected

edge whose coupling strength depends on the mean node unbalance between the two nodes. In addition, we augment the graph by adding an extra edge, connecting node i and node j if there is no directed link between them and their mean node unbalance is negative. Different weights are also associated with each path between any two nodes of the augmented undirected network, according to the mean node unbalance. Upper bounds on the coupling strength sufficient for synchronization in this augmented symmetrized network also guarantee global stability of synchronization in the original directed network. We show that the new augmented graph method is more effective than the Generalized Connection Graph method in sparse networks.

There are three steps in the new method. The differences are in symmetrizing the graph (Step 1), choosing the path (Step 2) and calculating the b_k for the inequality (Step 3).

Step 1. Symmetrize the graph by replacing each directed edge by an undirected edge with half the coupling strength and add quantity $\left|\frac{D_{ij}}{5}\right|$ to coupling strength, if $D_{ij} < 0$ and there is an edge in the symmetrized graph linking directly i and j.

New principal component of the Augmented Graph Stability method: If $D_{ij} < 0$ and there is no edge in the symmetrized graph linking directly i and j, then we add an edge in the graph (dotted red line in Fig. 1.1C). Then the quantity $\left|\frac{D_{ij}}{5}\right|$ assigns as the coupling strength to this augmented edge.

Step 2. Choose the same shortest path P_{ij} between any pair of nodes i, j of the symmetrized graph (Fig. 1.1C). The ingredient of the new method is that, when we add an edge, we choose it once to replace our previous choice where the node unbalance D_{ij} is negative. In this example, they are P_{13} and P_{14}. Our choice of paths is

$$
\begin{array}{llll}
P_{12} : e_{12} & \mathbf{P_{13} : e_{13}} & \mathbf{P_{14} : e_{14}} & P_{15} : e_{15} \\
P_{23} : e_{23} & P_{24} : e_{23}, e_{34} & P_{25} : e_{12}, e_{15} & \\
P_{34} : e_{34} & P_{35} : e_{34}, e_{45} & & \\
P_{45} : e_{45}. & & &
\end{array}
$$

Step 3. We recalculate the inequality when the edge is added, i.e. the quantity of b_k is re-calculated. The rest of calculations are same as in the previous method.

For e_{12} (link between nodes 1 and 2):

$$d_1 + D_1 = \frac{d}{2} + \frac{d}{10} > \frac{a}{5}\, b_k, \quad \text{where } b_k = \sum_{j>i;\, k\in P_{ij}}^{n} L(P_{ij}).$$

The chosen paths that pass through the e_{12} are $P_{12},\ P_{25}$. Their weighted lengths $L(P_{ij})$ are:

$$L(P_{12}) = |P_{12}| = 1 \text{ since } D_1^c + D_2^c < 0;$$
$$L(P_{25}) = |P_{25}|\chi(1 + \tfrac{D_{25}}{a}) = |P_{25}|(1 + \tfrac{d}{2a}) = 2(1 + \tfrac{d}{2a})$$

Summing up all the lengths, we obtain

$$\frac{d}{2} + \frac{d}{10} > \frac{a}{5}\ \left[1 + 2\left(1 + \frac{d}{2a}\right)\right].$$

Therefore, the synchronization condition for e_{12} decreases to $d > \frac{3a}{2}$.

For an additional edge e_{13} where the $b_k = 1$, we have: $\frac{d}{10} > \frac{a}{5} \cdot 1$. So $d > 2a$.

Exactly as for the e_{12} and e_{13}, we can calculate the synchronization bounds for the other edges. These bounds can be summarized as follows

$$\mathrm{e}_{12} : d > \tfrac{3a}{2} \quad \mathrm{e}_{13} : d > 2a \quad \mathrm{e}_{14} : d > 2a \quad \mathrm{e}_{15} : d > \tfrac{3a}{4}$$
$$\mathrm{e}_{23} : d > \tfrac{6a}{5} \quad \mathrm{e}_{34} : d > \tfrac{10a}{3} \quad \mathrm{e}_{45} : d > 3a.$$

Hence, according to the new method, the synchronization bottleneck for this entire network changes to e_{34}, where the maximum coupling strength reduced to $\frac{10a}{3}$.

1.4.4 Comparisons of the methods for other network configurations

To deepen our understanding of the new method, we apply the Augmented Graph Method to three other configurations. We find that both the previous Generalized Connection Graph method and our new method have their advantages when it comes to networks

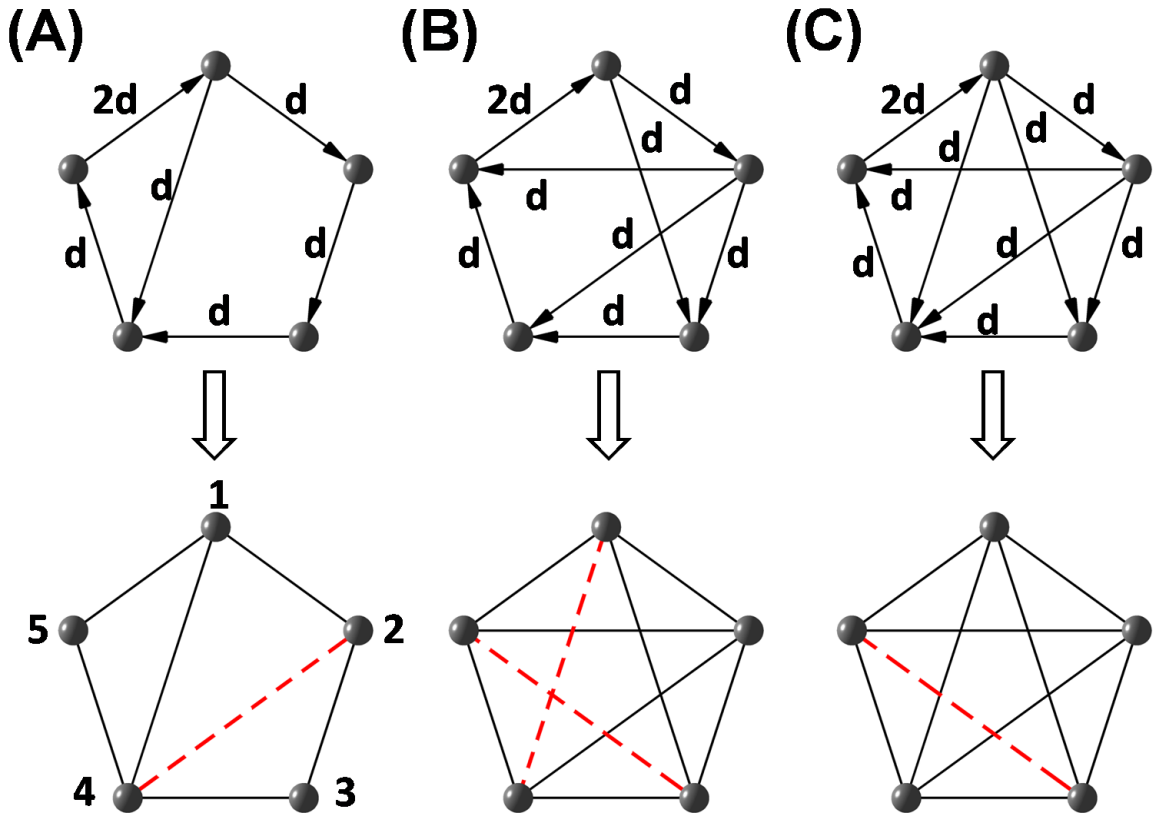


Figure 1.2. Three more configurations for methods comparison purpose.
(A) Sparser graph with six edges where the Augmented Graph method performs better: $d > \frac{10}{3}$. (B) Graph with eight edges where both methods give the same synchronization threshold: $d > 2a$. (C) Denser graph with nine edges where the previous method performs better: $d > \frac{6a}{5}$.

with different graph densities. Specifically speaking, in Table 1.1, we used both the Generalized Connection Graph method and the new Augmented Graph method to calculate the synchronization bounds for three networks of Fig. 1.2. This table demonstrates that the new method does give lower bound when the graph is sparser. For example, in Fig. 1.2A, adding an edge e_{24}, we can lower the load on edge e_{12}. However, in a denser graph of Fig. 1.2C, the new method can not lower the bound. Indeed, the added edge e_{35} with the new method, becomes a new bottleneck which increases the bound of the entire network. In between the sparse graph (Fig. 1.2A) and dense graph (Fig. 1.2C), there is a case that both methods yield the same bound (Fig. 1.2B row in Table 1.1). This occurs when the load of added edge (e_{14} or e_{35}) in the new method is equal to the bottleneck (e_{12}) of the old method.

Table 1.1. Comparison of the synchronization thresholds calculated using the Generalized Connection Graph method and Augmented Graph method in sparse and dense graphs.

	GCG[1]	BN[2]	AG[3]	BN[2]	EA[4]
Fig. 1.2A	$d > \frac{14}{3}a$	e_{12}	$d > \frac{10}{3}a$	e_{12}	e_{24}
Fig. 1.2B	$d > 2a$	e_{12}	$d > 2a$	e_{14}, e_{35}	e_{14}, e_{35}
Fig. 1.2C	$d > \frac{6}{5}a$	e_{13}	$d > 2a$	e_{35}	e_{35}

[1]Synchronization threshold calculated by using the Generalized Connection Graph;[2] Bottleneck: the edge where the maximum coupling strength is required to synchronize all oscillators of the network;[3] Synchronization threshold calculated by using the Augmented Graph method;[4] Edge Added according to the Augmented Graph method.

1.4.5 Computational algorithm and its application to larger irregular networks

Both of the old and the new methods have advantages on certain topologies. Typically, the two methods become more effective and give more correct information on the qualitative dependence of the synchronization thresholds on parameters of the network, while the number of oscillators composing the network increases. Unfortunately, the calculation of weighted path lengths can be quite a laborious task for larger networks with complicated

coupling schemes. Therefore, we have to develop pseudo-codes as an implementation of the algorithm for handling the computation. The algorithm first calculates the node unbalance and mean node unbalance for each node of the graph. It then augments the graph by adding an extra edge and connecting node i and node j if their mean node unbalance is negative. While re-weighing the graph similar to Fig. 1.1C, it chooses a shortest path P_{ij} between any pair of nodes i and j of the symmetrized graph. Finally, for each edge of the graph the algorithm determines the main inequality.

In the following implementation of pseudo codes, i and j represent the ith and jth nodes. k represents the kth edge. w_k represents the coupling strength of the kth edge. sw_k represents the coupling strength in the symmetrized graph. We require w_k to be sorted in an ascending order, according to the node's index. This is to guarantee that w_k and sw_k have the same order if no edge added to the graph and edge will be added, starting from $(m+1)$th element of sw_k. $|P_{ij}|$ represents the path length. $L(P_{ij})$ represents the weighted path length.

Input: Directed graph with various weights. **Output:** sc_k.

begin:

1. [initialize]

$l = 0$; $j = 0$; $sw_i = 0$; $k = 0$; compute node unbalance D_i^c and mean of node unbalance D_{ij} between node i and node j;

2. [symmetrize the graph, find the shortest path and compute weighted path length]

for node i from 1 to n

$j = i + 1$;

while $j \leq n$

find the shortest path between node i and j (i.e. using Dijkstra algorithm, please refer next subsection.);

$k = k + 1$;

if $D_{ij} < 0$

if there is an edge between node i and j

[re-assign a coupling strength and compute the weighted path length]

$sw_k = \frac{w_k}{2} + \frac{w_k}{2n}$; $L(P_{ij}) = |P_{ij}|$;

else [we augment the graph and change the shortest path between node i and j to this augmented edge]

$l = l + 1; sw_{m+l} = \frac{w_k}{2n}$; $L(P_{ij}) = 1$;

end if

else [half the coupling strength and compute the weighted path length there is an edge between i and j]

$sw_k = \frac{w_k}{2}$; $L(P_{ij}) = |P_{ij}|(1 + \frac{D_{ij}}{n})$;

end if

end while

end for

3. [compute the b_k and derive the inequality]

In case edge k from 1 to n

[count edge k's occurrence in the shortest path]

$$b_k = \sum_{j>i;\ k\in P_{ij}}^{n} L(P_{ij});$$

In case edge k from $n+1$ to $n+l$

$b_k = 1$;

solve the inequality $sw_k > \frac{a}{n} b_k$; record the solution as sc_k;

end.

Then we use this algorithm to compute the synchronization threshold for a randomly chosen directed graph. This graph (Fig. 1.3A) has 30 nodes, 37 edges and various coupling strength chosen from d, 2d and 3d. With this algorithm, we symmetrized the graph and adding new red dotted edges in the graph (Fig. 1.3B). The synchronization bottleneck for this network happened to be edge e_{89}. It is approximated to be $d > 25a$ by using the Augmented Graph method, while the old method yields $d > 44a$ for the same edge. This

shows that the Augmented Graph method does reduce the synchronization threshold in a sparse graph.

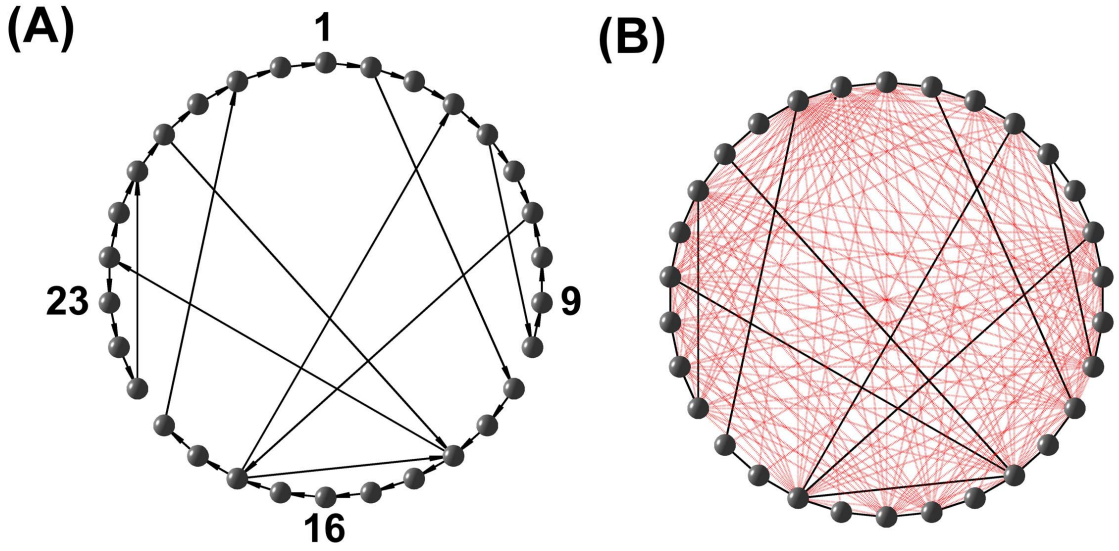


Figure 1.3. Calculating an upper bound for a sparse directed graph.
(A) A sparse directed graph with 30 nodes and 37 edges the weights are randomly assign from (d, 2d, 3d). The synchronization threshold using the old method is e_{89}: $d > 44a$. (B) Symmetrized graph with additional edges (red dotted line) obtained from the Augmented Graph method. The synchronization threshold is e_{89}: $d > 25a$.

1.4.6 Graph-based Stability method is path dependent

The proposed new algorithm is a path dependent method. That means that the choice of the path can yield different bottlenecks because the different selection of paths can lead to the load change on each edge. For example, in Fig. 1.4, our choice of paths is

$$P_{12} : a \qquad P_{13} : b \qquad \mathbf{P_{14} : b, c}$$
$$P_{23} : e \qquad P_{24} : d$$
$$P_{34} : c$$

However, one can also change the choice of P_{14} from {b, c} to {a, d}. In this case, one chooses to redistribute the load from {b, c} to {a, d}. Thus, the equation 1.25 has to be re-evaluated regarding these four edges. This implies that the synchronization threshold for the whole network has to be re-calculated. This observation has been illustrated in previous publications [44].

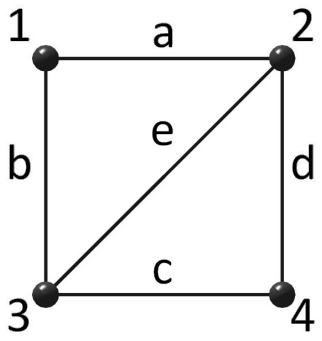


Figure 1.4. Augmented Graph Stability method is path dependent.

1.4.7 Augmented Graph Stability method can utilize the method for finding Shortest Path (SP)

It has been pointed out that choosing the shortest path may not give the minimum value of b_k [44]. However, we suggest that one may use the shortest path (SP) as path choices for calculating the synchronization threshold for the two following reasons. First, for a larger network, the derivation of the thresholds by hand is time-consuming as the size of network increases. One has to use an automated and well-developed method to handle this derivation. How to find the SP is one of classic combinatorial problems and it has been extensively studied, many efficient methods have been developed and are public accessible. Furthermore, in some cases, the rules/methods regarding how to choose the path (not necessarily the SP) for the b_k calculation may not be at hand immediately. The SP could be a first try.

The SP problem has a mathematical expression. Given a directed graph $G(V, E)$ where V and E are vertex set and edge set of G correspondingly. A constant c_{ij} represents fixed costs from the initial (source) vertex i to terminal (target) vertex j. A binary variable v_{ij} is equal to "1" only if there is an edge from vertex i to j; "0" otherwise. Therefore, the shortest path problem can be formulated as a linear integer programming problem:

$$\begin{aligned} \text{Minimize} \quad & \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}v_{ij} \\ \text{subject to} \quad & \sum_{j=1,j\neq i}^{n} v_{ij} - \sum_{j=1,j\neq i}^{n} v_{ji} = \phi_i \end{aligned} \tag{1.26}$$

where

$$\phi_i = \begin{cases} 1 & \text{if vertex } i \text{ is initial vertex;} \\ -1 & \text{if vertex } i \text{ is terminal vertex;} \\ 0 & \text{otherwise.} \end{cases}$$

Different algorithms used for shortest path selection may result in running time difference. We use the Dijkstra algorithm to select the shortest path for two reasons. First, it is an efficient algorithm and can be easily found in many textbooks (i.e. [45]). Second, it is easy to combine the re-weight procedure together with the shortest path algorithm. Of course, one may prefer other algorithms such as Dynamic Programming [46], A-star [47], etc. to substitute the Dijkstra. The Dijkstra algorithm runs in $O(m\log(n))$ time where m and n represents the number of edges and number of nodes in the graph respectively [45]. Therefore, if the proposed implementation utilizes the shortest path calculated by the Dijkstra methods will have a running time of $O(n \cdot (n-1) \cdot m \cdot \log(n))$ approximately.

1.4.8 Graph-based Stability method may also utilize k^{th} shortest path (KSP)

If the shortest path is not available for some reasons, the second can be used and the synchronization threshold have to be recalculated. If the second one is not available either, the third shortest path will be used, so on and so forth. This case may not lead us to consider the synchronization problem alone but to consider the KSP problem as an embedded problem. Therefore, the study of the SP-like problem and even the KSP problem itself might be helpful to further understand the Graph-based Stability Methods. Please refer to Appendix B for the details.

1.5 Conclusions

In this Chapter we have addressed an important question, regarding network synchronization: What is the stability criterion for synchronization in networks of identical (or nearly identical) oscillators stable, especially in regard to network topology and coupling strengths? This general question had been widely discussed, and powerful stability methods for network synchronization had been developed. The most popular approaches include the Master Stability function and the Connection Graph method. Both methods, originally developed for undirected networks, have been generalized to handle networks with directed connections. In this Chapter, we have developed a modification of the generalized Connection Graph method that gives tighter bounds on the coupling strength required for the onset of stable synchronization in sparse directed networks. We showed how the directed network can be turned into an augmented undirected network with weighted connections. The stability conditions for synchronization in this augmented directed network also ensure stable synchronization in the original directed network.

We hope that this method not only inspires research on complex networks, but may have applications to the synchronization phenomena in biology and engineering. We also hope this method can contribute to deepening our understanding towards neurological disorders

caused by synchronization of neurons at a network level. In the next Chapter, we will develop a graph-based method to solve disease related problems at a molecular level.

CHAPTER 2

MATHEMATICS IN PROTEOMICS

2.1 Background

Ca^{2+}, a secondary messenger in cellular signal transduction, plays an important role in many biological processes, including the regulation of cell division, differentiation, and apoptosis in the cell life cycle [48–51]. Ca^{2+}-binding proteins are significantly related to serious diseases such as Alzheimer's disease [3], heart disease [4], diabetes [4] , leukemia [5, 6], and cancers [7–10]. From a molecular perspective, mutations in close proximity to the Ca^{2+}-binding sites often alter a protein's ability to bind Ca^{2+}, a malfunction which is sometimes the primary cause of diseases [52–54]. Therefore, identifying Ca^{2+}-binding sites in proteins is a crucial step towards understanding the molecular basis of diseases related to Ca^{2+}-binding proteins. As illustrated in Fig. 2.1A, the coordination of Ca^{2+} utilizes various classes of oxygen atoms from carboxyl groups (Asp, Glu), carboxamide groups (Asn, Gln), and hydroxyl groups (Ser, Thr) in side chains, carbonyl oxygen atoms of most residues in the main chain, and from cofactors and water molecules. The majority of all Ca^{2+}-binding ligands originate from turn/loop regions [55–58]. Previous studies have revealed that Ca^{2+} is coordinated by 3-8 oxygen ligand atoms [57, 59–61] with an average of 6 ligands for all Ca^{2+}-binding sites, or 7 ligands for only EF-hand sites [56]. These hydrophilic oxygen atoms are embedded within multiple, concentric shells of hydrophobic carbon atoms [62]. A majority of Ca-O bond lengths fall within the range 2.2-2.9$\mathring{A}$ and Ca-C bond lengths fall within the range 2.4-4.6$\mathring{A}$ in Ca^{2+}-loaded X-ray structures [63].

Computational methods to predict Ca^{2+}-binding sites have been actively pursued using various approaches [51, 64–66]. Most of the published structure-based Ca^{2+}-binding site prediction algorithms, including FEATURE [67], Fold-X [68], and the approaches by Nayal et al. [59] and Yamashita et al. [62], rely on the spatial coordinates of ligand oxygen atom-

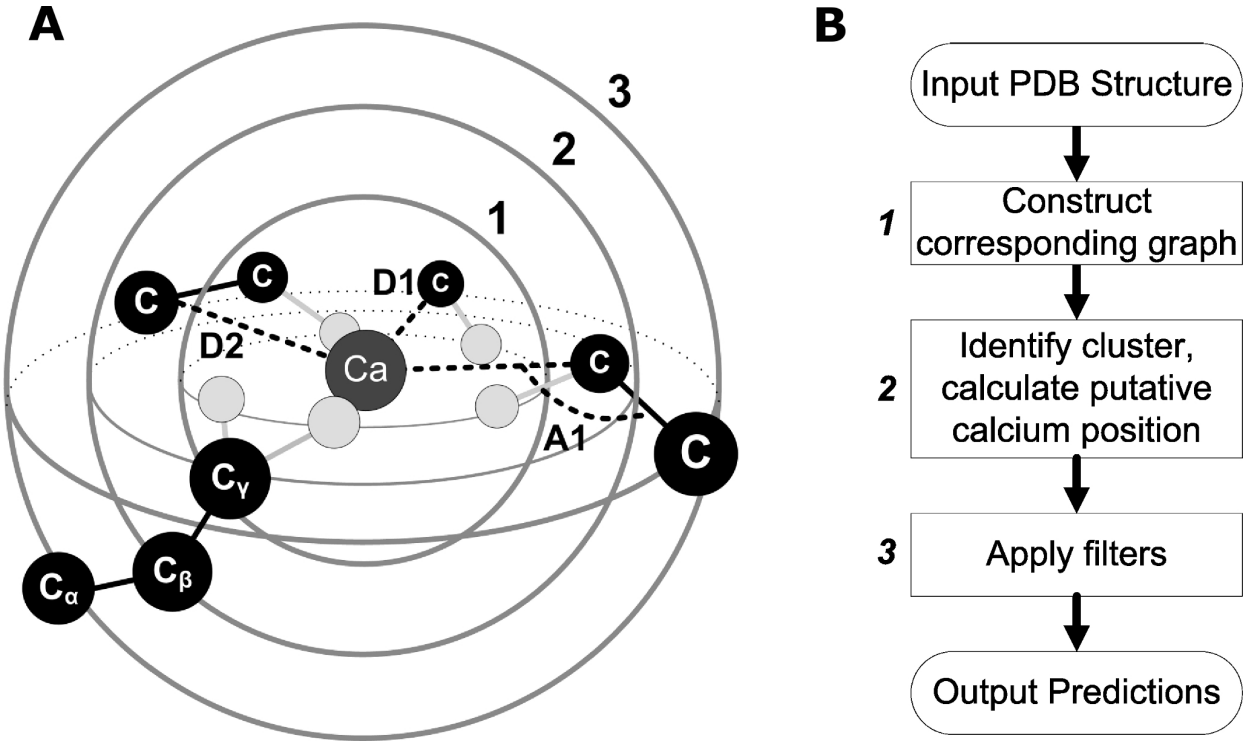


Figure 2.1. Definition of shells and algorithm workflow.
(A) The central Ca^{2+} is coordinated by the first shell of oxygen atoms (light gray), which is concentrically embedded into two other shells of carbon atoms (black). Depending on the length of the alkyl side chain, an atom of the second or third shell has a covalent bond with an atom from the first or second shell. D1 represents the distance between Ca^{2+} and second shell carbon atoms. D2 is the distance between Ca^{2+} and third shell carbon atoms. A1 stands for the angle formed by Ca^{2+} and the second and third shell carbon atoms, respectively (Ca-C1-C2). (B) Workflow of MUGC.

s. Previous work has led to the development of two algorithms, GG [69] and MUG [70], for predicting Ca^{2+}-binding sites by constructing a corresponding graph for each protein with a graph theoretic algorithm to identify oxygen atom clusters [69, 70]. These analyses of binding site geometry have been based mainly on X-ray structures deposited in the Protein Data Bank (PDB), and the prediction approaches derived from them have been tested mostly on X-ray structures with high resolutions. Unfortunately, Ca^{2+}-binding sites with weak affinity (0.05-2 mM) often remain unidentifiable or "invisible" in crystal X-ray structures due to low occupancy and conformational ensembles. For example, although extracellular Ca^{2+} is known to regulate family C of GPCR, Ca^{2+} was not observed in more than 20 X-ray structures of metabotropic glutamate receptor (mGluR) [71, 72]. Further prediction of Ca^{2+} binding sites in X-ray structures of low resolution and homology models requires the capability to overcome large errors and incorrect assignments of the side-chain oxygen atoms [73, 74]. As a complementary technique of structural elucidation, NMR offers us additional insights into Ca^{2+}-binding proteins [75, 76]. NMR structures differ from X-ray structures in that, typically, a whole ensemble of low energy conformations satisfying the experimental constraints is obtained from the structural calculations. These structures represent the dynamic nature of the protein in solution, in contrast to the static state of a crystal structure. However, the Ca^{2+} ions cannot be directly observed in NMR experiments, but rather are positioned in the structure based on indirect effects exhibited by chemical shifts and constraint-based assumptions. A barrier to identifying Ca^{2+}-binding sites in protein structures derived by NMR is that the geometric coordination of Ca^{2+}-binding sites cannot be determined by direct observation of Ca^{2+}, and this difficulty is compounded by the fact that the positions of the oxygen atom ligands that fix the Ca^{2+} position are not directly determined either, but extrapolated from templates of their residues, because the isotopically-abundant ^{16}O has an intrinsic zero nuclear spin. The previous work detailed the development of several graph theoretic algorithms to predict Ca^{2+} binding sites in proteins based on identification and refinement of oxygen clusters [70]. Results of these studies further suggested that the algorithm could be extended to observe the carbon atoms asso-

ciated with the oxygen binding ligands which would allow us to predict Ca^{2+}-binding sites in proteins where the Ca^{2+} ion may not be directly observable (e.g., low resolution structures, weak affinity binding sites, and NMR structures). We therefore hypothesized that the second, hydrophobic shell of carbon atoms enclosing a Ca^{2+}-binding site could sufficiently determine the site's location in either X-ray or NMR structures. To test this, we developed a new algorithm, MUG^C, which is capable of predicting Ca^{2+}-binding sites by pinpointing the Ca^{2+} ion position using carbon clusters (i.e., concentric rings of carbon atoms surrounding a ring of oxygen atoms chelating the Ca^{2+}, Fig. 2.1A), and applying filters based on the centers of mass of side-chain and main-chain oxygen atoms. We have applied MUG^C to delineate Ca^{2+}-binding sites in both X-ray and NMR protein structures without reference to explicit side-chain oxygen ligand atoms. The metal selectivity of MUG^C has been further evaluated by analyzing three additional protein datasets containing Mg^{2+}, Zn^{2+}, and Pb^{2+} binding sites. Additionally, MUG^C was evaluated with a negative control dataset consisting of protein structures not known to bind Ca^{2+} or other metal ions. Our results demonstrate not only that the Ca^{2+}-binding sites in NMR and X-ray structures can be identified based on geometric arrangement of the second-shell carbon cluster, but that this approach with Ca^{2+}-optimized selection parameters, can also selectively differentiate between Ca^{2+} and other relevant divalent cations. We further anticipate that application of this algorithm will enable us to identify previously-unknown Ca^{2+}-binding sites, deepen our understanding of structural characteristics of Ca^{2+}-binding sites, and improve our ability to design Ca^{2+}-binding proteins with diversified functions [77].

2.2 Methods

2.2.1 Definition of carbon shells

As seen in Fig. 2.1A, the Ca^{2+} ion is bound by charge interactions to oxygen atoms either from side-chain residues (e.g., Glu or Asp) or main-chain carbonyl oxygen. These atoms, in turn, are covalently bound to carbon atoms, which constitute a second shell. A

third shell of carbon atoms can be defined as carbon atoms covalently bound to a second shell. The two concentric shells of carbon atoms, in our hypothesis, constitute a scaffold which determines the central binding site. A set of physical parameters describing the spatial relationship of the atoms comprising the binding site can be defined by the angle Ca-C1-C2 and the distance between Ca^{2+} and C1 (D1 in Fig. 2.1A) and by the distance between Ca^{2+} and C2 (D2 in Fig. 2.1A), where C1 and C2 are carbon atoms within the second and third shells, respectively. The binding site, which includes both the Ca^{2+} and oxygen atoms, is enclosed in a second shell defined by a particular carbon cluster. The Ca^{2+} position then can be calculated by geometric parameters related to the second and third shell carbon atoms.

2.2.2 General description of algorithm

In general, execution of this algorithm involves three major steps (Fig. 2.1B). In step 1, taking a PDB structure (i.e. 3CLN) as input, we construct the protein topological graph whose vertices are the carbon atoms with associated oxygen atoms. Two vertices share an edge if the distance between them is less than some defined threshold. In step 2, we search for all maximum cliques in the graph to identify carbon clusters, and tentatively position Ca^{2+} at the geometric center (Ca^{2+} center) of each cluster. These clusters are required to have at least four carbon atoms, ensuring a minimum of four oxygen atoms in the site available to chelate Ca^{2+} [57, 78]. In step 3, we apply three different filters to remove clusters that are not suitable for Ca^{2+}-binding. The remaining clusters, as well as the Ca^{2+} center of each cluster, are the predicted Ca^{2+}-binding sites. When using dynamic NMR structures for prediction, MUG^C screens the best-fit site among all members of the ensembles and uses more inclusive geometric parameters than when using X-ray structures.

2.2.3 The topological graph of protein carbon atoms

To localize the initial calculation of the Ca^{2+} position, we construct a graph representation of the protein. First, we extract all Cartesian coordinates of carbon atoms covalently

bonded to oxygen atom(s) and calculate the distances between all of these carbon atoms. Then we construct a graph G(V, E) where V is the vertex set and E is the edge set of G. A vertex in V represents one extracted carbon atom. An edge is assigned between two vertices if the distance between these two vertices (C-C distance) is smaller than a predetermined cutoff (7.5 $\AA$ for X-ray structures and 8.3 $\AA$ for NMR structures). The constructed graph is then recorded in an adjacent matrix (Table E.7). For example, calmodulin has four binding sites (Fig. 2A and Fig. 2B). It also has a total of 209 carbon atoms covalently binding to oxygen atoms. After we construct its topological graph (Fig. 2C and Fig. 2D), the four binding sites are clearly discernible as regions of dense convergence in the graph.

2.2.4 Center of mass

Proteins in solution, especially their flexible side chains, are in constant motion. To deal with this motion, we use the abstracted side-chain mass center (Fig. 2A) as the reference for predicting Ca^{2+} position. Side-chain center of mass is beneficial because it reduces sensitivity to errors in the specific locations of side-chain atoms.

2.2.5 Ca^{2+} localization algorithm

After preparation of the topological graph and side-chain center of mass for a given protein, we first search all maximum cliques in a graph constructed from the carbon atoms. Finding all maximal cliques of a general graph is an NP-hard problem, [79] requiring more than polynomial computation time to process. Fortunately, in the generated carbon atom graph, the size of any maximal clique never exceeds ten. This ceiling is not a theoretical one, but a pragmatic consequence of our considering only carbon atoms which are covalently bonded with oxygen atoms. These carbon atoms maintain some distance from each other due to the charge repulsion from the attached oxygen atoms. Based on these properties, we apply a well-established algorithm of Bron and Kerbosch [80] to produce all the maximal cliques efficiently. In our case, the maximal cliques are generated within O(n) time, where n is the number of vertices in graph G.

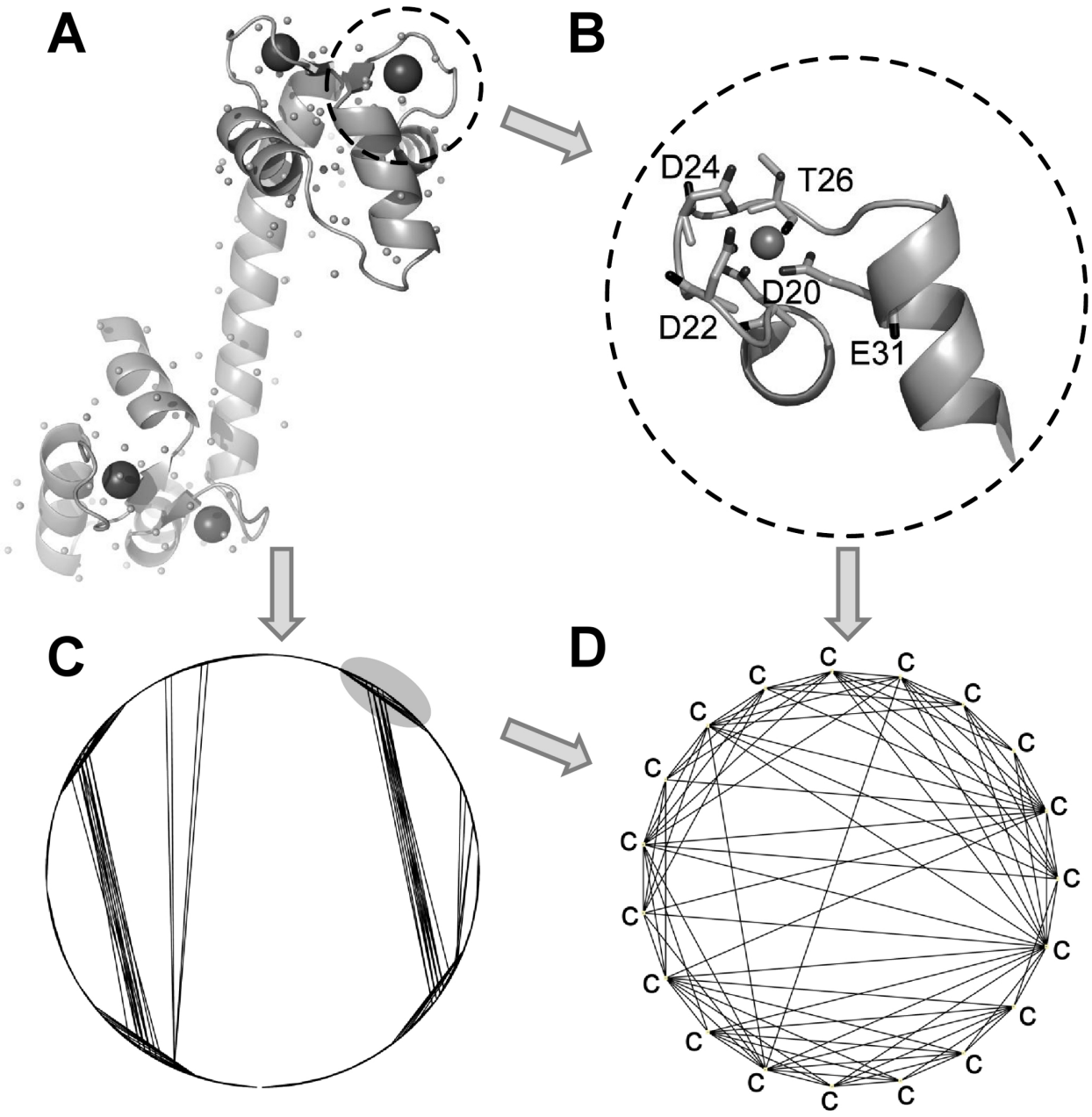


Figure 2.2. The structure of calmodulin (CaM) and topological graph of carbon atoms. (A) CaM with center of mass of side chain (the small dots). (B) Ca^{2+} binding site EF-I of CaM. (C) Topological graph of all carbon atoms in CaM associated with potential oxygen ligands (includes both side-chain and main-chain carbon atoms in putative binding residue). (D) The graph of CaM site EF-I loop.

2.2.6 Constraints and filters

We tentatively place Ca^{2+} in the geometric center of the carbon clusters, and then determine if they qualify based on constraints from various filters including the center of mass of side-chain, elimination of redundant predictions, van der Waals clashes, formal charge, and geometric constraints. Initial parameters were selected based on parameters used in previous studies and statistical analyses conducted [57, 69, 70, 78]. These parameters, including cutoff distances, were then optimized based on values for selectivity and sensitivity from analysis of the training dataset. These optimized parameters (Table E.8) were then applied to the test dataset. For example, the range of distance between Ca^{2+} and second shell carbon atom (D1 in Fig. 2.1A) is reported to be between 3.0 - 4.6$\mathring{A}$ for main-chain carbonyls [57]. The covalent bond length between second shell carbon atoms and its next-outer shell carbon is 1.54$\mathring{A}$. Therefore, we can estimate that the distance between a Ca^{2+} and the third shell carbon atoms may not exceed 6.14$\mathring{A}$ and should also be greater than D1. If a predicted Ca^{2+} position falls outside of this range, this position is not likely a correct prediction.

2.2.7 Performance evaluation on binding sites and binding residues

A Predicted True Site (PTS) is a true Ca^{2+}-binding site for which there is at least one Correct Hit (CH). Sensitivity (SEN) is applied to represent the percentage of PTS in all Documented Sites (DS). Selectivity (SEL) is applied to represent the percentage of Correct Hit (CH) in Total Predictions (TP). Sensitivity measures the proportion of actual binding sites which are correctly identified. Selectivity measures the proportion of predicted binding sites which are correct. Higher selectivity indicates fewer false positive predictions (=over-predicted sites). Higher sensitivity and selectivity are important for reducing the number of predictions and classification errors.

$$SEN = (PTS)/(DS) \times 100\%$$

$$SEL = (CH)/(TP) \times 100\%$$

As MUG^C predicts both Ca^{2+} position and binding residues, Correct Hit (CH) could be defined in two ways. In the first definition, a CH is a predicted position falling within a specific distance (here 3.5 $\mathring{A}$ [69, 70, 81]) of the documented Ca^{2+} position. In the second definition, a CH is a predicted cluster of residues that contains at least two true Ca^{2+}-binding residues [78]. In NMR, where Ca^{2+} is not observable, we measure the prediction performance by comparing the predicted residues to the holo X-ray crystal structures.

2.2.8 Algorithm implementations and computation time

The implementation language is mainly Java and Perl on URSA (a 576 core Super Computer based on the Power5+ processor and IBM's P series architecture). The original source codes are available upon request. Matlab, Mathematica and PyMOL were used for graphing and visualization. LPC/CSU online servers were used for identify binding ligand from holo structures [82]. The computation time of predicting one protein depends on the size of the protein and the proximity of carbon atoms to one another in the spatial structure, which affects the time required for our graph algorithm to search for all maximum cliques. In terms of CPU time, it may depend on what kind the computer we are using as well. For example, the protein calmodulin (3cln) has four classic Ca^{2+} binding sites and 148 residues. Analysis of this file on our supercomputer takes less than 15 seconds, while the same analysis on our personal computer (Intel Celeron M 370(1.5GHz) with RAM 512MB) takes close to two minutes. On the other hand, as we increase the number of pdb files to be processed during a single run of the algorithm, or evaluate pdb files for very large proteins, the processing time increases significantly. A supercomputer is desired as multiple protein structures can be processed simultaneously for the purpose of developing the parameters and evaluating the performance of MUG^C. Although we have not specifically calculated the rate and magnitude of this increase, we expect it to be exponential based on analyses of the previous algorithms performance [69, 70].

2.3 Results

2.3.1 Non-redundant X-ray dataset

To validate our hypothesis, we used two X-ray datasets: a training dataset (Table 2.3 and E.11), a testing dataset (Table 2.4 and E.12), and a negative control dataset (Table E.18). For the datasets we generated, "non-redundant" applies to sequence identity which means that we removed sequences with 90% similarity. For the published dataset, we made sure that no identical proteins were included within a single dataset. This also applied to NMR dataset.

The X-ray training dataset (Table E.11) was originally from Schymkowitz et al. [68] The X-ray testing dataset (Table E.12) was reproduced by incorporating the Ca^{2+}-binding proteins from Pidcock and Moore's datasets [56] and the validation structures for NMR testing dataset. We eliminated the redundant proteins in the datasets and revised the testing datasets to have at least one binding site in each protein coordinated by at least four binding ligand atoms. Binding sites with low coordination numbers (three or less) may be due to crystal packing or non-specific binding, imply reduced stability and lower binding affinity at best [78]. The X-ray training dataset contained 18 proteins with 45 documented Ca^{2+}. The testing dataset contained 43 proteins with 108 documented protein-coordinated Ca^{2+}. The X-ray training and testing datasets contained continuous (e.g. lactalbumin: 1B9O.pdb and calcineurin: 1AUI.pdb), semi-continuous (e.g. lipase: 1OIL.pdb and proteinase K: 2PRK.pdb), and discontinuous binding sites (e.g. thermitase: 1THM.pdb and penicillin acylase: 1AI4.pdb). The negative control dataset contained 24 proteins selected at random with resolution $\leq$ 2.0 $\AA$, less than 90% sequence homology, and no indication of metal binding sites in the selected structure or in related structures. All X-ray crystallography structures were obtained from the PDB.

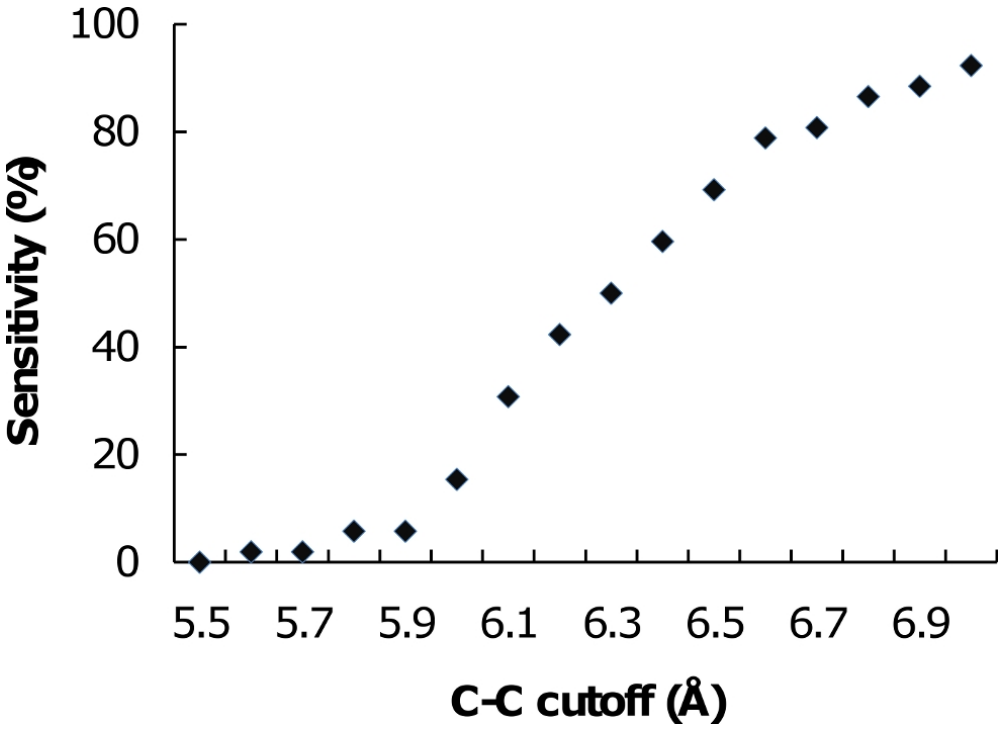


Figure 2.3. Performance in terms of sensitivity on X-ray dataset depending on C-C cutoff.

2.3.2 Sensitivity depending on C-C cutoff

Sensitivity of MUG^C was found to increase as the C-C cutoff increases (Fig. 3) on the X-ray training dataset. This is consistent with the previous finding that O-O cutoff is positively correlated with sensitivity [69]. We have used the larger $7.5\mathring{A}$ as cutoff, because this accommodates a distance twice the length of the combined Ca^{2+}-O and C-O bond lengths and we have developed effective methods to eliminate false positives within this range.

2.3.3 Eliminate false positive predictions with Filters

One of the concerns arising from not directly utilizing coordination atoms to predict Ca^{2+}-binding sites in proteins is the possibility of large number of false positive predictions. To reduce the number of reported false positive predictions, three types of filters were incorporated into the algorithm: 1). A charge filter, which requires that at least one

negatively-charged residue is present within the tentative binding site; 2). Geometric shell filters, which select the putative sites according to geometric relationships between the calculated Ca^{2+} position and the second and third shell carbon atoms; 3). Filters based on side-chain center of mass and van der Waals clashes. The side-chain center of mass is used in conjunction with main-chain oxygen atoms. If a main-chain oxygen atom is under consideration as the binding ligand, then the distance between the side-chain center of mass and Ca^{2+} must be greater than that of the Ca-O (carboxylic) distance in the X-ray structure. We use calmodulin (3CLN.pdb) from the X-ray training dataset to illustrate how these filters work. First, we used vertices representing 209 carbon atoms, using 7.5 $\mathring{A}$ as C-C cutoff, to construct a topological graph (see Methods). By searching all maximal cliques in the graph, 4626 non-redundant carbon clusters comprised of four or more carbon atoms were obtained. Among the 4626 clusters, 4589 are false positive predictions. The charge filter first eliminates 1639 carbon clusters. Next, the geometric shell filters eliminate an additional 2453 clusters, including 1405 clusters where the distance between Ca^{2+} center and third shell carbon atom is smaller than the distance between Ca^{2+} center and the second shell carbon atom, and another 1048 are eliminated based on previously-reported geometric parameters. [57] The third and final filter eliminates another 497 clusters. For example, we assume that the clash radius between Ca^{2+}-nitrogen is 2.55 $\mathring{A}$. If the distance between the Ca^{2+} center and each nitrogen atom is smaller than this value, we consider that there exists a clash and eliminate this cluster. Parameterization details are provided in appendix C. In calmodulin carbon clusters which sequentially passed all filters, are scored as firm predictions; this number is consistent with the documented binding sites. We also have applied the filters separately, to illustrate improved results obtained by sequential combination. The eliminated clusters are summarized in Table 2.1.

2.3.4 Performance on X-ray testing dataset

MUG^C was evaluated with the Ca^{2+}-loaded X-ray testing dataset (Table E.12). Out of the 108 documented protein-coordinated Ca^{2+} ions in the testing dataset, 99 are chelated

Table 2.1. False positive predictions remaining following applications of different filters in either consecutive sequence[a] or individually[b].

	Chg[c]	Geom[d]	COM[e]
Sequential	$\frac{2950}{4589}$	$\frac{497}{2950}$	$\frac{0}{497}$
Individual	$\frac{2950}{4589}$	$\frac{129}{4589}$	$\frac{267}{4589}$

[a]Filters were applied consecutively;[b]Each filter was applied individually;[c]Geometric filter;[d]Center of mass and clash filter. Numerator represents remaining false positive predictions.

by more than three binding residues. If we use the predicted Ca^{2+} position (CP) as a measure, MUG^C identified 102/104 sites with coordination numbers greater than three. Five of the binding sites in this dataset have only three binding residues each. In terms of binding residues (BR), MUG^C is able to identify 98/99 binding sites having more than three binding residues and 4/9 binding sites having three or fewer binding residues (Table 2.2 and Table E.12). The only binding site that was overlooked by MUG^C due to the fact that no negatively-charged residues are encountered in the binding site. This is discussed in greater detail in the Discussion section.

Table 2.2. Performance on 43 proteins with 108 Ca^{2+} in testing X-ray dataset, measured by CP[a] and BR[b].

	CP	BR
TDS[c]		
SEN[d]	94%	94%
SEL[d]	76%	43%
CN[f] $(n > 3)$		
SEN[d]	98%	98%
SEL[d]	76%	43%

[a]Prediction based on Ca^{2+} position;[b]Prediction based on binding residues;[c]Total documented sites;[d]Sensitivity;[e]Selectivity;[f]Coordination number.

For the negative control dataset comprised of proteins without known Ca^{2+}-binding sites, we define True Negative (TN) as any prediction which does not identify a Ca^{2+}-binding site, and False Negative (FN) as any prediction which does identify a Ca^{2+}-binding site. Based on these criteria, MUG^C correctly predicted 16/24 proteins as not being Ca^{2+}-binding proteins, with the remaining 8/24 proteins incorrectly identified as having Ca^{2+}-binding sites. A summary of predictions for this dataset is reported in Supplemental Table E.18. The prediction success rate (66%), while lower compared to values reported for sensitivity and selectivity with the testing dataset, still indicates that the majority of proteins were identified correctly, and we can further speculate that one or more of the 8 FN predictions may be Ca^{2+}-binding sites that remain to be identified as such. These results show that our hypothesis is valid on X-ray-derived Ca^{2+}-loaded structures.

2.3.5 Structural difference between X-ray crystallographic sites and NMR solution sites

Ca^{2+} binding sites with high affinity in X-ray structures are well defined due to direct observation of electron density of the metal and its coordinating oxygen atoms. For example, the static features of EF-hand Ca^{2+}-binding sites in proteins such as a troponin C exhibit structurally-similar pentagonal bipyramidal geometries (Fig. 2.4A). This geometry is well conserved in more than 10 X-ray structures of troponin C [60]. In contrast, Ca^{2+}-binding sites in NMR structures usually are not well defined due to lack of direct observable constraints and the dynamic ensembles. In addition, Ca^{2+}-binding sites are often located on the highly solvent-accessible surface, which reduces the possible connectivity that can be used to define the Ca^{2+}-binding site. For example, the high-resolution structure of troponin C (2TN4.pdb), determined in the presence of 10 mM of Ca^{2+}, has 23 structures in its NMR ensemble. Surprisingly, the third Ca^{2+}-binding site (D103, N105, D107, Y109 and E114) in the least-energy (first) structure of the ensemble cannot be recognized as a Ca^{2+} site by the criteria developed for static structures.

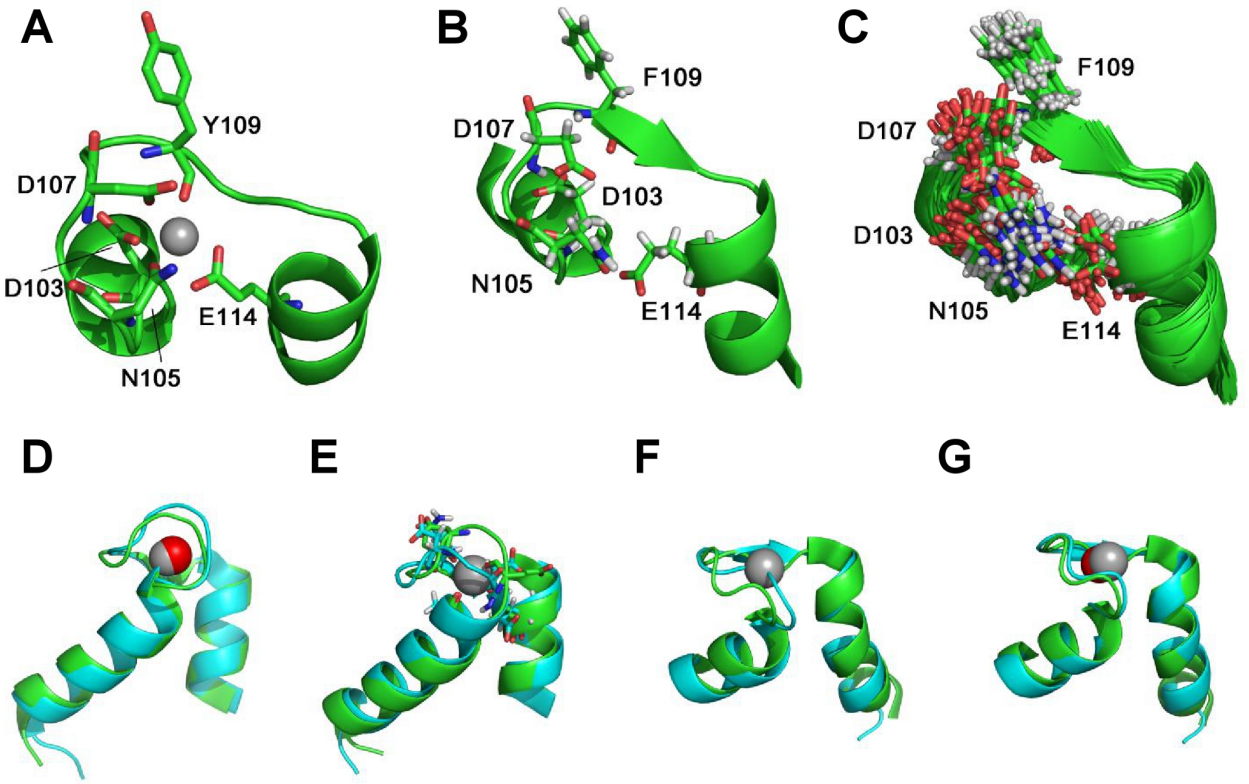


Figure 2.4. Structure comparison between X-ray holo and NMR structures. (A) X-ray structure of troponin C (2TN4.pdb) at a resolution of 2.00 $\AA$ (B) First ensemble of NMR troponin C (1TNW.pdb) determined without Ca^{2+} constraints. (C) All conformations in the NMR ensemble of troponin C (1TNW.pdb), determined without Ca^{2+} constraints. Sub-figures (D) through (G) indicate the alignments of the binding site in calbindin D9K NMR structures inferred without Ca^{2+} constraints (blue) and holo X-ray structure (green). Ca^{2+} in X-ray is gray and the geometric center of a carbon cluster in the NMR structure is red. (D) Ca^{2+} can be placed in the binding site formed by the loop A14-E27 in this first member of the ensemble. (E) The binding site formed by the loop D54-E65 of the first member of the ensemble does not appear to accommodate Ca^{2+}, though it is present in the X-ray structure (gray). (F) Similarly, the binding site formed by the loop A14-E27 of the second structure in the ensemble cannot accommodate Ca^{2+}, while (G), that formed by the loop D54-E65, can.

Figure 4B illustrates this lowest-energy structure, while Fig. 4C shows a composite of all structures in the ensemble. Dynamic motion of the Ca^{2+}-binding sites is implicit in the NMR ensemble, where an ideal binding conformation may exist only temporarily. Such observations motivated us to investigate the performance of algorithms on predictions of NMR structures.

2.3.6 Non-redundant NMR dataset

To validate our hypothesis on NMR structures, we used a published training dataset [78] (Table E.13) and constructed a testing dataset (Table E.14). The training NMR dataset (Table E.13) contains six, EF-hand-type Ca^{2+}-binding proteins with a total of 16 binding sites. In four of these the authors originally deposited structures for which they imposed Ca^{2+} constraints in determining the structures: calmodulin (2BBM.pdb), parvalbumin (2PAS.pdb), yeast frequenin (1FPW.pdb), and epidermal growth factor receptor pathway substrate 15 (1C07.pdb). It is not possible to project the original structures as they might have been constructed without invoking the Ca^{2+} constraints. In the other two cases (troponin C: 1TNW.pdb and calbindin D9K: 2BCB.pdb) the structures submitted were not modified based on Ca^{2+} constraints. We felt it important to include in the testing set only NMR structures which were calculated without use of Ca^{2+} constraints. The testing dataset (Table E.14) contains 11 NMR structures, all of which meet this criterion. Two additional criteria were imposed: i) The data corresponded to the holo forms of the proteins (i.e., all binding sites were occupied by Ca^{2+}; ii) The NMR structures had corresponding holo structures derived crystalographically, so that prediction results could be validated.

2.3.7 Analysis of C-C distance and geometric centers on a NMR training dataset

We analyzed the C-C distance of binding sites in the NMR structures with and without Ca^{2+} constraints added to the structural calculations. Each ensemble in the NMR training dataset was evaluated. If the total number of ensembles was greater than 20, we used only the first 20 ensembles in our training NMR dataset. This data reveals that in the NMR

structures with Ca^{2+} constraints, the second shell C-C distances are clustered from $4\mathring{A}$ to $7\mathring{A}$, and 90% of the distances fell below $8.3\mathring{A}$, which was used as cutoff for identification of the majority of the carbon atom clusters. The distribution of C-C distances in NMR binding sites exhibits a lower mean and smaller deviation in the constrained structures (Fig. 2.5A) as compared with structures lacking Ca^{2+} constraints (Fig. 2.5B). This is consistent with our intuition that the addition of Ca^{2+} to the structures pushes carbon clusters closer to each other in the binding sites, and therefore that the NMR structures should be close to their X-ray holo counterparts.

There exists at least one structure in the ensemble that is similar to the site conformation seen in models derived from X-ray diffraction of holo structures. Naturally, such sites are recognized as having canonical Ca^{2+}-binding geometry. For example, in the NMR structures of calbindin D9K (2BCB.pdb, derived without Ca^{2+} constraints), we observe that the geometric Ca^{2+} center determined by the main-chain carbon atoms of residues E17, D19, Q22, together with side-chain carbon of E27, is geometrically similar (within $0.55\mathring{A}$) to the Ca^{2+} center documented in the holo X-ray-derived structure (4ICB.pdb). Fig. 2.4D shows this NMR-observed binding loop superimposed on the X-ray structure. Similar congruity is seen between the geometric center fixed by side-chain carbon atoms from D54, N56, D58, E65 and main-chain carbon from E60 as seen in the holo X-ray structure and the second-ranked structure in the NMR ensemble (Fig. 2.4G). These observations encouraged us to use more inclusive parameters for the carbon clusters on NMR structures and predict Ca^{2+}-binding positions based on all ensembles.

2.3.8 Performance on NMR training dataset and testing dataset

For the training dataset (Table E.13), MUG^C identified all binding sites with a selectivity of 88%. For the testing dataset (Table E.14), MUG^C predicted 20 Ca^{2+}-binding sites out of the (X-ray authenticated) 21 binding sites with 95% sensitivity and 81% selectivity. These results show that using second shell carbon atoms can predict Ca^{2+} positions in the NMR structures calculated with or without Ca^{2+} constraints. Among NMR structures, the

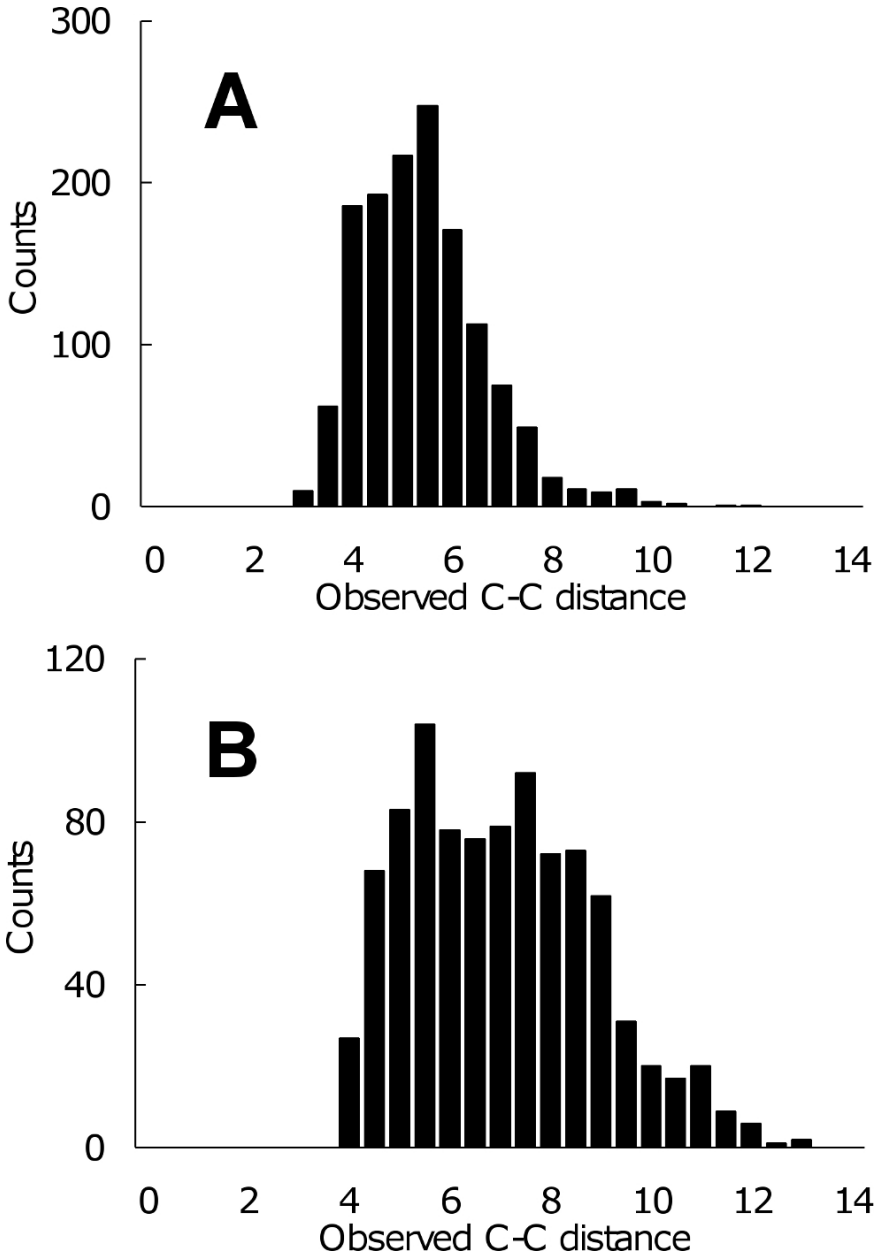


Figure 2.5. C-C distances analysis.
(A) four NMR structures from the training dataset with Ca^{2+} constraints (1C07.pdb, 1FPW.pdb, 2BBM.pdb and 2PAS.pdb). (B) Troponin C NMR structures without Ca^{2+} constraints (1TNW.pdb).

second binding site of the human centrin 2 (in complex with a 17 residue peptide (P1-XPC) from xeroderma pigmentosum group C protein) is missed because the binding site simply deviates too much from the site conformation seen in holo X-ray structures (RMSD of the loop is 2.594 $\mathring{A}$) [51].

2.3.9 Metal selectivity for Ca^{2+} over other divalent ions

Many proteins have well-documented binding sites for divalent metal ions other than Ca^{2+}. It becomes particularly relevant to ask whether the criteria we have developed to recognize Ca^{2+} sites from second- and third- shell carbon coordinates are able to discriminate sites known to bind other divalent metals of similar size; that is, how selective are these criteria for Ca^{2+} binding as opposed to other divalent metals. To address this question, we conducted additional research to determine whether the use of carbon shells in MUG^C could successfully discriminate between binding sites for Ca^{2+} as opposed to other divalent metals. Three additional testing datasets (Table E.15-E.17) comprised of X-ray structures of binding sites were evaluated for Mg^{2+} (52 sites), Zn^{2+} (51 sites) and Pb^{2+} (47 sites). Mg^{2+} and Zn^{2+} were selected for comparison due to their similar ionic radii (Mg^{2+} 0.72 $\mathring{A}$, Zn^{2+} 0.75 $\mathring{A}$) [83], and because they, along with Ca^{2+}, are the most abundant physiologically-relevant metals involved in biochemical reactions. Pb^{2+} was selected due to its similar ionic radius with Ca^{2+} (1.19 vs. 0.99 $\mathring{A}$) [83] and a volume of evidence indicating a close relationship between Pb^{2+} toxicity and Ca^{2+} metabolism [84–88]. For these analyses, a binding site was considered misclassified if a Ca^{2+}-binding site was predicted surrounding a non-Ca^{2+} ion (i.e., if it placed a Ca^{2+} ion within 3 $\mathring{A}$ of the documented other divalent metal [70, 78]), and if this predicted site is not known to be a true Ca^{2+}-binding site. Results of our analysis indicate that MUG^C does not misidentify Ca^{2+}-binding sites for 83%, 96% and 89% of Mg^{2+}, Zn^{2+}, and Pb^{2+} binding sites. Moreover, those binding sites classified as misidentifications may represent potential, unidentified Ca^{2+}-binding sites, or sites capable of binding multiple divalent ions, including Ca^{2+}. Several of the Mg^{2+} and Zn^{2+} sites evaluated exhibit atypical coordination geometries or utilized ligands that would be unusual for Mg^{2+} (e.g., carbonyl

oxygen atoms as seen in IKCZ.pdb) but not for Ca^{2+}, and for some of the Mg^{2+}-binding sites, very high concentrations of Mg^{2+} were added during crystallization (e.g., 250 mM [89] in 1OBW.pdb and 100 mM [90] in 1KCZ.pdb), so it is possible that the observed binding is representative of the crystallization conditions, but not necessarily of the proteins function in solution. If we remove these questionable misidentifications from our statistics (Identified as Other in Misclassified column in Tables E.18), our final results indicate that none of the remaining binding sites for proteins in the Mg^{2+} , Zn^{2+}, or Pb^{2+} datasets are identified by MUG^C as Ca^{2+}-binding sites, demonstrating excellent metal selectivity.

2.4 Discussion

2.4.1 Key factors for metal coordination

Our studies have revealed several key properties that are important for metal coordination. First, a second-shell of carbon clusters enclose the first shell atoms which directly coordinate Ca^{2+}. We hypothesize that the Ca^{2+} position within a Ca^{2+}-binding protein is determined as much by the positions of carbon atoms in the hydrophobic shells surrounding Ca^{2+} as by the immediate positions of the oxygen ligands comprising the actual binding site. A practical corollary to this hypothesis is that, in cases where the coordinates of ligand oxygens are poorly defined, the surrounding carbon shells can be relied upon to accurately predict the location of the Ca^{2+} center. Such cases are observed in crystallographically determined structures, where coordinates of side-chain oxygens may be poorly resolved because of their mobility. Limitations associated with positioning of oxygen atoms in NMR structures are also observed specifically because the naturally-abundant isotope of oxygen is spectroscopically silent in NMR. For backbone oxygen atoms, these reconstructed positions have higher precision, precisely because the geometry is fixed and there is no torsion angle involved. However, for sidechain oxygens, such as from the carboxylic groups of Asp and Glu, which are subject to torsional rotations, there are substantial uncertainties in the positions. The present work represents the first attempt to exploit the relative placement

of the carbon atoms and to pinpoint Ca^{2+} centers without reference to the locations of the directly ligated oxygen atoms, particularly involving those from side-chain. From the structural perspective of binding sites, the first (hydrophilic) oxygen shell in the binding sites permits the protein's exposure to water and ionic Ca^{2+}. This immediate binding scaffold is supported by a second (hydrophobic) shell of carbon atoms, which may restrict flexibility within the site and thereby ameliorate the decrease in binding-associated entropy [91]. In order to exercise the regulatory role of Ca^{2+} in the cell, binding sites in proteins must be able to bind and release Ca^{2+} within a physiological range of Ca^{2+} concentrations. This implies not only the existence of a "pre-organized" site, but also restricted structural flexibility within that site [60, 62, 91], as well as the stable positioning of carbon atoms oriented in such a way to facilitate formation of the hydrophilic oxygen shell which coordinates the Ca^{2+} directly. Our earlier studies demonstrated that the oxygen shell in the Ca^{2+}-binding site has an identifiable geometry (i.e., four or more oxygen atoms in the site, all separated from each other by an oxygen-oxygen distance $6\mathring{A}$) [69, 70]. Our current studies, described here, suggest that this structural regularity must be supported by the associated C-O bonds, implying an appropriately arranged geometry for the surrounding carbon shell - an arrangement which should also be identifiable. Second, we have shown that the vast majority of Ca^{2+}-binding sites have at least one negatively-charged residue within the tentative binding site. This observation justifies the utility of applying a charge filter, which improves selectivity in predicting various classes of Ca^{2+}-binding sites in the protein data bank [57]. In its X-ray structure analysis, the MUG^C algorithm missed only one site in the complex formed between proteolytically-generated lactoferrin fragment and proteinase K (1BJR.pdb) - an exception to the rule in that there is no negatively-charged binding residue in this binding site which has a coordination number of four. The Ca^{2+}-binding sites was composed of residues R12, S15, N257 and A273 [92]. It is likely that this binding site does not have strong Ca^{2+}-binding affinity. Third, our analysis of calmodulin has also shown that it is important to ensure that the predicted Ca^{2+} positions contain neither van der Waals clashes nor over-lapping side-chain centers of mass. The concept of side-chain center of mass

(SC-CoM) has been previously used in protein structural prediction [81]. In this work we present a novel application for the use of SC-CoM as an aid to predict Ca^{2+}-binding sites. In a sense, side-chain center of mass is used here as a surrogate for poorly-resolved ligand oxygen coordinates.

2.4.2 Implications for metal selectivity

From Table E.15-E.17, we can conclude that MUG^C does not mis-classify other metal binding sites as Ca^{2+}-binding sites in most cases. There are two key designs in MUG^C to distinguish Ca^{2+}-binding sites from non-Ca^{2+}-binding sites. First, carbon clusters utilized by MUG^C are restricted to those with associated oxygen atoms and were required to have at least four carbon atoms. Differences in coordination numbers between Ca^{2+} and the other metals, as well as variations in ion solvation result in different ions having different numbers of carbon atoms associated with binding. For example, Mg^{2+} tends to be more highly-solvated than Ca^{2+}, and the presence of more water molecules results in fewer carbon atoms within the microenvironment of the binding site. Additionally, both Zn^{2+} and Pb^{2+} typically utilize fewer binding ligands than Ca^{2+}, and utilize different ligand types [93]. As a hard Lewis acid, Ca^{2+} binds preferentially with oxygen atoms whereas both Zn^{2+} and Pb^{2+}, considered borderline Lewis acids, may bind with either hard or soft bases, utilizing both nitrogen and sulfur ligands, in addition to oxygen. Due to the smaller number of oxygen-based ligands for these metals, MUG^C selectively eliminates those sites as potential Ca^{2+}-binding sites. The second key design for identification of Ca^{2+}-binding sites relates to ionic radius, which is one factor by which proteins discriminate between divalent ions [94]. For example, Mg^{2+} is 28% smaller than Ca^{2+}, and this smaller VDW radius alters the geometry of the binding site which then may not accommodate the larger Ca^{2+} ion. After carefully calibrating the geometric parameters in MUG^C with respect to Ca^{2+} radius and the spatial relationships of binding ligands in Ca^{2+}-binding sites, MUG^C can distinguish Ca^{2+}-binding sites from those of other metals. Our results indicate that the algorithmic approach of MUG^C provides a useful tool for delineating metal binding sites. This differentiation

is achieved by carefully tuning the geometric and chemical parameters of MUG^C based on analysis of empirical data associated with Ca^{2+}-binding, and parameter optimization. Furthermore, we anticipate that this work will form the basis for incorporating additional prediction parameters derived from molecular dynamics simulations.

2.4.3 Comparison of MUG^C with other algorithms

The comparison among MUG, MUG^C, SitePredict and WebFeature (the web-based implementation of FEATURE) is based mainly on an NMR testing dataset. The MUG web-server does not accept NMR ensembles, so we submitted each member of the ensemble one by one; WebFeature and SitePredict do accept ensembles of structures. In these NMR structures there are no documented Ca^{2+} ions. The prediction results are summarized in Tables 2.3 and 2.4.

Table 2.3. Identification of Ca^{2+} positions on NMR structures by MUG^C, MUG and FEATURE.

	MUG^C	MUG	FEATURE	MUG^C+MUG
PTS[a]	20	19	7	21
DS[b]	21	21	21	21
CH[c]	330	284	21	610
TP[d]	403	451	21	859
SEN[e]	95%	90%	33%	100%
SEL[f]	81%	63%	100%	71%

[a]Predicted True Sites;[b]Documented Sites;[c]Correct Hits;[d]Total Predictions;[e]Sensitivity;[f]Selectivity.

To compare results with FEATURE, whose output is the predicted Ca^{2+} positions in the structures, we calculated the sensitivity and selectivity by mapping the Ca^{2+} position into the binding residues. If we observe at least one documented binding residue within $4\mathring{A}$ of the predicted position, then we count this position as correct prediction. Failure to meet these criteria results in a false positive. FEATURE predicted 7/21 binding sites with 33% sensitivity and 100% selectivity. Despite this algorithms advantage in selectivity, however,

Table 2.4. MUG^C and SitePredict predictions based on binding residues in NMR structures.

	MUG^C	SitePredict
PTS[a]	20	7
DS[b]	21	21
CH[c]	89	12
TP[d]	327	34
SEN[e]	95%	33%
SEL[f]	26%	35%

[a]Predicted True Sites;[b]Documented Sites;[c]Correct Hits;[d]Total Predictions;[e]Sensitivity;[f]Selectivity.

it fails to identify a significant proportion of sites in the dataset. This observation illustrates the persistent tradeoff between sensitivity and selectivity. Most of the published algorithms designed to predict Ca^{2+}-binding sites are based on optimal ligand geometry deduced from high-resolution X-ray static structures and thus rely heavily on the accuracy of the placement of ligand oxygen atoms. In contrast, MUG^C and SitePredict deliberately avoid use of specific side-chain and ligand coordinates in an effort to desensitize the method to vagaries in the location of ligands typical in low-resolution or homology-modeled structures. To compare our results with SitePredict, whose output is a list of residues involved in binding, we used such residues as a measurement of correctness of the prediction. According to its web-server (dated current as of Dec. 14, 2010) a default cutoff of 4 is used for predictions in binding residues (scores greater than 4 are considered as binding residues). We first compared MUG^C with SitePredict in NMR structures. SitePredict predicted 7/21 binding sites. Our data have shown that MUG^C exhibits significantly better performance in terms of sensitivity than SitePredict under conditions where they have almost comparable selectivity. The performance comparison with FEATURE and SitePredict, underscores the inadequacy of site-recognition algorithms informed by static structures to recognize sites in dynamic situations [64]. We also compared MUG^C with SitePredict in X-ray structures. Similar to testing on NMR structures, we considered that SitePredict is able to predict a binding site,

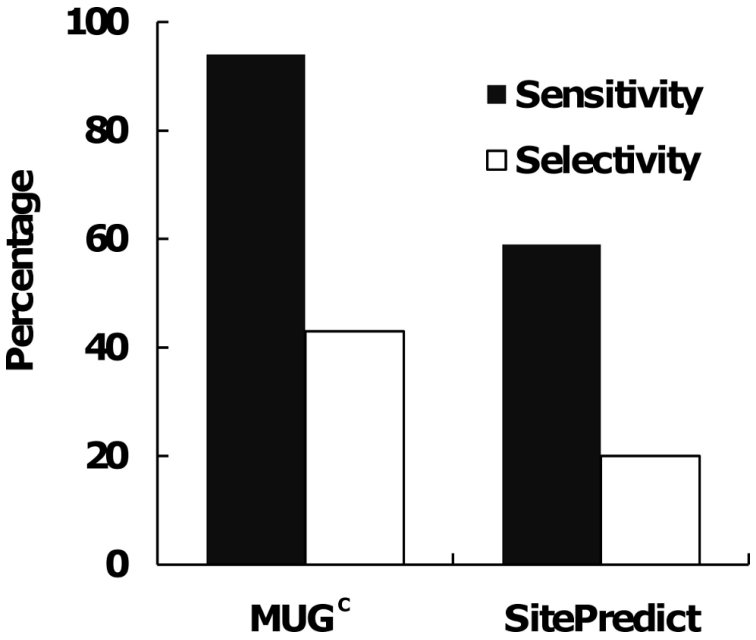


Figure 2.6. Comparison between MUG^C and SitePredict based on residues on testing X-ray dataset.

if it is able to identify at least one binding residue in an authentic site. For our comparative analysis, we applied a more stringent definition for the MUG^C's true-positive prediction sites and required that there be at least two binding residues predicted in authentic sites. If the predicted residue is not a binding residue, then it is a false positive residue. In the case that one binding residue appeared in two sites (thermolysin: 1HYT.pdb), we counted it twice for SitePredict's true positive residue, but once for MUG^C's true positive. The results show that, using these criteria, MUG^C has a sensitivity of 94%, while detecting the binding sites at a selectivity rate of 43% (Fig. 2.6). On the same dataset, SitePredict has a sensitivity of 59% and a selectivity of 20%. These results suggest that the performance of predicting binding residues could be improved by using second shell carbon atoms.

A comparison between MUG^C and our previously-reported MUG algorithm indicated little difference in results analyzing the static X-ray structure dataset, with MUG^C exhibiting 89% sensitivity and 76% selectivity compared to 91% sensitivity and 73% for MUG. However MUG^C results showed improvement compared to MUG with the testing NMR dataset: MUG^C has better sensitivity of 95%, and fewer false positive predictions,

although the selectivity of 81% leaves room for improvement. The results for MUG indicated sensitivity of 90% with a selectivity of 61%. MUG^C's superior performance with NMR datasets (Table 2.3), however, is somewhat muted by the fact that these datasets are small. The PDB contains many fewer NMR structures than X-ray structures, and very few Ca^{2+}-binding proteins in NMR structures inferred without Ca^{2+} constraints. Manually combining the two algorithms resulted in 100% sensitivity and 71% selectivity.

2.4.4 Challenges in algorithm evaluations

In this work, several statistical measurements were applied to assess the quality of our predicted results and estimate errors. First, we evaluated prediction error based on the difference in distance between the predicted and documented Ca^{2+} centers [67, 68]. Second, we evaluated a classification error based on ligand residues predicted to be involved in binding versus documented binding ligand residues (See Table 2.2). Third, we evaluated a negative control dataset comprised of proteins not currently known to bind Ca^{2+} or other metal ions. The challenge for evaluating the accuracy of predicting Ca^{2+}-binding sites stems from the fact that no consensual standard of quality has emerged from previous studies. Earlier works, such as those of Yamashita et al. [62] and Di Cera et al. [59], listed the prediction results but did not include statistical evaluations of the results. Glazer et al. applied sensitivity and selectivity to compare the performance of FEATURE with results reported by Nayal and Di Cera [64], however Schymkowitz et al. argued that the Fold-X algorithm was better at placing the Ca^{2+} position compared with FEATURE [68]. Babor et al. [95] later noted a large number of false positive predictions associated with Fold-X, and also suggested that its force field optimization step is very sensitive to small changes of position due to the electrostatic nature of the interactions. Quality evaluation is further complicated, as seen in this study, when the "structure" is in fact an ensemble of structures. A concise quality measurement over the ensemble is problematic. Yet another challenge comes from the definition of a false positive. We take as the most rigorous standard the position of Ca^{2+} explicitly observed by X-ray diffraction of holo proteins. But X-

ray models are not infallible; absence of Ca^{2+} at a physiologically functional binding site, especially a low affinity one, may simply mean that Ca^{2+} failed to crystallize at that site. Ironically, one might argue that the most exquisite use of prediction algorithms would be to reveal sites not visualized to contain crystallized Ca^{2+}, but subsequently proved to be bona fide sites. To predict sites of Ca^{2+} binding in proteins where the site may be indeterminate because of invisibility in X-ray and NMR structures, we have developed a graph-based, site-recognition algorithm which relies on carbon shell and side-chain center of mass information. This work shows that using information from carbon atoms, with formal ionic charges and center of mass as additional filters, can accurately identify Ca^{2+}-binding sites in X-ray holo structures with accurate performance. The binding sites in four holo NMR structures, computed with Ca^{2+} constraints, could be identified easily by this algorithm. Additionally, by testing 21 NMR binding sites that do not utilize Ca^{2+} constraints, we have demonstrated improved prediction results with NMR structures using carbon atoms comprising second and third concentric shells surrounding the binding sites. Finally, our results also demonstrate that the new algorithm is optimized for prediction of Ca^{2+}-binding sites, and able to discriminate Ca^{2+} from other divalent metal ions such as Mg^{2+}, Zn^{2+} and Pb^{2+}. The successful identification of Ca^{2+} positions by using the carbon shell deepens our understanding of the structure of Ca^{2+}-binding sites, thus further enhancing our capability to design Ca^{2+}-binding proteins [96–98]. This new algorithm may be applied advantageously to unrefined homology models, low-resolutions models and NMR structures.

To conclude, this chapter develops a graph-based method which has implications at a molecular level to identify Ca^{2+}-binding sites in proteins which may be related to diseases. Next chapter, we will develop an optimization method and algorithm which solve a disease prevention problem at a population level.

CHAPTER 3

APPLIED MATHEMATICS IN HEALTH CARE MANAGEMENT

3.1 Introduction

Chlamydia trachomatis (CT) and Neisseria gonorrhoeae (GC) are the two most commonly reported sexually transmitted diseases (STDs) in the United States. Most infections are asymptomatic and would not be detected without asymptomatic screening, especially for women. In 2008, 1,210,523 cases of chlamydia were reported to the Centers for Disease Control and Prevention (CDC) in the United States. This case count corresponds to a rate of 401.3 cases per 100,000 population, an increase of 9.2% compared with the rate in 2007 [99]. In 2008, 336,742 cases of gonorrhea were reported to CDC in the United States, corresponding to a rate of 111.6 per 100,000 population [99].

Many cases of CT and GC diseases are screened and treated by publicity funded clinics. In reality, these clinics may not have sufficient budgets to screen all eligible women with the most effective CT/GC tests and to offer these infected ones with more expensive, single-dose treatment that optimizes compliance. To effectively use limited resources, CT and GC control programs usually provide selective screening based on defined guidelines. For example, CDC recommends annual screening for CT and GC for sexually active adolescents and young women [100]. The U.S. Preventive Services Task Force (USPSTF) recommends screening all sexually active women, including those who are pregnant, for gonorrhea infection if they are at increased risk for infection [101].

For CT and GC control programs, however, identifying which subpopulations to screen for CT and/or GC infections is just one part of a complicated problem. The availability of several testing assays with various performances and costs presents a challenge for screening strategies: newer diagnostic tests that are less invasive and more sensitive offer increased opportunities for screening, but at a greater cost. In other words, the problem is whether

it is better to use a more sensitive and expensive test to screen fewer patients, or to use a relatively cheaper and less sensitive test to screen a greater number of patients. To further complicate the situation, test manufacturers market combination tests or bundled test at prices that are more lucrative than the price of a single-pathogen test. This situation encourages the testing for GC even when its prevalence in the population is extremely low.

3.1.1 Overview of Creating Resource Allocation Models for STDs

There are not a lot of resource allocation (optimization) models regarding the control of CT and GC infections. But many efforts have been made to develop models to investigate and evaluate HIV prevention and control programs [102–105]. To correlate with the practical relevance to CT infections, researchers initially developed a resource allocation model to determine the optimal strategy for curing CT infections among asymptomatic women at clinics [106]. Two years later, researchers proposed a mixed-integer program to model re-screening women who test positive for CT infections [107]. These two optimization models are able to offer simple guidelines for clinics on the selection of test and treatment for certain populations. However, these models are not able to manage two or more infections (e.g. CT and GC) at the same time at given clinics.

3.1.2 Overview of algorithms for solving STDs resource allocation models

Many health care researchers rely on an existing resource allocation model software to solve their proposed models because some software applications are easy to use [106, 108–110]. However, these applications sometimes may not provide the best outcomes due to the complexity of proposed models and the limitations of algorithms used in the software. For example, the resource allocation models used in the previous STD studies were nonlinear programming models and the optimal outcomes generated by the algorithm were never verified.

With respect to the nature of the resource allocation models that are typical nonlinear models, the algorithms for these models in general could be divided into two categories:

exact algorithms (e.g. dynamic programming or branch-and-bound [102, 111, 112]) and approximation algorithm (e.g. generalized reduced gradient method ([106, 108–110]). The exact algorithms, which are exhaustive and are guaranteed to find an optimal solution with a small number of variables, may run in exponential time [112]. In most cases, approximation algorithms, which may calculate near-optimal solutions, have to be used to speed up the computation time. This is the case for current commercial software applications, such as Excel Solver, MPL and Lingo [46]. When the resource allocation models become more complicated and various algorithms may lead to different outcomes, the knowledge of the limitations of various algorithms regarding computation accuracy and time is critical to the researchers. Unfortunately, comparing the computational accuracy and time of exact and approximation solutions to real-life STD models has not yet been examined or published. In other words, we do not know how well an approximation algorithm could perform on real-life health care data. This may be due to the fact that the optimization modelers tend to focus on sophistication of the mathematical formulation rather than the practical relevance [113] and algorithm accuracy and time [106, 109].

3.1.3 Our Research Objectives

We have three main objectives in this study. First, to improve CT and GC control and prevention in the United States, we created a resource allocation model which is a cubic binary programming model to consider two STDs. The model is designed to recommend an optimal strategy for identifying at-risk groups with a certain screening assay and treating those with positive results under a fixed budget. Second, to solve this resource allocation model, we developed a two-step branch-and-bound algorithm. Because our model had two diseases and a limited number of constraints and our two-step branch-and-bound algorithm is an exact algorithm rather than approximation algorithm. Our approach will always provide the optimal outcomes. Finally, we compared our computation results to those obtained by Excel Solver in terms of optimal outcome and computation time. The comparison can

help us to better understand the characteristics of the resource allocation models and the advantages and disadvantages of the algorithms used to solve the model [114].

3.2 Mathematical Model

3.2.1 Description of the model

The object of the model is to *maximize the number of cured infections (cured cases) among women under a fixed budget*. One patient with both infections cured is counted as two cured cases. We assume that patients would be tested and treated according to one of the four options below using a nonrapid test that would require women with positive tests to be recalled for treatment. Option 1). Single screening test and single treatment for CT only. A CT screening test is given to women and then CT treatment is given to those who had positive tests. Option 2). Single screening test and single treatment for GC only. Similar to Option 1. Option 3). Sequence screening tests that tested for CT and then GC if a positive CT result. A CT screening test is performed and then a GC test is performed on those women who had positive CT tests; and CT treatment is given to those who had positive CT tests and GC treatment is given for those who had positive GC tests. Option 4). Combo screening test for both CT and GC. Women are screened for both CT and GC at the same time using a combo test. CT or GC or both are treated if patients had positive tests for CT or GC or both, respectively.

There are other options in theoretical situations, such as screening patients for GC first, then testing those with positive GC results for CT, or screening patients for CT and presumptively treat patients for GC if they have positive CT results. However, the options are not listed in this model because they are not realistic for use in the United States due to the much lower prevalence of GC than CT and concerns about GC drug resistance [115, 116]. We do not count uninfected women who test positive at screening ("false positive") and get treated as a cured case. But, we do include the additional costs for treatment and the treatment visit of false positive results.

Realistically, we assume that the same test and treatment are offered to all women in each group. The reason to make this assumption is that strategy involving more than one test or treatment may be more complicated to implement in routine clinic practice, although it may cure more women at a fixed budget level. For example, clinics will face the challenges related to specimen handling, storage, transport, and billing for each test, and providers may need additional training to explain test performance issues to women in each group [106, 107]. This assumption complicates the mathematical formulations [106]. Two more simplifying assumptions are made in the model. First, we assume that all sexually active women who visit the clinic and are infected with CT or GC or both have no symptoms of infection. Second, no patients receive more than one test or treatment for the same infection at any one visit.

3.2.2 Data used in the model

As seen in Table 3.1, our model divides a theoretical cohort of 10,000 sexually active female patients into three age groups (younger than 20 years, 20-24 years, and 25-34 years) and four race/ethnicity groups (White, Black, Hispanic and Other). The age groups analyzed are similar to those classified in the CDC for screening and treating for CT infections [106]. The prevalence rate of each group is referenced from [117, 118]. Our model includes two CT tests (Pace 2 CT[1] and BD ProbeTec CT[2]), one GC test (culture), three combo tests (Pace 2C Combo[1], BD ProbeTec CT/GC, and APTIMA CT/GC[1]), two CT treatments (doxycycline and azithromycin), and two GC treatments (ceftriaxone and cefixime). These data were obtained from various published sources [116, 119–126]. The test sensitivity and specificity are shown in Table 3.2. All costs and other parameters are shown in Table 3.3.

[1]Gen-Probe, Inc., San Diego, CA
[2]Becton, Dickinson and Company, Franklin Lakes, NJ

Table 3.1. Population distribution characteristics of theoretical cohort of 10,000 women

	Race/ethnicity			
Age	White	Black	Hispanics	Others
Number of patients				
<20 years	2010	480	360	150
20-24 years	2680	640	480	200
25-34 years	2010	480	360	150
CT prevalence[1]				
<20 years	3.8%	15.6%	9.2%	10.7%
20-24 years	2.5%	14.4%	6.3%	7.5%
25-34 years	1.2%	11.8%	2.5%	3.3%
GC prevalence[2]				
<20 years	0.1%	1.9%	0.1%	0.2%
20-24 years	0.2%	2.2%	0.1%	0.2%
25-34 years	0.2%	1.8%	0.1%	0.2%

[1,2] All prevalence rates are referenced from [117] and [118].

3.2.3 The model

Several mathematical notations in the model and then the formulas are introduced here. More details are in the appendix D.

1. **Population notations.** The patient population is divided into 12 groups. For each group i, let x_i be a binary variable such that $x_i = 1$ if all the patients in the group i is identified for screening and $x_i = 0$ otherwise. Let $P_t(i)$ and $P_g(i)$ be the prevalence of the group i with CT and GC, respectively. Pop_i is the number of patients in ith group.

2. **Screening notations.** There are 6 available screening assays. Let y_j be a binary variable such that $y_j = 1$ if the screening assay j is used and $y_j = 0$ otherwise. For each assay j ($1 \leq j \leq 6$), let $Sn_t(j)$ and $Sp_t(j)$ be the sensitivity and specificity for CT; let $Sn_g(j)$ and $Sp_g(j)$ be the sensitivity and specificity for GC; let $Bc(j)$ and $Ac(j)$ be the unit-based costs and additional costs of the jth test. Let Vc be the costs per patient for the visit in which screening was done.

Table 3.2. Sensitivity and specificity of test assays and effectiveness of treatment regimens for chlamydia and gonorrhea

	CT	GC
Test Sensitivity		
Pace 2 CT	0.716 [119]	N/A
BD ProbeTec CT	0.928 [120]	N/A
Culture	N/A	0.848 [121]
Pace 2C Combo	0.716 [119]	0.781 [122]
BD ProbeTec CT/GC	0.928 [120]	0.966 [120]
APTIMA CT/GC	0.942 [123]	0.992 [123]
Test Specificity		
Pace 2 CT	0.995 [119]	N/A
BD ProbeTec CT	0.981 [120]	N/A
Culture	N/A	1.000 [121]
Pace 2C Combo	0.995 [119]	0.991 [122]
BD ProbeTec CT/GC	0.981 [120]	0.994 [120]
APTIMA CT/GC	0.995 [124]	0.995 [127]
Treatment Effectiveness		
Doxycycline	0.92 [125, 128]	N/A
Azithromycin	0.95 [126, 128]	N/A
Ceftriaxone	N/A	0.988 [116]
Cefixime	N/A	0.975 [116]

3. **Treatment notations.** There are 4 available treatment regimens. Let $z_{(k,l)}$ be a binary variable such that $z_{(k,l)} = 1$ where $k > 0$ and $l > 0$ if the regimen k is used for treating CT together with regimen l used for treating GC. $z_{(0,l)} = 1$ is that only the GC treatment regimen l is selected and CT will not be treated, and $z_{(k,0)} = 1$ is that only CT is treated. For each CT treatment regimen k ($k = 1, 2$), let $E_t(k)$ be the effectiveness of the kth regimen. Similarly, let $E_g(l)$ be the effectiveness of the lth ($l = 1, 2$) regimen for GC treatment. We also denote costs of drugs $Dc_t(k)$ for CT and $Dc_g(l)$ for GC, respectively. Let Tc be the costs per patient for the visit in which a treatment was done.

4. **Number of cured cases and unit costs.** Let Cur_{ijkl} and $Cost_{ijkl}$ be the rate of cured infection cases and costs correspondingly over the population of the ith group using the jth screening test and being treated with kth and/or lth treatment regi-

Table 3.3. Costs[1] related to CT and GC test and treatment and other parameters

	Baseline	
Test Cost		
Pace 2 CT	18.50	[129–131]
BD ProbeTec CT	29.79	[129–131]
GC Culture	9.26	[129–131]
Pace 2C Combo[2]	35.16	[129–131]
BD ProbeTec CT/GC	59.00	[129–131]
APTIMA CT/GC	61.67	[129–131]
Treatment Cost		
Doxycycline	8.12	[132]
Azithromycin	28.78	[132]
Ceftriaxone[3]	25.74	[129, 132]
Cefixime	10.06	[132]
Test Visit Cost	14.00	[130]
Treatment Visit Cost	28.43	[133]
PID Cost	2772	[134]
Probability of PID	0.20	[135]
Prob. of return for treatment	0.86	[136]

[1]All costs in 2006 US dollar values (adjusted with medical CPI where needed)[137].
[2] For the Pace 2C, a positive test (indicating either CT or GC, but not which organism) is followed by two separate supplemental tests.
[3] For ceftriaxone, the baseline price includes the drug plus the fee for intramuscular injection [129].

men(s). So the corresponding number of cases cured and costs are $Pop(i) \cdot Cur_{ijkl}$ and $Pop(i) \cdot Cost_{ijkl}$ for group i. Cur_{ijkl} under Option 1 ("Single screening test and single treatment for CT") is given as following:

$$Cur_{ijkl} = P_t(i) \cdot Sn_t(j) \cdot E_t(k) \cdot P_r \tag{3.1}$$

where P_r is the "probability of return for treatment" in Table 3.3.

$Cost_{ijkl}$ under Option 1 is given as following:

$$\begin{aligned} Cost_{ijkl} &= Bc(j) + Vc + [P_t(i) \cdot Sn_t(j) \\ &\quad + (1 - P_t(i)) \cdot (1 - Sp_t(j))] \cdot (Dc_t(k) + Tc) \cdot P_r \end{aligned} \tag{3.2}$$

For treatment, $P_t(i) \cdot Sn_t(j) + (1 - P_t(i)) \cdot (1 - Sp_t(j))$ gives the probability of a person having a positive test result, where $P_t(i) \cdot Sn_t(j)$ is the probability of a person having CT infection and testing positively, and $(1 - P_t(i)) \cdot (1 - Sp_t(j))$ is the probability of a person not having CT infection but having a (false) positive test result. The detailed calculations on Cur_{ijkl} and $Cost_{ijkl}$ for other options can be found in the appendix D.

5. **Objective function and constraints.** The objective function is to maximize the cured cases with available screening assays and treatment regimens for given patient groups.

$$\text{Max} \sum_{i,j,k,l} Pop_i \cdot Cur_{ijkl} \cdot x_i y_j z_{(k,l)} := \sum_{i=1}^{12} \sum_{j=1}^{6} \sum_{k=0}^{2} \sum_{l=0}^{2} Pop_i \cdot Cur_{ijkl} \cdot x_i y_j z_{(k,l)} \tag{3.3}$$

Subject to funding availability

$$\sum_{i,j,k,l} Pop_i \cdot Cost_{ijkl} \cdot x_i y_j z_{(k,l)} \leq b \tag{3.4}$$

which means the screening and treatment costs for identified groups should be smaller than or equal to the annual available funding b to a clinic. Furthermore, according to the realistic assumption that the same screening assay and the same treatment must be applied for all patients served at the clinic, only one assay among the six screening assays will be used:

$$\sum_{j=1}^{6} y_j = 1 \quad \text{and} \tag{3.5}$$

Only one treatment regimen for each infection among the two CT treatment regimens and two GC treatment regimens will be used:

$$\sum_{k,l}^{2,2} z_{(k,l)} = 1, \tag{3.6}$$

where k, l and $z_{(k,l)}$ are defined in treatment notations.

In the previous published mixed-integer programming model [107], we used x_{ijk} as a binary variable to select the choice of the best combination of i, j, k for single CT infection. In order to model the realistic assumption of applying same screening assay and the same treatment for all patients, we had to introduce two auxiliary binary variables[3] to control the combinatorial relationships among constraints [107]. These variables make the formulation tedious and limit our model on only two screening regimens and two treatment regimens. In this work, we keep the number of binary variables and simplify the formulation by defining x_i, y_j and $z_{(k,l)}$. Now, the new model is able to consider CT and GC infections together with more than two screening regimens and two treatment regimens.

3.3 Two-step Branch-and-bound Algorithm

The model 3.3-3.6 is a cubic binary programming problem. Except for exhaustive methods, there is no efficient algorithm to solve this problem [46]. The exhaustive (exact) algorithm runs in exponential time, which can only produce solutions for the problems with a small number of variables. To solve this model in general, approximate algorithms have to be used by commercial software applications. We do know that approximate algorithms generally are unable to distinguish between a local maximum and a global maximum [46], but we don't know much these local maximum will deviate from a global maximum in this practical case. Therefore, knowing the global optimal solutions rather than approximations

[3]Auxiliary binary variables representing the yes-or-no decisions are introduced to reduce the problem to a mixed-integer programming (MIP) [46]. In the MIP model these variables can be viewed as contingent decisions, i.e., decisions that depend upon previous decisions.

to the global optimal solutions (global maximum) is very important here. In this study, we define that *a global optimal solution to our model under a fixed budget is the strategy which guarantees the maximal value of cured cases*. Base on this real-life model, we established a two-step algorithm to calculate global optimal solutions. In the first step, by using an exhaustive algorithm, we reduced the original model to several classic 0-1 knapsack problems which is an NP-hard problem in combinatorial optimization [138, 139]. In the second step, we used the branch-and-bound method to solve each knapsack problem and select the best strategy among each knapsack problem. The following details the algorithm.

Because (3.5) and (3.6) show that there is only one possible j such that $y_j = 1$ and there is only one possible (k, l) such that $z_{(k,l)} = 1$, we initially identify how many combinatorial strategies for screening assays and treatment regimens exist in the model by using an exhaustive algorithm. In the real-life case, there are 26 possible screening and treatment strategies. Option 1). Single screening test and single treatment for CT only. There are a total of 4 $(= 2 \cdot 2)$ combinations using a single screening test (Pace 2 CT or BD ProbeTec CT) and a single treatment for CT (doxycycline or azithromycin). Option 2). Single screening test and single treatment for GC only. Similar to Option 1, there are 2 combinations using the culture test and a single treatment for GC (ceftriaxone or cefixime). Option 3). Sequence screening tests that tested for CT and then GC if a positive CT result. There are a total of 8 $(= 2 \cdot 1 \cdot 2 \cdot 2)$ combinations: two screening assays for CT (Pace 2 CT or BD ProbeTec CT), one GC screening assay (culture), two CT treatment regimens (doxycycline or azithromycin) and two GC treatment regimens (ceftriaxone or cefixime). Option 4). Combo screening test for both CT and GC. There are a total of 12 $(= 3 \cdot 2 \cdot 2)$ combinations: three combo screening assays (Pace 2C Combo, BD ProbeTec CT/GC, or APTIMA CT/GC), two treatments for CT (doxycycline or azithromycin) and two treatments for GC (ceftriaxone or cefixime).

After all possible strategies are exhausted, the original model was decomposited into 26 classical 0-1 knapsack problems. In general, a 0-1 knapsack problem is defined in the following way. Given a set of items, each with a weight and a value, determine the number

of each item to include in a collection so that the total weight is less than a given limit and the total value is as large as possible [46]. As for our case, the 12 population groups are "number of items"; the costs $Pop_i \cdot Cost_{ijkl}$ are the "weights"; the cured cases $Pop_i \cdot Cur_{ijkl}$ are "values", and the budget is the "given limit". We then applied the classical Horowitz-Sahni's branch-and-bound method [112] to find which group should be identified to go through the screening and treatment strategy. This method is further discussed in appendix D. After applied the method, we were able to record the corresponding numbers of the cured cases and the costs within each knapsack problem. These results are the optimal results for each of the 26 knapsack problems. Finally, the global optimal result to the whole model is reported as the one with maximum cured cases among the 26 optimal results.

3.4 Results and Discussion

3.4.1 Algorithms and application

The two-step branch-and-bound algorithm is our primary algorithm. It is an exact algorithm because both of first and second steps are exhaustive methods. Because in reality there are limited number of combined screening and treatment regimens, the first step could be enumerated quickly. (In this case, there are only 26 combinations.) In the second step, the Horowitz-Sahni's branch-and-bound method to each knapsack problem is also exhaustive [112]. This method consists of a systematic enumeration of all at-risk population groups, where large subsets of candidate groups are discarded *en masse*, by using upper and lower estimated bounds of the cured cases being optimized. We selected this method because it is one of the most effective, structured, and easiest to implement [112]. Within each knapsack problem, the global optimal result could be obtained because the number of at-risk population groups is small according to a realistic division of patients [106]. The selection of the maximal number of cured cases among each optimal solution to the 26 knapsack problems guarantees the global optimal solution to the original model. In summary, the

global optimal solution is obtained because the algorithm is exhaustive and the complexity of this real-life model is very reasonable.

As a general approach to the knapsack problem, dynamic programming is an alternative exact method [111] could be used to solve a global optimization problem. It has been proposed to solve a HIV resource allocation model [102] and epidemic control model [140]. We think it is interesting to see how the dynamic programming performs in the real-life case of controlling CT and GC infections. Therefore, we implemented this method by replacing the branch-and-bound method as the second step for the two-step algorithm. The dynamic programming method we used is based on the classical Bellman recursion [112, 141] due to its ease for implementation. The time and the space complexity of this method is O(nc), where n is the number of population groups and c is the budget in our case. As the budget increases, the time and space consumption increase, limiting the performance of the algorithm. The pseudocodes from [112] (p.38-39) were used in this study. The two-step algorithm is implemented in Java.

Microsoft Excel has been used in STD research as a convenient tool [106, 110]. For comparison purposes, Excel Solver was applied to solve our original model as oppose to the two-step algorithm. Excel Solver uses the Generalized Reduced Gradient (GRG2) algorithm for optimizing nonlinear problems and it is an approximate algorithm [108]. We tested and compared the performance of the two-step algorithm and Excel Solver under the different budget levels, in terms of running time, the number of cured cases and the screening and treatment strategy identified. From time to time, the solutions calculated with Excel Solver do not always give the maximal value of the objective function. The following computational results were run on an Intel Celeron M 1.6GHz processor and a RAM of 512MB.

3.4.2 Numerical results

Optimal strategy results under selected budget levels are presented in Table 3.4. The two-step branch-and-bound algorithm has a faster running time and provides the global optimal results rather than the approximate solutions generated by the Solver's GRG2 al-

gorithm, although Excel Solver can solve our original model directly. For example, at the budget level of $17,350, Solver's algorithm suggested to screen the group with a prevalence rate of 11.8% while the global optimal strategy screened the group with the highest prevalence rate of 15.6%. Using the optimal strategy calculated by two-step branch-and-bound algorithm, 10 more patients could be cured compared to Solver's algorithm. At the budget level of $30,000, Solver's GRG2 algorithm recommended to treat CT alone with azithromycin. However, a better optimal strategy suggested that two more patients could be cured if we treat CT and GC together for the same groups. At the budget level of $100,000 and $200,000, Solver's GRG2 algorithm screened groups different from ours, and it also selected different treatment regimens. As the result, Solver's GRG2 algorithm cured two and five fewer patients respectively than our algorithm suggested. These results indicate that the accuracy of Solver's solutions is improved by the proposed algorithm.

When the budget is low, the dynamic programming could identify the global optimal solution. However, when the budget is high, the dynamic program has a longer running time and it might run out of computer memory before it reaches its results.

Health care researchers [106, 107, 115, 142–145] rely on software applications to solve their proposed models and implement their methodologies because some software applications are easy to use. However, these applications sometimes may not be accurate. Therefore, researchers need to understand the limitations of the software applications. Software applications' performance and the accuracy of results may vary due to the complexity of proposed models and the differences among available algorithms. As discussed here, we know that the Excel Solver's GRG2 algorithms could generate an approximately optimal solution to our model. However, we only know how much these results will deviate from the global optimal solution in the real-life case by using our algorithm. Our proposed two-step branch-and-bound algorithm finds the global optimal strategy for the underlying model. Compared with our global optimal solution, we are now able to tell how good those approximate solutions are.

3.4.3 The model

The proposed cubic binary model could be widely used to manage budgets beyond the situation shown here with a fixed number of groups, screening assays and treatment regimens for CT and GC. It can be easily modified to solve the problem with different numbers of population groups, screening assays, and treatment regimens. It can also be modified to solve problems which have the characteristics of two or more major infections or diseases. For example, the model can be used to screen and treat patients for infectious diseases and chronic diseases.

There are still some improvements that could be done to the proposed model. In our current model, the side effect of the tests and treatments are not considered beyond the costs associated with treating false positive. For example, treating patients with false positive test results impose not only additional costs, but also medical side effects (such as gastrointestinal distress following azithromycin treatment [146]) which are difficult to value in monetary terms. Also, a false positive diagnosis can impose stress on a given patient and on her relationship with her sexual partner [147, 148]. For such concern, we could add some punishment components in the objective function. In other words, if there are too many mistreated cases, then a heavy punishment could be considered while selecting the optimal strategy.

We have restricted our work to one clinic. Government funding agencies, such as CDC, may need to optimize their funds for many clinics. Different clinics may use different screening and treatment regimens with more complicated constrains, which makes the optimization problem much more complicated. A future goal is to tackle this generalization problem.

3.5 Conclusion

We have designed a new mathematical model of screening and treatment for the two most common STDs in the US. It would be simple for program managers to use to optimize

their prevention and control programs. Benefits of this model include: it not only considers CT and GC together, but also is able to use different prevalence and costs parameters. Furthermore, the model can be expanded to provide optimal strategy for different number of population groups, screening assays, and treatment regimens, and for two or more infections or diseases. Meanwhile, tailor-made for the model, the new two-step branch-and-bound algorithm showed its improvements towards calculating a global optimal solution on the real-life data as compared to Excel Solver.

Table 3.4. Optimal strategy results for screening and treating 10,000 female patients for chlamydia and gonorrhea under the selected budget levels by three different algorithms

	Two-Step Algorithm		Solver's
	BnB[1]	DP[2]	GRG2[3]
Budget= $\$17,350$			
Optimal Strategy			
Cured cases	42	42	32
Costs($)	17,348.92	17,348.92	16,941.28
Costs($) per case cured	413.07	413.07	529.42
Screening	Pace 2 CT	same	same
Treatment	DXC[4]	same	same
Running time (second)	<1	1	3
Budget= $\$30,000$			
Optimal Strategy			
Cured cases	65	65	63
Costs($)	29,151.61	29,151.61	29,654.71
Costs($) per case cured	448.49	448.49	471.71
Screening	Pace 2 CT	same	same
Treatment	DXC+CFX[5]	same	ATM[6]
Running time (second)	<1	9	4
Budget= $\$100,000$			
Optimal Strategy			
Cured cases	198	198	196
Costs($)	97,389.52	97,389.52	98,864.24
Costs($) per case cured	491.87	491.87	504.41
Screening	BD ProbeTec CT	same	same
Treatment	DXC+CFIX[7]	same	ATM+ CFX
Running time (second)	<1	44	1
Budget= $\$200,000$			
Optimal Strategy			
Cured cases	267	267	262
Costs($)	197,104.2	197,104.2	166,227.57
Costs($) per case cured	738.22	738.22	634.46
Screening	BD ProbeTec CT	same	same
Treatment	DXC+CFX	same	ATM+CFX
Running time (second)	<1	13	4
Budget= $\$500,000$			
Optimal Strategy			
Cured cases	393	out of memory	393
Costs($)	476,323.83	n/a	476,323.83
Costs($) per case cured	1,212.02	n/a	1,212.02
Screening	BD ProbeTec CT	n/a	same
Treatment	ATM+CFX	n/a	same
Running time (second)	<1	n/a	1

[1]branch-and-bound method; [2]dynamic programming; [3]generalized reduced gradient method; [4]doxycycline [5]ceftriaxone; [6]azithromycin; [7]cefixime.

MAJOR FINDINGS AND SIGNIFICANCE

This dissertation has successfully developed three mathematical methods for solving complex problems in dynamical networks, proteomics, and disease prevention. The Augment Graph Stability method calculates the upper bounds for global synchronization in directed networks. This method extends the Connection Graph method by transforming a directed network into a symmetrized-and-undirected network and then augmenting the transformed network. The synchronization criterion for the augmented symmetrized-and-undirected network also guarantees global stability of synchronization in the original directed network. With this method, bottlenecks for synchronizing each node in networks can be identified. Results show that this method outperforms the previous Connection Graph method in sparse graphs. The new approach can be applied to study the synchronization in any network such as engineering and biological networks. In particular, the method can potentially be used to analyze the emergence of abnormal synchronized rhythms, associated with epilepsy and Parkinson's disease, caused by changes in network connectivity at a multi-cellular level.

With respect to the disease at a molecular level, the success of the graph theory algorithm to predict Ca^{2+}-binding site in proteins in Chapter 2 validates the hypothesis that the second, hydrophobic shell of carbon atoms enclosing a Ca^{2+}-binding site could sufficiently determine the site's location in either X-ray or NMR structures. This new algorithm allows us to predict Ca^{2+}-binding sites in proteins where the Ca^{2+} ion may not be directly observable (e.g., low resolution structures, weak affinity binding sites, and NMR structures). Results regarding running the algorithm on datasets containing Mg^{2+}, Zn^{2+}, and Pb^{2+} binding sites, demonstrate not only that the Ca^{2+}-binding sites in NMR and X-ray structures can be identified based on geometric arrangement of the second-shell carbon cluster, but also that this approach with Ca^{2+}-optimized selection parameters, can also selectively differentiate between Ca^{2+} and other relevant divalent cations. The application of this algorithm will enable us to identify previously-unknown Ca^{2+}-binding sites, deepen our understand-

ing of structural characteristics of Ca^{2+}-binding sites, and improve our ability to design Ca^{2+}-binding proteins with diversified functions.

The proposed combinatorial optimization model solves a real clinical issue of limited budget at a population level. This model can be widely used to manage budgets beyond the situation (discussed in Chapter 3) with a fixed number of groups, screening assays and treatment regimens for CT and GC. It can be easily modified to solve the problem with different numbers of population groups, screening assays, and treatment regimens. It can also be modified to solve problems which have the characteristics of two or more major infections or diseases. For example, the model can be used to screen and treat patients for infectious diseases and chronic diseases. Running on real-life data, a proposed algorithm to solve the model calculates the optimal solution within a very short time. The new algorithm improves the accuracy of an approximate solution obtained by Excel Solver. This study has shown that a resource allocation model and algorithm might have a significant impact on real clinical issues. Finally, the innovations of the three mathematical methods will hopefully inspire further studies for mathematical methods regarding problems in complex networks, biology, chemistry, health and diseases.

REFERENCES

[1] S. Farmer, "Neural rhythms in parkinsons disease." *Brain*, vol. 125, no. 6, pp. 1175–76, 2002.

[2] V. Belykh, I. Belykh, and M. Hasler, "Connection graph stability method for synchronized coupled chaotic systems," *Physica D*, vol. 195, no. 1-2, pp. 159–187, 2004.

[3] P. Vito, E. Lacana, and L. D'Adamio, "Interfering with apoptosis: Ca(2+)-binding protein alg-2 and alzheimer's disease gene alg-3," *Science*, vol. 271, no. 5248, pp. 521–5, 1996.

[4] M. J. Berridge, M. D. Bootman, and H. L. Roderick, "Calcium signalling: dynamics, homeostasis and remodelling," *Nat Rev Mol Cell Biol*, vol. 4, no. 7, pp. 517–29, 2003.

[5] B. Calabretta, L. Kaczmarek, W. Mars, D. Ochoa, C. W. Gibson, R. R. Hirschhorn, and R. Baserga, "Cell-cycle-specific genes differentially expressed in human leukemias," *Proc Natl Acad Sci U S A*, vol. 82, no. 13, pp. 4463–7, 1985.

[6] P. Hocker and P. Reizenstein, "Calcium and potassium disturbances in acute leukemia," *Blut*, vol. 29, no. 6, pp. 398–406, 1974.

[7] L. P. Slomnicki, B. Nawrot, and W. Lesniak, "S100a6 binds p53 and affects its activity," *Int J Biochem Cell Biol*, vol. 41, no. 4, pp. 784–90, 2009.

[8] R. Mamillapalli, J. VanHouten, W. Zawalich, and J. Wysolmerski, "Switching of g-protein usage by the calcium-sensing receptor reverses its effect on parathyroid hormone-related protein secretion in normal versus malignant breast cells," *J Biol Chem*, vol. 283, no. 36, pp. 24 435–47, 2008.

[9] C. Wang, T. Chen, N. Zhang, M. Yang, B. Li, X. Lu, X. Cao, and C. Ling, "Melittin, a major component of bee venom, sensitizes human hepatocellular carcinoma cells to

tumor necrosis factor-related apoptosis-inducing ligand (trail)-induced apoptosis by activating camkii-tak1-jnk/p38 and inhibiting ikappabalpha kinase-nfkappab," *J Biol Chem*, vol. 284, no. 6, pp. 3804–13, 2009.

[10] H. Yang, S. Murthy, F. H. Sarkar, S. Sheng, G. P. Reddy, and Q. P. Dou, "Calpain-mediated androgen receptor breakdown in apoptotic prostate cancer cells," *J Cell Physiol*, vol. 217, no. 3, pp. 569–76, 2008.

[11] S. H. Strogatz, "Exploring complex networks." *Nature*, vol. 410, p. 268276, 2001.

[12] L. Fabiny, P. Colet, R. Roy, and D. Lenstra, "Coherence and phase dynamics of spatially coupled solid-state lasers." *Phys. Rev. A*, vol. 47, pp. 4287–4296, 1993.

[13] D. Mills, "Precision synchronization of comuter network clocks." *ACM Computer Communication Review*, vol. 24, pp. 28–42, 1993.

[14] C. M. Gray and W. Singer, "Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex." *Proc. Natl. Acad. Sci.*, vol. 86, pp. 1698–1702, 1989.

[15] R. Stoop, K. Schindler, and L. Bunimovich, "Neocortical networks of pyramidal neurons." *Nonlinearity*, vol. 13, pp. 1515–1529, 2000.

[16] K. Persson and Rekling, "Population calcium imaging of spontaneous respiratory and novel motor activity in the facial nucleus and ventral brainstem in newborn mice." *Journal of Physiology*, vol. 15, no. 589, pp. 2543–58, 2011.

[17] I. Belykh, E. Lange, and M. Hasler, *Physical Review Letters*, vol. 94, p. 188101, 2005.

[18] T. Koeniga, L. Prichepb, T. Dierksa, D. Hubla, L. Wahlundd, E. John, and V. Jelicd, "Decreased eeg synchronization in alzheimer disease and mild cognitive impairment." *Neurobiology of Aging*, vol. 26, no. 2, pp. 165–171, 2005.

[19] K. Zhao and I. Belykh, "Augmented graph method for synchronization in directed networks," *submitted*.

[20] H. Fujisaka and T. Yamada, "Stability theory of synchronized motion in coupled-oscillator systems." *Prog. Theor. Phys.*, vol. 69, no. 2, p. 32, 1983.

[21] V. Afraimovich, S. Chow, and J. Hale, "Synchronization in lattice of coupled oscillators." *Physica D*, vol. 103, pp. 442–451, 1997.

[22] L. Pecora and T. Carroll, "Synchronization in chaotic systems." *Phys. Rev. Lett*, vol. 64, pp. 821–824, 1990.

[23] Y. Kuramoto, "in international symposium on mathematical problems in theoretical physics," *Lecture Notes in Physics*, vol. 39, p. 420, 1975.

[24] N. Kopell and G. Ermentrout, "Coupled oscillators and the design of central pattern generators," *Math. Biosci*, vol. 90, p. 87, 1988.

[25] S. Watanabe, S.and Strogatz, "Integrability of a globally coupled oscillator array," *Phys. Rev. Lett*, vol. 70, no. 16, p. 2391, 1993.

[26] S. Strogatz and R. Mirollo, "Phase-locking and critical phenomena in lattices of coupled nonlinear oscillators with random intrinsic frequencies," *Physica D*, vol. 31, no. 2, pp. 143–168, 1988.

[27] D. Somers and N. Kopell, "Waves and synchrony in networks of oscillators of relaxation and non-relaxation type," *Physica D*, vol. 89, no. 1-2, p. 169, 1995.

[28] V. Belykh, N. Verichev, L. Kocarev, and L. Chua, *Chua's Circuit: A Paradigm for Chaos.* Singapore: World Scientific., 1993.

[29] J. Heagy, T. Carroll, and L. Pecora, "Synchronous chaos in coupled oscillator systems," *Phys. Rev. E*, vol. 50, no. 3, p. 1874, 1994.

[30] J. Heagy, L. Pecora, and T. Carroll, "Short wavelength bifurcations and size instabilities in coupled oscillator systems," *Phys. Rev. Lett.*, vol. 74, p. 4185, 1994.

[31] C. Wu and L. Chua, "Synchronization in an array of linearly coupled dynamical systems," *IEEE Trans. Circuits Syst., I: Fundam. Theory Appl.*, vol. 43, p. 161, 1996.

[32] V. Afraimovich, S. Chow, and J. Hale, "Synchronization in lattices of coupled oscillators," *Physica D*, vol. 103, no. 1-4, p. 442, 1997.

[33] J. Hale, "Diffusive coupling, dissipation and synchronization," *J. Dyn. Diff. Eq.*, vol. 9, p. 1, 1997.

[34] L. Pecora and T. Carroll, "Master stability functions for synchronized coupled systems," *Phys. Rev. Lett.*, vol. 80, no. 10, p. 2109, 1998.

[35] K. Josić, "Synchronization of chaotic systems and invariant manifolds," *Nonlinearity*, vol. 13, no. 4, p. 1321, 2000.

[36] A. Pogromsky and H. Nijmeijer, "An observer point of view on synchronization of discrete-time systems," *IEEE Trans. Circuits Syst., I: Fundam. Theory Appl.*, vol. 48, p. 152, 2001.

[37] C. Wu, *IEEE Trans. Circuits Syst., I: Fundam. Theory Appl.*, vol. 50, p. 294, 2003.

[38] J. Jost and M. Joy, "Spectral properties and synchronization in coupled map lattices," *Phys. Rev E*, vol. 65, no. 1, p. 16201, 2001.

[39] G. Rangarajan and M. Ding, "Stability of synchronized chaos in coupled dynamical systems," *Phys. Lett. A*, vol. 296, no. 4-5, p. 204, 2002.

[40] V. Belykh, I. Belykh, M. Hasler, and K. Nevidin, "Cluster synchronization in three-dimensional lattices of diffusively coupled oscillators," *Int. J. Bifurcat. Chaos*, vol. 13, p. 756, 2003.

[41] I. Belykh, V. Belykh, and M. Hasler, "Generalized connection graph method for synchronization in asymmetrical networks," *Physica D*, vol. 224, pp. 42–51, 2006.

[42] V. Belykh, I. Belykh, and M. Hasler, "Hierarchy and stability of partially synchronous oscillations of diffusively coupled dynamical systems." *Phys. Rev. E*, vol. 62, p. 6332, 2000.

[43] I. Belykh, V. Belykh, K. Nevidin, and M. Hasler, "Persistent clusters in lattices of coupled nonidentical chaotic systems." *Chaos*, vol. 13, pp. 165–178, 2003.

[44] I. Belykh, M. Hasler, M.and Lauret, and H. Nijmeijer, "Synchronization and graph topology," *Int. J. Bifurcat. Chaos*, vol. 15, no. 11, pp. 3423–43, 2005.

[45] T. H. Cormen, C. Stein, R. L. Rivest, and C. E. Leiserson, *Introduction to Algorithms*. MIT: McGraw-Hill Higher Education., 2001.

[46] F. S. Hillier and G. J. Lieberman, *Introduction to operations research*, 7th ed. New York, US. 654-720: MaGraw-Hill, 2001.

[47] J. Pearl, *Heuristics: intelligent search strategies for computer problem solving*. Addison-Wesley., 1984.

[48] D. K. Ermak G, "Calcium and oxidative stress: from cell signaling to cell death." *Mol Immunol*, vol. 38, pp. 713–721, 2002.

[49] L. P. Berridge MJ, Bootman MD, "Calcium - a life and death signal." *Nature*, vol. 395, pp. 645–648, 1998.

[50] N. P. Orrenius S, Zhivotovsky B, "Regulation of cell death: the calcium-apoptosis link." *Nat Rev Mol Cell Biol*, vol. 4, pp. 552–565, 2003.

[51] Y. Zhou, W. Yang, M. Kirberger, H. Lee, G. Ayalasomayajula, and J. Yang, "Prediction of ef-hand calcium-binding proteins and analysis of bacterial ef-hand proteins." *Proteins*, vol. 65, pp. 643–655, 2006.

[52] N. Wopfner, O. Dissertori, F. Ferreira, and P. Lackner, "Calcium-binding proteins and their role in allergic diseases," *Immunol Allergy Clin North Am*, vol. 27, no. 1, pp. 29–44, 2007.

[53] E. D. Chrysina, K. Brew, and K. R. Acharya, "Crystal structures of apo- and holo-bovine alpha-lactalbumin at 2. 2-a resolution reveal an effect of calcium on inter-lobe interactions," *J Biol Chem*, vol. 275, no. 47, pp. 37 021–9, 2000.

[54] M. B. Pepys, P. N. Hawkins, D. R. Booth, D. M. Vigushin, G. A. Tennent, A. K. Soutar, N. Totty, O. Nguyen, C. C. Blake, C. J. Terry, and et al., "Human lysozyme gene mutations cause hereditary systemic amyloidosis," *Nature*, vol. 362, no. 6420, pp. 553–7, 1993.

[55] J. P. Glusker, "Structural aspects of metal liganding to functional groups in proteins," *Adv Protein Chem*, vol. 42, pp. 1–76, 1991.

[56] E. Pidcock and G. R. Moore, "Structural characteristics of protein binding sites for calcium and lanthanide ions," *J Biol Inorg Chem*, vol. 6, pp. 479–489, 2001.

[57] M. Kirberger, X. Wang, H. Deng, W. Yang, G. Chen, and J. J. Yang, "Statistical analysis of structural characteristics of protein ca2+-binding sites," *J Biol Inorg Chem*, vol. 13, no. 7, pp. 1169–81, 2008.

[58] J. A. Davis, P. A. Handford, and C. Redfield, "The n1317h substitution associated with leber congenital amaurosis results in impaired interdomain packing in human crb1 epidermal growth factor-like (egf) domains," *J Biol Chem*, vol. 282, no. 39, pp. 28 807–14, 2007.

[59] M. Nayal and E. Di Cera, "Predicting ca2+-binding sites in proteins," *Proc Natl Acad Sci U S A*, vol. 91, pp. 817–821, 1994.

[60] C. A. McPhalen, N. Strynadka, and M. N. James, "Calcium-binding sites in proteins: a structural perspective," *Adv. Protein Chem.*, vol. 42, pp. 77–144, 1991.

[61] R. H. Kretsinger, "Calcium coordination and the calmodulin fold: divergent versus convergent evolution," *Cold Spring Harb Symp Quant Biol*, vol. 52, pp. 499–510, 1987.

[62] M. M. Yamashita, L. Wesson, G. Eisenman, and D. Eisenberg, "Where metal ions bind in proteins," *Proc Natl Acad Sci U S A*, vol. 87, pp. 5648–5652, 1990.

[63] T. Dudev, Y. L. Lin, M. Dudev, and C. Lim, "First-second shell interactions in metal binding sites in proteins: a pdb survey and dft/cdm calculations," *J Am Chem Soc*, vol. 125, no. 10, pp. 3168–80, 2003.

[64] D. S. Glazer, R. J. Radmer, and R. B. Altman, "Combining molecular dynamics and machine learning to improve protein function recognition," *Pac Symp Biocomput*, vol. 13, pp. 332–343, 2008.

[65] A. J. Bordner, "Predicting small ligand binding sites in proteins using backbone structure," *Bioinformatics*, vol. 24, no. 24, pp. 2865–71, 2008.

[66] J. S. Sodhi, K. Bryson, L. J. McGuffin, J. J. Ward, L. Wernisch, and D. T. Jones, "Predicting metal-binding site residues in low-resolution structural models," *J Mol Biol*, vol. 342, no. 1, pp. 307–320, 2004.

[67] L. Wei and R. B. Altman, "Recognizing protein binding sites using statistical descriptions of their 3d environments," pp. 497–508, 1998.

[68] J. W. H. Schymkowitz, F. Rousseau, I. C. Martins, J. Ferkinghoff-Borg, F. Stricher, and L. Serrano, "Prediction of water and metal binding sites and their affinities by using the fold-x force field," *Proc Natl Acad Sci U S A*, vol. 102, no. 29, pp. 10 147–10 152, 2005.

[69] H. Deng, G. Chen, W. Yang, and J. J. Yang, "Predicting calcium-binding sites in proteins - a graph theory and geometry approach," *Proteins*, vol. 64, pp. 34–42, 2006.

[70] X. Wang, M. Kirberger, F. Qiu, G. Chen, and J. J. Yang, "Towards predicting ca2+-binding sites with different coordination numbers in proteins with atomic resolution," *Proteins*, vol. 75, no. 4, pp. 787–798, 2009.

[71] Y. Huang, Y. Zhou, W. Yang, R. Butters, H. W. Lee, S. Y. Li, A. Castiblanco, E. Brown, and J. Yang, "Identification and dissection of ca2+-binding sites in the extracellular domain of ca2+-sensing receptor," *J Biol Chem*, vol. 282, no. 26, pp. 19 000–19 010, 2007.

[72] N. Kunishima, Y. Shimada, Y. Tsuji, T. Sato, M. Yamamoto, T. Kumasaka, S. Nakanishi, H. Jingami, and K. Morikawa, "Structural basis of glutamate recognition by a dimeric metabotropic glutamate receptor," *Nature*, vol. 407, no. 6807, pp. 971–7, 2000.

[73] M. Babor, S. Gerzon, B. Raveh, V. Sobolev, and M. Edelman, "Prediction of transition metal-binding sites from apo protein structures," *Proteins*, vol. 70, no. 1, pp. 208–217, 2008.

[74] M. J. Betts and M. J. Sternberg, "An analysis of conformational changes on protein-protein association: implications for predictive docking," *Protein Eng*, vol. 12, no. 4, pp. 271–83, 1999.

[75] P. A. Handford, M. Baron, M. Mayhew, A. Willis, T. Beesley, G. G. Brownlee, and I. D. Campbell, "The first egf-like domain from human factor ix contains a high-affinity calcium binding site," *EMBO J*, vol. 9, no. 2, pp. 475–80, 1990.

[76] K. A. McClintock and G. S. Shaw, "A novel s100 target conformation is revealed by the solution structure of the ca2+-s100b-trtk-12 complex," *J Biol Chem*, vol. 278, no. 8, pp. 6251–7, 2003.

[77] K. Zhao, X. Wang, H. Wong, M. Kirberger, R. Wohlhueter, G. Chen, and J. Yang, "Predicting ca2+-binding sites using refined carbon clusters," *submitted*.

[78] X. Wang, K. Zhao, M. Kirberger, H. Wong, G. Chen, and J. J. Yang, "Analysis and prediction of calcium-binding pockets from apo-protein structures exhibiting calcium-induced localized conformational changes," *Protein Sci*, vol. 19, no. 6, pp. 1180–90, 2010.

[79] E. Tomita, A. Tanaka, and H. Takahashi, *The worst-case time complexity for generating all maximal cliques*. Heidelberg: Springer Berlin, 2004, vol. 3106.

[80] C. Bron and J. Kerbosch, "Algorithm 457: finding all cliques of an undirected graph." *Communications of the ACM*, vol. 16, no. 9, pp. 575–579, 1973.

[81] Y. Zhang, A. Kolinski, and J. Skolnick, "Touchstone ii: a new approach to ab initio protein structure prediction," *Biophys J*, vol. 85, no. 2, pp. 1145–64, 2003.

[82] V. Sobolev, A. Sorokine, J. Prilusky, E. E. Abola, and M. Edelman, "Automated analysis of interatomic contacts in proteins," *Bioinformatics*, vol. 15, no. 4, pp. 327–32, 1999.

[83] R. D. Shannon, "Revised effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides," *Acta Cryst*, vol. A32, pp. 751–767, 1976.

[84] T. J. Simons and G. Pocock, "Lead enters bovine adrenal medullary cells through calcium channels," *J Neurochem*, vol. 48, no. 2, pp. 383–9, 1987.

[85] P. L. Goering, "Lead-protein interactions as a basis for lead toxicity," *Neurotoxicology*, vol. 14, no. 2-3, pp. 45–60, 1993.

[86] T. L. Dowd, J. F. Rosen, C. M. Gundberg, and R. K. Gupta, "The displacement of calcium from osteocalcin at submicromolar concentrations of free lead," *Biochim Biophys Acta*, vol. 1226, no. 2, pp. 131–7, 1994.

[87] H. Godwin, "The biological chemistry of lead," *Current Opinion in Chemical Biology*, vol. 5, pp. 223–227, 2001.

[88] W. D. Atchison, "Effects of toxic environmental contaminants on voltage-gated calcium channel function: from past to present," *J Bioenerg Biomembr*, vol. 35, no. 6, pp. 507–32, 2003.

[89] E. H. Harutyunyan, V. Y. Oganessyan, N. N. Oganessyan, S. M. Avaeva, T. I. Nazarova, N. N. Vorobyeva, S. A. Kurilova, R. Huber, and T. Mather, "Crystal structure of holo inorganic pyrophosphatase from escherichia coli at 1.9 a resolution. mechanism of hydrolysis," *Biochemistry*, vol. 36, no. 25, pp. 7754–60, 1997.

[90] M. Asuncion, W. Blankenfeldt, J. N. Barlow, D. Gani, and J. H. Naismith, "The structure of 3-methylaspartase from clostridium tetanomorphum functions via the common enolase chemical step," *J Biol Chem*, vol. 277, no. 10, pp. 8306–11, 2002.

[91] A. R. Means, *Calcium regulation of celluar function*. Academic Press, 1994, vol. 30.

[92] T. P. Singh, S. Sharma, S. Karthikeyan, C. Betzel, and K. L. Bhatia, "Crystal structure of a complex formed between proteolytically-generated lactoferrin fragment and proteinase k," *Proteins*, vol. 33, no. 1, pp. 30–8, 1998.

[93] M. Kirberger and J. J. Yang, "Structural differences between pb(2+)- and ca(2+)-binding sites in proteins: Implications with respect to toxicity," *J Inorg Biochem*, 2008.

[94] J. J. Falke, S. K. Drake, A. L. Hazard, and O. B. Peersen, "Molecular tuning of ion binding to calcium signaling proteins," *Q Rev Biophys*, vol. 27, no. 3, pp. 219–90, 1994.

[95] M. Babor, H. M. Greenblatt, M. Edelman, and V. Sobolev, "Flexibility of metal binding sites in proteins on a database scale," *Proteins*, vol. 59, no. 2, pp. 221–30, 2005.

[96] W. Yang, A. L. Wilkins, S. Li, Y. Ye, and J. J. Yang, "The effects of ca2+ binding on the dynamic properties of a designed ca2+-binding protein," *Biochemistry*, vol. 44, no. 23, pp. 8267–73, 2005.

[97] W. Yang, A. L. Wilkins, Y. Ye, Z. R. Liu, S. Y. Li, J. L. Urbauer, H. W. Hellinga, A. Kearney, P. A. van der Merwe, and J. J. Yang, "Design of a calcium-binding protein

with desired structure in a cell adhesion molecule," *J Am Chem Soc*, vol. 127, no. 7, pp. 2085–93, 2005.

[98] J. Zou, A. M. Hofer, M. M. Lurtz, G. Gadda, A. L. Ellis, N. Chen, Y. Huang, A. Holder, Y. Ye, C. F. Louis, K. Welshhans, V. Rehder, and J. J. Yang, "Developing sensors for real-time measurement of high ca2+ concentrations," *Biochemistry*, vol. 46, no. 43, pp. 12 275–88, 2007.

[99] C. for Disease Control and Prevention, *Sexually transmitted disease surveillance 2008*, http://www.cdc.gov/std/stats08/trends.htm, retrieved Sept 1, 2010.

[100] ——, *Recommends screening of all sexually active women 25 and under.*, http://www.cdc.gov/std/infertility/default.htm, retrieved Sept 16, 2010.

[101] U. P. S. T. Force, *Screening for Gonorrhea*, http://www.uspreventiveservicestaskforce.org/uspstf/uspsgono.htm, retrieved Sept 13, 2010.

[102] E. H. Kaplan and H. Pollack, "Allocating HIV prevention resrouces," *Socio-Econ. Plann. Sci.*, vol. 4, pp. 257–263, 1998.

[103] A. Lasry and G. S. Zaric, "Multi-level resource allocation for HIV prevention: A model for developing countries." *European Journal of Operational Research*, vol. 180, pp. 786–799, 2007.

[104] M. L. Brandeau and G. S. Zaric, "Optimal investment in HIV prevention programs: more is not always better." *Health Care Management Science*, vol. 12, pp. 27–37, 2009.

[105] P. Sendi and M. J. Al, "Revisiting the decision rule of cost-effectiveness analysis under certainty and uncertainty." *Social Science & Medicine*, vol. 57, pp. 969–974, 2003.

[106] G. Tao, B. K. Abban, T. L. Gift, G. Chen, and K. L. Irwin, "Applying a mixed-integer program to model re-screening women who test positive for *C.trachomatis* infection," *Health Care Manag Sci*, vol. 7, pp. 134–144, 2004.

[107] G. Tao, T. L. Gift, C. M. Walsh, K. L. Irwin, and W. J. Kassler, "Optimal resource allocation for curing *Chlamydia trachomatis* infection among asymptomatic women at clinics operating on a fixed budget," *Sex Transm Dis*, vol. 29, pp. 703–709, 2002.

[108] Microsoft, *"Microsoft Excel Solver User's Guide" for Windows*, http://support.microsoft.com/kb/82890, retrieved Jan 20, 2008.

[109] A. Lasry, M. W. Carter, and G. S. Zaric, "S4hara: System for HIV/AIDS resource allocation." *Cost Effectiveness and Resource Allocation*, vol. 6, p. 7, 2008.

[110] M. Rauner, S. Brailsford, and S. Flessa, "Use of discrete-event simulation to evaluate strategies for the prevention of mother-to-child transmission of HIV in developing countries." *Journal of the Operational Research Society*, vol. (56), pp. 222–233, 2005.

[111] S. Martello, D. Pisinger, and P. Toth, "New trends in exact algorithms for the 0-1 knapsack problem." *European Journal of Operation Research*, vol. 123, pp. 325–332, 2000.

[112] S. Martello and P. Toth, *Knapsack problems*, ser. Wiley-Interscience Series in Discrete Mathematics and Optimization. Chichester: John Wiley & Sons Ltd., 1990, algorithms and computer implementations.

[113] A. Lasry, A. Richter, and F. Lutscher, "Recommendations for increasing the use of HIV/AIDS resource allocation models." *BMC Public Health*, vol. (9(Suppl 1)), p. S8, 2009.

[114] K. Zhao, G. Chen, G. Thomas, and G. Tao, "Optimization model and algorithm help to screen and treat sexually transmitted diseases," *International Journal of Computational Models and Algorithms in Medicine*, vol. 1, no. 4, pp. 1–18, 2010.

[115] T. L. Gift, C. Walsh, A. Haddix, and K. L. Irwin, "A cost-effectiveness evaluation of testing and treatment of *Chlamydia trachomatis* infection among asymptomatic

women infected with *Neisseria gonorrhoeae*." *Sex Transm Dis*, vol. 29, pp. 542–551, 2002.

[116] L. M. Newman, J. S. Moran, and K. A. Workowski, "Update on the management of *gonorrhea* in adults in the United States," *Clin Infect Dis*, vol. 44(Suppl 3), pp. S84–S101, 2007.

[117] L. W. Dicker, D. Mosure, W. Levine, C. Black, and B. S.M., "Impact of switching laboratory tests on reported treads in chlamydia trachomatis infections." *American Journal of Epidemiology*, vol. 151, pp. 430–435, 2000.

[118] W. Miller, C. Ford, M. Morris, M. Handcock, J. Schmitz, M. Hobbs, M. Cohen, K. Harris, and J. Udry, "Prevalence of chlamydial and gonococcal infections among young adults in the united states." *Journal Of American Medical Association*, vol. 291, pp. 2229–36, 2004.

[119] C. M. Black, J. Marrazzo, R. E. Johnson, E. W. H. III, R. B. Jones, T. A. Green, J. Schachter, W. E. Stamm, G. Bolan, M. E. S. Louis, and D. H. Martin, "Head-to-head multicenter comparison of DNA probe and nucleic acid amplification tests for *Chlamydia trachomatis* infection in women performed with an improved reference standard," *J Clin Microbiol*, vol. 40, pp. 3757–3763, 2002.

[120] B. Van Der Pol, D. V. Ferrero, L. Buck-Barrington, E. r. Hook, C. Lenderman, T. Quinn, C. A. Gaydos, J. Lovchik, J. Schachter, J. Moncada, G. Hall, M. J. Tuohy, and R. B. Jones, "Multicenter evaluation of the BD PROBETEC ET System for detection of *Chlamydia trachomatis* and *Neisseria gonorrhoeae* in urine specimens, female endocervical swabs, and male urethral swabs," *J Clin Microbiol*, vol. 39, pp. 1008–16, 2001.

[121] D. H. Martin, C. Cammarata, B. Van Der Pol, R. B. Jones, T. C. Quinn, C. A. Gaydos, K. Crotchfelt, J. Schachter, J. Moncada, D. Jungkind, B. Turner, and C. Peyton,

"Multicenter evaluation of AMPLICOR and automated COBAS AMPLICOR CT/NG tests for *Neisseria gonorrhoeae*," *J Clin Microbiol*, vol. 38, pp. 3544–9, 2000.

[122] E. H. Koumans, R. E. Johnson, J. S. Knapp, and M. E. St Louis, "Laboratory testing for *Neisseria gonorrhoeae* by recently introduced nonculture tests: a performance review with clinical and public health considerations." *Clin Infect Dis*, vol. 27, pp. 1171–80, 1998.

[123] C. A. Gaydos, T. C. Quinn, D. Willis, A. Weissfeld, E. W. Hook, D. H. Martin, D. V. Ferrero, and J. Schachter, "Performance of the APTIMA Combo 2 assay for detection of *Chlamydia trachomatis* and *Neisseria gonorrhoeae* in female urine and endocervical swab specimens," *J Clin Microbiol*, vol. 41, pp. 304–309, 2003.

[124] J. Schachter, J. M. Chow, H. Howard, G. Bolan, and J. Moncada, "Detection of *Chlamydia trachomatis* by nucleic acid amplification testing: our evaluation suggests that CDC-Recommended approaches for confirmatory testing are ill-advised," *J Clin Microbiol*, vol. 44, pp. 2512–2517, 2006.

[125] M. Augenbraun, L. Bachmann, T. Wallace, L. duBouchet, W. McCormack, and E. Hook, "Compliance with doxycycline therapy in sexually transmitted diseases clinics." *Sex Transm Dis*, vol. 25, pp. 1–4, 1998.

[126] D. H. Martin, T. F. Mroczkowski, Z. A. Dalu, J. McCarty, R. Jones, S. Hopkins, and R. Johnson, "A controlled trial of single dose azithromycin for the treatment of *Chlamydia* urethritis and cervicitis." *N Engl J Med*, vol. 327(13), pp. 921–925, 1992.

[127] M. R. Golden, J. P. Hughes, L. E. Cles, K. Crouse, K. Gudgel, J. Hu, P. D. Swenson, W. E. Stamm, and H. H. Handsfield, "Positive predictive value of Gen-Probe APTIMA Combo 2 testing for *Neisseria gonorrhoeae* in a population of women with low prevalence of *N.gonorrhoeae* infection," *Clin Infect Dis*, vol. 39, pp. 1387–90, 2004.

[128] W. M. Geisler, "Management of uncomplicated *Chlamydia trachomatis* infections in adolescents and adults: evidence reviewed for the 2006 Centers for Disease Control

and Prevention sexually transmitted diseases treatment guidelines." *Clin Infect Dis*, vol. 44, pp. S77–S83, 2006.

[129] C. for Medicare and M. Services, *CMS programs and information.*, http://www.cms.hhs.gov/, retrieved Sept 14, 2010.

[130] M. Howell, R. M. A., T. C. Quinn, W. Brthwaite, and C. A. Gaydos, "Screening women for *Chlamydia trachomatis* in family planning clinics: The cost-effectiveness of DNA amplification assays." *Sex Transm Dis*, vol. 25(2), pp. 108–117, 1998.

[131] A. of Public Health Laboratories, *2001 Sexually Transmitted Diseases Laboratory Test Method Survey.*, http://www.aphl.org/, retrieved Sept 14, 2010.

[132] T. Healthcare, *Drug Topics Red Book*, 2008th ed. Montvale, NJ: Thomson Healthcare, 2008.

[133] C. E. Begley, L. McGill, and P. B. Smith, "The incremental cost of screening, diagnosis, and treatment of *Gonorrhea* and *Chlamydia* in a family planning clinic." *Sex Transm Dis*, vol. 16, pp. 63–7, 1989.

[134] J. Yeh, E. Hook, and S. J. Goldie, "A refined estimate of the average lifetime cost of pelvic inflammatory disease." *Eval Program Plann*, vol. 30, pp. 369–378, 2003.

[135] A. E. Washington, R. E. Johnson, and L. L. J. Sanders, "*Chlamydia trachomatis* infections in the United States. what are they costing us?" *JAMA*, vol. 257, pp. 2070–2, 1987.

[136] L. H. Bachmann, C. M. Richey, K. Waites, J. R. Schwebke, and E. W. Hook, "Patterns of *Chlamydia trachomatis* testing and follow-up at a university hospital medical center," *JAMA*, vol. 26, pp. 496–499, 1999.

[137] B. of Labor Statistics, *Consumer price index-all urban consumers*, http://www.bls.gov/cpi/, retrieved Aug 14, 2010.

[138] H. Kellerer, U. Pferschy, and D. Pisinger, *Knapsack problems*, 1st ed. Berlin: Springer-Verlag, 2004.

[139] C. Wilbaut, S. Hanafi, and S. Salhi, "A survey of effective heuristics and their application to a variety of knapsack problems." *IMA Journal of Management Mathematics*, vol. (19), p. 227, 2008.

[140] S. Blount, A. Galambosi, and S. Yakowitz, "Nonlinear and dynamic programming for epidemic intervention." *Applied Mathematics and Computation*, vol. 123-136, pp. 325–332, 1997.

[141] R. Bellman, *Dynamic programming*. Mineola, N.Y.: Dover Publications Inc., 2003, reprint of the sixth (1972) edition, With an introduction by Eric V. Denardo.

[142] D. M. Faissol, J. L. Swann, P. M. Griffin, and T. L. Gift, "The role of bathhouses and sex clubs in HIV transmission: Findings from a mathematic model," *J Acquir Immune Defic Syndr*, vol. 44, pp. 386–394, 2007.

[143] T. L. Gift and K. L. Irwin, "Factors that influence the cost effectiveness of *Gonorrhea* screening in emergency departments." *Sex Transm Dis*, vol. 32, pp. 437–438, 2005.

[144] G. R. Burstein, M. H. Snyder, D. Conley, D. R. Newman, C. M. Walsh, G. Tao, and K. L. Irwin, "*Chlamydia* screening in a health plan before and after a national performance measure introduction," *Obstet Gynecol*, vol. 106, pp. 327–334, 2005.

[145] D. Hu, E. W. Hook, and S. J. Goldie, "Screening for *Chlamydia trachomatis* in women 15 to 29 eyars of age: a cost-effectiveness analysis," *Ann Intern Med*, vol. 141, pp. 501–513, 2004.

[146] C. Y. Lau and A. K. Qureshi, "Azithromycin versus doxycycline for genital chlamydial infections-a meta-analysis of randomized clinical trials," *Sex Transm Dis*, vol. 29(9), pp. 497–502, 2002.

[147] L. A. Shrier, S. K. Harris, and W. R. Beardslee, "Temporal associations between depressive symptoms and self-reported sexually transmitted disease among adolescents," *Arch Pediatr Adolesc Med*, vol. 156(6), pp. 599–606, 2002.

[148] L. M. Niccolai, K. A. Livingston, F. F. H. Teng, and M. M. Pettigrew, "Behavioral intentions in sexual partnerships following a diagnosis of *Chlamydia trachomatis*," *Prev Med*, vol. 46(2), pp. 170–176, 2008.

[149] B. Smith, D. Gunther, B. Rao, and R. Ratliff, *Interfaces*, vol. 31, no. 2, pp. 37–55, 2001.

[150] J. Hopfield and D. Tank, *Biol Cybern*, vol. 52, no. 3, pp. 141–152, 1985.

[151] F. Araujo, B. Ribeiro, and L. Rodrigues, *IEEE Transactions on Neural Networks*, vol. 12, no. 5, pp. 1067–1073, 2001.

[152] C. Ahn, R. Ramakrishna, C. Kang, and I. Choi, *Electronics Letters*, vol. 37, no. 19, pp. 1176–1178, 2001.

[153] J. Wang, *IEEE Transactions on Circuits and Systems - Part I: Fundamental Theory and Applications*, vol. 43, no. 6, pp. 482–486, 1996.

[154] X. Wang and Y. Qu, H.and Zhang, *Neurocomputing*, vol. 72, pp. 3028–3033, 2009.

[155] Y. Wang, G. Wu, L.and Wei, and S. Wang, *Digital Signal Processing*, vol. 21, pp. 517–521, 2001.

APPENDICES

Appendix A: Synchronization threshold for two coupled Lorenz systems

In this Appendix, we follow the steps in the previous study [2] to review the calculation of the stability parameter a in the requirement (1.9) for a two-node network (1.1) of x-coupled Lorenz systems.

Consider the following coupled system

$$\begin{cases} \dot{x}_i = \sigma(y_i - x_i) + \sum_{j=1}^{n} \varepsilon_{ij}(t) x_j, \\ \dot{y}_i = r x_i - y_i - x_i z_i \\ \dot{z}_i = -b z_i + x_i y_i, \quad i = 1, ..., n \end{cases} \tag{E.1}$$

where the vector (x_i, y_i, z_i) is the vector x_i from (1.1). σ, r, and b are standard parameters of the individual Lorenz system.

We follow the steps in the previous study to prove that the requirement (1.9) is true for the network (E.1). To do so, we need to prove the eventual dissipativeness of the individual Lorenz system. It has previously been done by finding an appropriate level of the Lyapunov function (see, for example, [43]); the Lorenz system is indeed eventually dissipative and has an absorbing domain

$$B = \{x^2 + y^2 + (z - r - \sigma)^2 < b^2(r + \sigma)^2/4(b - 1)\}.$$

Therefore, the orbits of the attractor of the individual Lorenz system are bounded by

$$|\varphi| < b(r + \sigma)/2\sqrt{b - 1}, \quad \varphi = x, y, (z - r - \sigma). \tag{E.2}$$

It has also been shown [43] that the upper bounds (E.2) are valid for each oscillator of the coupled system (E.1).

The auxiliary system (1.7) from the requirement (1.9) reads

$$\begin{cases} \dot{X}_{ij} = \sigma(Y_{ij} - X_{ij}) - aX_{ij} \\ \dot{Y}_{ij} = \left(r - U_{ij}^{(z)}\right) X_{ij} - Y_{ij} - U_{ij}^{(x)} Z_{ij} \\ \dot{Z}_{ij} = U_{ij}^{(y)} X_{ij} + U_{ij}^{(x)} Y_{ij} - bZ_{ij}, \quad i,j = 1,...,n, \end{cases} \tag{E.3}$$

where $U_{ij}^{(\xi)} = (\xi_i + \xi_j)/2$ for $\xi = x, y, z$ represent the corresponding sum variables, and $-aX_{ij}$ is the extra term, coming from the addition of matrix $A = aP$.

The cross terms in the system (E.3) can be eliminated by using

$$\xi_j \eta_j - \xi_i \eta_i = U^{(\eta)}(\xi_j - \xi_i) + U^{(\xi)}(\eta_j - \eta_i).$$

The Lyapunov functions (1.8), that we use to prove the global stability of the origin of system (E.3), read

$$W_{ij} = X_{ij}^2/2 + Y_{ij}^2/2 + Z_{ij}^2/2, \ i,j = 1,...,n. \tag{E.4}$$

Their derivatives with respect to the system (E.3) are calculated as follows

$$\dot{W}_{ij} = -\left[(a+\sigma)X_{ij}^2 + (U^{(z)} - r - \sigma)X_{ij}Y_{ij} + Y_{ij}^2 - U^{(y)}X_{ij}Z_{ij} + bZ_{ij}^2\right]. \tag{E.5}$$

We apply the Silvester's criterion for proving negative definiteness of the quadratic forms (E.5) to get three conditions [2]: $a + \sigma > 0$,

$$\begin{vmatrix} a+\sigma & \frac{U^{(z)}-r-\sigma}{2} \\ \frac{U^{(z)}-r-\sigma}{2} & 1 \end{vmatrix} > 0, \quad \text{and} \quad \begin{vmatrix} a+\sigma & \frac{U^{(z)}-r-\sigma}{2} & -\frac{U^{(y)}}{2} \\ \frac{U^{(z)}-r-\sigma}{2} & 1 & 0 \\ -\frac{U^{(y)}}{2} & 0 & b \end{vmatrix} > 0. \tag{E.6}$$

Plugging the estimate (E.2) for $U^{(y)}$ and $U^{(z)}$ into (E.6), we get the condition under which the origin of system (1.7) is globally stable:

$$a > a^* = \frac{b(b+1)(r+\sigma)^2}{16(b-1)} - \sigma. \tag{E.7}$$

This implies that the requirement (1.9) is fulfilled for networks of x-coupled Lorenz systems. At the same time, it implies that the two-node network of x-coupled Lorenz systems synchronizes completely and globally as long as the coupling $\varepsilon_{12} = \varepsilon_{21}$ exceeds the value $a^*/2$.

Appendix B: A Neural-network Algorithm for All k Shortest Path Problem

All k Shortest Path (KSP) Problem is one of classic combinatorial problem. Many real-life problem can be converted into the k shortest path problem. One of these applications is flight search algorithm. The travel agencies in airline industry need to find the lowest air fares from departure city to destination. Consider each city as a vertex, if there is an air fare in between two cities, then we can construct an edge, the cost will be the weights on that edge and the direction of that edge is determined by the availability from one to the another.

Various travel agencies utilized different algorithms to solve a KSP-like problem while pricing the itinerary. For example, Sabre Inc. the owner of Travelocity.com who holds 44.7% market share in US, has utilized different algorithm such as dynamic programming and Dijkstra-type algorithm for flight searching algorithm [149]. After Travelport acquired Worldspan, it controls a 46.3% market share in the US using 2002 airline booking data. This company empowers Priceline, adopted A* algorithm together with Breath-first search for an initial fare estimate. Amadeus who owned Expedia and others travel agencies used similar search engine [149]. ITA Software who is supporting Orbitz's pricing tool and search engine, uses different flight search algorithms according to Carl de Marcken, a founder and scientist at the company. Unfortunately, due to the confidential nature of the subject, the specific techniques used in these search engine are not published.

This study provides a generalized neural network form with a simple example. Secondly, this study will propose a new graph evolution algorithm based on the mechanism of neural network. Finally, the performance of the new algorithm will be analyzed.

Coupled neural network in continuous time and KSP problem

An overview

Coupled neural networks in continuous time has been suggested as an effective method [150] [9] [151] for solving shortest path problem. In these methods, decision variables v_{ij}

(or edges in a graph) are represented by the activation states of neurons which are further modeled by a system of differential equations. A Lyapunov (energy) function is defined to drive each neuron into its stable state [152]. Furthermore, these neurons are defined to interact with each other by chemical (pulsed) coupling as opposed to electrical (linear) coupling, while both forms of coupling are observed in nature though [153][152].

Unlike defining neurons as edges in a graph, recent techniques represent neurons as vertices in the graph instead [154] [155]. The utilized neurons in a network fire through the chemical coupling. The dynamics of each neurons are modeled by differential equations, which are designed to realize that the smaller coupling strength (e.g. connection weights in a graph) lead to earlier firing times. In other words, after an excitation from the initial vertex, the signal will spread (like a wave propagation) based on the graph (network) topology and individual dynamics. By constructing a proper neural network, a tracked signal travels from an initial vertex to terminal vertex through the shortest path. An advantage of the method is that the spreading time of the signal (wave) is independent of the number of vertices in the graph but only determined by the path length from the initial vertex to the terminal vertex. Unfortunately, in many realistic neural models due to their non-linear nature, e.g. Hodgkin-Huxley model, Hindmarsh-Rose model, leech model and etc [2], solutions can not be found analytically. Simulations have to be conducted for integrating the system, which could be time-consuming.

Next, we will provide a generalized form in a continuous time for the existing neural models which have not be found in previous publication. It offers theoretical foundation for a new neural KSP algorithm.

General form of coupled neural network in continuous time

To solve the KSP problem, we can construct a network of n interacting linear/non-linear l-dimensional dynamical system (neurons) and denote them as $x_i = (x_i^1,\ x_i^2,\ ...,\ x_i^l),\ i = 1, ..., n$, then we can define the following general framework for the coupled network:

$$\dot{x}_i = F(x_i) - \sum_{j=1,\ j\neq i}^{n} d_{ij}(t)\tau(x_j)$$

where $F(x_i)$ define each individual system, $\tau(x_j)$ is a activation (sigmoid) function $\tau(x_j) = \frac{1}{1+e^{-\lambda(x_j-\theta)}}$. It represents that the ith neuron is extincted by jth neuron while the potential of jth neuron exceed a synaptic threshold θ. $d_{ij}(t)$ is the coupling strength (weights/costs) depend on time t from vertex i to vertex j and is a positive number.

The advantage of offering this form is that the existing methods which model vertices in a graph as neurons, can be included under this form.

A Neural Network algorithm for KSP problem

The ideal case is that we can solve the system in a closed form. However, if we look into the general form, the sigmoid function which makes the system non-linear complicates the problem. We may not able to derived an analytical solutions. Therefore, when computer handles this problem, it needs to break the complex problem in continuous time into discrete time.

Simple example in discrete time

We present how a neuron network can help to find the shortest path using a simple example. This is important for understanding that why the underlying neural KSP algorithm is an exact algorithm. By "Exact algorithm", we mean finding the global optimal solution if exists. Let's consider a simple asymmetric directed graph in Fig. E.1.

Results

The results are listed below from node one to the ending node:

1-4: 13

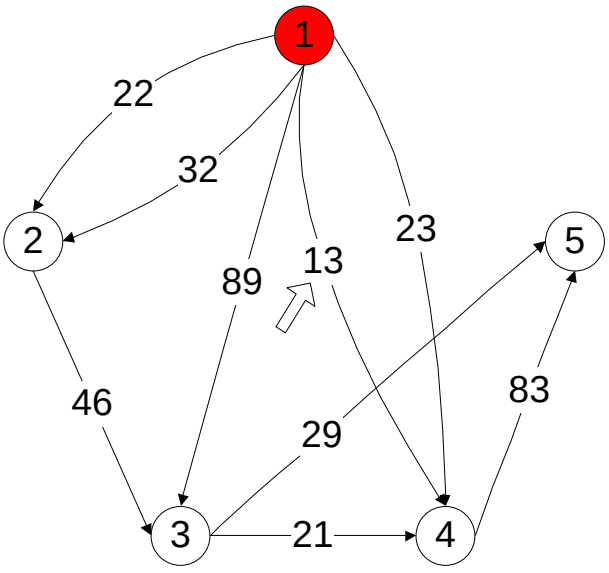
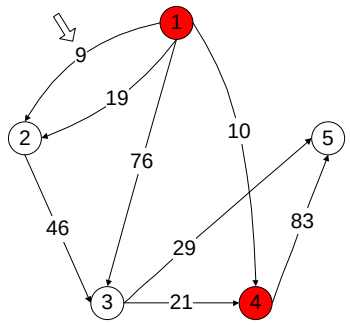


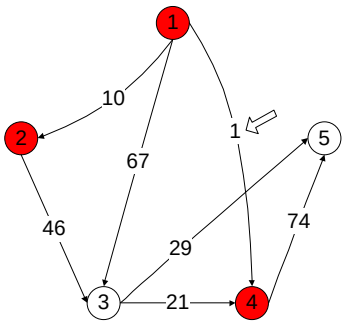
Figure E.1. Example of Neural KSP algorithm. Starting state.

1-4: 13
1-2: 9



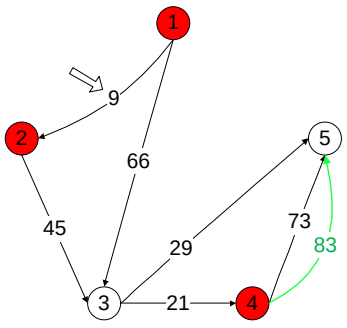
(a) State 2

1-4: 13
1-2: 9
1-4: 1



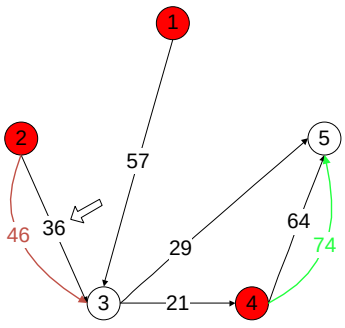
(b) State 3

1-4: 13
1-2: 9
1-4: 1
1-2: 9



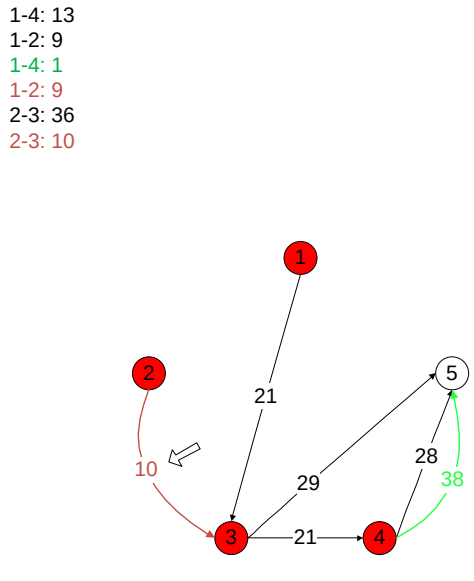
(c) State 4

1-4: 13
1-2: 9
1-4: 1
1-2: 9
2-3: 36

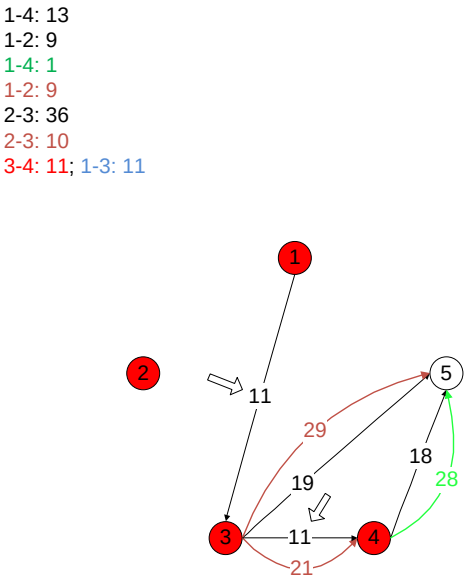


(d) State 5

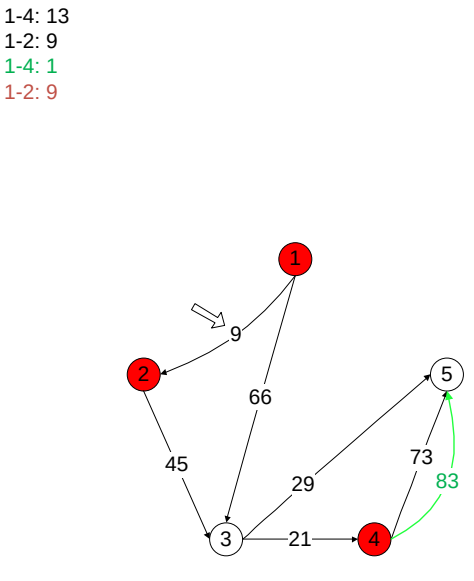
Figure E.2. Example of Neural KSP algorithm. Intermediate states.



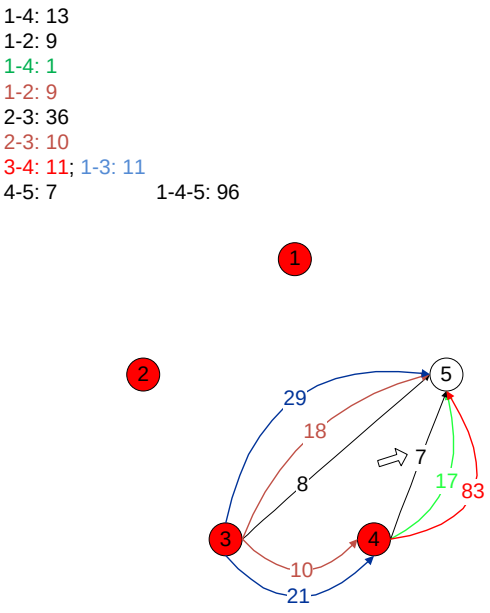
(a) State 6



(b) State 7

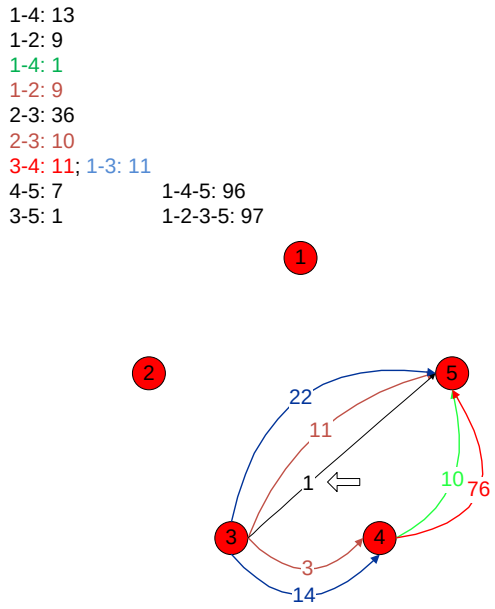


(c) State 4

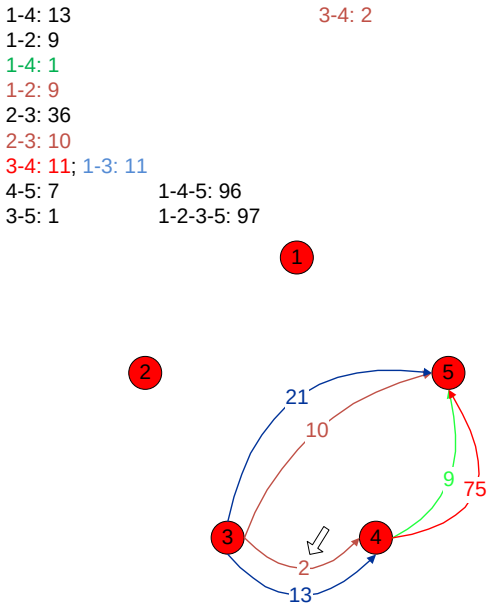


(d) State 8

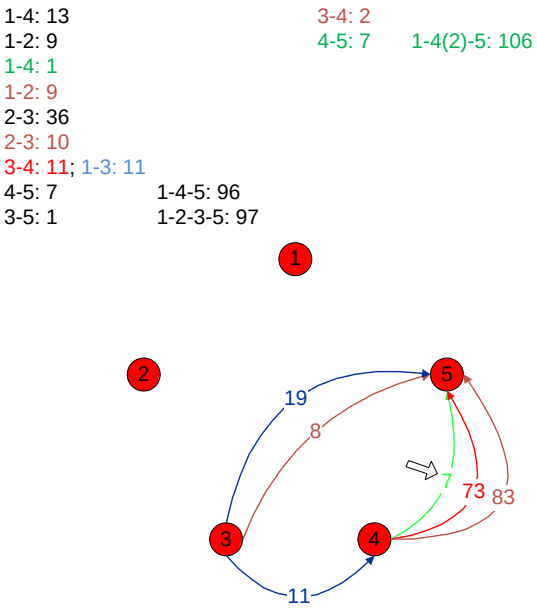
Figure E.3. Example of Neural KSP algorithm. Intermediate states.



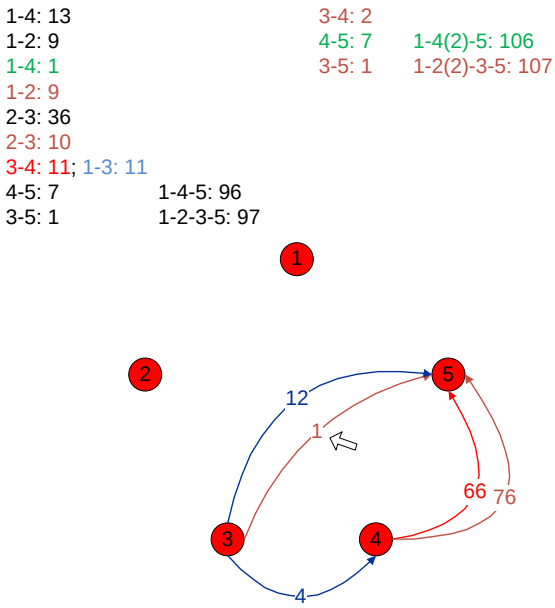
(a) State 9



(b) State 10

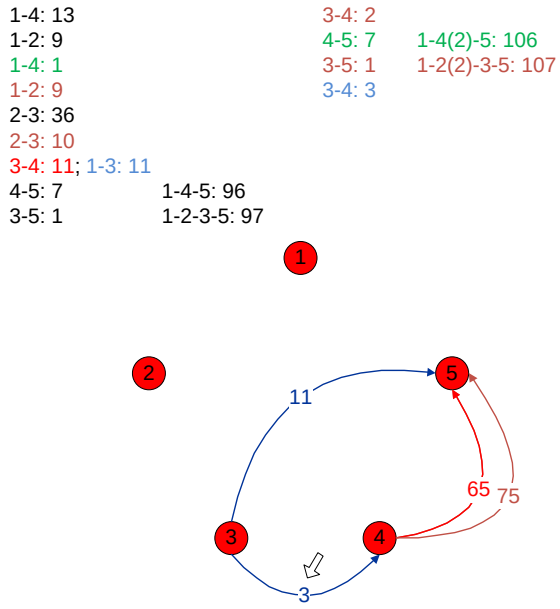


(c) State 11

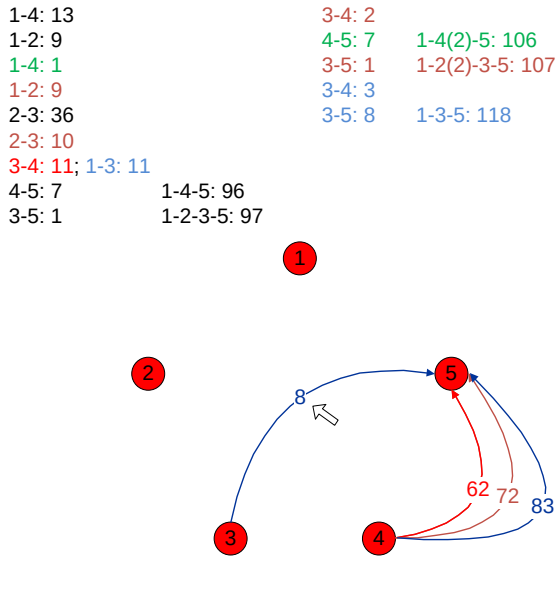


(d) State 12

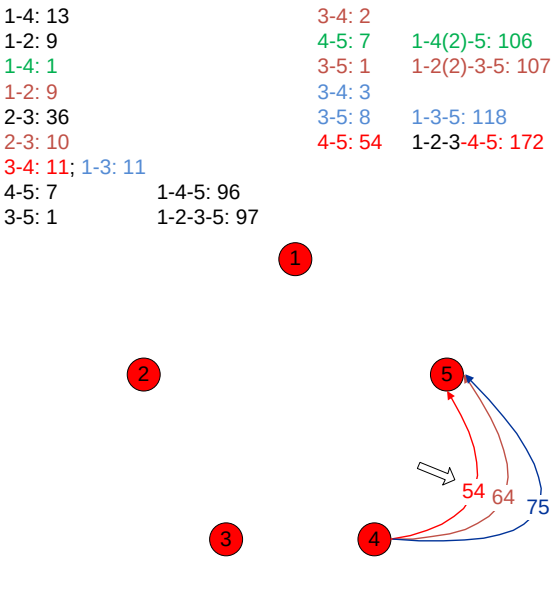
Figure E.4. Example of Neural KSP algorithm. Intermediate states.



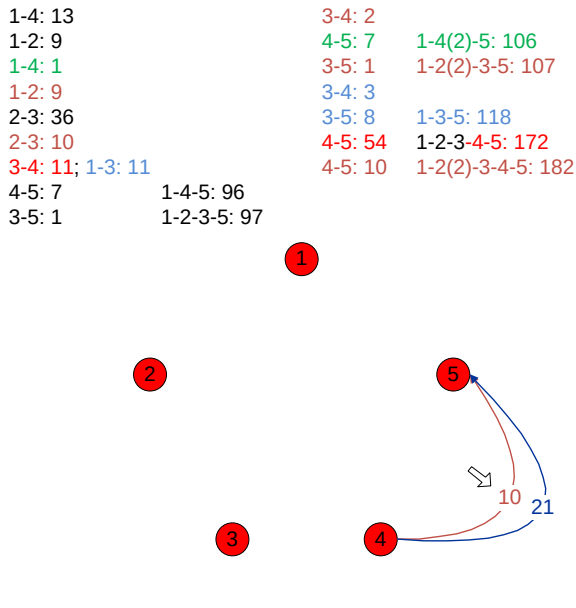
(a) State 13



(b) State 14



(c) State 15



(d) State 16

Figure E.5. Example of Neural KSP algorithm. Intermediate states.

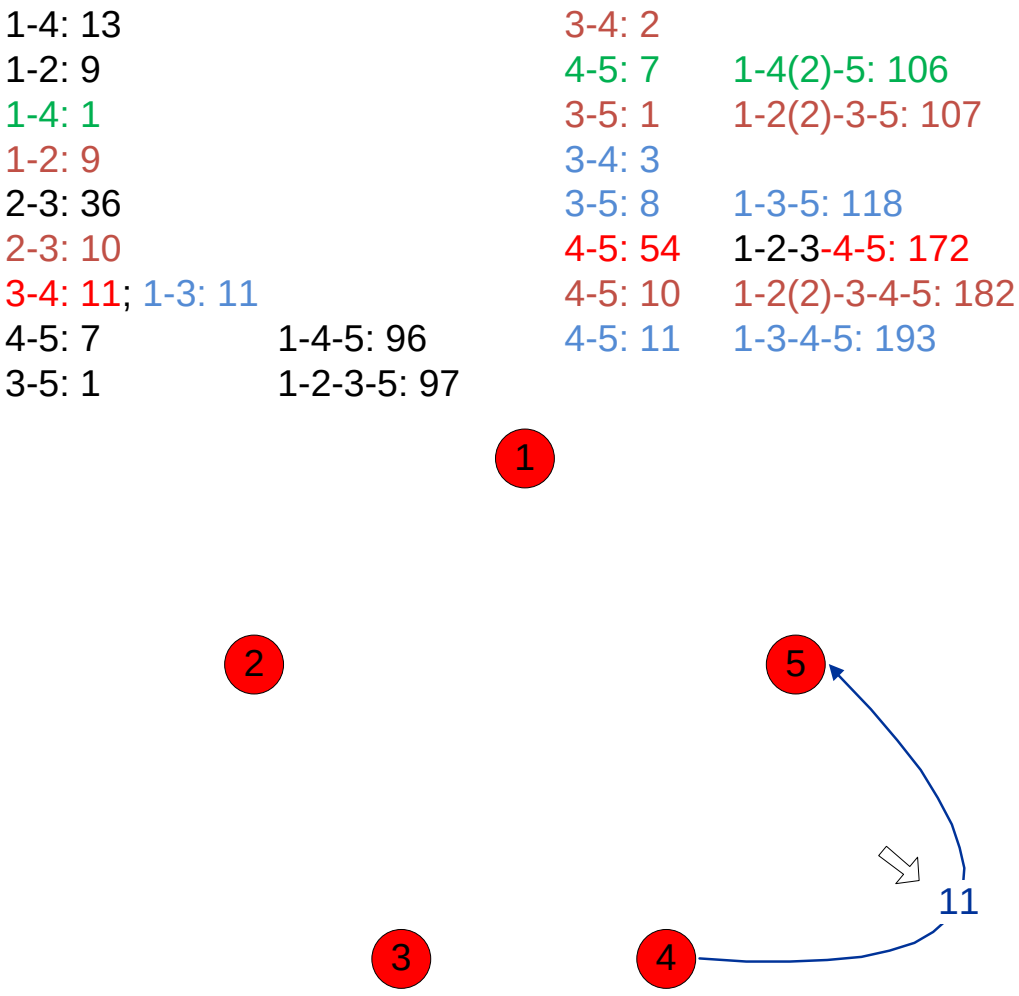


Figure E.6. Example of Neural KSP algorithm. Final state.

Table E.1. Performance on four simple graphs.

K	Run times	Nodes	Edges	Reading	Near KSP	Neural KSP	Path length
1	10	5	7	0	0	0	2
		100	138	0	0	1.5	3
		500	125022	1263	733	32	2
		1000	499205	5974	3432	78	2
5	10	5	7	0	0	1.5	2, 3, 4
		100	138	0	0	1.5	3, 4, 5, 7
		500	125022	1263	764	187	2, 3, 5
		1000	499205	5974	3432	234	2, 3
10	10	5	7	0	-	-	-
		100	138	0	7	16	3, 4, 5, 6, 7
		500	125022	1263	765	452	2, 3, 5
		1000	499205	5974	3495	3557	2, 3, 4, 5, 6
20	10	5	7	0	-	-	-
		100	138	0	15	16	3, 4, 5, 6, 7
		500	125022	1263	780	2964	2, 3, 4, 5, 6, 8, 9
		1000	499205	5974	4041	23290	2, 3, 4, 5, 6, 7, 8

Table E.2. Performance on four multigraphs.

K	Run times	Nodes	Edges	Reading[1]	Near KSP[2]	Neural KSP[3]	PL[4]	PC[5]
1	10	5	14	0	0	1.5	2	96
		100	149	0	0	1.5	3	49
		500	500084	38266	35565	2574	2	9
		1000	1996820	124040	138239	13572	2	9, 9
5	10	5	14	0	1.5	1.6	3	97, 99, 106, 107
		100	149	0	15	16	4	64, 65, 66, 67
		500	500084	32619	32479	8768	2	10, 10, 11, 11, 11
		1000	1996820	127140	138934	24445	2, 3	10, 10, 10, 10, 10
10	10	5	7	0	4	5	2, 3, 4	109, 118, 120, 172, 182
		100	149	0	15	31	4, 5	68, 69, 77, 78, 79
		500	500084	33266	33954	15444	2	12, 12, 12, 12, 12
		1000	1996820	131407	138544	41496	2, 3	11_{10}
15	10	5	7	0	15	16	3	193
		100	149	0	30	40	4	80, 81, 82, 87,88
		500	500084	33266	33954	26098	2	13, 13, 13, 13, 13
		1000	1996820	131407	138544	41496	-	-

[1]CPU reading times in millisecond; [2]Near KSP algorithm in terms of CPU time in millisecond; [3] The Proposed Neural KSP algorithm in terms of CPU time in millisecond; [4]Path Length; [5]Path Costs.

Table E.3. Performance on four multigraphs.(Continued)

K	Run times	Nodes	Edges	Reading[1]	Near KSP[2]	Neural KSP[3]	PL[4]	PC[5]
20	10	5	7	0	62	67	-	-
		100	149	0	30	40	4, 5	89, 90, 90, 90, 91
		500	500084	33443	34154	42713	2	14, 14, 14, 14, 14
		1000	1996820	129940	138796	80418	2,3	12_{18}
25	10	5	7	0	62	67	-	-
		100	149	0	31	47	4, 5, 6	92, 113, 118, 119, 120
		500	500084	33443	32885	49171	2	15, 15, 15, 15, 15
		1000	1996820	129940	138796	80418	Same	Same
30	10	5	7	0	62	67	-	-
		100	149	0	40	60	6	121, 122, 123
		500	500084	33443	34154	108692	2, 3	16, 16, 16, 16
		1000	1996820	129940	138796	80418	Same	Same
35	10	5	7	0	62	67	-	-
		100	149	0	40	60	-	-
		500	500084	33443	33103	148692	2, 3	17, 17, 17, 17, 17
		1000	1996820	129940	138796	80418	Same	Same

[1]Reading times; [2]Near KSP algorithm; [3] The Proposed Neural KSP algorithm; [4]Path Length; [5]Path Costs.

Table E.4. Performance on 100 nodes multigraphs.(Continued II).

K	Run times	Nodes	Edges	Reading[1]	Near KSP[2]	Neural KSP[3]	PL[4]	PC[5]
1	10	100	158	0	1	1.5	1	10
			149	0	0	1.5	3	49
5	10	100	158	0	15	16	1	15, 20, 25, 30
			149	0	15	16	4	64, 65, 66, 67
10	10	100	158	0	16	32	1, 3	35, 40, 45, 49, 50
			149	0	15	31	4	68, 69, 77, 78, 79
15	10	100	158	0	31	42	4	64, 65, 66, 67, 68
			149	0	30	40	4	80, 81, 82, 87, 88
20	10	100	158	0	31	47	4	69, 77, 78, 79, 80
			149	0	30	40	4, 5	89, 90, 90, 90, 91

[1] Reading times; [2]Near KSP algorithm; [3] The Proposed Neural KSP algorithm; [4]Path Length; [5]Path Costs.

Note: Small scale of network. We added 9 edges (10, 15, 20, 25, 30, 35, 40, 45 and 50 from vertex 1 to vertex 99 directly) to make path length 1 as the shortest path.

Table E.5. Performance on 1000 nodes multigraphs.(Continued III)

K	Run times	Nodes	Edges	Reading[1]	Near KSP[2]	Neural KSP[3]	PL[4]	PC[5]
1	10	1000	1996825	132003	132399	1139	1	6
		*	1996820	129940	138239	13572	2	9, 9
			1996812	137295	137779	10858	3	11
			1996804	135695	141251	67279	4	18
			1996804	151158	168468	101868	5	19
			1996804	135695	141251	67279	6	18
			1996800	146295	147092	420359	7	21
5	10	1000	1996825	142322	142249	14882	1, 2	6, 7, 8, 9, 9
		*	1996820	127140	138934	24445	2, 3	10, 10, 10, 10, 10
			1996812	141897	146063	27207	2, 3	12, 12, 12, 12
			1996804	139095	142855	216484	4	18, 19, 19, 19, 19
			1996800	138559	144862	1183217	7	22_7
10	10	1000	1996825	130899	146360	25818	1, 2, 3	10_6
		*	1996820	131407	138544	41496	2, 3	11_{10}
			1996812	1410525	149104	271659	4	20_{10}
			1996804	138495	147716	561180	4	18, 19, 19, 19, 19
			1996800	-	-	-	7	23_{28}

[1]Reading times; [2]Near KSP algorithm; [3] The Proposed Neural KSP algorithm; [4]Path Length; [5]Path Costs.

Table E.6. Performance on 1000 nodes multigraphs.(Continued IV)

K	Run times	Nodes	Edges	Reading[1]	Near KSP[2]	Neural KSP[3]	PL[4]	PC[5]
15	10	1000	1996825	133143	149174	49449	2, 3	11_{10}
		*	1996820	129940	138796	41496	same	same
			1996812	143145	149293	97235	2, 3	14_{13}
			1996804	same	same	same	same	same
			1996800	-	-	-	7	23_{28}
20	10	1000	1996825	133143	149174	49449	same	same
		*	1996820	129940	138796	80418	2, 3	12_{18}
			1996812	same	same	same	same	same
			1996804	141897	153099	638006	4	21_{21}
			1996800	-	-	-	7	23_{28}

[1]Reading times; [2]Near KSP algorithm; [3] The Proposed Neural KSP algorithm; [4]Path Length; [5]Path Costs.

Note: Large scale of network. We added 5 edges (6, 7, 8, 9, and 10 from vertex 1 to vertex 1000 directly) to make path length 1 as the shortest path. We deleted 8 edges (9, 9, 10, 10, 10, 10, 10, 10) to make the path length 3 as the shorest path. We deleted additional 12 edges to make the path length 7 as the shorest path. Based on *, we generate all other graphs by adding or removing some edges.

Appendix C: Supporting Information Tables

Table E.7. Adjacent matrix for carbon atoms graph representing binding loop of D20-E31 from calmodulin (3CLN.pdb)

C ID	D20C	D20CG	K21C	D22C	D22CG	G23C	D24C	D24CG	G25C	T26C	T26CB	I27C	T28C	T28CB	T29C	T29CB	K30C	E31C	E31CD
D20C	1	1	1	1	1	1	0	1	0	1	0	0	0	0	0	0	0	1	1
D20CG	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	1
K21C	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1
D22C	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	1
D22CG	1	1	1	1	1	1	1	1	0	1	0	0	0	1	0	0	0	0	1
G23C	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	1
D24C	0	1	0	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
D24CG	1	1	0	1	1	1	1	1	1	1	1	0	0	1	0	0	0	0	1
G25C	0	1	0	0	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0
T26C	1	1	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	1
T26CB	0	1	0	0	0	0	1	1	1	1	1	1	0	1	0	0	0	0	1
I27C	0	1	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1
T28C	0	0	0	0	0	0	0	0	0	1	0	1	1	1	1	1	1	1	1
T28CB	0	0	0	0	1	0	0	1	0	1	1	1	1	1	1	1	1	1	1
T29C	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
T29CB	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	0
K30C	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
E31C	1	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
E31CD	1	1	1	1	1	1	0	1	0	1	1	1	1	1	1	0	1	1	1

Binding loop of D20-E31 from calmodulin (3CLN.pdb) is used as example to illustrate how the adjacent matrix is constructed for the Figure 2d in the Methods section. In the Table S1, "D20C" represents the mainchain carbon from the 20th Asp while the "D20CG" represents the sidechain gamma carbon from the same residues. "1" represents the distance between two carbon atoms is smaller than 7.5Å; "0" otherwise. In this example, the distance cutoff of C-C is 7.5Å.

Terminology

In coordination chemistry, a ligand is an ion or molecule that binds to a central metal atom to form a coordination complex. (from Wikipedia)

Table E.8. The parameters used in the dataset for MUG^C in X-ray and NMR.

	X-ray	NMR
Cutoff of maximum clique	7.5Å	8.3Å
dist(ca,c1)	(2.5Å ,4.5Å)	(1.74Å ,4.9Å)
dist(ca,c2)	> dist(ca,c1)	> dist(ca,c1)-0.5
angle(ca,c1,c2)	(>90)	(>70)
Center of Mass		
R-Ca (sidechain O)	(<4.3)	(<4.5)
Ca-O (mainchain O)	(<Ca-R)	(<Ca-R)
Clash	Van der waals radius	Van der waals radius
Ca-N	(>2.55)	(>2.55)
Ca-C	(>1.74 [a]/2.7 [b])	(>1.74 [a]/2.7 [b])
Ca-O	(>1.6)	(>1.6)

[a, b]: 1.74 for monodentate and 2.7 for bidentate

Table E.9. Summary of X-ray training dataset.

	X-ray	NMR
Cutoff of maximum clique	7.5Å	8.3Å
dist(ca,c1)	(2.5Å ,4.5Å)	(1.74Å ,4.9Å)
dist(ca,c2)	> dist(ca,c1)	> dist(ca,c1)-0.5
angle(ca,c1,c2)	(>90)	(>70)
Center of Mass		
R-Ca (sidechain O)	(<4.3)	(<4.5)
Ca-O (mainchain O)	(<Ca-R)	(<Ca-R)
Clash	Van der waals radius	Van der waals radius
Ca-N	(>2.55)	(>2.55)
Ca-C	(>1.74 [a]/2.7 [b])	(>1.74 [a]/2.7 [b])
Ca-O	(>1.6)	(>1.6)

[a, b]: 1.74 for monodentate and 2.7 for bidentate

Table E.10. Summary of X-ray testing dataset.

PDB[a]	Res[b]	Protein[c]	Chain[d]	Size[e]
1BJR	2.44	Complex:lactoferrin fragment and proteinase K	E, I	289
1JDA	2.20	Maltotetraose forming exo-amylase	A	418
1FZC	2.30	Fibrin	A,B,C,D,E,F,G,H,I,J	1382
1SBH	1.80	Subtilisin	A	275
1OBR	2.30	Carboxypeptidase T	A	323
1EGZ	2.30	Cellulase	A, B, C	873
1ESL	2.00	E-Selectin	A	157
1AI4	2.35	Penicillin acylase	A, B	763
1ATL	1.80	Atrolysin C	A, B	500
1AX0	1.90	Lectin	A	239
1B9Z	2.10	β-Amylase	A	516
1BF2	2.00	Pseudomonas Isoamylase	A	750
1CE5	1.90	β-Trypsin	A	230
1CLX	1.80	Xylanase	A, B, C, C	1380
1GCG	1.90	Galactose binding protein	A	309
1HYT	1.70	Thermolysin	A	316
1IAG	2.00	Adamalysin (II)	A	201
1IRB	1.70	Carboxylic ester hydrolase	A	123
1JS4	2.00	Endoexocellulase E4	A, B	1210
1KBC	1.80	Neutrophil collagenase	A, B	328
1KIT	2.30	Hydrolase	A	757
1KVX	1.90	Carboxylic ester hydrolase	A	123
1MMQ	1.90	Matrilysin	A	165
1NBC	1.75	Cellusomal scaffolding Protein A	A, B	310
1OAC	2.00	Amine oxidase	A, B	1443
1OIL	2.10	Lipase	A, B	640
1SBF	2.43	Soybean agglutinin	A	234
1SRA	2.00	Calcium biniding protein	A	151
1TCM	2.20	Cyclodextrin glycosyl Transferase	A, B	1372
1TN3	2.00	Tetranectin	A	137
2FIB	2.01	Fibrogen	A, B	254
2TEP	2.50	Peanut lectin	A, B, C, E	928
4LIP	1.75	Lipase	D, E	638
1C9M	1.67	Bacillus lentus subtilisin	A	269
2SCP	2.00	Sarcoplasmic Ca(2+)-binding protein (SCP)	A, B	174
1TVG	1.6	HSPC034	A	153
2GGM	2.35	Human centrin 2 xeroderma pigmentosum group C protein complex	A,B,C,D	172
3FIA	1.45	Human intersectin-1 protein	A	121
1K9K	1.76	calcium bound human S100A6	A, B	90
1DAN	2.00	Complex of active site inhibited human blood coagulation factor via with human recombinant soluble tissue factor	L, H, T, U, C	152, 254, 80, 121, 4
1MHO	2.00	S100B from bovine brain	A	88
1EDH	2.00	E-cadherin domains 1 and 2 in complex with calcium	A, B	226
2EGD	1.8	human S100A13	A, B	98

[a] PDB code. [b] PDB resolution. [c] Protein name. [d] Chain number. [e] Number of residues.

Table E.11. Prediction results on the X-ray training dataset.

PDB ID	Ca# [a]	Documented Ligands	Predicted Ligands	R[b]
1ALA	2505	M28, G32, T37, E72	M28, G32, E72	3/4
	2506	I100, G102, G104, E144	I100, G102, G104, E144	4/4
	2507	M259, G261, G263, D303	M259, G261, G263, D303	4/4
1ALV	3425	A107, D110, E112, E117	A107, D110, E112, E117	4/4
	3426	D150, D152, T154, K156, E161	D150, D152, T154, K156, E161	5/5
	3427	D180, D182, S184, T186, E191	D180, D182, S184, T186, E191	5/5
	3428	D135, D223, D225, N226	D135, D223, D225, N226	4/4
1AUI	4393	D30, D32, S34, S36, E41	D30, D32, S34, S36, E41	5/5
	4394	D62, D64, N66, E68, E73	D62, D64, N66, E68, E73	5/5
	4395	D99, D101, D103, Y105, E110	D99, D101, D103, Y105, E110	5/5
	4396	D140, D142, D144, R146, E151	D140, D142, D144, R146, E151	4/4
1AVS	1266	D30, D32, D36, E41	D30, D32, D36, E41	4/4
	1267	D66, D68, S70, T72, E77	D66, D68, S70, T72, E77	3/3
	1268	D30, D32, D36, E41	D30, D32, D36, E41	3/3
	1269	D66, D68, S70, T72, E77	D66, D68, S70, T72, E77	4/4
1B9O	1032	K79, D82, D84, D87	K79, D82, D84, D87	4/4
1EXR	1469	D20, D22, D24, T26, E31	D20, D22, D24, T26, E31	5/5
	1470	D56, D58, N60, T62, E67	D56, D58, N60, T62, E67	5/5
	1471	N129, D131, D133, H135, E140	N129, D131, D133, H135, E140	5/5
	1472	E47	-	
	1473	D93, D95, N97, L99, E104	D93, D95, N97, L99, E104	5/5
1FJ3	2461	D138, E177, D185, E187, E190	D138, E177, D185, E187, E190	4/4
	2462	E177, D185, E190	E177, D185, E190	3/3
	2463	D57, D59, G61	-	0/3
	2464	Y193, T194, I197, D200	Y193, T194, I197, D200	4/4
1GLG	2362	D134, N136, D138, Q140, Q142, E205	D134, N136, D138, Q140, Q142, E205	6/6
1K96	715	S20, E23, D25, T28, E33	S20, E23, D25, T28, E33	5/5
	716	D61, N65, D65, E67, E72	D61, N65, D65, E67, E72	5/5
1NLS	1875	D10, Y12, N14, D19	D10, Y12, N14, D19	4/4
1PSH	2788	Y27, G29, G31, D48	Y27, G29, G31, D48	4/4
	2789	Y27, G29, G31, D48	Y27, G29, G31, D48	4/4
	2790	Y27, G29, G31, D48	Y27, G29, G31, D48	4/4
1SCD	1922	D41, L75, N77, T79, V81	D41, L75, N77, T79, V81	5/5
	1923	A169, Y171, V174	-	4/4
1SNC	1184	D21, D40, T41	D21, T41	2/3
1THM	2005	D5, D47, V82, N85, T87, I89	D5, D47, V82, N85, T87, I89	4/4
	2006	D57, D62, T64, Q66	D57, D62, Q66	3/4
2PRK	2019	P175, V177, D200	P175, V177, D200	3/3
	2020	T16, D260	-	2/2
3EST	1824	E70, N72, Q75, N77, E80	E70, N72, Q75, N77, E80	5/5
4ICB	641	A14, E17, D19, Q22, E27	A14, E17, D19, Q22, E27	5/5
	642	D54, N56, D58, E60, E65	D54, N56, D58, E60, E65	5/5
5PAL	843	D90, D92, D94, K96, E101	D90, D92, D94, K96, E101	4/4
	844	D51, D53, S55, F57, E59, E62	D51, D53, S55, F57, E59, E62	6/6

[a]: metal identification number in PDB file. [b]: the correctly predicted ligands over documented ligands

Table E.12. Prediction results on the X-ray testing dataset.

PDB ID	Ca# [a]	Documented Ligands	Predicted Ligands	R[b]
1BJR	2090	P175, V177, D200	P175, V177, D200	3/3
	2089	R12, S15, N257, A273	-	0/4
1JDA	3299	N116, D151, D154, D162, G197	N116, D151, D154, D162, G197	5/5
	3300	D1, Q2, H13, D16, E17	D1, Q2, H13, D16, E17	5/5
1FZC	11170	D381, D383, W385	D381, D383, W385	3/3
	11171	D318, D320, F322, G324	D318, D320, F322, G324	4/4
	11172	D381, D383, W385	D381, D383, W385	3/3
	11173	D318, D320, F322, G324	D318, D320, F322, G324	4/4
1SBH	1942	Q2, D41, L75, N77, V81	Q2, D41, L75, N77, V81	5/5
	1943	A169, Y171, V174	-	0/3
1OBR	2584	D56, E57, E61, E104	D56, E57, E61, E104	4/4
	2585	S50, D51,E57,E59	S50, D51,E57,E59	4/4
	2586	D51,E59, N101	D51,E59, N101	3/3
	2587	S7, Y9, E14	S7, Y9, E14	3/3
1EGZ	6810	G121, D158, D160, N161	G121, D158, D160, N161	4/4
	6811	G121, D158, D160, N161	G121, D158, D160, N161	4/4
	6812	G121, D158, D160, N161	G121, D158, D160, N161	4/4
1ESL	1267	E80, N82, N105, D106	E80, N82, N105, D106	4/4
	1268	E33, E36	-	0/2
	1270	Q20, Y23	-	0/2
1AI4	6074	E152, D73, V75, D76, P205, D252	E152, D73, V75, D76, P205, D252	6/6
1ATL	3260	E9, D93, C197, N200	E9, D93, C197, N200	4/4
	3262	E9, D93, C197, N200	E9, D93, C197, N200	4/4
1AX0	1996	D129, F131, N133, D136	D129, F131, N133, D136	4/4
1B9Z	4310	D56, D60, Q61, E141, E144	D56, D60, Q61, E141, E144	5/5
1BF2	5737	D128, E229, T230, N232, D259	D128, E229, T230, N232, D259	5/5
1CE5	1631	E70, N72, V75, E80	E70, N72, V75, E80	4/4
1CLX	10801	N253, D256, N258, N261, D262	N253, D256, N258, N261, D262	5/5
	10802	N253, D256, N258, N261, D262	N253, D256, N258, N261, D262	5/5
	10803	N253, D256, N258, N261, D262	N253, D256, N258, N261, D262	5/5
	10804	N253, D256, N258, N261, D262	N253, D256, N258, N261, D262	5/5
1GCG	2895	D134, N136, D138, K140, Q142, E205	D134, N136, D138, K140, Q142, E205	6/6
1HYT	2440	D138, E177, D185, E187, E190	D138, E177, D185, E187, E190	5/5
	2441	E177, N183, D185, E190	E177, N183, D185, E190	4/4
	2442	D57, D59, N61	-	0/3
	2443	Y193, T194, I197, D200	Y193, T194, I197, D200	4/4
1IAG	1623	E9, D93, C197, N200	E9, D93, C197, N200	4/4
1IRB	951	Y28, G30, G32, D49	Y28, G30, G32, D49	4/4
1JS4	9586	S210, G211, D214, E215, D261	S210, G211, D214, E215, D261	5/5
	9587	T504, D506, D571, N574, D575	T504, D506, D571, N574, D575	5/5
	9588	S210, G211, D214, E215, D261,	S210, G211, D214, E215, D261,	5/5
	9589	T504, D506, D571, N574, D 575	T504, D506, D571, N574, D 575	5/5
1KBC	2591	D137, G169, G171, D173,	D137, G169, G171, D173,	4/4
	2592	D154, G155, N157, I159, D177, E180	D154, G155, N157, I159, D177, E180	6/6
	2595	D137, G169, G171, D173	D137, G169, G171, D173	4/4
	2596	D154, G155, N157, I159, D177, E180	D154, G155, N157, I159, D177, E180	6/6
1KIT	5861	A253, N256, D289, T313	A253, N256, D289, T313	4/4
	5862	D621, D682, A683	D621, D682, A683	3/3
1KVX	956	Y28, G30, G32, D49	Y28, G30, G32, D49	4/4
1MMQ	1272	D175, G176, G178, T180, D198, E201	D175, G176, G178, T180, D198, E201	6/6
	1273	D158, G190, G192, D194	D158, G190, G192, D194	4/4
1NBC	2437	T44, D46, T122, N125, D126	T44, D46, T122, N125, D126	5/5

	2438	T44, D46, T122, N125, D126	T44, D46, T122, N125, D126	5/5
1OAC	11388	D533, L534, D535, D678, A679	D533, L534, D535, D678, A679	5/5
	11389	E573, Y667, D670, E672	E573, Y667, D670, E672	4/4
	11391	D533, L534, D535, D678, A679	D533, L534, D535, D678, A679	5/5
	11392	E573, Y667, D670, E672	E573, Y667, D670, E672	4/4
1OIL	4677	D242, D288, Q292, V296	D242, D288, Q292, V296	4/4
	4678	D242, D288, Q292, V296	D242, D288, Q292, V296	4/4
1SBF	1735	D126, F128, N130, D133	D126, F128, N130, D133	4/4
1SRA	1264	D222, P225, D227, Y229, E234	D222, P225, D227, Y229, E234	4/4
	1265	D257, D259, D261, Y263, E268	D257, D259, D261, Y263, E268	5/5
	1266	P241, I243, E246	-	0/3
1TCM	10513	D27, N29, N32, N33, G51, D53	D27, N29, N32, N33, G51, D53	4/4
	10514	N139, I190, D199, H233	N139, I190, D199, H233	4/4
	10515	D27, N29, N32, N33, G51, D53	D27, N29, N32, N33, G51, D53	5/5
	10616	N139, I190, D199, H233	N139, I190, D199, H233	4/4
1TN3	1068	D116, E120, G147, E150, N151,	D116, E120, G147, E150, N151,	5/5
	1069	Q143, D145, E150, D165	Q143, D145, E150, D165	4/4
2FIB	2036	D318, D320, F322, G324	D318, D320, F322, G324	4/4
2TEP	7081	D123, Y125, N127, D132	D123, Y125, N127, D132	4/4
	7083	D123, Y125, N127, D132	D123, Y125, N127, D132	4/4
	7085	D123, Y125, N127, D132	D123, Y125, N127, D132	4/4
	7087	D123, Y125, N127, D132	D123, Y125, N127, D132	4/4
4LIP	4669	D242, D288, Q292, V296	D242, D288, Q292, V296	4/4
	4670	D242, D288, Q292, V296	D242, D288, Q292, V296	4/4
1C9M	1897	G2, D41, L75, N77, I79, V81	G2, D41, L75, N77, I79, V81	6/6
	1898	A169, Y171, A174, G195 D197	A169, A174, D197	3/5
2SCP	2739	D16, D18, D20, A22, D27	D16, D18, D20, A22, D27	5/5
	2740	D104,N106, D108, N110, E115	D104,N106, D108, N110, E115	5/5
	2741	D138, N140, D142, L144, E149	D138, N140, D142, L144, E149	5/5
	2742	D16, D18, D20, A22, D27	D16, D18, D20, A22, D27	5/5
	2743	D104,N106, D108, N110, E115	D104,N106, D108, N110, E115	5/5
	2744	D138, N140, D142, L144, E149	D138, N140, D142, L144, E149	5/5
1TVG	1086	N29, D32, N34, T37, H130	N29, D32, N34, T37, H130	5/5
2GGM	2682	D114, D116, T118, K120, N125	D114, D116, T118, K120, N125	5/5
	2683	D150, D152, D154, E156, E161	D150, D152, D154, E156, E161	5/5
	2684	D114, D116, T118, K120, N125	D114, D116, T118, K120, N125	5/5
	2685	D150, D152, D154, E156, E161	D150, D152, D154, E156, E161	5/5
3FIA	780	D66, N68, D70, R72, E77	D66, N68, D70, R72, E77	5/5
1K9K	1417	S20, E23, D25, T28, E33	S20, E23, D25, T28, E33	5/5
	1418	D61, D63, D65, E67, E72	D61, D63, D65, E67, E72	5/5
	1423	S20, E23, D25, T28, E33	S20, E23, D25, T28, E33	5/5
	1424	D61, D63, D65, E67, E72	D61, D63, D65, E67, E72	5/5
1DAN*	4723	D46, G47, N49, D63, N64	D46, G47, N49, D63, N64	5/5
	4724	CGU	-	
	4725	CGU	-	
	4726	CGU	-	
	4727	CGU	-	
	4728	CGU	-	
	4729	CGU	-	
	4730	CGU	-	
	4731	D70, D72, E75, E80	D70, D72, E75, E80	4/4
1MHO	713	S18, E21, D23, K26, E31	S18, E21, D23, K26, E31	5/5
	714	D61, D63, D65, E67, E72	D61, D63, D65, E67, E72	5/5
1EDH	3230	E11, E69, D100, Q101, D103	E11, E69, D100, Q101, D103	5/5
	3231	E11, D67, E69, D103	E11, D67, E69, D103	4/4

	3232	E11, N12, D67, E69, D103	E11, N12, D67, E69, D103	5/5
	3234	E11, E69, D100, Q101, D103	E11, E69, D100, Q101, D103	5/5
	3235	E11, D67, E69, D103	E11, D67, E69, D103	4/4
	3236	E11, N12, D67, E69, D103	E11, N12, D67, E69, D103	5/5
2EGD	1386	A24, E27, R29, S32, E37	A24, E27, R29, S32, E37	5/5
	1387	D64, N66, D68, E70, E75	D64, N66, D68, E70, E75	5/5
	1388	A24, E27, R29, S32, E37	A24, E27, R29, S32, E37	5/5
	1389	D64, N66, D68, E70, E75	D64, N66, D68, E70, E75	5/5

[a]: metal identification number in PDB file. [b]: the correctly predicted ligands over documented ligands

Table E.13. Prediction results on the NMR training dataset.

Protein	ID	M[a]	L[b]	Real Ligands	Predicted Ligands	R[c]
Epidermal growth factor receptor pathway substrate 15	1C07	20	95	D28, D30, D32, F34, E39	D28, D30, D32, F34, S36, E39	5/5
Calcium-binding protein NCS-1	1FPW	20	190	D73, D75, N77, F79, E84	D73, D75, N77, F79, E84	5/5
				D109, N111, D113, Y115, E120	D109, N111, D113, Y115, E120	5/5
				D157, N159, D161, Y163, E168	D157, N159, D161, Y163, E168	5/5
Troponin C	1TNW	23	162	D30, D32, G34, D36, E41	D30, G33, G34, D36, E41	5/5
				D66, D68, S70, T72, E77	D66, T72, D74, E77	3/5
				D106, N108, D110, F112, E117	D106, N108, D114, E117	4/5
				D142, N144, D146, R148, E153	D142, D146, R148, E150	4/5
Calmodulin	2BBM	1	148	D20, D22, D24, T26, E31	D24, T26, T28, E31	4/5
				D56, D58, N60, T62, E67	D56, D58, T62, E67	4/5
				D93, D95, N97, Y99, E104	D93, D95, N97, Y99, E104	5/5
				N129, D131, D133, D135, E140	N129, D131, D133, E140	4/5
Calbindin D9K	2BCB	32	75	A14, E17, D19, Q22, E27	A14, E17, D19, Q22, E27	5/5
				D54, N56, D58, E60, E65	D56, N58, E60, E65	4/5
Parvalbumin	2PAS	9	109	D51, D53, S55, F57, E62	D51, D53, S55, F57, E62	5/5
				D90, D92, D94, K96, E101	D90, D92, D94, K96, E101	5/5

[a]: number of structures in the ensembles. [b]: number of the residues in proteins. [c]: number of correctly predicted ligands over number of documented ligands.

Table E.14. Prediction results on the NMR testing dataset.

Protein	X-ray	Chain	NMR	Chain	Identity	M[a]	L[b]	Real Ligands	Predicted Ligands	R[c]
SERINE PROTEASE PB92	1C9M	A	1AH2	A	98%	18	269	G2, D40, L75, N77, I79, V81	S3, D40, L75, G78, I79	4/6
								A163, Y165, A168, G189, D191	A163, R164, Y165, A168, D191	4/5
Nereis diversicolor sarcoplasmic calcium-binding protein (NSCP)	2SCP	B	1Q80	A	100%	17	174	D16, D18, D20, A22, D27	D16, D18, A22, D27	4/5
								D104, N106, D108, N110, E115	D104, N106, D108, N110, E115	5/5
								D138, N140, D142, L144, E149	D138, N140, L144, E149	4/5
human protein HSPCO34	1TVG	A	1XPW	A	100%	20	153	N29, D32, N34, T37, H129	N29, D32, N34, T37, H129	5/5
the human centrin 2 in complex with a 17 residues peptide (P1-XPC) from xeroderma pigmentosum group C protein	2GGM	A	2A4J	A	100%	20	96	D114, D116, T118, K120, N125	D114, T118, K120	3/5
								D150, D152, D154, E156, E161	-	0/5
human intersectin-1 protein**	3FIA	A	2KHN	A	99%	20	121	D76, N78, D80, R82, E87	D76, N78, R82, E87	4/5
Staphylococcal nuclease	1SNC	A	1JOQ	A	99%	30	149	D21, D40, T41, THP151	D21, D40, T41	3/3
Human blood coagulation FVII	1DAN	L	1F7E	A	100%	20	46	D46, G47, N49, D63, N64	D46, N49, D63	3/5
S100B	1MHO	A	1UWO	A, B	96%	20	91	S18, E21, D23, K26, E31	S18, E21, D23, K26	4/5
								D61, D63, D65, E67, E72	D61, D63, E67, E72	4/5
								S18, E21, D23, K26, E31	E21, D23, K26, E31	4/5
								D61, D63, D65, E67, E72	D61, D63, E67, E72	4/5
Epithelial cadherin	1EDH	A	1SUH	A	100%	20	146	E11, E69, D100, Q101, D103	E11, E69, Q101, D103	4/5
								E11, D67, E69, D103	E11, N12, D67, E69	3/4
S100A13	2EGD	A	2K8M	B,C	100%	20	98	A24, E27, R29, S32, E37	A24, G28, R29, S32, E37	5/5
								D64, N66, D68, E70, E75	D64, D68, E70, E75	4/5
								A24, E27, R29, S32, E37	A24, G28, R29, S32, E37	5/5
								D64, N66, D68, E70, E75	D64, D68, E70, E75	4/5

[a]: number of structures in the ensembles. [b]: number of the residues in proteins. [c]: number of correctly predicted ligands over number of documented ligand.

Table E.15. Testing on Mg^{2+}-binding proteins (X-ray structures).

PDB[a]	Res[b]	Protein[c]	Chain[d]	Mg#[e]	Mis-classified[f]
1CMC	1.8	Met repressor (metj)	A,B	1693	No
				1721	No
1EBH	2.2	Enolase	A,B	6631	No
				6633	No
1XLB	2.3	D-xylose isomerase	A	6055	No
1CHN	1.6	Chey	A	968	No
1EO3	1.9	Restriction enzyme ecoRV	A,B	4229	No
				4230	No
				4235	No
				4236	No
1VSD	1.9	Integrase	A	1129	No
1MUS	2.5	Adenine phosphoribosyltransferase	A,B	4435	No
				4436	No
1QB7	1.9	Xanthine-guanine phosphoribosyltransferase	A	1857	No
1EYJ	2.1	Fructose-1,6-bisphosphatase	A,B	5011	No
				5056	No
2UAG	1.7	D-glutamate ligase	A	3247	No
			A	3248	No
3PRN	1.9	Porin	A	2203	No
1HBN	1.1	Methyl-coenzyme m reductase	B	19432	No
			D	19557	No
			E	19570	No
2TCT	2.1	Tetracycline repressor	A	1575	No
1LUC	1.5	Bacterial luciferase	A	5096	No
			B	5106	No
1KQP	1.0	Nh(3)-dependent nad(+) synthetase	B	8739	No
1NG1	2.0	Signal sequence recognition protein FFH	A	2279	No
1BL3	2.0	Integrase	B	3445	No
1NUL	1.8	Xanthine-guanine phosphoribosyltransferase	A	2159	No
2UAG	1.7	D-glutamate ligase	A	3247	No
1IDE	2.5	Isocitrate dehydrogenase	A	3881	No
1JIV	2.0	DNA beta-glucosyltransferase	A	2871	No
			A	2872	No
1DOZ	1.8	Ferrochelatase	A	2490	No
1G8T	1.1	Nuclease sm2 isoform	A	3878	No
1A73	1.8	Intron 3 (i-ppo) encoded endonuclease	A	3353	No
1FWK	2.1	Homoserine kinase	D	9162	No
1JKK	2.4	Death-associated protein kinase	A	2247	No
1LDF	2.1	Glycerol uptake facilitator	A	1936	No
1OBW	2.1	Inorganic pyrophosphatase	A,B,C	4141	Other
				4142	No
				4143	No
				4144	Other
				4145	No
				4146	No
				4147	Other
1KCZ	1.9	Beta-methylaspartase	A,B	6431	Other
				6440	Other
1RK2	1.8	Ribokinase	A,B,C,D	8992	Other
				9035	Other
				9078	Other
				9121	Other

[a] PDB code. [b] PDB resolution. [c] Protein name. [d] Chain number. [e] metal identification number in PDB file. [f] mistakenly classified Mg^{2+}-binding site as Ca^{2+}-binding site.

Table E.16. Testing on Zn^{2+}-binding proteins (X-ray structures).

PDB[a]	Res[b]	Protein[c]	Chain[d]	Zn#[e]	Mis-classified[f]
1FWZ	2.3	Diphtheria toxin repressor	A	1588	No
1CY5	1.3	Apoptotic protease activating factor 1	A	749	No
				750	No
				752	No
1WEJ	1.8	E8 antibody	A	4170	No
1E67	2.1	Azurin	A,B,C,D	3901	No
				3906	No
				3907	No
				3908	No
1GS8	1.9	Nitrite reductase	A	2590	No
				2591	No
1F5F	1.7	Sex hormone-binding globulin	A	1369	No
				1370	No
1GI4	1.3	Beta-trypsin	A	3369	No
2CBA	1.5	Carbonic anhydrase	A	2081	No
1F3Z	1.9	Glucose-specific phosphocarrier	A	1109	No
1C8Y	2.0	Endo-beta-n-acetyl-glucosaminidase H	A	2015	No
4ENL	1.9	Enolase	A	3291	No
1I6N	1.8	Loli protein	A	2231	No
1IM5	1.6	Pyrazinamidase	A	1439	No
1VSH	1.9	Integrase	A	1129	No
				1130	No
				1131	No
1NOY	2.2	DNA polymerase	A	5953	No
2CTB	1.5	Carboxypeptidase A	A	2452	No
1TOA	1.8	Periplasmic binding protein	A,B	4295	No
				4302	No
1A2P	1.5	Barnase	A,B,C	2628	No
				2629	No
				2630	No
1EU3	1.6	Superantigen Smez-2	A,B	3419	No
				3436	No
1EWC	1.9	Enterotoxin H	A	1733	No
1EU4	2.5	Superantigen spe-H	A	1668	No
1AST	1.8	Astacin	A	1593	No
1ZFP	1.8	Growth factor receptor binding protein	E	870	No
1K4P	1.0	3,4-dihydroxy-2-butanone 4-phosphate synthase	A	1643	No
1K9Z	1.5	Halotolerance protein HAL2	A	2731	No
				2732	No
				2735	No
				2733	No
1CNQ	2.2	Fructose-1,6-bisphosphatase	A	2572	No
1KSP	2.3	Klenow fragment	A	4817	No
3IVE	2.0	Immunoglobulin	A	893	No
1M5E	1.4	Glutamate receptor 2	A	6150	No
1L7O	2.2	Phosphoserine phosphatase	B	3208	No
8RNT	1.8	Ribonuclease T1	A	779	No
1XLL	2.5	D-xylose isomerase	A,B	6057	Other
				6058	No
				6059	Other
				6060	No

[a] PDB code. [b] PDB resolution. [c] Protein name. [d] Chain number. [e] metal identification number in PDB file. [f] mistakenly classified Zn^{2+}-binding site as Ca^{2+}-binding site.

Table E.17. Testing on Pb^{2+}-binding proteins (X-ray structures).

PDB[a]	Res[b]	Protein[c]	Chain[d]	Pb#[e]	Mis-classified[f]
1E9N	2.20	DNA-lyase	A,B	4339	No
				4340	No
				4341	No
				4342	No
1FJR	2.3	Methuselah ectodomain	A,B	3120	No
				3121	No
				3169	No
				3170	No
1NA0	1.60	Designed protein CTPR3	A,B	1969	No
				1970	No
				1975	No
				1976	No
				1977	No
1QNV	2.5	5-aminolaevulinic acid dehydratase	A	2548	No
				2549	No
1SN8	2.00	Ribonuclease E	A,B	1330	No
				1331	No
1SYY	1.7	Ribonucleoside-diphosphate reductase	A	2617	No
1XXA	2.20	Arginine repressor	A~F	3245	No
				3246	No
				3259	No
				3284	No
1ZHY	1.6	KES1 protein	A	3516	No
				3517	No
2CH7	2.5	Methyl-accepting chemotaxis protein	A,B	4633	No
				4634	No
2FJ9	1.6	Acyl-CoA-Binding protein	A	710	No
2FP1	1.55	Chorismate mutase	A,B	2705	No
				2706	No
2OQ1	1.9	Tyrosine-protein kinase	A,B	2200	No
2QD5	2.3	Ferrochelatase	A,B	5845	No
				5846	No
				5965	No
				5966	No
2QKL	2.3	Hydrolase	A,B	1772	No
3EC8	2.6	FLJ10324	A	1083	No
				1084	No
3FHH	2.6	Outer membrane heme receptor ShuA	A	4778	No
				4779	No
				4780	No
				4781	No
1HD7	1.95	DNA-lyase	A	2072	Other
2G0A	2.35	Cytosolic 5'-nucleotidase III	A,B	4655	Other
				4671	Other
2O3C	2.30	APEX nuclease 1	A,B	6648	Other
				6649	No
				6650	Other

[a] PDB code. [b] PDB resolution. [c] Protein name. [d] Chain number. [e] metal identification number in PDB file. [f] mistakenly classified Pb^{2+}-binding site as Ca^{2+}-binding site.

Table E.18. Testing on a negative control dataset (X-ray structures).

PDB[a]	Protein[b]	FN[c]
1DTS	Dethiobiotin synthase	0
1L68	Lysozyme	0
1PTX	Scorpion toxin II	0
1VCC	DNA topoisomerase I	0
1WBA	Winged bean albumin 1	0
2ENG	Endoglucanase V	0
2YLE	Human spir-1 kind fsi domain in complex with the fsi peptide	0
3O5F	Fk1 domain of FKBP51	0
3OQ7	Multidrug-Resistant Clinical Isolate 769 HIV-1 Protease Variants	0
1IQR	DNA photolyase	0
1IUG	Aspartate aminotransferase which belongs to subgroup IV	0
1IZ0	Quinone Oxidoreductase	0
1J27	Hypothetical protein, TT1725	0
1J3M	Conserved hypothetical protein TT1751	0
1JJF	Feruloyl esterase domain of the cellulosomal xylanase z of clostridium thermocellum	0
1TCA	Lipase	2
2OLB	Oligo-peptide binding protein	4
1TTB	Transthyretin	2
1BDM	Malate Dehydrogenase	5
1K4N	Protein EC4020	3
2AQJ	Tryptophan 7-halogenase (PrnA)	5
1ISO	Isocitrate dehydrogenase	2
1SGV	Trna psi55 pseudouridine synthase (trub)	3

[a] PDB code. [b] Protein name. [c] Number of False Negative predictions.

Appendix D: Formula in Resource Allocation Model

1. Three other options calculations:

(1). Single screening and treating for GC only.

$$Cur_{ijkl} = P_g(i) \cdot Sn_g(j) \cdot E_g(l) \cdot P_r \tag{E.8}$$

Similar to (3.2), we have

$$\begin{aligned} Cost_{ijkl} &= Bc_g(j) + Vc + [P_g(i) \cdot Sn_g(j) \\ &\quad +(1 - P_g(i)) \cdot (1 - Sp_g(j))] \cdot (Dc_g(l) + Tc) \cdot P_r \end{aligned} \tag{E.9}$$

(2). Sequence screening tests that tested for CT and then GC if a positive CT result.

$$\begin{aligned} Cur_{ijkl} &= Cur_{ijkl} \; in \; (3.1) + P_t(i) \cdot Sn_t(j) \cdot P_{g|t}(i) \cdot Sn_g(j) \cdot E_g(l) \cdot P_r \\ &\quad +(1 - P_t(i)) \cdot (1 - Sp_t(j)) \cdot P_{g|\bar{t}}(i) \cdot Sn_g(j) \cdot E_g(l) \cdot P_r \end{aligned} \tag{E.10}$$

- $P_t(i) \cdot Sn_t(j)$ gives the rate over the population of group i tested positively by using the jth CT screening test;
- $P_t(i) \cdot Sn_t(j) \cdot P_{g|t}(i) \cdot Sn_g(j) \cdot E_g(l) \cdot P_r$ gives the rate of the cured number of the GC patients infected by both of CT and GC and tested both positively.
- $(1 - P_t(i)) \cdot (1 - Sp_t(j))$ is the rate of those who are not infected with CT but who test positive. So $(1 - P_t(i)) \cdot (1 - Sp_t(j)) \cdot P_{g|\bar{t}}(i) \cdot Sn_g(j)$ gives the percentage of patients who are in a "stroke of good luck" case. In this case, patients only have GC and were accidentally diagnosed as having CT with the jth test firstly and were caught with the second GC test finally, which in turn shows that $(1 - P_t(i)) \cdot (1 - Sp_t(j)) \cdot P_{g|\bar{t}}(i) \cdot$

$Sn_g(j) \cdot E_g(l) \cdot P_r$ is the percentage of the cured number of GC patients in the case of the "stroke of good luck".

$$\begin{aligned} Cost_{ijkl} &= Cost_{ijkl} \; in \; (3.2) + [P_t(i) \cdot Sn_t(j) + (1 - P_t(i)) \cdot (1 - Sp_t(j))] \\ &\cdot Bc_g(j) + [P_t(i) \cdot Sn_t(j) \cdot P_{g|t}(i) + (1 - P_t(i)) \cdot (1 - Sp_t(j)) \\ &\cdot P_{g|\bar{t}}(i)] \cdot Sn_g(j) \cdot (Dc_g(l) + Tc) \cdot P_r \end{aligned} \tag{E.11}$$

- $[P_t(i) \cdot Sn_t(j) + (1 - P_t(i)) \cdot (1 - Sp_t(j))] \cdot Bc_g(j)$ represents the rate over the population of GC testing costs for the patients testing positive for CT.
- $[P_t(i) \cdot Sn_t(j) \cdot P_{g|t}(i) + (1 - P_t(i)) \cdot (1 - Sp_t(j)) \cdot P_{g|\bar{t}}(i)]$ represents the rate over the population of those testing positive on CT and then positive on GC.
- $[P_t(i) \cdot Sn_t(j) \cdot P_{g|t}(i) + (1 - P_t(i)) \cdot (1 - Sp_t(j)) \cdot P_{g|\bar{t}}(i)] \cdot Sn_g(j) \cdot (Dc_g(l) + Tc) \cdot P_r$ is the rate over the population of treatment costs for curing these patients with a positive GC test.

(3). Combo screening test for both CT and GC.

$$Cur_{ijkl} = Cur_{ijkl} \; in \; (3.1) + Cur_{ijkl} \; in \; (E.8) \tag{E.12}$$

$Cost_{ijkl} \; in \; (3.2) + Cost_{ijkl} \; in \; (E.9)$ will give the basic count except the visit costs for the screening test is counted twice and the treatment costs for those testing positive on both CT and GC are counted twice. Subtracting them, we obtain the following.

$$\begin{aligned} Cost_{ijkl} &= Cost_{ijkl} \; in \; (3.2) + Cost_{ijkl} \; in \; (E.9) \\ &- V_c - P_t(i) \cdot P_{g|t}(i) \cdot Tc \cdot P_r \end{aligned} \tag{E.13}$$

Note: For a combo assay, there is an additional cost which is calculated slightly different from (E.13). Thus, we added the extra costs to the previous formula and it is

$$\begin{aligned} Cost_{ijkl} &= Cost_{ijkl} \; in \; (E.13) + [P_t(i) \cdot Sn_t(j) \\ &\quad + P_g(i) \cdot Sn_g(j) - P_t(i) \cdot P_{g|t}(i)] \cdot Ac(j) \end{aligned} \tag{E.14}$$

2. Useful formula:

Let $P_{g|t}(i)$ be the conditional probability of a CT patient in group i having GC and $P_{t|g}(i)$ be the conditional probability of a GC patient in group i having CT. From Bayes' law, we obtain

$$P_{t|g}(i) = \frac{P_t(i) \cdot P_{g|t}(i)}{P_g(i)}. \tag{E.15}$$

Therefore, if $P_{g|t}(i)$ is given, $P_{t|g}(i)$ can be calculated by the above equation.

Let $P_{g|\bar{t}}(i)$ be the conditional probability of GC infection in a patient without CT infection. The following equations are useful while calculating costs.

$$P_{g|\bar{t}}(i) = \frac{P_g(i) - P_t(i) \cdot P_{g|t}(i)}{1 - P_t(i)} \tag{E.16}$$

Because $P_g(i) = P_t(i) \cdot P_{g|t} + (1 - P_t(i)) \cdot P_{g|\bar{t}}(i)$.

Similarly, we need $P_{t|\bar{g}}(i)$ in the cost estimates. $P_{t|\bar{g}}(i)$ can be presented as the following:

$$P_{t|\bar{g}}(i) = \frac{P_t(i) - P_g(i) \cdot P_{t|g}(i)}{1 - P_g(i)} \tag{E.17}$$

3. The Horowitz-Sahni branch-and-bound method.

In general, this algorithm has two moves. The descriptions and pseudocodes are published [112]: "*A forward move consists* of inserting the largest possible set of new consecutive items into the current solutions. *A backtracking move* consists of removing the last inserted item from the current solution. Whenever a forward move is exhausted, the upper bound corresponding to the current solutions is computed and compared with the best solution

so far, in order to check whether further forward moves could lead to a better one; if so, a new forward move is performed, otherwise, a backtracking follows." (p.30-31). In this algorithm, items initially are sorted according to decreasing rates of the values per unit weight. The pseudocodes we used [112] are: $(\hat{x}_j)$=current solution; $\hat{z}$=current solution value $(= \sum_{j=1}^{n} p_j \hat{x}_j)$; $\hat{c}$=current residual capacity $(= c - \sum_{j=1}^{n} w_j \hat{x}_j)$; (x_j)=best solution so far; z= value of the best solution so far $(= \sum_{j=1}^{n} p_j x_j)$.

input: n, c, p_j, w_j; **output**: z, x_j;

begin:

1:[initialize]

z=0; $\hat{z} = 0$; $\hat{c} = c$; $p_{n+1} = 0$; $w_{n+1} = +\infty$; $j = 1$.

2:[compute upper bound U_1]

find $r = \min\{i : \sum_{k=j}^{i} w_k > \hat{c}\}$; $u = \sum_{k=j}^{r-1} p_k + \left\lfloor (\hat{c} - \sum_{k=j}^{r-1} w_k) p_r / w_r) \right\rfloor$;

if $z \geq \hat{z} + u$ then go to 5;

3:[perform a forward step]

while $w_j \leq \hat{c}$ do $\hat{c} = \hat{c} - w_j$; $\hat{z} = \hat{z} + p_j$; $\hat{x}_j = 1$; $j = j + 1$;

if $j \leq n$ then $\hat{x}_j = 0$; $j = j + 1$;

if $j < n$ then go to 2; if $j = n$ then go to 3;

4:[update the best solution so far]

if $\hat{z} > z$ then $z = \hat{z}$; for $k = 1$ to n do $x_k = \hat{x}_k$;

$j = n$;

if $\hat{x}_n = 1$ then $\hat{c} = \hat{c} + w_n$; $\hat{z} = \hat{z} - p_n$; $\hat{x}_n = 0$;

5:[backtrack]

find $i = \max\{k < j : \hat{x}_k = 1\}$;

if no such i then return;

$\hat{c} = \hat{c} + w_i$; $\hat{z} = \hat{z} - p_i$; $\hat{x}_i = 0$; $j = i + 1$; go to 2;

end.