ALIGNMENT OF PPI NETWORK USING NOVEL TOPOLOGICAL MEASURES

ASIYA JAHANGIR

03-243222-003

Supervisor: Dr. Ansar Siddique

A thesis submitted in fulfilment of the

requirements for the award of the degree of

Master of Science (Computer Science)

Department of Computer Sciences

BAHRIA UNIVERSITY LAHORE CAMPUS

Nov 2024

Approval of Examination

Scholar's Name: Asiya Jahangir Registration No: 78663

Programme of Study: Master of Science in Computer Science

Thesis Title: Alignment of PPI Network Using Novel Topological Measures.

It is to certify that the above student's thesis has been completed to my satisfaction and, to my belief, its standard is appropriate for submission for Evaluation. I have also conducted a plagiarism test of this thesis using HEC-prescribed software and found a similarity index at 5 % that is within the permissible limit set by the HEC for the MS/MPhil degree thesis. I have also found the thesis in a format recognized by the BU for the MS/MPhil thesis.

Principal Supervisor's Signature: _____

Date: 7/11/2024 Name: Dr. Ansar Siddique

Author's Declaration

I, Asiya Jahangir hereby state that my MS thesis titled "Alignment of PPI Network Using Novel Topological Measures" is my work and has not been submitted previously by me for taking any degree from this university Bahria University Lahore Campus or anywhere else in the country/world.

At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw/cancel my MS degree.

Name of student: Asiya Jahangir

Date: 7/Nov/2024

Plagiarism Undertaking

I, solemnly declare that the research work presented in the thesis titled "Alignment of PPI Network Using Novel Topological Measures" is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero-tolerance policy of the HEC and Bahria University towards plagiarism. Therefore, I as an Author of the above-titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred to/ cited.

I undertake that if I am found guilty of any formal plagiarism in the above-titled thesis even after the award of the MS degree, the university reserves the right to withdraw/revoke my MS degree and that HEC and the University have the right to publish my name on the HEC / University website on which names of students are placed who submitted plagiarized thesis.

Student / Author's Sign: _____

Name of the Student: Asiya Jahangir

DEDICATION

This thesis is wholeheartedly dedicated to my parents for their moral, spiritual, emotional, and financial support. Most importantly, I dedicate this thesis to Allah Almighty for His fruitful help, guidance, and power and for giving me strength to complete this study.

ACKNOWLEDGMENTS

First, I would like to thank my supervisor Dr. Ansar Siddique for giving me time, ideas, and encouragement. Thank you, Sir, for always being humble, calm, and cooperative throughout the study.

I want to show my gratitude to Dr. Umair Ayub for being my technical support throughout the research process. I am thankful to you for providing me with time, guidance, directions, and valuable assistance during the thesis.

I am grateful to all the people who helped me to complete my questionnaires and believed in me through all the way.

ABSTRACT

Experimental Protein-Protein Interaction (PPI) Network data has been collected using high-throughput PPI profiling techniques. PPI network analysis aids in the molecular-level understanding of the proteins. The PPI network alignment can reveal the relationships between multiple species, which improves our knowledge of biological systems and helps to transfer knowledge across the species. PPI network alignment's primary goal is to build a combined network that helps research intricate pathways to identify the roles of unidentified proteins. It aids in the identification of biological processes and molecular-level function understanding of the proteins across species. Through topological and biological similarity, network alignment offers a means of identifying comparable sections across various species and can facilitate the transmission of biological knowledge between them. Several strategies for network alignment have been developed, but it is still difficult to achieve high AFS (Average Functional Similarity) and Coverage. Moreover, the topological methods did not produce quality alignments in terms of average functional similarity. Similarly, the biological information did not guarantee high topological performance. This thesis presents a PPI network alignment algorithm and reviews the existing studies (in terms of semantic similarity and coverage). This thesis investigates different topological measures in addition to biological information to improve the topological and biological performance of the aligners. This thesis presents a novel topological approach that maximizes the AFS, coverage, and overall topological quality.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

PPI	-	Protein-Protein Interaction
HINT	-	High-quality INTeractromes
AFS	-	Average Functional Similarity
MF	-	Molecular Function
BP	-	Biological Process
GO	-	Gene Ontology
PSI-BLAST	-	(Position-Specific Iterated BLAST)

CHAPTER 1

INTRODUCTION

Proteins are composed of amino acids that combine to form a chain of sequences. Proteins are large and complex molecules that perform different functionalities in living organisms. There are 20 different kinds of amino acids. Every protein has a different amino acid sequence that results in different 3D structures and different functions. Different species can have different types of proteins. Proteins interact with other proteins to carry out their various functions [1]. Interaction between proteins enhances biological pathways. The interaction between proteins with other large bio-molecules or other proteins leads to the formation of complex pathways that reveal protein functions. The interactions between proteins are modeled as a graph or a network. In this network, nodes represent proteins, and the interaction between them is represented by edges [2]. The network is known as the Protein-Protein Interaction network. The PPI network is comparable to a mathematical depiction of the interaction between proteins. These are unique and have a particular biological importance [3]. The alignment of networks facilitates understanding of the interactions between distinct species and increases understanding of biological processes. Finding homologous proteins and their interspecies-conserved connections can be aided by comparing the PPI networks [4]. Drug design and the study of illness and health states can both benefit from PPI Networks [5]. Through topological and biological similarity, network alignment offers a means of identifying comparable sections across various species and can facilitate the transmission of biological knowledge between them. Numerous alignment techniques with high biological similarity and coverage in PPI networks are developed for this reason [6].

1.1. Background of the research

Proteins are large bio-molecules that perform many operations in living organisms and control any cell life cycle [7]. Proteins by interacting with other proteins, perform various functions – forming a network known as a Protein Protein Interaction network. PPI network is a graph-based representation that facilitates the identification of fundamental biological knowledge, including functionally conserved proteins and their interaction, and preserving evolutionary pathways among several species. The alignment of networks is PPI network alignment [8].

The network of protein-protein interactions known as the PPI network is crucial in modeling and analysis of biological processes. Research on functional modules derived from PPI networks enhances our comprehension of biological systems [9]. The enormous volume of PPI network data needs to be analyzed using experimental and computational techniques.

1.1.1. Experimental Techniques

The experimental techniques to identify the interactions include Yeast two Hybrid screening and Affinity Purification Mass Spectrometry [10]. The yeast 2-hybrid experiment is a tried-and-true method for discovering the interactions of proteins. For researchers, this is a potent tool that they frequently employ in conjunction with one or two other techniques to investigate the many interactions that occur within cells [11]. Affinity Purification Mass Spectrometry is another experimental technique that uses the proximity biotinylation technique [12]. Affinity Purification Mass Spectrometry may be utilized to look at protein complexes more broadly at the interactome level or to investigate individual protein-protein interactions within protein complexes. Other potential techniques are Intragenic Complementation [13], and NAPPA (Nucleic Acid Programmable Protein Array) [14].

1.1.2. Computational Techniques

An enormous amount of biological data has been created by computational methods. A thorough knowledge of bio-mechanism and evolution depends on the effective analysis and mining of this complicated material. The computational techniques include Computational Prediction of Protein-Protein Interaction, Genomic

Context Method, Text Mining Method, and Machine Learning Method [15].

1.1.3. Model of PPI Network

Proteins are modeled as nodes in a graph and a pair of interacting proteins are connected by undirected, potentially weighted edges [16]. It is a typical method of modeling protein-protein interaction networks. Reliability data related to the respective interactions can include using edge weights. It is noteworthy to mention that a more complicated model would be more appropriate for representing protein-protein interaction networks since interactions between proteins are frequently discovered through protein complex detection rather than as binary interactions. Protein complexes may be modeled using hyper-graphs rather than ordinary graphs, as each hyper-edge includes all the proteins in the complex [17].

1.2. Significance of the Research

The PPI network alignment using biological and topological similarities introduced ways to identify and facilitate the transmission of biological knowledge across different species. The study of proteins in the group is more informative as compared to their study in isolation as group study will reveal interaction mechanisms (such as pathways understanding) in addition to protein-independent properties (such as functions) [18]. The alignment of PPI networks can reveal hidden patterns of the group of proteins, and it can help study biological processes and biological pathways.

1.3. Aims and Objectives

The aims and Objectives of this thesis are:

- Investigation of different topological measures in addition to biological information so that AFS, Coverage, and Topological performance can be enhanced.
- To achieve ASF, Coverage, and concurrently alignments' Topological Quality.

1.4. Research Gap

Although many PPI network algorithms (aligners) have been introduced (to date) to align PPI networks, the performance of these aligners regarding functional similarity, coverage, and topological quality has not been achieved yet. Several network alignment algorithms perform well in terms of functional similarity, but these algorithms failed to achieve high coverage. Similarly, several algorithms performed well in terms of topological quality, but their functional similarity is low. BioAlign is the best algorithm in this domain, producing alignments with maximum functional similarity and coverage, but it still lacks in producing alignments with topological quality. The goal of this thesis is to determine which topological measure can enhance functional similarity, coverage, and topological quality. The alignment of networks facilitates understanding of the interactions between distinct species and increases understanding of biological processes.

1.5. Motivation

The alignment of networks facilitates understanding of the interactions between distinct species and increases understanding of biological processes. Drug design and the study of illness and health states can both benefit from PPI Networks. To date, an algorithm has not been developed that produces the best alignment quality and results.

1.6. Problem Statement

Several alignment algorithms have been introduced to improve alignment quality. Some of the alignment algorithms result in high functional similarity and/or high coverage while some result in better topological quality. The alignment done based on biological similarity has high functional similarity and low coverage while the alignment done based on topological quality has high coverage and low functional similarity. It is hard to achieve functional similarity, coverage, and topological quality at the same time.

1.7. Key Research Questions

Following are the key research questions this study aims to solve:

- Which algorithms are better in terms of AFS, Coverage, and/or Topological Quality?
- Which topological measure is more efficient in producing alignments with high AFS, Coverage, and Topological Quality?

1.8. Thesis organization

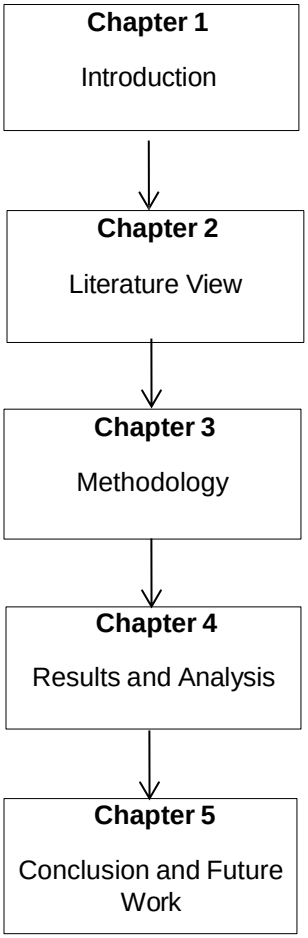


Figure 1.1 Thesis Organization

Figure 1.1 illustrates the organization of the thesis. **Chapter 2** represents the

literature review of PPI network alignment algorithms. It also discusses the relevant studies and gives an overview of the comparison between different alignment techniques. **Chapter 3** contains the methodology used in this study. It contains the procedure used to create an alignment algorithm. **Chapter 4** gives a detailed analysis of the results and evaluation. It shows the comparison between the new alignment technique with the existing aligners. The last is **Chapter 5** which shows the summary of the thesis and future work related to this study. This will give new pathways for researchers in this study.

CHAPTER 2

LITERATURE REVIEW

Proteins interact with other proteins in order to perform their functions – which in turn forms a network like structure, known as a PPI network. In this network, the modeling of proteins are represented by nodes and the interaction between proteins is modeled as edges, thus forming a network-like structure [19]. A more complicated model would be more appropriate for representing protein-protein interaction networks since interactions between proteins are frequently discovered through protein complex detection rather than as binary interactions. The complexity of protein is modeled as hyper-graphs rather than ordinary graphs. Each hyper-edge includes all proteins in the complex.

Experimental PPI network data is collected using high-throughput PPI profiling techniques [20]. PPI network analysis aids in the molecular-level understanding the proteins and their interactions. The alignment can reveal relationships among multiple species, which improves our knowledge of biological systems and helps to transfer knowledge across the species [21]. The objective of the PPI alignment is to create a combined network that facilitates finding out the role of unidentified proteins. The data of the PPI Network are kept in different datasets including BioGRID, IntAct, HINT, etc. This data is stored as a network where all the possible interactions of proteins related to a specific species are contained within a single network [22]. Studying the whole network of proteins gives more insight into biological activities rather than studying them in isolation.

The comparison between two PPI networks is known as pair-wise network alignment. The interaction between one network to the larger portion of the network is the alignment of that network. There are two categories of alignment, Global NA and Local NA. Some of the PPI algorithms use GNA and some use LNA. Both types are further described in this chapter.

An extensive summary of previously suggested algorithms is provided in this

chapter. The methods, advantages, and disadvantages of the previously suggested PPI network alignment algorithms are described here:

ModuleAlign is a unique global network alignment technique that defines a module-based homology score by utilizing local topology data [23]. Using a hierarchical clustering of proteins that have comparable functions that are included in similar module. It produces good results in the case of coverage but its complexity is high and has low functional similarity. HubAlign, using the greedy approach, aligns the important protein nodes first, using topology, after which alignment is extended to entire network [24].

MONACO is a unique and adaptable network alignment method which iteratively optimizes 'local' neighborhoods around focal nodes [25]. Since it adapts effectively to the expanding size of the network which makes it highly efficient. PROPER, to align the nodes, uses protein sequence similarity after that extends the aligned nodes by the use of topology [26]. PROPER aligns a few nodes (low-scoring nodes are not aligned), which is considered incomplete alignment.

IBNAL uses a unique clique-based index through which the alignment process is sped up which results in aligning big networks in a matter of seconds [27]. It outperforms other cutting-edge aligners in terms of topological and biological quality. MAGNA++, aligns nodes based on sequence similarity by using a genetic algorithm [28]. It concurrently maximizes any chosen node's edge conservation measure which as a result increases alignment quality. NETAL employs a greedy approach on the basis of a scoring matrix, which is generated using topological and biological data of input networks [29]. These aligners mainly use a greedy algorithm meaning they align nodes that have high scores which is also considered as incomplete alignment.

SAlign utilizes structural data along with topological and sequence information [30]. This technique results in increased semantic similarity and coverage. SAlign reveals the effects of topological and biological information. As it uses a significant amount of topological data which in turn decreases the ontological similarity. BioAlign, a multi-stage algorithm, uses additional biological information. This biological information includes sequence and 3D-structure similarity along with sequence similarity, secondary structure motifs, and remote homology and neighborhood expansion for topological alignment of nodes [31]. Compared to the current methods, BioAlign yields noticeably best results regarding functional

similarity and coverage. BEAMS adds values to the network edges according to sequence similarity and for the alignment, it uses a clustering method [32]. It outperforms many algorithms regarding AFS but the coverage of BEAMS is low as compared to HubAlign.

The comparison of prior aligners based on features, alignment heuristics, topological measures, and datasets, as well as their benefits and drawbacks, is displayed in Table 2.1.

Table 2.1 Previously Proposed Aligners.

S. No	**Aligners**	**Year**	**Topological Method**	**Alignment Heuristic**	**Limitations**
i.	ModuleAlign	2016	Min. Degree Heuristic+Clustering Similarity Scores	Hungarian algorithm	Low AFS and High Complexity
ii.	HubAlign	2014	Minimum degree heuristics	Greedy algorithm	Low AFS
iii.	MONACO	2021	Iterative Optimal matching of local neighborhood	Greedy algorithm	Low Coverage
iv.	IBNAL	2018	Clique-Degree Similarity	Clique-Size based Greedy algorithm	Used Go Annotations
v.	MAGNA	2015	-	Genetic Algorithm	Low Semantic Similarity Time Complexity
vi.	NETAL	2013	Local Topology Measure	Greedy Algorithm	Low AFS and coverage
vii.	PROPER	2016	Local Topology Measure	Neighborhood Expansion	Low Coverage
viii.	Salign	2020	Alignment score+Min. Degree Heuristic	Greedy algorithm	Low semantic similarity
ix	BioAlign	2022	3D structure+Remote homology+Predicted secondary structure motif	Greedy Algorithm	Extra biological information
x	BEAMS	2014	Local topology measure	clustering-based algorithm	Low Coverage

From above mentioned aligners, it is observed that the aligners that performed well in terms of functional similarity have low coverage while the aligners that perform well in terms of coverage have low functional similarity. MONACO, BEAMS, and

BioAlign performed better in comparison to the other algorithms. BioAlign outperforms every other aligner regarding AFS and coverage. However, despite its advantages, BioAlign has a notable limitation. It relies on additional biological information while incorporating very little topological information. Based on available information, the algorithm with maximum AFS, Coverage, and Topological Quality is still not developed.

2.1. PPI Networks

Proteins are large bio-molecules that perform many operations in living organisms and control any cell life cycle. Proteins interact with other proteins in order to perform their functions. This interaction enables them to perform complex biological processes [33]. The interaction of proteins with other proteins forms a network known as the PPI Network. The nodes represent proteins and interacting protein pairs are connected by non-directed, potentially weighted edges [34]. For example, N1= (P1, I1) represents a network where P is several proteins and I is used for interaction/edges.

2.2. PPI Network Alignment

The PPI alignment refers to the aligning of two different PPI networks [35]. For example, the network N1= (P1, I1), align with the network N2= (P2, I2). The alignment algorithms are of two types; 1) Global network alignment and 2) Local network alignment [36]. These are described below.

2.2.1. Global Network Alignment

The one-on-one mapping of nodes of a network is known as Global Network Alignment (GNA). The main goal of GNA algorithms is to align a greater possible number of proteins with the other networks' proteins [37]. This aligner uses one-one mapping between two networks of proteins i.e. each protein of network 1 is aligned with only one protein of network 2.

2.2.2. Local Network Alignment

The Local alignment uses many-many mapping of nodes [38]. LNA works in

two stages, 1) mining the networks/communities and 2) merging the network/communities. Secondly, the alignment is done by many-many mapping. For example, one group of proteins of network 1 is aligned with multiple groups of proteins of network 2.

Table 2.2 GNA and LNA Aligners.

GNA	Aligning Approach	Biological information	LNA	Aligning Approach	Biological information
BioAlign	Multi-Stage alignment	3D structure + Remote homology + Predicted secondary structure motif	MOPHY	Mine and Merge	Sequence Similarity
SAlign	Alignment score+ Min. Degree Heuristic	Sequence Similarity + 3D structure similarity	Jancura	Mine and Merge	Orthology Relationships
ModuleAlign	Hungarian Algorithm	Sequence Similarity	MODULA	Mine and Merge	Sequence Similarity
HubAlign	Greedy Algorithm	Sequence Similarity	AlignNemo	Merge and Mine	Sequence Similarity
$SAlign_{mc}$	Monte-Carlo approach	Sequence Similarity + 3D structure Similarity	NetworkBlast	Merge and Mine	Sequence Similarity
MAGNA++	Genetic Algorithm	Sequence Similarity	AlignMCL	Merge and Mine	Sequence Similarity
METAL	Topological based alignment	None	MaWISH	Merge and Mine	Sequence Similarity
PROPER	Neighborhood Expansion	Sequence Similarity	NetAligner	Merge and Mine	Sequence Similarity

Table 2.2 shows the GNA and LNA based algorithhms along with their aligning approach and biological information. Mostly alignment algorithms of PPI networks are related to GNA (Global Network Alignment). GNA algorithms make sure to align a maximum number of nodes that are semantically or biologically related. These alignment algorithms are BioAlign, SAlign, $SAlign_{mc}$, ModuleAlign, HubAlign, MAGNA, NETAL, PROPER, etc. SAlign, a GNA algorithm, uses biological data and network-related information. The effects of network-related and biological (sequence and structure) information were revealed through SAlign. Because of the use of an excessive amount of topological information resulted in decreased semantic similarity. In contrast, HubAlign incorporates network-related information to align the proteins and does not consider the biological context. ModuleAlign consists of an iterative technique that is used to find the alignment. However, it does not produce biologically relevant alignments, which means its Average Functional Similarity (AFS) is low. $SAlign_{mc}$ is a variant of SAlign, which uses the Monte Carlo approach rather than the greedy approach. $SAlign_{mc}$ was proposed to overcome the issues related to the SAlign alignment algorithm. $SAlign_{mc}$ produces distinct alignments of two PPI networks with functional similarity and coverage [30]. In MAGNA++, the alignment of the nodes is done by a genetic algorithm approach [39]. Its semantic similarity is very low and cannot handle large-scale networks. Another limitation of MAGNA++ is time complexity. NETAL uses a self-serving approach based on a matrix score [29]. Since it uses a greedy approach, meaning it only aligns the high-scoring nodes, which significantly decreases the completeness of alignment. NETAL not only has low coverage but its AFS is also low. PROPER just like NETAL, aligns a few nodes, which is considered incomplete alignment. Thus to overcome these limitations, BioAlign was developed. BioAlign is a GNA algorithm which is comprised of three stages. On the first stage, BioAlign produces homology. The second stage deals with the alignment of nodes by using biological information. In the third stage, alignment is done using the neighborhood expansion method. Although BioAlign advanced in integrating diverse biological data it still faces challenges as it relies heavily on biological information and uses very little topological information.

LNA algorithms are divided into two main parts: 1) Extract and Combine strategy, and 2) Combine and Extract strategy. Algorithms that use mine and merge strategy are MORPHY, Jancura, and MODULA. MORPHY highly relies on sequence similarity which causes it to fail to identify biologically meaningful communities [40]. Jancura maps the complexity using orthology relationships which may not capture relevant interactions [41]. MODULA, similar to MORPHY, beyond sequence

similarity, may not identify meaningful communities [42]. The algorithms, AlignNemo, NetworkBlast, AlignMCL, MaWISH, and NetAligner uses the combine and extract strategy. AlignNemo uses distance-based edge formulation [43]. It does not perform well with multi-interaction proteins. AlignMCL is similar to AlignNemo, alignment graphs lead to many false positives [44]. An edge-based merging technique, NetworkBlast, results in a complex graph with a significant number of false positive interactions [45]. NetAligner consists of a dual-phase process. At first, it generates highly conserved interaction, and then in the second step, it extends the interaction of nodes [46]. It misses the large connected components. MaWISH uses the maximum induced sub-graph technique [47]. It reduces false positives but struggles with identifying large connected components. GSLAlign [48], a most recent developed local network alignment algorithm which consists of two steps. Firstly, it detects specific communities using GraphSAGE algorithm and in second stage it aligns the detected communities based on sequence similarity. GSLAlign results in much better AFS and coverage in comparison with LePrimAlign and MaWish.

Some of the alignment algorithms are related to both categories (LNA and GNA). These are LePrimAlign, GLAlign, and UAlign. LePrimAlign is a hybrid of both alignment, GNA and LNA, but it highly depends on global aligning and does not fully overcome the limitations of LNA [49]. GLAlign is another hybrid approach, it does not resolve the issues related to GNA or LNA [50]. On the other hand, UAlign is a unified approach. It leverages the strength of multiple aligners to improve the overall alignment quality. Due to its unified nature, UAlign allows it to balance the trade-off between global and local alignments, which leads to more biologically meaningful alignments [51]. Table 2.3 represents the alignment algorithms that use both Global and Local network alignment.

Table 2.3: Hybrid Approach

Aligners	**Alignment Approach**	**Biological Information**
LePrimAlign	Hybrid (local + Global alignment)	Sequence Similarity + Global alignment score
GLAlign	Hybrid (local + Global alignment)	Sequence Similarity + Global alignment score
UAlign	Unified Alignment	Sequence Similarity + Structural Similarity

Although various alignment algorithms have been developed, related to either GNA or LNA or both, none of them has achieved maximum functional similarity, high coverage, and better topological quality at the same time. This thesis introduces a novel method, $BioAlign_c$, that combines both LNA and GNA. It effectively balances biological information and topological information.

2.3. Biological Features

2.3.1. Protein Sequence Similarity

Proteins are made up of amino acids that combine to form a chain of sequences. These amino acids have 20 different types. Each protein has a different amino acid sequence that results in different 3D structures and functions. The resemblance between sequences of different proteins is used to identify the relationship between proteins. Numerous tools have been developed to find the similarity between two proteins i.e. BLAST [52].

2.3.2. Protein 3D Structure Similarity

The amino acids of proteins fold into distinct 3D structures. In a 3D coordinate system, each amino acid is represented as a point [53]. There are several tools developed to find 3D structure similarity between two proteins i.e. TMAlign [54].

2.3.3. Protein Secondary Structure

There are three types of secondary structures of amino acids of proteins, H, S, or L. These letters are used to identify various patterns at the level of secondary structure. Secondary structure occurs due to local interactions [55].

2.3.4. Remote Homology

Remote homology represents the connection of proteins that share similar functions but limited sequence similarity. The remotely homologous proteins play an important role in studying protein structures and their functions [56]. Remote Homology is predicted by various biological-based heuristic models using machine learning, and deep learning.

2.4. Evaluation of PPI Network

The main objective of GNA-based algorithms is to match as many biologically comparable proteins as possible. The Network 1 alignment with Network 2 should be optimized for both coverage and semantic similarity. Coverage refers to the fraction of nodes from network 1 that have corresponding alignments with nodes of network 2. Semantic similarity means that the nodes from network 1 have biological similar parts in network 2. The Average Functional Similarity (AFS) metric is used to check the similarity (biologically) between two aligned nodes. Within the scope of PPI networks, functional similarity measures the similarity based on their biological context, rather than their sequence or structure. This context includes categories such as BP and MF. The high functional similarity of alignments suggests that the matched protein pairs are comparable regarding MF and BP. Another metric, GOC, evaluates the biological similarity among two aligned PPI networks. Coverage determines the completeness of network [31].

2.5. Average Functional Similarity

To calculate the average functional similarity, we use the Resnik method, an IC-based approach that depends on specific databases, leading to varying functional similarity outcomes. Despite this, Resnik is effective in assessing protein similarity

based on their biological context. The method is implemented using GOATOOLS, which is well-suited for this type of analysis [57]. Additionally, we incorporate GO-FReG MSA, a tool that integrates Gene Ontology (GO) information with Multiple Sequence Alignment (MSA) data. This combination ensures a comprehensive analysis by leveraging both biological context and sequence alignment. The average functional similarity is computed using the equation 2.1:

$$AFS_c = \frac{1}{min(N_1, N_2)} \times \sum_{u,v \epsilon a} FS_c(u, v) \tag{2.1}$$

Here, AFS_C represents the complete alignment regarding Molecular Function (MF) or Biological Process (BP). FS_C denotes the functional similarity of a pair based on the Resnik method. Node u and v belong to the two different networks, respectively. N1 and N2 represent the number of nodes.

2.6. GO Term Consistency

Another metric used for the measurement of functional similarity is the Go-Term Consistency (GOC) [30]. The GO terms are extracted from the GO database. GOC is calculated as the proportion of the common GO terms to the total GO terms across both sets, as shown in Eq. 2.2:

$$GOC_{u,v} = \frac{GO(u) \cap GO(v)}{GO(u) \cup GO(v)} \tag{2.2}$$

In this formula, GO(*u*) and GO(*v*) are GO terms associated with pairs of proteins *u* and *v*, respectively.

A set or group of protein pairs is said to be the alignment of two networks. To determine the NGOC (Normalized GOC) for the entire alignment, the GOC values for all pairs in the finalized alignment are summed and then divided by the size of the network, as described in Eq. 2.3:

$$NGOC = \frac{1}{min(N_1, N_2)} \times \sum_{u,v \epsilon a} GOC(u,v) \quad (2.3)$$

In this formula, *a* denotes the whole group of pairs of proteins in the alignment. The maximum alignment size is determined by the size of the smaller network.

CHAPTER 3

METHODOLOGY

The methodology involves Biological similarity and topological similarity. Biological similarity includes the information related to proteins (sequence similarity and structural similarity). The biological scoring matrix is calculated based on protein sequence and/or structure similarities. Topological similarity holds the information about the structure of the network. The importance of the nodes is measured by topological features. The results from biological and topological similarity scores are combined to calculate the final alignment metric. The nodes of the networks are aligned based on the Greedy algorithm (nodes with maximum scores will be aligned first). Evaluation of the final alignment involves the assessment of the aligner regarding AFS, Coverage, and Topological Quality. Figure 3.1 represents the methodology used for this thesis.

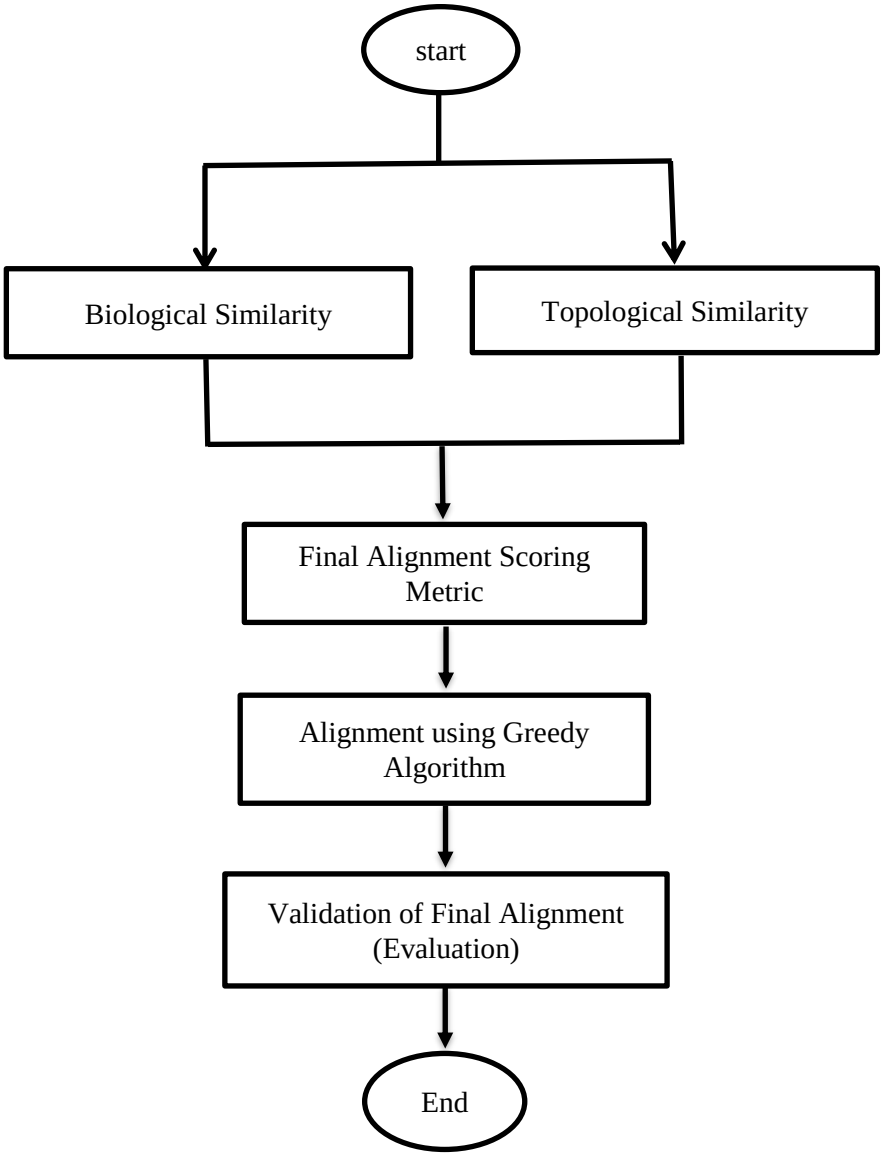


Figure 3.1: Research Methodology

We chose to build on BioAlign since it outperformed the rest of the algorithms, particularly concerning Average Functional Similarity (AFS) and Coverage. While BioAlign succeeds in these areas, it falls short in terms of topological quality because it relies heavily on biological information and integrates only a small amount of topological data. To close this gap, we modified BioAlign and included several topological approaches to improve its overall performance. Our objective was to enhance BioAlign such that it retains its high AFS and Coverage while also maximizing topological quality. By combining several methods and topological measures, we want to produce a better version of BioAlign that excels in all essential alignment metrics.

3.1. BioAlign$_c$

BioAlign, a multi-stage algorithm, initially involves aligning a maximum number of comparable proteins using 3D structural similarity, global sequence similarity, and local sequence similarity. After aligning the highly related proteins, the remaining proteins are aligned with remote homology and predicted secondary structure motifs. In the last stage, BioAlign includes topological information via Neighborhood Expansion. When aligning low-scoring nodes, BioAlign incorporates numerous biological and topological sources of information to guarantee that semantic similarity is not compromised [31].

BioAlign originally implemented Neighborhood Expansion to enhance the alignment process. However, in our approach, we explored different topological measures to get the highest Average Functional Similarity (AFS), Coverage, and topological quality. We examined degree matching and clustering coefficients as novel topological measurements. By comparing these three methods—neighborhood expansion, clustering coefficient, and degree matching—we aimed to discover which strategy was best for optimizing AFS, Coverage, and the alignments' topological quality. As a result, the clustering coefficient was used along with Neighborhood Expansion to refine the alignment while preserving the network's functional and structural integrity.

Stages of BioAlign$_c$

BioAlign$_c$ comprises three stages. Each performs their respective tasks. These stages are further discussed below:

3.2. Stage 1

The first stage of the BioAlign$_c$ approach focuses on protein alignment utilizing three crucial factors: 3D structural similarity, global sequence similarity, and local sequence similarity. To represent pairwise similarities, three matrices were generated: Matrix M1 for 3D structural similarity, Matrix M2 for global sequence similarity, and Matrix M3 for local sequence similarity. The TMAlign tool was used to calculate 3D structural similarity, which includes parameters such as average length normalization, specific transformations, and termination settings. The BLAST tool was used to calculate

global sequence similarity, which is represented as a bit-score and includes features such as gap penalties and special scoring matrices. The SWAlign tool was used to determine local sequence similarity, using a custom equation normalizing the similarity scores based on factors such as sequence identity and length.

In Stage 1, the matrices are sorted by descending similarity scores, and BioAlign_c aligns the nodes that are the most similar. The top alignment list includes node pairings with a 3D structure similarity of at least 0.8, a global sequence similarity of more than 200 (bit-score), and a local sequence similarity of more than 4. These very comparable node pairings are considered biologically significant and are used for further analysis. During the second phase, the alignment is extended to nodes with lower similarity scores, with thresholds of 0.5 for 3D structure similarity, 50 for global sequence similarity, and 2 for local sequence similarity. Furthermore, for local alignments, the sequence length must surpass the average length of 35. This thorough selection method, led by threshold optimized by grid search, guarantees that only functionally related proteins are aligned, reducing the possibility of matching functionally divergent proteins based only on high local sequences.

Grid search approaches are used to adjust the sequence and structure similarity thresholds, which range from 0.3 and 0.8 for 3D structure similarity, 30 to 80 for global sequence similarity, and 1.0 to 4.0 for local sequence similarity. This technique guarantees that the chosen thresholds are tight enough to identify functionally meaningful alignments, as indicated by the density plot analysis, which revealed a considerable reduction in candidate node pairs as a result of strict thresholds. The final alignment list is created using these criteria, guaranteeing that only the most biologically and functionally related protein pairings are selected. Algorithm 1 outlines the stage 1 mechanism.

```
procedure SEED_GENERATION
    Similarity Matrices: M1, M2, and M3                                  ▷ Input
    Similarity Thresholds: str.t, seq.t and lseq.t                       ▷ Input
    TopNodes and Seeds                                                   ▷ Output
    Seeds: Seeds ← []
    Sort M1, M2, M3 on the basis of scores
    for all node pairs (ui, uj) do
        if ui ∉ Seeds and uj ∉ Seeds and M1[ui, uj] ≥ 0.8 then
            Seeds.append(ui, uj)
        end if
    end for
    for all node pairs (ui, uj) do
        if ui ∉ Seeds and uj ∉ Seeds and M2[ui, uj] ≥ 200 then
            Seeds.append(ui, uj)
        end if
    end for
    for all node pairs (ui, uj) do
        if ui ∉ Seeds and uj ∉ Seeds and M3[ui, uj] ≥ 4.0 then
            if Alignment_Length(ui, uj) ≥ 35 then
                Seeds.append(ui, uj)
            end if
        end if
    end for
    TopNodes = Seeds                                  ▷ Top-Nodes are separated
    for all node pairs (ui, uj) do
        if ui ∉ Seeds and uj ∉ Seeds and M1[ui, uj] ≥ str.t then
            Seeds.append(ui, uj)
        end if
    end for
    for all node pairs (ui, uj) do
        if ui ∉ Seeds and uj ∉ Seeds and M2[ui, uj] ≥ seq.t then
            Seeds.append(ui, uj)
        end if
    end for
    for all node pairs (ui, uj) do
        if ui ∉ Seeds and uj ∉ Seeds and M3[ui, uj] ≥ lseq.t then
            if Alignment_Length(ui, uj) ≥ 35 then
                Seeds.append(ui, uj)
            end if
        end if
    end for
end procedure
```

Algorithm 1: Seed Generation based on Biological Scoring Matrices

3.3. Stage 2

Stage 2 of BioAlign$_c$ focuses on improving the initial alignments generated during Stage 1 by including additional layers of biological information. This step plays an essential role in improving alignment since it takes into consideration proteins that were not directly aligned in Step 1 due to weak or indirect similarities. Stage 2 is divided into three distinct sub-stages, each concentrating on a different component of biological similarity: remote homology, secondary structure motifs, and, optionally,

network-related information.

3.3.1. Stage 2-i: Remote Homology

In BioAlign$_c$ Stage 2-I, the alignment of proteins based on remote homology is carried out using particular methods and settings designed to successfully discover and align remotely related proteins. PSI-BLAST is the primary tool for detecting remote homologs, and it is essential for this process to be performed. PSI-BLAST identifies homologous proteins by repeatedly refining the search depending on the alignments identified in each round [58].

PSI-BLAST is customized with several essential parameters to improve its performance for the detection of Remote Homology. The open penalty gap is set at 5, which limits the chance of introducing a gap into the alignment. The extension penalty gap is set to 2, which determines the cost of extending gaps once they are introduced. The search results are filtered using an E-value threshold of 10, which allows for the inclusion of weaker but still biologically significant matches. The word size is set to one, influencing the sensitivity of the initial search and ensuring that even short and weak similarities are considered.

Following the identification of remote homologs, protein pairs are sorted and aligned according to the number of common homologs. This phase guarantees that proteins with a greater number of shared homologs are aligned first, hence closing gaps during the preliminary alignments. By adding these parameters and tools, BioAlign$_c$ improves the alignment process, capturing evolutionary interactions that go beyond direct sequence or structure similarity. Algorithm 2 shows the mechanism of Remote Homology.

```
1: procedure ALIGN_REMOTE_HOMOLOGS
2:     Seeds                                          ▷ Input, output of Algorithm 3
3:     N1, N2: Unique nodes of both networks                                  ▷ Input
4:     Files contain homologous proteins for each protein                     ▷ Input
5:     Seeds: ExtendedListofSeeds                                             ▷ Output
6:     for all i in N1 do
7:         L1_i = get the list of homologous proteins from File_i
8:     end for
9:     for all j in N2 do
10:        L2_j = get the list of homologous proteins from File_j
11:    end for
12:    for all i in N1 do
13:        for all j in N2 do
14:            C_{i,j} = common(L1_i, L2_j)                                   ▷ Common Homologs
15:        end for
16:    end for
17:    sort C_{i,j} on the basis of Common Homologs
18:    for all node pairs (u_i, u_j) do
19:        if u_i ∉ Seeds and u_j ∉ Seeds and C_{i,j} ≥ 1 then
20:            Seeds.append(u_i, u_i)
21:        end if
22:    end for
23: end procedure
```

Algorithm 2: Alignment using Remote Homology

3.3.2. Stage 2-ii: Secondary Structure Motifs

Stage 2-II of the BioAlign$_c$ method aligns proteins by comparing predicted secondary structure motifs, such as α-helices, β-sheets, loops, and turns. This method is crucial when proteins have low sequence similarity but share important structural motifs required for their function. The process starts with secondary structure prediction, which uses tools like PSIPRED or DSSP to analyze protein sequences and determine the types and locations of secondary structure elements [59]. Proteins with similar secondary structural motifs are then matched and aligned according to their presence and order. This sub-stage allows for the alignment of proteins that have developed divergently at the sequencing level yet retain conserved structural motifs, indicating functional commonalities.

```
procedure ALIGNMENT_USING_MOTIFS
    Seeds                                          ▷ Input, Output of Algorithm 4
    N1, N2: Unique Nodes of Both Networks                                  ▷ Input
    SS = Predicted Secondary Structure of All Proteins                     ▷ Input
    Seeds: ExtendedListofSeeds                                             ▷ Output
    for all i in N1 do
        L1_i = count_motifs from SS_i                          ▷ Motifs = HTH and HLH
    end for
    for all j in N2 do
        L2_j = count_motifs from SS_j                          ▷ Motifs = HTH and HLH
    end for
    for all i in N1 do
        for all j in N2 do
            D_{i,j} = Calculate_Difference(L1_i, L2_j)
                          ▷ Proteins of similar counts of Motifs produce small differences
        end for
    end for
    sort D_{i,j} on the basis of differences
    for all node pairs (u_i, u_j) do
        if u_i ∉ Seeds and u_j ∉ Seeds and L1_i > 0, L2_j > 0 then
            Seeds.append(u_i, u_j)
        end if
    end for
end procedure
```

Algorithm 3: Alignments using Secondary Structure Motif

3.4. Stage 3

Topological Information is a refinement step that integrates network topological features like "Clustering Coefficient" and "Neighborhood Expansion" to improve protein alignments. Because its inclusion depends upon the particular objectives of the alignment and the significance of network topology within the study's context, this sub-stage is regarded as optional. However, when structural and sequence-based alignments have been carried out and the results need to be refined by considering the locations and interactions of the proteins within their networks, it becomes quite useful.

3.4.1. Neighborhood Expansion

Network alignment techniques such as Neighborhood Expansion are used by considering the local topological environment around each node in a network to improve protein matching. In this technique, proteins that share similar local

neighborhoods, defined by direct connections or immediate neighbors, are more likely to align. This implies that proteins that interact with similar protein sets within a network are probably going to carry out similar tasks, regardless of how different their fundamental sequences or structures are. The alignment procedure catches more biologically significant interactions that could be missed by purely sequence-based methods by expanding the alignment to include proteins with comparable neighborhoods.

3.4.2. Clustering Coefficient

Another topological technique used in network alignment to determine the degree of interconnectedness within a protein's local neighborhood is the Clustering Coefficient. It measures to what extent proteins are connected to one another. A protein with a high clustering coefficient has neighbors that are highly interrelated, resulting in a tightly-knit cluster that is often associated with specific functional modules or complexes within a biological network. More significant biological insights are formed by preserving these functional modules across the aligned networks by aligning proteins with similar clustering coefficients. Clustering Coefficient is calculated as the proportion of number of actual connections and the number of possible connections. This measure indicates proteins that are part of densely connected sub-networks, that are often functionally important. We employ the clustering coefficient during the alignment process to uphold structural and functional coherence. By utilizing this approach, proteins can be aligned not only based on direct interactions but also taking into account the broader context of interactions.

Utilizing these topological characteristics helps in maintaining the functional structure of the network. Neighborhood Expansion aligns proteins that share similar interaction partners, and the Clustering Coefficient sustains the cohesion of functional modules. Consequently, these metrics yield a thorough alignment that encompasses both local and global network properties of the network, resulting in more precise and biologically relevant outcomes.

Algorithm 4: Topological Alignment - BioAlign$_c$

1: Input:

Pre-aligned pairs // Protein pairs aligned from Stage 1 and Stage 2
Remaining unaligned proteins // Unaligned proteins from both networks

2: Output:

Topologically Aligned Pairs // Final set of aligned proteins based on topological features.

3: Steps:

```
Initialize Topologically Aligned Pairs = {}

For each protein p1 in Remaining_unaligned_proteins:
For each protein p2 in Remaining_unaligned_proteins:
If (p1 and p2 are neighbors of aligned pairs in Pre-aligned pairs):
Calculate the Neighborhood Expansion score for p1 and p2
If (score is above threshold):
Add (p1, p2) to Topologically Aligned Pairs

For each (p1, p2) in Topologically Aligned Pairs:
Compute clustering coefficient C1 for p1
Compute clustering coefficient C2 for p2
If (|C1 - C2| < Threshold):
Retain (p1, p2) in Topologically Aligned Pairs
Else:
Remove (p1, p2) from Topologically Aligned Pairs

For each remaining unaligned protein u1:
For each remaining unaligned protein u2:
Check their neighborhood overlap with already aligned proteins
If (overlap exceeds threshold):
Align (u1, u2)
Add (u1, u2) to Topologically Aligned Pairs

Return Topologically Aligned Pairs
```

3.5. Dataset

Networks were collected from the HINT (High-quality INTeractromes) database. This database contains interactions that are experimentally verified [60]. BioAlign$_c$ utilizes networks that have more than 1000 interactions. Table 3.1 shows the species names, no. of edges and nodes, and the percentage of structures.

Table 3.1 HINT Dataset

Species Names	Mouse	Human	Yeast	Worm	Fly
No. of Nodes	744	10791	5036	4486	4798
No. of Edges	1229	47427	19085	11496	25679
Structural %	17	43	29	2	3

3.6. Research Strategy

The research strategy employed for this thesis comprises of multi-stage approach in order to analyze the protein interaction networks systematically. The first step is to explore the machine learning techniques to determine which of the techniques is suitable for the alignment of the network topologically. A high volume of data related to different species is collected. This contains all the homologous proteins and their interactions. This data is available in the HINT dataset. The second step is the alignment of proteins of species 1 with the proteins of species 2. This alignment is done based on biological information as well as network-related information. Refinement of the alignment is done using the neighborhood expansion technique. According to institute policy, the timeframe being followed is given below. It ensures efficient progress in developing an alignment algorithm that achieves maximum AFS, coverage, and topological quality at the same time.

3rd SEMESTER ACTIVITY: Phase 1

Activity	WEEKS															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Literature review	■	■	■	■					Mid Term						Exam Preparation	Exam week
Data collection					■	■	■									
Data preprocessing								■		■	■					
Analysis and predicting												■	■	■		

4th SEMESTER ACTIVITY: Phase 2

Activity	WEEKS															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Proposed model building	■	■						Midterm preparation	Mid Term					Exam Preparation	Exam Preparation	Exam week
Testing and Validation			■	■	■											
Model evaluation						■	■			■						
Final documentation											■	■	■			

3.7. Tools Used in this Research

- Visual Studio (VS) code was used to implement and evaluate alignment algorithms. This code is also accessible through Google Collab, Kaggle, GitHub, or any other application software that supports the development environment.
- MS Word is used for Documentation purposes.
- Mendeley used for Reference Management.
- MS Visio of Charts or Flow Diagrams

3.8. Research Ethics

3.8.1. Privacy and Confidentiality

Research in biological domains, especially those that involve protein-protein interaction networks, requires attention to the ethical use of that data. In this research, we used the HINT dataset available for public use. It has all the experimentally verified PPIs. It contains no private or sensitive information, making it appropriate for research purposes.

3.8.2. Informed Consent

The author of the PPI network algorithm (BioAlign) allowed to use their data for research purposes with appropriate citations and references.

3.8.3. Transparency Report

Table 3.2 shows the complete transparency report of this research study. It shows the method, procedure, and result of this research study.

Table 3.2: Report of Transparency

Problem	**Literature Review**	**Data Collection**	**Method**	**Result**
The problem discussed is that no alignment algorithm simultaneously gives better AFS, coverage, and topological quality.	The literature review gives brief information about the PPI network. It also shows the comparison and performance of 21 PPI alignment algorithms.	The dataset used for this study is HINT which was taken from its official website (hint.yulab.org).	A new alignment technique is introduced ($BioAlign_c$) using the PPI network algorithm (BioAlign). The new method uses different topological measures (neighborhood expansion and clustering coefficient) to align the proteins.	The result shows the superiority of the new alignment technique ($BioAlign_c$) over other aligners. Its performance showed better results than the rest of the aligners.

CHAPTER 4

RESULTS AND ANALYSIS

This section represents the final results of BioAlign and the variant of BioAlign (BioAlign$_c$). This section also contains the comparison of the results of the new variant (BioAlign$_c$) with existing aligners. These aligners include ModuleAlign, HubAlign, BEAMS, PROPER, MONACO, IBNAL, NETAL, MAGNA, Twadn and SANA. The comparison is done based on average functional similarity and coverage concerning MF and BP.

4.1. Comparison Between the Results of BioAlign and BioAlign$_c$

We named the new version as BioAlign$_c$ as it incorporated neighborhood expansion with clustering coefficient to align proteins. To differentiate it from the original BioAlign approach, this new variant, called BioAlign$_c$, indicates the incorporation of these advanced topological measures; clustering coefficient and neighborhood expansion.

The species pairs to evaluate the Average Functional Similarity (AFS) and coverage of BioAlign and BioAlign$_c$ were Mouse-Human, Mouse-Yeast, Yeast-Human, Mouse-Fly, and Mouse-Worm. For the Mouse-Human pair, BioAlign generated an AFS of 0.78 for MF (Molecular Function) and an AFS of 0.67 for BP (Biological Process). BioAlign$_c$ improved these metrics to 0.87 for AFSMF and 0.77 for AFSBP and, generated nodes alignment of 96% in both MF and BP. Similarly, for the Mouse-Yeast pair, BioAlign generated an AFSMF of 0.46 and AFSBP of 0.31 with nodes alignment of 73% for Mf and 88% for BP, however, BioAlign$_c$ showed a slight decrease in AFS at 0.43 for MF and increased AFS of 0.39 for BP which in turn showed a remarkable increase in nodes alignment of 97% for MF and 99% for BP. For the Yeast-Human pair, BioAlign generated an ASFMF of 0.55 and ASFBP of 0.41 with alignment of nodes of 63% for MF and 74% for BP. BioAlign$_c$ showed a similar

ASFMF of 0.55 but slightly decreased for ASFBP of 0.37, however, it improved the alignment of nodes up to 96% for MF and 97% for BP. For the Mouse-Fly pair, BioAlign generated an AFSMF of 0.67 and ASFBP of 0.45, with alignment of nodes of 79% for MF and 83% for BP. BioAlign$_c$ showed better results in ASFMF of 0.80 and ASFBP of 0.56 with aligned nodes of 96% in MF and 98% in BP. For the Mouse-Worm pair, BioAlign resulted in an ASFMF of 0.62 and 0.45 in ASFBP, with alignment of nodes of 73% for MF and 68% for BP. BioAlign$_c$ outperformed these results with 0.77 ASF for MF and 0.62 with ASFBP, with significantly higher alignment of nodes of 93% for MF and 96% for BP. Table 4.1 demonstrates the comparison of BioAlign and BioAlign$_c$ along with species pairs and their AFS and coverage (nodes MF and BP).

Table 4.2 demonstrates the superiority of the performance of BioAlign$_c$ over BioAlign regarding AFS (Average Functional Similarity) and coverage. BioAlign generated an average of 0.62 for AFSMF and 0.47 for AFSBP while successfully aligning 75% nodes for MF and 81% nodes for BP. BioAlign$_c$ outperformed these results. The average result of BioAlign$_c$ concerning AFSMF is 0.68 and for AFSBP is 0.54 while aligning nodes of 96% for MF and 97% for BP.

The results of these analyses show that BioAlign$_c$ is a more effective approach to network alignment than BioAlign, consistently outperforming BioAlign in terms of both functional similarity and the percentage of nodes across these different species pairings.

Table 4.1: Comparison of BioAlign and its variant BioAlign$_c$.

Pairs	Evaluation criteria	BioAlign	BioAlign$_c$
Mouse-Human	AFS_{MF}	0.78	**0.87**
	AFS_{BP}	0.67	**0.77**
	$Nodes_{MF}$	88	**96**
	$Nodes_{BP}$	92	**96**
Mouse-Yeast	AFS_{MF}	**0.46**	0.43
	AFS_{BP}	0.31	**0.39**
	$Nodes_{MF}$	73	**97**
	$Nodes_{BP}$	88	**99**
Yeast-Human	AFS_{MF}	**0.55**	**0.55**
	AFS_{BP}	**0.41**	0.37
	$Nodes_{MF}$	63	**96**
	$Nodes_{BP}$	74	**97**
Mouse-Fly	AFS_{MF}	0.67	**0.80**
	AFS_{BP}	0.45	**0.56**
	$Nodes_{MF}$	79	**96**
	$Nodes_{BP}$	83	**98**
Mouse-Worm	AFS_{MF}	0.62	**0.77**
	AFS_{BP}	0.45	**0.62**
	$Nodes_{MF}$	73	**93**
	$Nodes_{BP}$	68	**96**

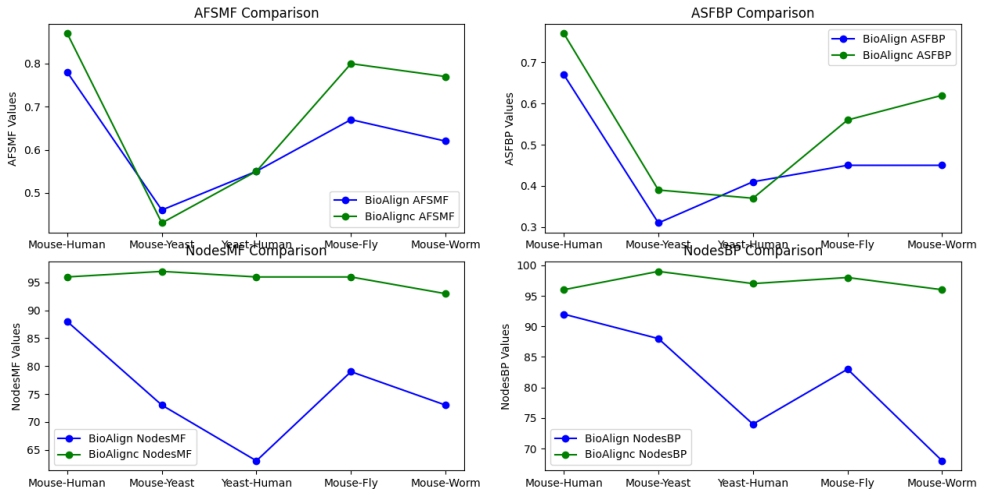


Figure 4.1: AFS and coverage of BioAlign and BioAlign$_c$.

Table 4.2: Average result of BioAlign and BioAlign$_c$.

Method	**AFS$_{MF}$**	**AFS$_{BP}$**	**Nodes$_{MF}$**	**Nodes$_{BP}$**
BioAlign	0.62	0.47	75	81
BioAlign$_c$	0.68	0.54	96	97

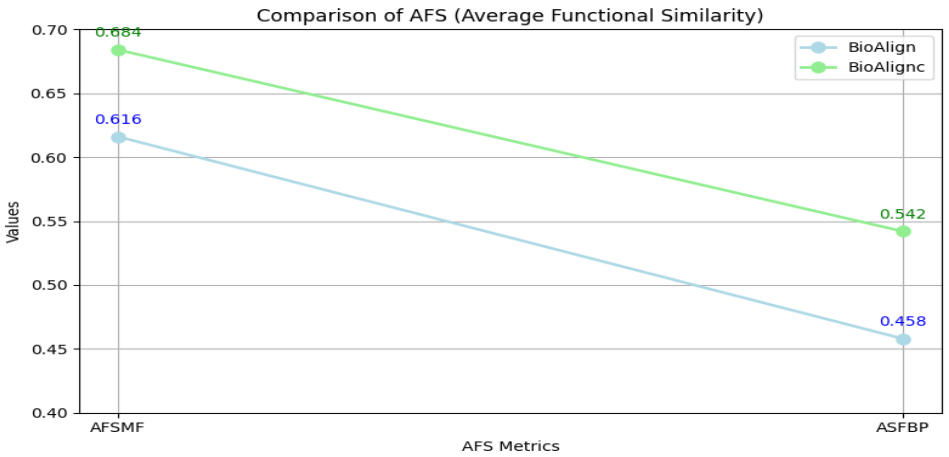


Figure 4.2: AFS (MF and BP) of BioAlign and BioAlign$_c$**.**

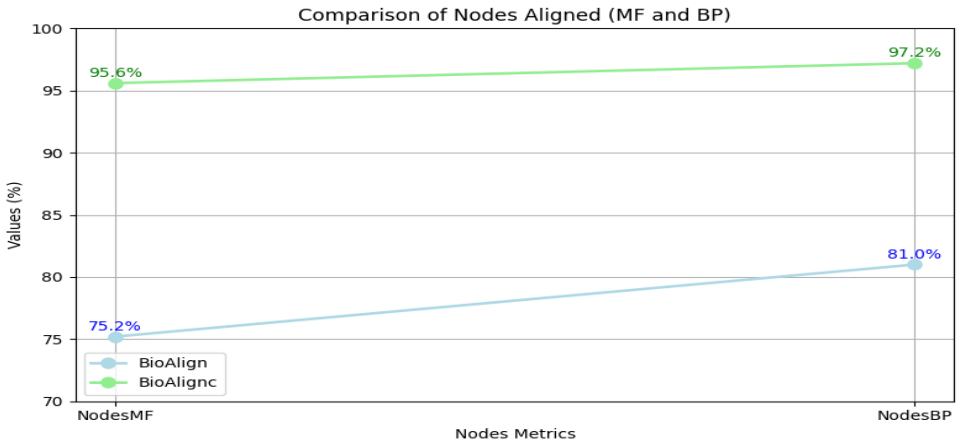


Figure 4.3: Coverage (Nodes MF and BP) of BioAlign and BioAlign$_c$.

4.2. Comparison of BioAlign$_c$ with the Existing Aligners

Table 4.3 shows the results of BioAlign$_c$ with the rest of the algorithms. We evaluated the performance of BioAlign$_c$ and existing algorithms regarding the terms AFS and coverage concerning Molecular Function and Biological Process. The existing aligners include ModuleAlign, HubAlign, BEAMS, PROPER, MONACO, IBNAL, NETAL, MAGNA, Twadn and SANA. While comparing the performance of the various network alignment algorithms in terms of AFS and coverage, BioAlign$_c$ stands out as the best method. BioAlign$_c$ achieved an AFSMF of 0.68 and an AFSBP of 0.54 which is significantly higher than other aligners. It also excels in coverage with node alignment of 96 for MF and 97 for BP. These results demonstrate that BioAlign$_c$ has higher functional alignment accuracy and comprehensive alignment with respect to topological measures.

With an AFSMF of 0.61 and an AFSBP of 0.46, BEAMS follows BioAlign$_c$, although its node-based metrics (NodesMF: 68, NodesBP: 72) are less. While ModuleAlign performs somewhat better in topological alignment than some other approaches (NodesMF: 68, NodesBP: 76), it performs weaker in functional alignment, with an AFSMF of 0.35 and an AFSBP of 0.27. Both PROPER and HUBALIGN exhibit moderate success in balancing functional and topological alignments, with PROPER obtaining an AFSMF of 0.52 and an AFSBP of 0.40, and HUBALIGN at 0.44 and 0.32, respectively.

On the lower end, AFSMF and AFSBP values of around 0.30 are shown by techniques such as IBNAL, NETAL, SANA [61], and MAGNA, showing inferior functional alignment capabilities. They also maintain relatively low node-based measurements, indicating a less efficient topological alignment.

Despite outperforming most, Twadn's AFSMF of 0.57 and AFSBP of 0.42 still fall short of BioAlign$_c$‘s results [62]. With identical metrics of AFSMF 0.50 and AFSBP 0.37 as well as comparable node measures (MONACO having 69 nodesMF and 75 nodesBP and SAlign having 73 nodesMF and 80 nodesBP), MONACO and SAlign demonstrate balanced performance, however it is not as good as BioAlign$_c$. All facts considered, BioAlign$_c$ is the better option when compared to the other approaches since it is the most efficient approach in both functional and topological metrics.

Table 4.3: Comparison of BioAlign$_c$ with existing aligning algorithms.

Pairs	Evaluation metrics	BA_C	BE	MO	SA	MA	HA	TW	IB	SN	PR	MG	NE
Mouse-Human	AFS_{MF}	**0.87**	0.79	0.56	0.58	0.42	0.48	0.73	0.35	0.34	0.76	0.36	0.33
	AFS_{BP}	**0.77**	0.70	0.43	0.43	0.30	0.34	0.63	0.26	0.25	0.66	0.26	0.24
	$Nodes_{MF}$	96	85	82	82	74	78	86	72	75	87	76	73
	$Nodes_{BP}$	96	87	87	84	81	82	88	82	82	91	82	82
Mouse-Yeast	AFS_{MF}	0.43	**0.47**	0.43	0.40	0.31	0.36	0.44	0.29	0.30	0.38	0.29	0.31
	AFS_{BP}	**0.39**	0.32	0.31	0.27	0.23	0.25	0.31	0.21	0.22	0.25	0.21	0.22
	$Nodes_{MF}$	97	64	67	72	71	71	69	63	67	56	67	64
	$Nodes_{BP}$	99	73	77	91	88	90	79	76	84	72	83	83
Yeast-Human	AFS_{MF}	**0.55**	0.54	0.48	0.48	0.26	0.46	0.50	0.30	0.27	0.42	0.26	0.26
	AFS_{BP}	0.37	**0.40**	0.36	0.35	0.22	0.34	0.36	0.24	0.23	0.32	0.22	0.22
	$Nodes_{MF}$	96	55	59	64	60	63	60	58	61	57	59	60
	$Nodes_{BP}$	97	63	70	76	72	74	70	70	73	67	70	72
Mouse-Fly	AFS_{MF}	**0.80**	0.67	0.55	0.50	0.36	0.42	0.62	0.33	0.32	0.55	0.37	0.32
	AFS_{BP}	**0.56**	0.49	0.40	0.37	0.28	0.31	0.46	0.24	0.23	0.41	0.28	0.23
	$Nodes_{MF}$	96	74	73	73	66	67	77	58	67	69	63	57
	$Nodes_{BP}$	98	79	80	80	74	76	82	58	76	77	62	60
Mouse-Worm	AFS_{MF}	**0.77**	0.58	0.49	0.56	0.41	0.49	0.56	0.30	0.32	0.52	0.31	0.29
	AFS_{BP}	**0.62**	0.41	0.34	0.41	0.30	0.37	0.41	0.25	0.25	0.39	0.35	0.24
	$Nodes_{MF}$	93	67	66	75	71	73	68	62	59	61	64	62
	$Nodes_{BP}$	96	63	62	69	66	67	64	70	57	56	76	72

Figure 4.4: Comparison of aligners with respect to AFS and coverage.

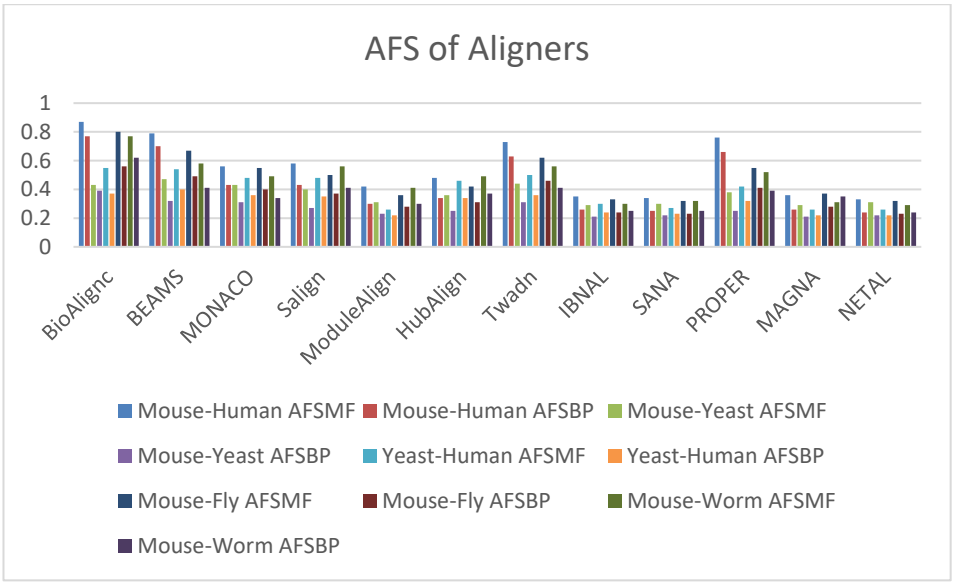


Figure 4.5: Average Functional Similarity (AFS) of alignment algorithms.

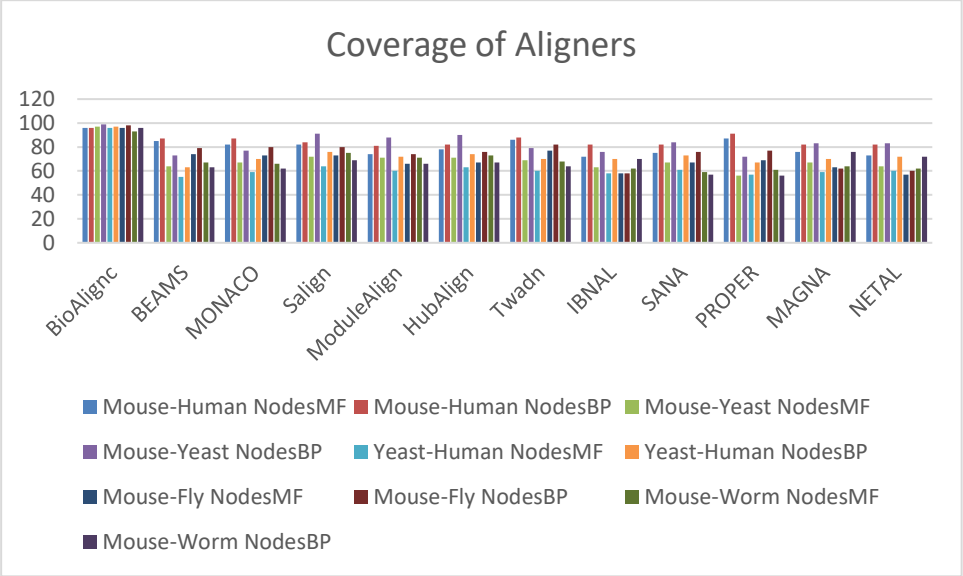


Figure 4.6: Coverage of the alignment algorithms.

Table 4.4: Average result of aligners.

Method	AFS_{MF}	AFS_{BP}	$Nodes_{MF}$	$Nodes_{BP}$
$BioAlign_c$	0.68	0.54	96	97
SAlign	0.50	0.37	73	80
BEAMS	0.61	0.46	68	72
MODULEALIGN	0.35	0.27	68	76
PROPER	0.52	0.40	66	73
HUBALIGN	0.44	0.32	70	77
IBNAL	0.31	0.24	63	71
NETAL	0.30	0.23	63	74
SANA	0.31	0.24	66	74
MAGNA	0.32	0.24	66	75
Twadn	0.57	0.42	72	77
MONACO	0.50	0.37	69	75

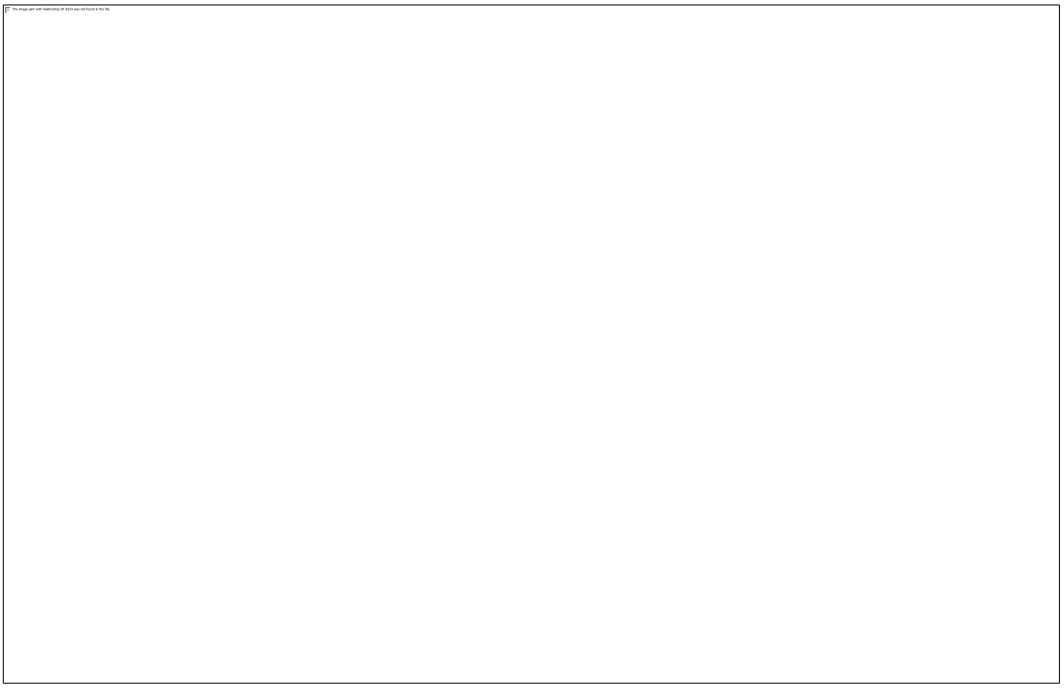


Figure 4.7: Comparison of BioAlign$_c$ with other aligners.

4.3. Enrichment Analysis

As previously mentioned, the best alignment algorithm typically produces alignments with high Average Functional Similarity (AFS) and coverage. To evaluate the excellence of these alignment results by determining the percentage of nodes aligned whose AFS is greater than these two limits, 0.50 and 0.75. The percentage of aligned nodes attaining AFS above these limits is shown in Figures 4.8 and 4.9.

Compared to other aligners, results show that BioAlign$_c$ successfully aligned more nodes with AFS above 0.50 and 0.75, suggesting greater accuracy in protein pair alignment. Results from BioAlign$_c$, BEAMS, and PROPER for the Mouse-Human pair were roughly comparable. BioAlign$_c$ marginally outperformed BEAMS and Twadn in the Mouse-Yeast pair. Figures 4.10 and 4.11 show only aligners with greater AFS than 0.50 and 0.75. These aligners whose AFS is greater than 0.50 that provide maximum coverage are BioAlign$_c$, BEAMS, PROPER, and twadn. Similarly, aligners that have greater AFS than 0.75 with maximum coverage are BioAlign$_c$, BEAMS, and PROPER. BioAlign$_c$ performed noticeably better than the current aligners for every other pair. When taking into account an AFS of more than 0.90,

BioAlign$_c$ displayed a similar pattern and aligned a greater number of nodes with an AFS over 0.75.

These results indicate that, in comparison to other methods, BioAlign$_c$ consistently aligned a higher number of nodes with high AFS. This suggests that BioAlign$_c$ can greatly assist in identifying functionally similar proteins among various species and is especially successful in identifying highly similar groups of proteins across different networks.

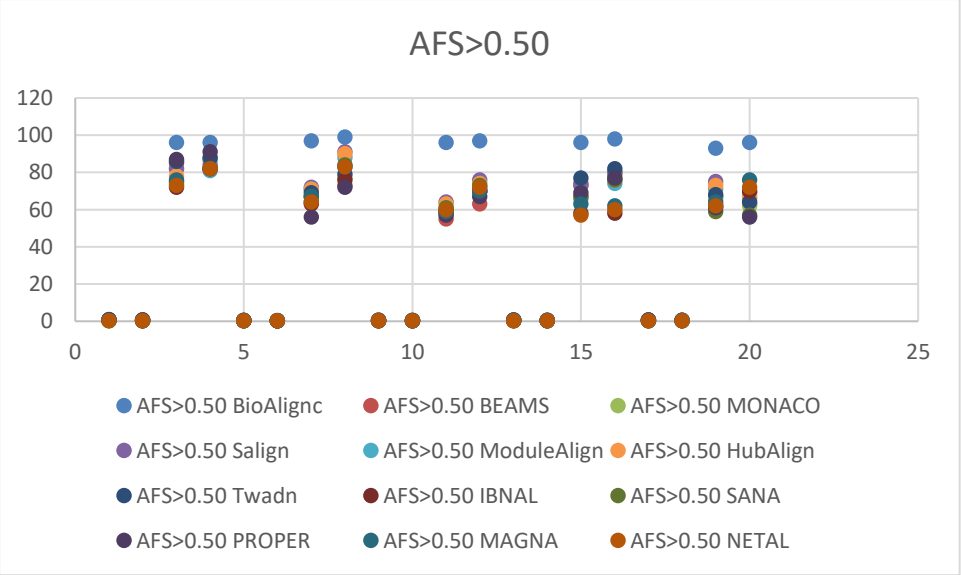


Figure 4.8: Alignment algorithms whose AFS is greater than 0.50.

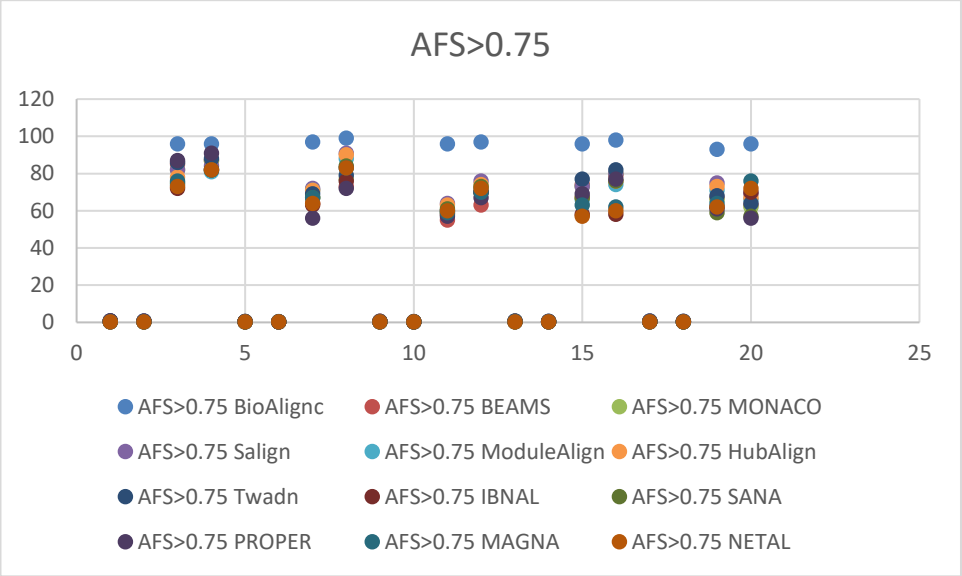


Figure 4.9: Alignment algorithms whose AFS is greater than 0.75.

Figure 4.10: Aligners whose AFS is greater than 0.50 for coverage.

Figure 4.11: Aligners whose AFS is greater than 0.75 for coverage.

4.4. Setting of Parameter

A range of thresholds for 3D structure similarity, global sequence similarity, and local sequence similarity were examined in order to maximize the quality of the alignment. While looser standards had the opposite impact, stricter thresholds improved semantic similarity while decreasing the percentage of aligned nodes. After extensively investigating, it was discovered that a 3D structure similarity threshold greater than 0.5, a global sequence similarity threshold greater than 50 (bit-score), and a local sequence similarity threshold greater than 2.0, produced the best balance between semantic similarity and coverage. Tested criteria for 3D structural similarity ranged from 0.3 to 0.8. Tests were conducted on global and local sequence similarity levels, ranging from 30 to 80 and 1.0 to 4.0, respectively.

4.5. Execution time of Aligners

Table 4.12 represents the total time of execution of all aligning algorithms based on five pairs of networks. The execution time of $BioAlign_c$ is similar to BioAlign. As compared to PROPER, MONACO, and twadn, $BioAlign_c$ execution time was slightly higher as it uses more sources of information. Meanwhile, $BioAlign_c$

execution time was less in comparison with BEAMS, HubAlign, SAlign, SANA, ModuleAlign, and MAGNA.

Figure 4.12: The Execution time of Aligners

In above figure 4.12, the average execution time of aligners with respect to five pairs are given. The execution time of PROPER is 3 seconds, Twadn is 5 seconds, and ModuleAlign is 26 seconds. As this figure shows, the execution time of BioAlign$_C$ and BioAlign is similar and the total time is 48 seconds. MONACO is 30 seconds. The execution time of MAGNA, SANA, and BEAMS is 58, 60, and 54 seconds. At the last, HubAlign and SAlign execution time is 74 and 88 seconds.

CHAPTER 5

CONCLUSION AND FUTURE WORK

Conclusion

Numerous PPI network alignment techniques have been developed to date, but it has been difficult to achieve a balance between topological quality, functional similarity, and coverage. While some current algorithms prioritize topological quality over functional similarity, several excel in functional similarity but fall short in coverage. The most innovative approach called BioAlign, has achieved great progress by generating alignments with high functional similarity and coverage. It is still not able to provide alignments with the best topological quality, though.

In order to increase overall alignment quality, this thesis introduces BioAlign$_c$, an improved version of BioAlign that incorporates further topological measurements. Specifically, BioAlign$_c$ was created to address the shortcomings of earlier algorithms by integrating a multi-stage alignment procedure that integrates topological and biological information. BioAlign$_c$ accomplishes a more balanced and thorough alignment by utilizing 3D structure, global and local sequence similarities, remote homology, predicted secondary structure motifs, and topological factors like neighborhood expansion and clustering coefficient.

By greatly improving topological quality in addition to optimizing functional similarity and coverage, BioAlign$_c$ outperforms the previous version, BioAlign. With the inclusion of carefully selected topological metrics, BioAlign$_c$ produces alignments that are more accurate, robust, and reflective of the actual biological interactions between proteins in various species.

Overall, BioAlign$_c$ effectively addresses the trade-offs between functional similarity, coverage, and topological quality, representing significant improvements in PPI network alignment. This makes it an effective technique for revealing functional modules that are conserved across species and expanding our

knowledge of complex biological networks.

Future Work

This thesis demonstrates how the integration of novel topological measures can lead to more reliable and insightful alignments, paving the pathway forward in this domain. Although BioAlign$_c$ is a state-of-the-art algorithm, there are still several areas that need further exploration.

1. Disease Analysis and Protein Function Prediction

BioAlign$_c$ can be used to identify the proteins that are involved in various diseases by applying PPI network models of organisms with human networks. Such as aligning yeast or mouse networks with human networks which helps in predicting the unknown proteins in human diseases, diseases like cancer, or any other autoimmune diseases.

2. Cross-Species Study

Diseases that transfer from animals to humans, such as COVID-19 or Ebola, are a significant threat to human beings. BioAlign$_c$ could be applied to study the relation between the species. BioAlign$_c$ can align the PPI networks between humans and species like bats, rodents, or birds, which could provide critical insight into how proteins involved in viral infections behave across species.

3. Addition of other topological factors

There are further topological factors like edge density, betweenness centrality, etc. that need to be explored for the improvement of network structure. Advanced topological factors need to be studied. With the integration of new and advanced topological measures, the alignment process could be enhanced and its topological quality will be significantly better than the others.

4. Refinement of BioAlign$_c$

For future work on this research, it is essential to enhance BioAlign$_c$ capacity so that it could handle more complex data. Future work could focus on optimizing the BioAlign$_c$ performance on a larger dataset. It is one of the drawbacks of these PPI network algorithms that if they are performing well with respect to accuracy, they lack in efficiency and when they have maximum efficiency, they lack in providing maximum accuracy. BioAlign$_c$ is a method that provides maximum accuracy and efficiency as compared to the other algorithms. Meanwhile, there are still a lot of ways that it can improve its performance even more.

REFERENCES

[1] X. Meng, J. Xiang, R. Zheng, F. X. Wu, and M. Li, "DPCMNE: Detecting Protein Complexes From Protein-Protein Interaction Networks Via Multi-Level Network Embedding," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 19, no. 3, pp. 1592–1602, 2022, doi: 10.1109/TCBB.2021.3050102.

[2] M. Menor-Flores and M. A. Vega-Rodríguez, "Boosting-based ensemble of global network aligners for PPI network alignment," *Expert Syst. Appl.*, vol. 230, no. May, p. 120671, 2023, doi: 10.1016/j.eswa.2023.120671.

[3] M. Milano, C. Zucco, M. Settino, and M. Cannataro, "An Extensive Assessment of Network Embedding in PPI Network Alignment," *Entropy*, vol. 24, no. 5, 2022, doi: 10.3390/e24050730.

[4] G. I. López-Cortés *et al.*, "The Spike Protein of SARS-CoV-2 Is Adapting Because of Selective Pressures," *Vaccines*, vol. 10, no. 6, 2022, doi: 10.3390/vaccines10060864.

[5] M. R. Hasan, B. K. Paul, K. Ahmed, and T. Bhuyian, "Design protein-protein interaction network and protein-drug interaction network for common cancer diseases: A bioinformatics approach," *Informatics Med. Unlocked*, vol. 18, p. 100311, 2020, doi: 10.1016/j.imu.2020.100311.

[6] J. E. Tomkins and C. Manzoni, "Advances in protein-protein interaction network analysis for Parkinson's disease," *Neurobiol. Dis.*, vol. 155, no. March, p. 105395, 2021, doi: 10.1016/j.nbd.2021.105395.

[7] E. N. A, K. B. B, M. R. C, and M. D. D, "A novel link prediction algorithm for protein-protein interaction networks by attributed graph embedding," *A Nov. link Predict. algorithm protein-protein Interact. networks by Attrib. graph Embed.*, 2021, [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0010482521005667

[8] S. S. & H. S. Kanchan Jha, "Prediction of protein–protein interaction using graph neural networks," *Predict. protein–protein Interact. using graph neural networks*, 2022.

[9] K. B. Giulia Muzio, Leslie O'Bray, "Biological network analysis with deep learning," *Biol. Netw. Anal. with Deep Learn.*, 2020.

[10] Y.-D. C. Steven Wang, Runxin Wu, Jiaqi Lu, Yijia Jiang, Tao Huang, "Protein-protein interaction networks as miners of biological discovery," *Protein-protein Interact. networks as miners Biol. Discov.*, 2020.

[11] P. C. P. castel@ucsf. ed. · A. H.-M. · Y. K. · B. P. S. · F. McCormick1, "DoMY-Seq: A yeast two-hybrid–based technique for precision mapping of protein–protein interaction motifs," *DoMY-Seq A yeast two-hybrid–based Tech. Precis. Mapp. protein–protein Interact. motifs*, 2021.

[12] T. M. 1ORCID andRocco S. 4 Rosa Terracciano 1,*, Mariaimmacolata Preianò 2, Annalisa Fregola 2, Corrado Pelaia 3ORCID, "Mapping the SARS-CoV-2–Host Protein–Protein Interactome by Affinity Purification Mass Spectrometry and Proximity-Dependent Biotin Labeling: A Rational and Straightforward Route to Discover Host-Directed Anti-SARS-CoV-2 Therapeutics," *Mapp. SARS-CoV-2–Host Protein–Protein Interactome by Affin. Purif. Mass Spectrom. Prox. Biot. Labeling A Ration. Straightforward Route to Discov. Host-Directed Anti-SARS-CoV-2 Ther.*, 2021.

[13] M. K. 1ORCID andKai K. K. 1 Maximilian Zeidler 1, Alexander Hüttenhofer 2, "Intragenic MicroRNAs Autoregulate Their Host Genes in Both Direct and Indirect Ways—A Cross-Species Analysis," *Intragenic MicroRNAs Autoregulate Their Host Genes Both Direct Indirect Ways—A Cross-Species Anal.*, 2020.

[14] X. Y. Q. D. P. Q. B. V. T. W. L. L. Dai, “CETSA Feature Based Clustering for Protein Outlier Discovery by Protein-to-Protein Interaction Prediction,” *CETSA Featur. Based Clust. Protein Outlier Discov. by Protein-to-Protein Interact. Predict.*, 2022.

[15] B. L. Zhourun Wu, Qing Liao, “A comprehensive review and evaluation of computational methods for identifying protein complexes from protein–protein interaction networks,” *A Compr. Rev. Eval. Comput. methods identifying protein complexes from protein–protein Interact. networks*, 2019.

[16] K. B. A, E. N. B, R. P. mohammadiani C, and Y. L. A, “Spectral clustering on protein-protein interaction networks via constructing affinity matrix using attributed graph embedding,” *Spectr. Clust. protein-protein Interact. networks via Constr. Affin. matrix using Attrib. graph Embed.*, 2021.

[17] R. Tang, “Improved Dynamic PPI Network Construction and Application of Data Mining in Computer Artificial Intelligence Systems,” *Improv. Dyn. PPI Netw. Constr. Appl. Data Min. Comput. Artif. Intell. Syst.*, 2022.

[18] 2ORCID Marianna Milano 1, 2,†ORCID, Giuseppe Agapito 2, 3,*,†ORCID andMario Cannataro 1, “Challenges and Limitations of Biological Network Analysis,” *Challenges Limitations Biol. Netw. Anal.*, 2022.

[19] J. Z. Z. X. X. W. W. Guan and J. Guan3*, “Protein Function Prediction Based on PPI Networks: Network Reconstruction vs Edge Enrichment,” *Protein Funct. Predict. Based PPI Networks Netw. Reconstr. vs Edge Enrich.*, 2021.

[20] L. H. S. Y. X. L. H. Y. K. S. M. Zhou, “A Distributed Framework for Large-scale Protein-protein Interaction Data Analysis and Prediction Using MapReduce,” *A Distrib. Framew. Large-scale Protein-protein Interact. Data Anal. Predict. Using MapReduce*, 2022.

[21] J. B. Edwin K. Silverman, Harald H. H. W. Schmidt, Eleni Anastasiadou, Lucia Altucci, Marco Angelini, Lina Badimon, Jean-Luc Balligand, Giuditta Benincasa, Giovambattista Capasso, Federica Conte, Antonella Di Costanzo, Lorenzo Farina, Giulia Fiscon, Laurent Gat, “Molecular networks in Network Medicine: Development and applications,” *Mol. networks Netw. Med. Dev. Appl.*, 2020.

[22] B. Khorsand, A. Savadi, and M. Naghibzadeh, “SARS-CoV-2-human protein-protein interaction network,” *SARS-CoV-2-human protein-protein Interact. Netw.*, 2020.

[23] S. Hashemifar, J. Ma, H. Naveed, S. Canzar, and J. Xu, “ModuleAlign: Module-based global alignment of protein-protein interaction networks,” *Bioinformatics*, vol. 32, no. 17, pp. i658–i664, 2016, doi: 10.1093/bioinformatics/btw447.

[24] S. Hashemifar and J. Xu, “HubAlign: An accurate and efficient method for global alignment of protein-protein interaction networks,” *Bioinformatics*, vol. 30, no. 17, pp. 438–444, 2014, doi: 10.1093/bioinformatics/btu450.

[25] H. M. Woo and B. J. Yoon, “MONACO: Accurate biological network alignment through optimal neighborhood matching between focal nodes,” *Bioinformatics*, vol. 37, no. 10, pp. 1401–1410, 2021, doi: 10.1093/bioinformatics/btaa962.

[26] E. Kazemi, H. Hassani, M. Grossglauser, and H. Pezeshgi Modarres, “PROPER: Global protein interaction network alignment through percolation matching,” *BMC Bioinformatics*, vol. 17, no. 1, pp. 1–16, 2016, doi: 10.1186/s12859-016-1395-9.

[27] A. E. A. M. J. Kalita, “Index-Based Network Aligner of Protein-Protein Interaction Network,” *Index-Based Netw. Aligner Protein-Protein Interact. Networks*, 2018, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7575732

[28] V. Vijayan, V. Saraph, and T. Milenković, “MAGNA11: Maximizing accuracy in global network alignment via both node and edge conservation,” *Bioinformatics*, vol. 31, no. 14, pp. 2409–2411, 2015, doi: 10.1093/bioinformatics/btv161.

[29] B. Neyshabur, A. Khadem, S. Hashemifar, and S. S. Arab, “NETAL: A new graph-based method for global alignment of protein-protein interaction networks,” *Bioinformatics*, vol. 29, no. 13, pp. 1654–1662, 2013, doi: 10.1093/bioinformatics/btt202.

[30] U. Ayub, I. Haider, and H. Naveed, “SAlign–a structure aware method for global PPI network alignment,” *BMC Bioinformatics*, vol. 21, no. 1, pp. 1–18, 2020, doi: 10.1186/s12859-020-03827-5.

[31] U. Ayub and H. Naveed, “BioAlign: An Accurate Global PPI Network Alignment Algorithm,” *Evol. Bioinforma.*, vol. 18, 2022, doi: 10.1177/11769343221110658.

[32] F. Alkan and C. Erten, “BEAMS: Backbone extraction and merge strategy for the global many-to-many alignment of multiple PPI networks,” *Bioinformatics*, vol. 30, no. 4, pp. 531–539, 2014, doi: 10.1093/bioinformatics/btt713.

[33] 2 and Ali Asghar Peyvandi Nahid Safari-Alighiarloo, 1 Mohammad Taghizadeh, 2 Mostafa Rezaei-Tavirani, 1 Bahram Goliaei, "Protein-protein interaction networks (PPI) and complex diseases," *Protein-protein Interact. networks complex Dis.*, 2014.

[34] P. S. D. K. B. J. K. Kalita, "Centrality analysis in PPI networks," *Cent. Anal. PPI networks*, 2016.

[35] M. L. Y. L. J. W. F.-X. W. Y. Pan, "A Topology Potential-Based Method for Identifying Essential Proteins from PPI Networks," *A Topol. Potential-Based Method Identifying Essent. Proteins from PPI Networks*, 2016.

[36] M. Huang, N. D. Shah, and L. Yao, "Evaluating global and local sequence alignment methods for comparing patient medical records," *BMC Med. Inform. Decis. Mak.*, vol. 19, no. Suppl 6, pp. 1–13, 2019, doi: 10.1186/s12911-019-0965-y.

[37] C.-Y. M. A and C.-S. Liao, "A review of protein–protein interaction network alignment: From pathway comparison to global alignment," *A Rev. protein–protein Interact. Netw. alignment From Pathw. Comp. to Glob. alignment*, 2020.

[38] M. M. P. H. Guzzi, "Improving the Robustness of Local Network Alignment: Design and Extensive Assessmentof a Markov Clustering-Based Approach," *Improv. Robustness Local Netw. Alignment Des. Extensive Assessmentof a Markov Clust. Approach*, 2016.

[39] V. Saraph and T. Milenković, "MAGNA: Maximizing Accuracy in Global Network Alignment," *Bioinformatics*, vol. 30, no. 20, pp. 2931–2940, 2014, doi: 10.1093/bioinformatics/btu409.

[40] J. L. & M. K. Sinan Erten, Xin Li, Gurkan Bebek, "Phylogenetic analysis of modularity in protein interaction networks," *Phylogenetic Anal. Modul. protein Interact. networks*, 2009.

[41] E. C. S. P. & E. M. Pavol Jancura, Eleftheria Mavridou, "A methodology for detecting the orthology signal in a PPI network at a functional complex level," *A Methodol. Detect. orthology signal a PPI Netw. a Funct. complex Lev.*, 2012.

[42] P. H. G. P. V. S. R. J. K. Kalita, "MODULA: A network module based local protein interaction network alignment method," *Modul. A Netw. Modul. based local protein Interact. Netw. alignment method*, 2015.

[43] C. G. Giovanni Ciriello , Marco Mina , Pietro H. Guzzi, Mario Cannataro, "AlignNemo: A Local Network Alignment Method to Integrate Homology and Topology," *AlignNemo A Local Netw. Alignment Method to Integr. Homol. Topol.*, 2012.

[44] M. M. P. H. Guzzi, "AlignMCL: Comparative analysis of protein interaction networks through Markov clustering," *AlignMCL Comp. Anal. protein Interact. networks through Markov Clust.*, 2012.

[45] R. S. Maxim Kalaev, Mike Smoot, Trey Ideker, "NetworkBLAST: comparative analysis of protein networks," *NetworkBLAST Comp. Anal. protein networks*, 2008.

[46] R. A. Pache, A. Céol, and P. Aloy, "NetAligner-a network alignment server to compare complexes, pathways and whole interactomes," *Nucleic Acids Res.*, vol. 40, no. W1, pp. 157–161, 2012, doi: 10.1093/nar/gks446.

[47] and A. G. Mehmet Koyutürk, Yohan Kim, Umut Topkara, Shankar Subramaniam, Wojciech Szpankowski, "Pairwise Alignment of Protein Interaction Networks," *Pairwise Alignment Protein Interact. Networks*, 2006.

[48] U. A. & H. Naveed, "GSLAlign: community detection and local PPI network alignment," *GSLAlign community Detect. local PPI Netw. alignment*, 2024.

[49] S. Maskey and Y. R. Cho, "LePrimAlign: Local entropy-based alignment of PPI networks to predict conserved modules," *BMC Genomics*, vol. 20, no. Suppl 9, pp. 1–12, 2019, doi: 10.1186/s12864-019-6271-3.

[50] M. M. P. H. G. M. Cannataro, "GLAlign: A Novel Algorithm for Local Network Alignment," *GLAlign A Nov. Algorithm Local Netw. Alignment*, 2019.

[51] N. P. N Malod-Dognin, K Ban, "Unified alignment of protein-protein interaction networks," *Unified alignment protein-protein Interact. networks*, 2017.

[52] Y.-H. C. Cameron L M Gilchrist, Thomas J Booth, Bram van Wersch, Liana van Grieken, Marnix H Medema, "cblaster: a remote search tool for rapid identification and visualization of homologous gene clusters," *cblaster a Remote search tool rapid Identif. Vis. Homol. gene Clust.*, 2021.

[53] V. A. Felix Gabler, Seung-Zin Nam, Sebastian Till, Milot Mirdita, Martin Steinegger, Johannes Söding, Andrei N. Lupas, "Protein Sequence Analysis Using the MPI Bioinformatics Toolkit," *Protein Seq. Anal. Using MPI Bioinforma. Toolkit*, 2020.

[54] T. I. Shuichiro Makigaki, "Sequence alignment using machine learning for accurate template-based

protein structure prediction," *Seq. alignment using Mach. Learn. accurate template-based protein Struct. Predict.*, 2020.

[55] Y. Wang, H. Mao, and Z. Yi, "Protein secondary structure prediction by using deep learning method," *Protein Second. Struct. Predict. by using Deep Learn. method*, 2017.

[56] B. L. Junjie Chen, Mingyue Guo, Xiaolong Wang, "A comprehensive review and comparison of different computational methods for protein remote homology detection," *A Compr. Rev. Comp. Differ. Comput. methods protein Remote Homol. Detect.*, 2018.

[57] P. F. & H. T. D. V. Klopfenstein, Liangsheng Zhang, Brent S. Pedersen, Fidel Ramírez, Alex Warwick Vesztrocy, Aurélien Naldi, Christopher J. Mungall, Jeffrey M. Yunes, Olga Botvinnik, Mark Weigel, Will Dampier, Christophe Dessimoz, "GOATOOLS: A Python library for Gene Ontology analyses," *GOATOOLS A Python Libr. Gene Ontol. Anal.*, 2018.

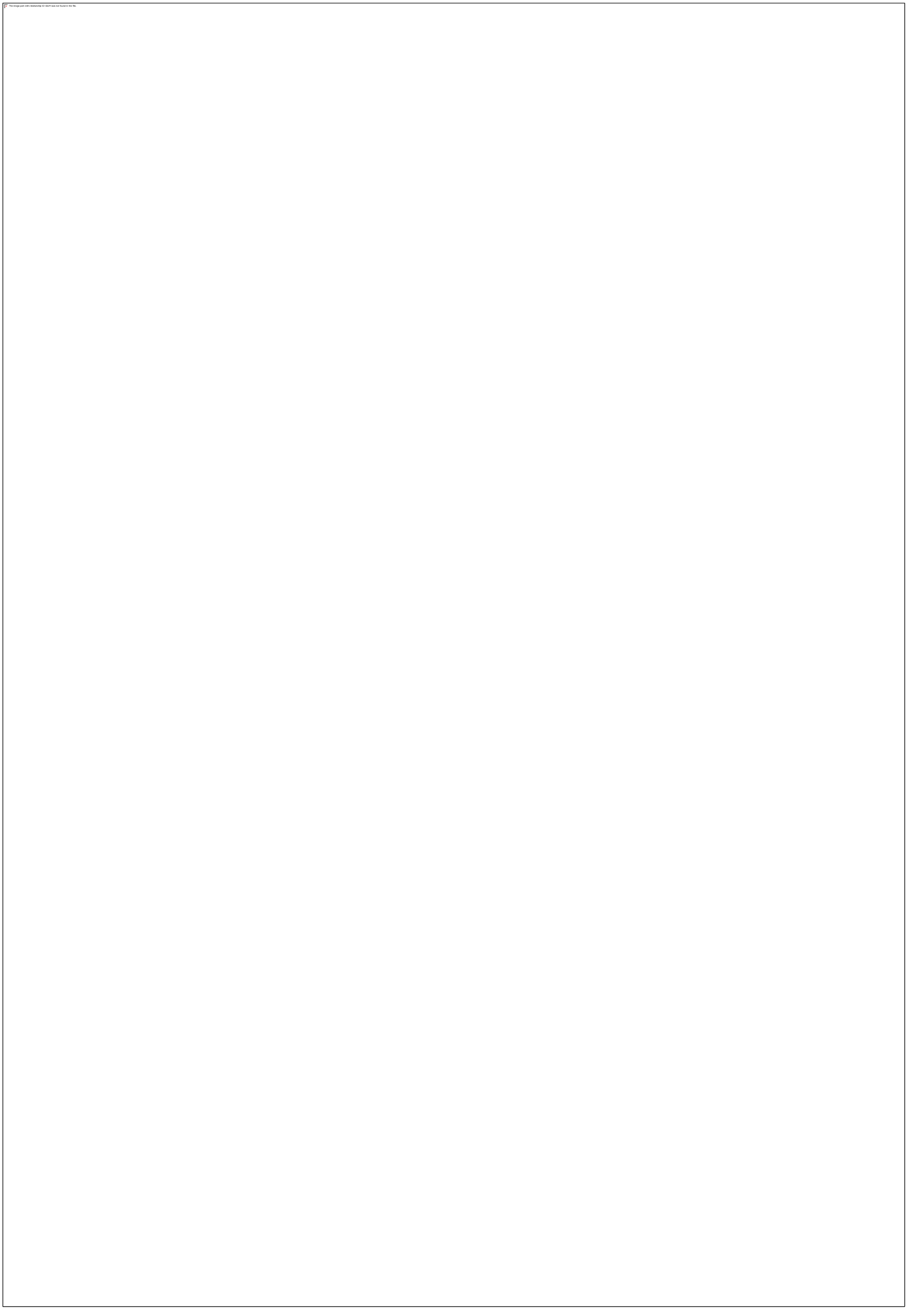
[58] H. A. N. J. J. S. M. Shovan, "Improved Identification Performance of Lysine Glycation PTM using PSI-BLAST," *Improv. Identif. Perform. Lysine Glycation PTM using PSI-BLAST*, 2020.

[59] A. G. de Brevern, "Impact of protein dynamics on secondary structure prediction," *Impact protein Dyn. Second. Struct. Predict.*, 2020.

[60] J. D. & H. Yu, "HINT: High-quality protein interactomes and their applications in understanding human disease," *HINT High-quality protein interactomes their Appl. Underst. Hum. Dis.*, 2012.

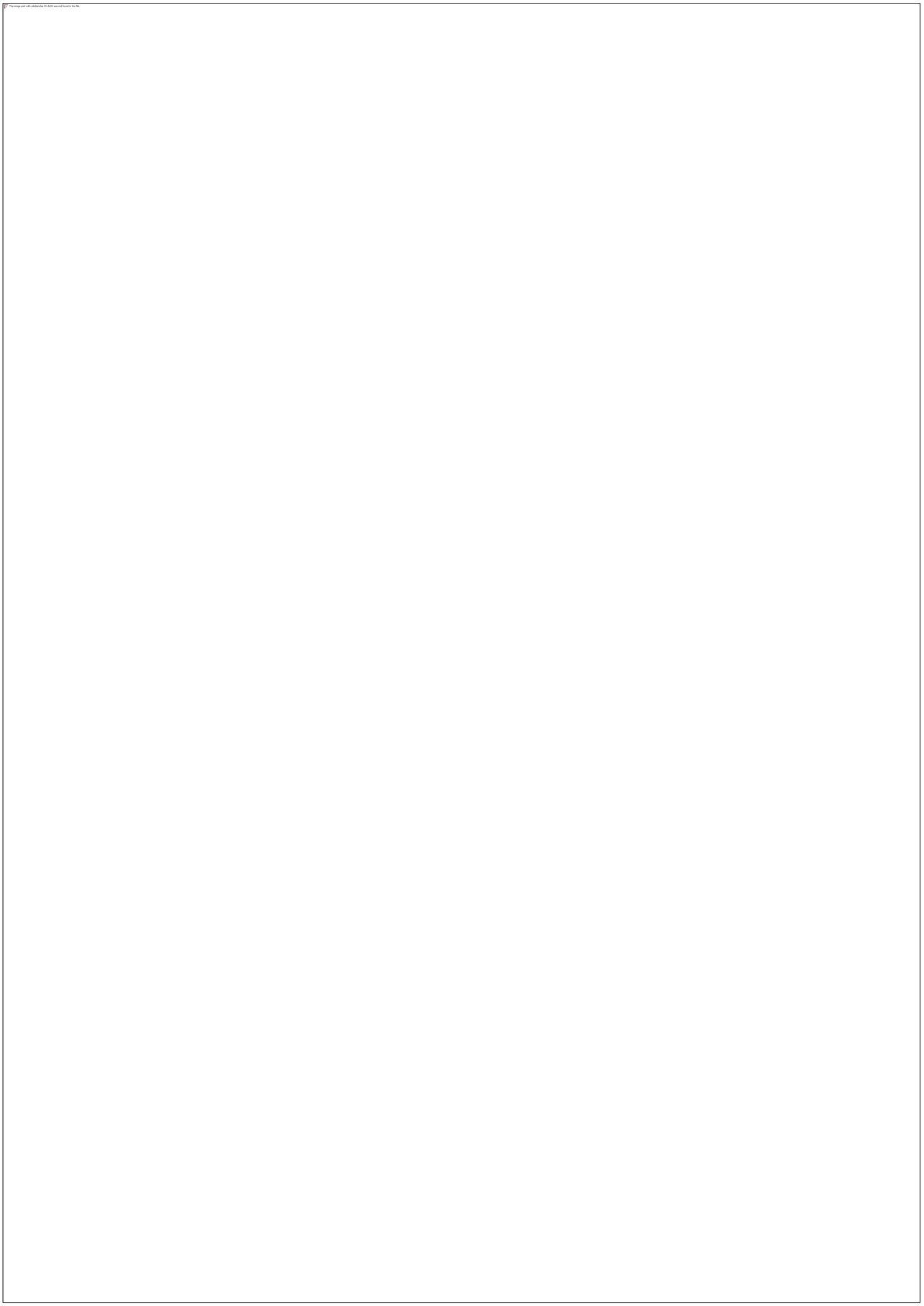
[61] S. Wang, G. R. S. Atkinson, and W. B. Hayes, "SANA: cross-species prediction of Gene Ontology GO annotations via topological network alignment," *npj Syst. Biol. Appl.*, vol. 8, no. 1, 2022, doi: 10.1038/s41540-022-00232-x.

[62] X. S. & J. H. Yuanke Zhong, Jing Li, Junhao He, Yiqun Gao, Jie Liu, Jingru Wang, "Twadn: an efficient alignment algorithm based on time warping for pairwise dynamic networks," *Twadn an Effic. alignment algorithm based time warping pairwise Dyn. networks*, 2020.


The image part with relationship ID rId24 was not found in the file.


The image part with relationship ID rId24 was not found in the file.


The image part with relationship ID rId24 was not found in the file.


The image part with relationship ID rId24 was not found in the file.