NABII: NETWORK ALIGNMENT BASED ON INTERACTIONS INTENSITY

by
Fadilh Alhubail
B.A., King Saud University, 2010

A thesis submitted to the Graduate Faculty of the University of Colorado at Colorado Springs in partial fulfillment of the requirements for the degree of Master of Science Department of Computer Science 2015
© Copyright Fadilah Alhubail 2015

Permission is granted to copy, distribute and/or modify this document under the terms of the GNU.
This thesis for the Master of Science degree by
Fadilah Alhubail
has been approved for the
Department of Computer Science
by

Professor Jugal Kalita, Chair

Professor Jia Rao

Professor Lisa Hines
This thesis studies protein-protein interaction (PPI) network alignment, one of the fundamental problems in biological network analysis that maps the proteins of one network in correspondence with the proteins of another network with respect to their topological and biological similarities. Many algorithms and techniques have been introduced to solve this problem, but it is considered NP hard and the perfect alignments are impossible to obtain. Here, we introduce our algorithm, Network Alignment Based on Interaction Intensity (NABII), which aligns PPI networks based on interactions intensity; it works based on the topological and biological similarity. NABII aligner produces high quality alignments with high biological quality. Moreover, it outperforms the existing aligners in performance since it is 80% faster than most existing aligners.
Dedicated to my parents, Zahra and Mousa, and to my husband Haydar. Without you nothing will be possible, thanks for your support.
Acknowledgements

We thank Jugal Kalita for supervising this work and providing helpful advice. In addition, we thank Connor Clark and Michael Bankston for constant help with useful information. Also, we thank Vikram Saraph for his assistance with using the aligner MAGNA.
Contents

Abstract iii

Acknowledgements v

List of Figures viii

List of Tables ix

Abbreviations x

1 Introduction 1
   1.1 Objectives ............................................. 2
   1.2 Outline of Thesis ..................................... 3

2 Background 4
   2.1 Introduction ............................................. 4
   2.2 Protein-Protein Interaction Networks ...................... 6
   2.3 PPI Network Alignment .................................. 8
   2.4 Pairwise Global Network Alignment ....................... 8
   2.5 Pairwise Local Network Alignment ......................... 9
   2.6 Topological Similarity vs. Biological Similarity ........ 11

3 Related Work 12
   3.1 PathBLAST .............................................. 12
   3.2 IsoRank ................................................... 13
   3.3 GRAAL .................................................... 13
   3.4 H-GRAAL .................................................. 14
   3.5 MI-GRAAL ............................................... 14
   3.6 NATALIE ................................................ 15
   3.7 GHOST ................................................... 15
   3.8 HubAlign ................................................ 16
   3.9 PISwap ................................................. 16
Table of Contents

3.10 MAGNA .................................................. 17
3.11 Optnetalign ............................................. 17

4 Evaluation Methods And Test Data ........................................ 18
  4.1 Evaluation Methods For Global Alignment Quality ................. 18
  4.1.1 Edge Correctness (EC) ........................................ 19
  4.1.2 Node Correctness (NC) ........................................ 19
  4.1.3 Interaction Correctness (IC) ................................... 19
  4.1.4 Induced Conserved Structure (ICS) ............................. 20
  4.1.5 Gene Ontology Consistency (GOC) Evaluation .................. 21
  4.2 Test Data ...................................................... 21
    4.2.1 Synthetic PPI Network Dataset ............................... 21
    4.2.2 Real PPI Network Dataset ..................................... 23
  4.3 Conclusion ...................................................... 23

5 NABII Algorithm For PPI Network Alignment ................................ 25
  5.1 Introduction .................................................... 25
  5.2 NABII Algorithm ................................................ 26
    5.2.1 Calculating The Intensity Matrix .............................. 26
    5.2.2 Calculating The Difference Between Nodes ................... 28
    5.2.3 Seed-And-Extend Approach .................................. 30
  5.3 Results And Discussion ........................................... 32
    5.3.1 Biological Results .......................................... 33
    5.3.2 Topological Results ......................................... 36
    5.3.3 Time Performance ............................................ 40
  5.4 Conclusion ....................................................... 41

6 Conclusion ......................................................... 42

Bibliography .......................................................... 45
## List of Figures

2.1 An example of yeast protein network ........................................ 7  
2.2 Pairwise global alignment finds only one overall mapping from the first network to the second network, where $|V_1| \leq |V_2|$ .................. 9  
2.3 Pairwise local alignment aligns local portions of one network to local portions of another network to find regions of high similarity ........ 10  

5.1 Example for calculating interactions intensity for node A ............ 27  
5.2 The number of interactions around the node A from level 1 to level 3 28  
5.3 Example of the first network $G_1$ and the second network $G_2$ ...... 28  
5.4 Calculating the difference between node A, and the nodes from $G_2$ 29  
5.5 Seed-and-extend, m=4 .................................................... 30  
5.6 Seed-and-extend, m=8 .................................................... 31  
5.7 Pseudo code of NABII algorithm ............................................ 32  
5.8 GOC results of ISOBench dataset, ($n = 5, m = 200$) .................. 34  
5.9 GOC results of ISOBench dataset, ($n = 3, m = 100$) .................. 35  
5.10 GOC results of ISOBench dataset, ($n = 7, m = 150$) .................. 35  
5.11 ICS results of ISOBench dataset, ($n = 5, m = 200$) .................. 37  
5.12 ICS results of ISOBench dataset, ($n = 3, m = 100$) .................. 38  
5.13 ICS results of ISOBench dataset, ($n = 7, m = 150$) .................. 39  
5.14 ICS results for NAPAbench dataset ....................................... 40  
5.15 Time Performance ......................................................... 41
List of Tables

4.1 Number of edges in NAPAbench dataset . . . . . . . . . . . . . . . . . . . . . 22
4.2 Number of nodes and edges in IsoBase dataset . . . . . . . . . . . . . . . . . 23
Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NABII</td>
<td>Network Alignments Based on Interaction Intensity</td>
</tr>
<tr>
<td>PPI</td>
<td>Protein Protein Interaction</td>
</tr>
<tr>
<td>EC</td>
<td>Edge Correctness</td>
</tr>
<tr>
<td>NC</td>
<td>Node Correctness</td>
</tr>
<tr>
<td>IC</td>
<td>Interaction Correctness</td>
</tr>
<tr>
<td>ICS</td>
<td>Induced Conserved Structure</td>
</tr>
<tr>
<td>GOC</td>
<td>Gene Ontology Consistency</td>
</tr>
<tr>
<td>NAPAbench</td>
<td>Network Alignment Performance Assessment benchmark</td>
</tr>
<tr>
<td>ACO</td>
<td>Ant Colony Optimization</td>
</tr>
<tr>
<td>TSP</td>
<td>Traveling Salesman Problem</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

The design and development of new pharmaceutical drugs depend on the ability of scientists to understand biological nature of diseases and how medicine can be designed to cure a specific disease or reduce its symptoms. A biological mechanism to cure a disease requires understanding how proteins that are induced in the disease interact with one another to regulate their function [1]. In particular, interactions among proteins have received serious attention of researchers and the number of studies have increased in this area. Protein-Protein Interaction (PPI) networks are graphical representations of how the proteins interact with each protein representing a node in the network and each edge representing an interaction between two proteins. PPI network alignment is one technique that we can use to find an overall match between proteins from different species. Aligning two or more PPI networks is challenging due to the complexity and size of PPI networks. Although the number of known PPI networks is growing and we have a large amount of available PPI data, we still need to do more research in this area [2].
The PPI network alignment problem is a young area and it is impossible to find an exact solution for this problem because it is classified as an NP-hard problem, and researchers are still far from finding an approximate solution with perfect performance and alignment [3]. Moreover, the incomplete and noisy nature of the existing databases make this problem very difficult [3]. Previous work on PPI network alignment has focused on local network alignment problem more than global network alignment one. Research has also focused on pairwise alignments more than multiple alignments [4].

The recent activities in network alignment resulted on many aligners; all of which utilize improved methods for aligning these types of networks resulting in different designs where each has better results than the last.

1.1 Objectives

The objective of this thesis is to come up with a new algorithm that improves results of existing state-of-the-art algorithms. We do this by introducing a new way of performing alignment, based on measuring the similarity of interaction intensity around the node. Current network aligners have two major issues. The first one is that existing algorithms run slowly, especially in aligning very large PPI networks. They also consume a large amount of memory, and some of them crash because they run out of memory. The second issue is that the alignment accuracy is still low. This encourages us to develop a new method for aligning PPI networks to improve both biological results and computational efficiency. Our goal
is to implement a fast algorithm that produces high performance and high biological quality at the same time. We choose four existing recent aligners to compare our algorithm results with.

### 1.2 Outline of Thesis

This thesis is organized into six chapters. Chapter 1 contains the introduction and the objective of our thesis. In chapter 2, we provide the background to the problem of network alignment, describing PPI networks, PPI network alignment problem, and the difference between global and local alignments. We also discuss both biological and topological similarities. In chapter 3, we review the most notable alignment algorithms starting from older algorithms to the newest, explaining briefly how each aligner produces alignments. In chapter 4, we introduce the common topological evaluation methods that are used to evaluate the topological quality of the alignments. We also introduce a biological metric that evaluates the biological accuracy of an alignment. In chapter 5, we introduce Network Alignment Based on Interaction Intensity (NABII) algorithm. NABII is the main contribution of this thesis; it is an algorithm to align pairwise global PPI networks. We explain the algorithm in detail, and we show our results compared to some of other existing aligners. We compare our algorithm to GRAAL, MiGRAAL, MAGNA, and PINALOG aligners. Lastly, in chapter 6, we summarize what we did, and what we discovered in this work.
Chapter 2

Background

In this chapter, we introduce PPI networks. Then, we discuss the problem of global PPI network alignment and compare it to pairwise local alignment. We also introduce some evaluation methods that we can use to evaluate the quality of the alignment.

2.1 Introduction

A large amount of biological network data, including Protein-Protein Interaction (PPI) and gene networks, have been produced recently. This gives us the opportunity to perform extensive research and experiments to analyze biological networks. Analyzing two biological networks is challenging due to the complex structures of networks and interactions among them. Comparative analyses of biological networks are expected to have a large impact on the understanding of biology and disease, and it could help solve problems in systems biology as PPI
data keep increasing in amount [5]. In our research, we focus on the task of aligning pairwise PPI networks from different species. That is, given a pair of PPI networks, the alignment problem is finding a mapping from the nodes of one network to the nodes from the other network, to find an optimal mapping between the nodes of the two networks that best represents topological and biological functions [6]. Successful PPI network alignments have revealed large numbers of shared subnetworks between species as diverse as *S. cerevisiae* and *H. sapiens*, and reconstructed phylogenetic relationships between species based entirely on the amount of overlap discovered between their PPI networks [7].

PPI network analysis and alignment are crucial for transferring insights about and information known for one species to other related species, and adopt these insights to understand the mechanisms of human diseases [8]. Many methods have been introduced and developed to align PPI networks, both globally and locally, some using sequence similarity, network topology, or both in establishing equivalence. [9] The ultimate goal of PPI network alignment is to create algorithms that can automatically find similar subnetworks of two PPI networks that are common to the two species, and from there form hypotheses about protein function in species for which less information exists [3]. The main challenge in designing such an PPI network aligner is how to estimate the topological similarity between two nodes accurately, and how to combine that with sequence similarity to produce a good alignment [3].

An aligner can only come up with an approximate solution for the problem because exact solutions for global network alignment are unreachable due to NP-completeness of the underlying subgraph isomorphism problem [10]. Existing
aligners differ in the quality of the alignments and in the running time.

In this thesis, we focus on global alignment of pairwise PPI networks. Pairwise global alignment algorithms align two PPI networks. However, multiple alignment algorithms align two or more networks at the same time [3]. Global alignment provides a unique alignment from one network to another network considering all nodes and edges of both networks. In contrast, local alignment aims to find similar sub-networks in proteins of different species considering only small subset of nodes and edges from both networks; this kind of alignment can be ambiguous because the vertex from one network could be mapped to many vertices in the second network [11]. The incomplete and noisy nature of existing PPI network databases makes this problem more difficult [3]. Previous work on PPI network alignment has focused on the local network alignment problem more than the global network alignment one. In addition, research has focused on pairwise alignments more than multiple alignments [4].

### 2.2 Protein-Protein Interaction Networks

A Protein-Protein Interaction network is a graphical representation of proteins and interactions among these proteins in a given organism, where the nodes in the network represent proteins, and an edge between two proteins represents an interaction between them to perform a biological function. PPI networks are important because almost all functions in cells are governed by protein interactions. It has been proposed that all proteins in a given cell are connected through an extensive network, where non-covalent interactions are continuously forming and
Figure 2.1: An example of yeast protein network

Proteins interact, form physical bonds, and create a complex network to perform a specific function [11]. Understanding these kinds of networks is very hard due to the size and complexity of the networks. Figure 2.1 represents an example of yeast protein network [12].

Over the years, researchers have performed many experiments on proteins, and collected and created several large PPI network databases. The results are stored in datasets such as DIP [14], Isobase [15], and BIOGrid [16]. The Gene Ontology (GO) project, which has produced a large database that contains the proteins and their biological functions, is a collaborative effort in this area. The GO database is constantly being updated and revised when experiments are performed and new discoveries are made [3].
2.3 PPI Network Alignment

The network alignment problem is the problem of finding the best way to fit one network to another network. PPI network alignment is a way to find likely subnetworks of PPI networks of two species by using sequence similarity between proteins, topological similarity, or both sequence and topological similarities [3]. The purpose of PPI network alignment is creating an algorithm that can automatically find likely orthologous proteins between species. From this alignment we can extract information that enables us to understand how proteins interact biologically. An alignment obtains an aligned PPI network with regulated functions.

2.4 Pairwise Global Network Alignment

Global alignment provides a unique alignment from one network to another network. Thus every vertex in the first network is mapped to a vertex in the second network [11]. We are given two PPI networks $G_1(V_1, E_1)$ and $G_2(V_2, E_2)$, where the vertices $V_i$ represent proteins and the edges $E_i$ represent interactions between those proteins. We assume that $|V_1| \leq |V_2|$, and we assume that $f$ is a total injective function $f : V_1 \rightarrow V_2$. The alignment is total if it maps all vertices from the first network $G_1$ to some vertices in the second network $G_2$; and it is injective if every element is mapped to exactly one element in the second network [17]. Pairwise global alignment finds the best overall alignment between two networks; it finds only one alignment that maps each node from the first network to only one node from the second network. No two nodes from the smaller network can be mapped to the same node in the second network. See Figure 2.2.
Most aligners use two-stage algorithms to find global network alignment. First, for each pair of nodes in $V_1$, $V_2$, they compute the similarity, by either the local topology of the graph, or by their sequence similarity [3]. Second, they take the similarities between these nodes as weighted edges and match the nodes in both networks based on the weighting [3]. The global network alignment problem is NP-complete; so there is no efficient algorithm for it that is likely to be found [17].

### 2.5 Pairwise Local Network Alignment

Although we focus on global alignment of PPI network in this thesis, it is important to contrast it with local network alignment. Local network alignment procedures can detect multiple regions of high similarity between two networks. In particular, such an algorithm can match a given protein of a network to several proteins of the other network in different local matches [17]. This kind of alignment can be ambiguous because the vertex from one network could be mapped to many
vertices in the second network [18]. Several methods have been implemented to obtain pairwise local network alignment of PPI networks, i.e., to find subsets of matching pairs of proteins with conserved subgraphs of interactions [17]. See Figure 2.3.

![Figure 2.3: Pairwise local alignment aligns local portions of one network to local portions of another network to find regions of high similarity](image)

However, local alignment became less popular especially after it was found that the very large regions of apparent structural similarity could be found by global alignment instead [3, 11]. According to this discovery, we conclude that finding small local regions of similarity became less popular than seeking the best consistent matching across all nodes concurrently. Global network alignment is more difficult than local alignment since it must find a solution among all possible global matches [17].
2.6 Topological Similarity vs. Biological Similarity

There are two methods to measure the similarity between two nodes. The first one is topological similarity, which analyzes the shape of the networks around the two node using multiple metrics such as degree, eccentricity, betweenness, or the more recently developed graphlet degree [3, 19]. At this point, the nodes that appear to be in topologically similar regions of their respective networks are considered similar [3]. The second method is biological similarity and in particular, sequence similarity information. Using sequence similarity and BLAST E-value for protein sequence similarity are very useful, and have been popular measures of node similarity [3]. Some alignments use both topological and sequence similarity, and some use only one measure of similarity.
There are a number of existing algorithms for the PPI network alignment problem. We will review the most notable alignment algorithms in this section starting from older algorithms to the newest ones.

### 3.1 PathBLAST

The earliest local network alignment algorithm is PathBLAST, introduced by Kelley et al. [18]. It is a basic method that searches for high-scoring alignments between pairs of protein interaction paths, for which proteins of the first path are paired with putative orthologs occurring in the same order in the second path. PathBlast can discriminate between true- and false-positive interactions and allows for functional annotation of PPI pathways based on the similarity to the network of another, well-characterized species [18].
3.2 IsoRank

IsoRank was the first algorithm used for global network alignment and also is the most cited. IsoRank remains a popular baseline for evaluating the performance and quality of new algorithms. It was introduced by Singh et al. [20]. This aligner uses a two-stage approach. In the first stage, it associates a score with each possible match between the nodes of the two networks. In the second stage, it constructs the mapping by extracting high-scoring and consistent pairwise match. IsoRank provides many valuable insights, although it indicated a relatively low overlap between the yeast and fly networks. It generated an alignment with 1420 edges even though both networks (yeast and fly) have more than 25000 edges each.

3.3 GRAAL

GRAAL [7], introduced by Kuchaiev et al, is the first algorithm in the GRAAL series. It is the first algorithm that relies on topological similarity without the use of sequence similarity to extract network alignments. GRAAL is a pairwise network aligner; it aligns a pair of nodes from different networks based on a similarity measure of their local neighborhoods. GRAAL counts graphlets, small induced subgraphs, for each node in each network. Then, it computes the occurrence of the node in each graphlet and uses the computation results in constructing a graphlet degree signature. Finally, it uses a heuristic matching algorithm to align the two nodes that are most similar, and works outward to align their neighbors, until all nodes in $V_1$ have been aligned.
3.4 H-GRAAL

Milinkovic et al. introduced an algorithm called H-GRAAL (Hungarianalgorithm-based GRAph ALigner) [5]. H-GRAAL relies solely and explicitly on network topological similarity. It produces an alignment with a minimum total cost between networks, where the total cost is calculated by summing the cost of all aligned pairs, and the cost of an aligned pair of nodes is computed based on their node signature similarity. So this algorithm could be used to find an optimal alignment from the first graph to the second graph with respect to cost. H-GRAAL requires polynomial time to solve the linear assignment problem. The authors found that their algorithm is capable of producing high-quality alignments with high node correctness, edge correctness and interaction correctness. H-GRAAL extracts biologically relevant and statistically significant meaning using only network topology.

3.5 MI-GRAAL

Kuchaiev and Przulj introduced a network alignment algorithm called Matching-based Integrative GRAph ALigner (MI-GRAAL) [11]. MI-GRAAL can integrate any number and type of similarity measurements between network nodes (e.g. proteins), including any topological network similarity measurements, sequence similarity, functional similarity and structural similarity. This algorithm can align large functionally connected regions of PPI networks. It was able to discover that 77.7% of the proteins in the yeast high-confidence PPI network are linked to a connected sub-network that is fully contained in the human high-confidence PPI network. MI-GRAAL produces alignments with impressive topological quality in
comparison with alignments of the same networks with three other global network alignment algorithms: IsoRank, GRAAL and H-GRAAL.

3.6 NATALIE

El-Kebir et al. introduced the software tool called NATALIE 2.0 for the global network alignment problem [13]. Their algorithm combines both sub-gradient optimization and a dual descent method. It works by performing the sub-gradient method until termination and then switching over to the dual descent method; this is repeated many times. The results show that the Lagrangian relaxation approach is powerful and it results in good quality alignment. Its running time is comparable to those of GRAAL and IsoRank.

3.7 GHOST

Patro and Kingsford introduced GHOST, which is a framework for the global alignment of biological networks [21]. GHOST uses a two-phase approach to align two or more networks. The first phase of GHOST uses a seed-and-extend strategy. It seeds regions of an alignment with high scoring pairs of nodes for two or more different networks and then extends the alignments around the neighborhoods of these two nodes. The second phase uses a local search strategy to explore regions around the initial alignment to choose a better solution. This aligner yields alignments with a substantially higher topological quality than either IsoRank or Natalie 2.0.
3.8 HubAlign

Hashemifar and Xu introduced a global network alignment algorithm called HubAlign [22]. It makes use of both network topology and sequence homology information by depending on the observation that topologically important proteins in a PPI network are more likely to be aligned. HubAlign assigns a score, or weight to each node and edge of a PPI network using an iterative minimum-degree heuristic algorithm, and measures the topological and functional importance of a node and an edge in the PPI network with respect to the global network topology. This algorithm complexity is about $O(n^2 \log(n))$, which is good compared to other heuristic algorithms. They compare their results with other methods like MI-GRAAL, and IsoRank. The results show that their algorithm produces an alignment with much larger edge correctness, the largest common connected subgraphs and higher symmetric substructure scores than the other methods except NETAL.

3.9 PISwap

Chindelevitch et al. introduced a method to compute pairwise PPI network alignments [23]. The method has two phases. The first phase of the algorithm is finding a maximum weight match in the bipartite graph obtained by joining pairs of proteins in the two networks. This match can be obtained by the well-known Hungarian algorithm and costs polynomial time. In the second phase of the algorithm, the authors apply a local search method to iteratively improve the initial match, taking into account both the sequence score and the topology score.
of the match. The whole method costs $O(n^3)$ time, that could be reduced by a factor of 2 if the number of vertices in both pieces is roughly equal.

3.10 MAGNA

MAGNA is a genetic algorithm aligner, introduced by Saraph et al. [24]. It simulates a “population” of alignments that “evolves” over time. The initial population can contain random alignments or alignments produced by existing methods. After this, it chooses the “fittest” candidates, which are conserved with the most edges, to survive and proceed to the next generation. It uses a crossover operator that produces a new alignment from two parents by computing a permutation that is halfway between the two parents and swaps elements of their permutations. It repeats this step until the number of conserved edges cannot be optimized further.

3.11 Optnetalign

Clark and Kalita introduce a multiobjective memetic algorithm for global network alignment. Their algorithm generates a large number of alignments using crossover, mutation, and swap-based hill climbing. They provide many experimental results, showing that Optnetalign can produce a large number of highly diverse alignments [3], situated on a Pareto front.
Chapter 4

Evaluation Methods And Test Data

4.1 Evaluation Methods For Global Alignment Quality

There are many methods to evaluate the quality of the results of global alignment algorithms. It is important to measure the quality of an aligner’s output because the real and true alignments between two or more PPI networks are unknown, especially for real biological data. So we need to use metrics to evaluate our results since we cannot know the percentage of nodes that are mapped to their true orthologs [3]. To measure the topological quality of an alignment, we can use several scores or metrics. In this section, we review some of the scores that are used in this area.
4.1.1 Edge Correctness (EC)

Edge Correctness is a metric that measures the percentage of edges in the first graph that are correctly aligned to edges in the second graph. A high edge correctness score means that both networks share similar topologies [25]. This metric quantifies the similarity between two networks. One can maximize the EC score by aligning the maximum number of edges. EC is equal to 100% if and only if the second graph $G_2$ contains an isomorphic copy of $G_1$ [11]. Optimizing the EC score is an NP-hard problem. EC is defined as given below [11].

$$EC = \frac{|\{(u, v) \in E_1 \land (f(u), f(v)) \in E_2\}|}{|E_2|} \times 100\% \quad (4.1)$$

4.1.2 Node Correctness (NC)

Node Correctness is the percentage of nodes in the first network that are correctly aligned to nodes in the second network [5]. So, to measure the NC, we need to have correct node mappings. If we denote the correct node mapping by $f : V_1 \rightarrow V_2$. We define NC as given below [25].

$$NC = \frac{|\{u \in V_1 : f(u) = g(u)\}|}{|V_1|} \times 100\% \quad (4.2)$$

4.1.3 Interaction Correctness (IC)

Interaction Correctness is the percentage of interactions that are aligned with respect to the correct node mapping. We say that an interaction $A, B$ is correctly aligned if two connected vertices $A$ and $B$ from the first network $G_1$ are correctly

$$\text{Interaction Correctness} = \frac{|\{u \in V_1 : f(u) = g(u)\}|}{|V_1|} \times 100\%$$
aligned to their partners in the second network $G_2$ and if their partners interact in $G_2$ [5]. IC is stricter than EC, and it is hard to use it in a real application because we need to have the correct node mapping beforehand.

If we denoted by $g: V_1 \to V_2$ the correct node mapping of $G_1$ to $G_2$ and by $f: V_1 \to V_2$ an alignment produced by our algorithm, we compute IC as given below [25].

\[
IC = \frac{|\{(u, v) \in E_1 : (f(u), f(v)) \in E_2, f(v) = g(v)\}|}{|E_1|} \times 100\% \tag{4.3}
\]

### 4.1.4 Induced Conserved Structure (ICS)

The Induced Conserved Structure (ICS) score represents the ratio of the number of edges conserved by $f$ to the number of edges in the sub-network of $G_2$ induced on the nodes in $G_2$ that are aligned to the nodes in $G_1$ [24]. This score is an extension of the EC score with a further intuition. If some regions in $G_2$ are dense, we could map a sparse region of $G_1$ in many different ways. Aligning a sparse region of $G_1$ to a sparse region of $G_2$ is better than aligning a sparse region of $G_1$ to a dense region of $G_2$. So ICS score penalizes alignments that are mapped to a denser subgraph of $G_2$ [3]. We use ICS score to evaluate our algorithm and compare it with other algorithms since it is a good topological metric. ICS is defined as follows [24].

\[
ICS = \frac{|f(E_1)|}{|E(G_2[f(V_1)])|} \tag{4.4}
\]
4.1.5 Gene Ontology Consistency (GOC) Evaluation

The Gene Ontology (GO) is widely used to annotate molecular attributes of genes and gene products [26]. We use GO to evaluate the biological accuracy of the alignment since the literature on PPI network alignment makes use of protein annotations from the GO Database [3]. Many methods to measure GO term agreement that have been used in the literature. A common one that is used to test the biological quality of alignments is based on GO consistency of the aligned pairs of proteins. It measures the overlap of GO terms between aligned nodes. GO consistency (GOC) is defined as [27]:

\[
GOC(u, v) = \frac{|GO(u) \cap GO(v)|}{|GO(u) \cup GO(v)|}
\] (4.5)

GOC is calculated for each pair of aligned proteins from the alignment results. The bigger the GOC is, the more similar function these proteins have. If GOC equals the maximum value, which is 1, it means that these proteins have exactly the same function [28]. We report the sum of GOC over all proteins pairs for each resulted alignment.

4.2 Test Data

4.2.1 Synthetic PPI Network Dataset

To perform the evaluation of our algorithm, and the algorithms that we compare with, we make use of the Network Alignment Performance Assessment benchmark (NAPAbench) synthetic PPI network dataset [29]. This dataset was created
specifically for evaluating and comparing the performance of various network alignment algorithms. The NAPAbench dataset is generated by using state-of-the-art algorithms such that it is generated with biologically realistic properties. NAPAbench has been used to evaluate the quality of old and new PPI network alignment algorithms [3].

Because of the high time requirements of some algorithms that we compare our algorithm with, we use a subset of the standard NAPAbench dataset, consisting of twelve pairwise alignment problems. This subset contains three problems of the duplication with random mutation (DMR) model [dd], three problems of the duplication-mutation-complementation (DMC) model [82], and three problems of a crystal growth (CG) model, with all models recently proposed [38]. The number of nodes in the networks is between 3000 and 4000 nodes, but the number of interactions is different in each network due to the difference in the network topology. Table 4.1 presents the number of edges in each network [3].

<table>
<thead>
<tr>
<th>Network</th>
<th>$G_1$ Edges</th>
<th>$G_2$ Edges</th>
</tr>
</thead>
<tbody>
<tr>
<td>CG1</td>
<td>11986</td>
<td>15986</td>
</tr>
<tr>
<td>CG2</td>
<td>11987</td>
<td>15987</td>
</tr>
<tr>
<td>CG3</td>
<td>11984</td>
<td>15984</td>
</tr>
<tr>
<td>DMC1</td>
<td>6090</td>
<td>8112</td>
</tr>
<tr>
<td>DMC2</td>
<td>6361</td>
<td>8310</td>
</tr>
<tr>
<td>DMC3</td>
<td>6077</td>
<td>8347</td>
</tr>
<tr>
<td>DMR1</td>
<td>6017</td>
<td>8238</td>
</tr>
<tr>
<td>DMR2</td>
<td>5739</td>
<td>7602</td>
</tr>
<tr>
<td>DMR3</td>
<td>6457</td>
<td>8431</td>
</tr>
</tbody>
</table>

Table 4.1: Number of edges in NAPAbench dataset
4.2.2 Real PPI Network Dataset

The use of real datasets is also important to evaluate the accuracy of alignments because such data is noisy and incomplete. The Isobase dataset is popular and it has been used for evaluating alignment algorithms. So we use the Isobase dataset, which is an experimentally-derived PPI network dataset [15]. This dataset is supplemented with sequence similarity and GO annotation data [3]. The number of nodes in this dataset is much larger than in the NAPAbench dataset, between 2000 and 10000. We use four networks to evaluate our algorithm (C. Elegans, D. Melanogaster, H. Sapiens, S. Cerevisiae). The exact number of nodes and edges are shown in Table 4.2.

<table>
<thead>
<tr>
<th>Network</th>
<th>Nodes</th>
<th>Edges</th>
</tr>
</thead>
<tbody>
<tr>
<td>C. Elegans</td>
<td>2805</td>
<td>4495</td>
</tr>
<tr>
<td>D. Melanogaster</td>
<td>7518</td>
<td>25635</td>
</tr>
<tr>
<td>H. Sapiens</td>
<td>9633</td>
<td>34327</td>
</tr>
<tr>
<td>S. Cerevisiae</td>
<td>5499</td>
<td>31261</td>
</tr>
</tbody>
</table>

Table 4.2: Number of nodes and edges in IsoBase dataset

4.3 Conclusion

Many methods have been introduced to evaluate the quality of the results of global alignment algorithms. Measuring the quality of the alignment is important because the true alignments between two or more PPI networks are unknown, especially for real biological data. In our experiments in this thesis, we evaluate an alignment’s topological quality by computing the ICS score for the alignment of NAPAbench and IsoBase datasets, while we use the GOC score to evaluate an
alignment’s biological accuracy for the IsoBase dataset because it is supplemented with GO annotation data.
5.1 Introduction

Current global network aligners have two major issues. The first one is that existing algorithms they run slowly, especially in aligning very large PPI networks. They also consume a large amount of memory, and some of them crash because they run out of memory. The second issue is that the alignment accuracy is still low. This encourages us to develop a new method for global alignment of PPI global networks to improve both biological results and computational efficiency.
of interactions, i.e., the number of edges, around a node as a measure of topological similarity. We also make the use of biological similarity in our algorithm as well.

5.2 NABII Algorithm

We present a novel global network alignment algorithm, denoted as NABII, to align a PPI networks pairwise using both biological and topological similarity. Our algorithm is a seed-and-extend algorithm; it aligns the most similar pair and then aligns neighbors of already-matched nodes. Here, we explain the working of our algorithm.

5.2.1 Calculating The Intensity Matrix

We assume that we have two networks $G_1(V_1, E_1)$ and $G_2(V_2, E_2)$, where $V_i$ represents the set of nodes and $E_i$ represents the set of edges. To guarantee that each node in $G_1$ is mapped to exactly only one vertex in $G_2$, we assume that the number of nodes in $G_1$ is less than the number of nodes in $G_2$; so $|V_1| \leq |V_2|$.

The first step in our algorithm is calculating the intensity matrix. Our goal from this step is finding similar nodes topologically. The intensity matrix stores the number of interactions and edges around each node given $n$, the depth or level to which we look at. We view the sub-network up to depth $n$ around a node as a tree, with the node under focus as the root of the tree. Based on the level $n$, we calculate all interactions around the node. For example, if $n = 3$, we have to find all interactions that exist in the 3-hop neighborhood of the given node.
In Figure 5.1, to calculate the number of interactions around node A, we calculate the number of interactions that connect A directly to its neighborhood, which consists of four interactions. The first level interactions are simply represented by the direct edges A-B, A-C, A-D, and A-E. Then, we calculate the number of the interactions at the second level, i.e., interactions of neighbors of A. The neighbors of A are B, C, D, and E, and they are connected to the other nodes, for a total of five additional interactions. The second level interactions are represented by the edges B-D, B-E, B-C, E-G, and C-F, for a total of five. In every level we calculate only the new interactions, which were not considered in prior levels. At level 3, we have only one new interaction, represented by the edge G-H. Figure 5.2 represents the number of interactions around the node A from level 1 to level 3. We perform this calculation for every node in the two graphs $G_1$ and $G_2$, Figure 5.3.
5.2.2 Calculating The Difference Between Nodes

In this step, we choose a node from $G_1$ and we use the intensity matrix to align it to the most similar node from $G_2$. To find the most similar one, we calculate
the difference between this node and each node in $G_2$; then we choose the node with the least difference from $G_2$. We use Formula 5.1 to calculate the difference.

$$Diff(v_1, v_2) = \frac{\sum_{i=1}^{n}|i_n(v_1) - i_n(v_2)|}{n}$$

(5.1)

where $i_n$ represents the number of interactions in level $i$ at that node, and $n$ represents the number of levels. For example, in Figure 5.4, we compare the similarity between the node A from $G_1$ and every node in $G_2$. Then, we choose the node with the least difference value and align it to A. If we have more than one node that has similar difference values, we choose one of them randomly. In our example, we chose node B’, from $G_2$ to align it to A, from $G_1$ because it has the least difference value.

![Figure 5.4: Calculating the difference between node A, and the nodes from $G_2$](image)
5.2.3 Seed-And-Extend Approach

From the aligned nodes, we extend and align similar neighbors until we align \( m \) or fewer nodes, where \( m \) is the number of produced alignments by extending approach. In the example, we aligned A to B'; so in this step, we align the neighbors of A to the neighbors of B', Figure 5.5. We randomly choose a node from the neighbors of A, and find the most similar node from the neighbors of B' to align them. We use Formula 5.1 to calculate the similarity, like we did in the second step. After that, we continue extending and aligning nodes until we align \( m \) nodes. Then, we stop extending and go back to the first step; we choose another node and repeat the same steps until we finish aligning all nodes in \( G_1 \). In Figure 5.5, we suppose \( m=4 \), after aligning 4 nodes from \( G_1 \) to 4 nodes from \( G_2 \), we stop aligning and repeat starting from the first step.

\[ \text{Figure 5.5: Seed-and-extend, } m=4 \]
In case we could not produce $m$ alignments, Figure 5.6, we choose a random node from $G_1$, and we use GO annotation to find a similar node that has a similar function and align them together. From this step, we continue extending and aligning nodes until we reach $m$.

![Diagram](image.png)

**Figure 5.6:** Seed-and-extend, $m=8$

In this case we make use of sequence similarity or biological similarity, because it has been argued that relying only on topological similarity can be misleading, since actual complexes may appear disconnected in current noisy, incomplete datasets [3]. So sequence similarity information is necessary to produce the best possible alignments [30].
5.3 Results And Discussion

To test our algorithm, we use the NAPAbench synthetic dataset and the ISOBase real dataset. In these tests, we aligned PPI networks from *C. elegans, D. melanogaster, S. cerevisiae* and *H. sapiens*. We evaluate an alignment’s topological quality by using ICS score for the alignment of NAPAbench and IsoBase datasets, where as we use GOC score to evaluate the biological accuracy of IsoBase.
dataset alignments because it is supplemented with the GO annotation dataset. We compare our results with four existing aligners, GRAAL [7], Mi-GRAAL [11], MAGNA [24] and PINALOG [9]. We must note that GRAAL, and Mi-GRAAL crashed in the last three problems, and the missing bars in the figures mean that the corresponding program crashed or ran out of memory.

5.3.1 Biological Results

Here we report the results in terms of the GOC biological metric for ISOBase datasets. Figure 5.8 represents the results of NABII algorithm using parameters ($n = 5, m = 200$). NABII outperforms all existing aligners using GOC. The value of GOC found by NABII on each problem instance is comparable to or exceeds that of PINALOG, which was previously the best performer on this data set. The other aligners perform much less on biological quality than our aligner. GOC values of the other aligners fluctuate between 12.5 and 54, whereas the least GOC value in our results is about 157. So our algorithm gives us about 67% better results than the other aligners excluding PINALOG, and about 20% more than PINALOG. These results mean that our algorithm produces alignments with high biological quality, and that was our main goal to begin with.
Figure 5.8: GOC results of ISOBose dataset, \((n = 5, m = 200)\)

Figure 5.9 represents the experiment using parameters \((n=3, m=100)\). Here, we see that NABII manages astounding performance. GRAAL produced an alignment with low biological quality, and we think that is because GRAAL doesn’t use biological data such as GO annotation in the algorithm. It relies completely on the topological information. PINALOG gives results with high quality alignment compared to the other algorithms, excluding NABII. NABII is still outperforming all presented algorithms.
In Figure 5.10, we used parameters \((n = 3, m = 100)\). We see that our GOC results are much the same as before with the best alignments. The high level of GOC in our resulting alignments gives us some confidence in the quality of the alignments produced. By comparing the results of our algorithm and the other algorithms, we conclude that using biological information affects the results a lot and increases the quality of the produced alignment.

---

**Figure 5.9:** GOC results of ISOBase dataset, \((n = 3, m = 100)\)

**Figure 5.10:** GOC results of ISOBase dataset, \((n = 7, m = 150)\)
5.3.2 Topological Results

We used ICS score to evaluate our results and compared it with the other algorithms. We used the NAPAbench datasets and the Isobase real datasets for testing our algorithm. Here we report the results using ICS topological score for Isobase datasets (Figure 5.9). For these tests, we get a better range of results, and poorer results overall. We think that the variance of our results is due to the noise present in experimentally derived data. However, NABII outperforms or matches each of GRAAL, MAGNA, PINALOG in all problems, except in *C. elegans* and *H. sapiens*. Aligning these two networks produces an alignment with poorer ICS than GRAAL and MAGNA, and higher than PINALOG. Mi-GRAAL also produces an alignment with high ICS results. This could be due to the fact that, some algorithms like GRAAL are not trying to maximize both topological and sequence similarity, but instead try to maximize topological fit only.

NABII outperforms Mi-GRAAL in the problem of aligning *C. elegans* and *S. cerevisiae* since it gives about 0.22 while Mi-GRAAL gives about 0.19. In addition, NABII outperforms PINALOG in all problems.
Figure 5.11: ICS results of ISOBase dataset, \((n = 5, m = 200)\)

Figure 5.12 represents the experiment using parameters \((n=3, m=100)\). Here, we see that our ICS results are different from the prior experiment (Figure 5.11). In the first three problem, ICS results is slightly less in the problem of aligning *C. elegans* and *D. melanogaster*. In the other problem, our ICS performs an equal or better results than the other aligners, excluding the problem of *C. elegans* and *S. cerevisiae*. In this problem, Mi-GRAAL ICS result is about 20% more than NABII algorithm. The ICS of Mi-GRAAL is about 0.25, while the ICS of NABII is about 0.21. In general, this experiment produced an alignment with higher ICS results, but a slightly less GOC results.
The third experiment was performed using parameters \((n = 7, m = 150)\). Here, NABII performed better than PINALOG in all problems, except in the problem of aligning \textit{D. melanogaster} and \textit{H. sapiens} and the problem of aligning \textit{S. cerevisiae} and \textit{H. sapiens}. In general, this experiment’s results is less than the other experiments. We think it is better to choose \(n\), the depth or level of node’s neighbors, between 3 and 5 to get better topological results. We noticed that, the deeper levels give less useful information due to the complexity and noise of real networks.
For the NAPAbench datasets (Figure 5.14), we performed experiments with nine datasets, and we found that the average of topological results we obtain for Isobase dataset testing is better compared to results with the other aligners. We think the noise in the real datasets affected the results and the quality of the alignments that are obtained with the other aligners. On average, NABII outperforms MAGNA and GRAAL and produces alignments with much better results.
5.3.3 Time Performance

In our experimental tests, we observe that NABII is successful in aligning large PPI networks faster compared to other algorithms (Figure 5.15). Most published algorithms for network alignment are too slow. Our algorithm is still much faster than all the ones we compared with. The time efficiency outperforms GRAAL and Mi-GRAAL, being about 80% faster. Moreover, it is 20% to 50% faster than MAGNA and PINALOG. We ran these aligners for aligning PPI networks from the ISOBase dataset on a 2.5 GHz Linux system with 4 GB random access memory. As it is shown in Figure 5.15, running time of NABII is much less than that of the other algorithms.
In this chapter, we focus on the global network alignment problem, and we presented a powerful algorithm for computing the global alignment of two PPI networks. NABII algorithm is an algorithm that produces an alignment for two PPI networks relying on the interaction intensity around the node. We consider GO annotation in addition to topological information. Our algorithm outperform all presented algorithm biologically. In general, NABII produces high level of GOC, and this gives us some confidence in the quality of the produced alignments. We also noticed that NABII is much faster than the algorithm that we compared with.
Chapter 6

Conclusion

We presented a novel algorithm for pairwise global alignment of PPI networks. Although many algorithms have been published to align PPI networks, comparing their performance is still difficult and that is because we do not have the correct alignments for any real datasets. Our algorithm, NABII, outperforms existing state-of-the-art methods. It produces alignments based on the intensity of interactions in the network. It uses both sequence similarity and topological similarity to produce the alignment.

Current global network aligners have two major issues, the high computational cost and the low quality of the biological alignments. The NABII algorithm was developed to produce pairwise global alignment of PPI networks in an efficient time compared to other existing aligners. In addition, it produces alignments with high biological quality. We compared our algorithm with four existing aligners, GRAAL, Mi-GRAAL, PINALOG, and MAGNA. NABII outperforms these algorithms biologically.
In summary, first, we reviewed the problem of PPI network alignment and its purpose, showing the differences between global and local alignments, and the differences between biological and topological similarities. After that, we presented a few existing aligners, and explained each of them briefly. In the next section, we showed the common topological evaluation methods that are used to evaluate the topological quality of the alignments. We also introduced a biological metric that evaluates the biological accuracy of an alignment. In Chapter 5, we introduced the NABII algorithm. NABII algorithm produces an alignment for two PPI network based on interaction intensity of the nodes and GO annotations. It is a seed-and-extend algorithm; it aligns the most similar pair using intensity matrix and a formula to find the most similar node from the second network. Then, it aligns neighbors of already-matched nodes. We use GO annotation in a later step to complete aligning all nodes. The value of GOC found by NABII on each problem instance is comparable to or exceeds all algorithm that we compared with. We noticed that, the deeper levels of neighbors give less useful information due to the complexity and noise of real networks. So the depths between 3 and 5 give us better results. The NABII has a low computational cost compared to the other presented algorithms.

For future work, our algorithm can be improved by using both topological and biological information equally using weighted matrix. We also suggest improving the quality of evaluation metrics, because we noticed that the higher biological results, we get lower topological results and vice versa. It is hard to know the real alignments of real networks datasets, so the topological metrics are not reliable enough to measure the topological quality of an alignment. Moreover, we need to
collect more biological information since using biological sequences could help in increasing the quality of an alignment.
Bibliography


