MORUF ADEDEJI ADEAGBO

PGREMLIN: PREDICTION OF LIGANDS USING CONSERVED SUBSTRUCTURES IN PROTEIN-PROTEIN INTERACTION

Dissertation submitted to the Computer Science Graduate Program of the Universidade Federal de Viçosa in partial fulfillment of the requirements for the degree of *Magister Scientiae*.

Adviser: Sabrina de Azevedo Silveira

VIÇOSA - MINAS GERAIS
2022

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus

T
A228p
2022

Adeagbo, Moruf Adedeji, 1987-
pGREMLIN: prediction of ligands using conserved substructures in protein-protein interaction / Moruf Adedeji Adeagbo. - Viçosa, MG, 2022.
1 dissertação eletrônica (58 f.): il.

Texto em inglês.

Inclui apêndices.
Orientador: Sabrina de Azevedo Silveira
Dissertação (mestrado) - Universidade Federal de Viçosa, Departamento de Informática, 2022.
Referências bibliográficas: .
DOI: https://doi.org/10.47328/ufvbbt.2023.006
Modo de acesso: World Wide Web.

1. Aprendizado do computador; 2. Proteínas - Estrutura; 3. Ligações químicas; I. Silveira, Sabrina de Azevedo II. Universidade Federal de Viçosa. Departamento de Informática. Programa de Pós-Graduação em Ciência da Computação III. Título

CDD 22. ed. 006.31

Bibliotecário(a) responsável: EUZEBIO LUIZ PINTO CRB-6/3317

MORUF ADEDEJI ADEAGBO

PGREMLIN: PREDICTION OF LIGANDS USING CONSERVED SUBSTRUCTURE IN PROTEIN-PROTEIN INTERACTION INTERFACE

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, para obtenção do título de *Magister Scientiae.*

APROVADA: 21 de dezembro de 2022.

Assentimento:

Moruf Adedeji Adeagbo
Autor

Sabrina de Azevedo Silveira
Orientadora

I dedicate this to God Almighty for channeling my path thus far and showering me with good health, wisdom, knowledge, and understanding. Which of the blessings of Allah can I deny? None!!!!

This dedication also goes to those whom Allah used to make this journey easier, especially my dear wife Halimah Adebimpe and my wonderful boys AbdulHaqq Airolafin, AbdulHayy Ayinuola, and AbdulHasib Woleola, for their love, patience, prayers, support, and affection in every moment of my life endeavors. You were always by my side, even in the most difficult times. I sincerely appreciate you.

ACKNOWLEDGEMENTS

I thank God Almighty for channeling my path this far and showering me with good health, wisdom, knowledge, and understanding. Which of the blessings of Allah can I deny? None!!!!

My first appreciation goes to my parents, the late Alhaji Yekeen Adeagbo and Alhaja Silifat Adeagbo, for bringing me into the world, raising me with love and care, and channeling my path with prayers and support up to this moment. In addition, words can never express how grateful I am to my siblings for their love and constant encouragement in my pursuit of my goals. My deepest appreciation goes to those whom Allah used to make this journey smooth, especially my darling wife Halimah Adebimpe and my wonderful boys AbdulHaqq Airolafin, AbdulHayy Ayinuola, and AbdulHasib Woleola, for the love, patience, prayers, and affection shown in every moment of this stage in our lives. You were always by my side, even in the most difficult times. I really appreciate you. Also, my sincere appreciation goes to my father's in-law, Alhaji Muritala Agbaje, and his caring wife, Alhaja Agbaje, and the entire progeny of Muritala Agbaje. Your love, care, and support throughout this journey can never be quantified. This appreciation would be incomplete without thanking my wonderful and caring supervisor, Prof. Sabrina de A. Silveira, who has provided all of the necessary support, guidance, and mentorship throughout the duration of this work and shown me the path I deem fit to dedicate my research career to. I am immensely grateful. Also, I thank the management of First Technical University (TECH-U) for deeming me fit to embark on this wonderful and life-changing journey. Lastly, I thank everyone who contributed in different ways to make this event possible, especially my colleagues from Nigeria, especially Abdul, Shafiu, Emanuel, Ola, Henry, and Abimbola, and my colleagues in bioinformatics, especially Vinicius, Igor, Vagner, and Isabela.

ABSTRACT

ADEAGBO, Moruf Adedeji, M.Sc., Universidade Federal de Viçosa, December, 2022. **pGReMLIN: Prediction of Ligands Using Conserved Substructures in Protein-protein Interaction**. Adviser: Sabrina de Azevedo Silveira.

One of the sources of data for the protein dataset being explored in bioinformatics is the Protein Data Bank (PDB). Searching for similar conserved structures in biological problems was an early attempt and is still a popular method widely used by researchers. In this work, a graph-based approach named pGReMLIN is proposed, which predicts similar structures in protein-protein or protein-peptide complexes using the established conserved structural arrangements of Serine protease and BCL-2.
The protein-protein or protein-peptide complexes are downloaded from the PDB, and the interface between proteins at the atomic level is modeled as a graph, with atoms as nodes and the non-covalent interactions between atoms as edges. The computation of a similar pattern of structural arrangement in the data graph was achieved using a set of candidate pairs of data graph D(G) and query graph Q(G) based on state space representation and feasibility rules that embed equality of structure and attributes. pGReMLIN was able to compute, in the data graph, a conserved structural arrangement that represents highly conserved interactions at the specificity binding interface of trypsin and trypsin-like proteins. Also, our method was able to identify similar patterns with large pattern size in the data graph common to Serine protease and BCL-2 and as well as unique to Serine protease and BCL-2. Furthermore, pGReMLIN was able to present various large patterns of similar conserved patterns that can be further explored to determine their potency. Hence, the presence of similar patterns in our protein-protein dataset indicates the possibility that patterns found in our dataset have similar functions to the conserved structures.

Keywords: Protein-protein interaction. Graph-subgraph isomorphism. Ligands prediction.

RESUMO

ADEAGBO, Moruf Adedeji, M.Sc., Universidade Federal de Viçosa, dezembro de 2022. **pGReMLIN: Previsão de ligantes usando subestruturas conservadas na interação proteína-proteína.** Orientadora: Sabrina de Azevedo Silveira.

Uma das fontes de dados para o conjunto de dados de proteínas que está sendo explorado em bioinformática é o Protein Data Bank (PDB). A busca por estruturas conservadas semelhantes em problemas biológicos foi uma tentativa inicial e ainda é um método popular amplamente utilizado por pesquisadores. Neste trabalho, uma abordagem baseada em gráficos chamada pGReMLIN é proposta, que prediz estruturas similares em complexos proteína-proteína ou proteína-peptídeo usando os arranjos estruturais conservados estabelecidos de Serina protease e BCL-2.
Os complexos proteína-proteína ou proteína-peptídeo são baixados do PDB, e a interface entre as proteínas no nível atômico é modelada como um gráfico, com os átomos como nós e as interações não covalentes entre os átomos como arestas. O cálculo de um padrão semelhante de arranjo estrutural no gráfico de dados foi obtido usando um conjunto de pares candidatos de gráfico de dados D(G) e gráfico de consulta Q(G) com base na representação do espaço de estado e regras de viabilidade que incorporam a igualdade de estrutura e atributos . pGReMLIN foi capaz de calcular, no gráfico de dados, um arranjo estrutural conservado que representa interações altamente conservadas na interface de ligação de especificidade de tripsina e proteínas semelhantes a tripsina. Além disso, nosso método foi capaz de identificar padrões semelhantes com tamanho de padrão grande no gráfico de dados comum à serina protease e BCL-2 e também exclusivo para serina protease e BCL-2. Além disso, pGReMLIN foi capaz de apresentar vários grandes padrões de padrões conservados semelhantes que podem ser mais explorados para determinar sua potência. Portanto, a presença de padrões semelhantes em nosso conjunto de dados proteína-proteína indica a possibilidade de que os padrões encontrados em nosso conjunto de dados tenham funções semelhantes às estruturas conservadas.

Palavras-chave: Interação proteína-proteína. Isomorfismo grafo-subgrafo. previsão de ligantes.

LIST OF FIGURES

LIST OF TABLES

SUMMARY

Chapter 1

Background

Proteins are regarded as macromolecules that play major biological roles in the cell (Rasheed et al., 2020). Proteins are made up of 20 different types of amino acids attached to one another to make a long chain (Medlineplus, 2021). This culminates in the consideration of proteins as bio-molecular devices that have the tendency to recreate in the laboratory. Proteins, in general play functional, enzymatic and structural roles (Day, 2009) such as immunity, biosynthesis, transport and photosynthesis to reach the development and healthful requests of developing seedlings (Rasheed et al., 2020)**.**

1.1. Introduction

Proteins rarely act alone (Laskowski & Thornton, 2015) as their functions have the tendency to be regulated. The establishment of physical contacts between numbers of protein molecules with high specificity resulting from biochemical signals (Qvit & Crapster, 2014) is regarded as Protein–Protein Interactions (PPIs). This biochemical event is the effect of interactions between hydrophobic and hydrogen bonding, as well as electrostatic forces. The roles of PPIs in many biological processes are highly essential and can never be underestimated. PPIs regulate cells' mechanism, motivate intercellular communication, regulate the expression of genes and signal transduction (Edwards et al., 2002). Also, the information on the structuring of biological pathways and coordination from each protein function are also made possible by the PPI network (Ding & Kihara, 2019).

As system biology tends towards PPIs as one of its main objectives (Yu & Kong, 2022) (Rao et al., 2014), many approaches have been deployed by researchers to investigate PPIs. Among those methods are Biochemical, Biophysical and Theoretical, Genetic and Computational methods. Each of these methods has its own weaknesses and strengths particularly when considering the specificity and sensitivity of the method. Among those approaches is the computational method that employs network analysis using graph theory or statistical methods for the analysis of network interactions to unravel not just the individual interactions but the interactions nature of the cell or pathway (Barabasi & Oltvai, 2004). PPI networks are in the form of protein networks heterogeneously joined by their interactions which computationally analyze

PPI networks based on the arrangement (Rao et al., 2014). This arrangement mathematically takes the form of a graph comprising nodes and edges. Proteins are represented as nodes and the connection of adjacent nodes, that is, the interactions are represented as edges (Rehman et al., 2018). This approach has been acclaimed to be effective and has lower experimental cost (Yan et al., 2011). Graph theory is essential (Koutrouli et al., 2020) and advantageous in the prediction of protein structures, side-chain cluster identification in protein structures and various associated problems of protein structure identification. (Queiroz et al., 2020) employed the graph theory to model protein-protein interfaces as graphs at the atomic level to detect conserved structural arrangements that represent highly conserved interactions at the specificity binding pocket of trypsin and trypsin-like proteins from serine protease dataset. The structural arrangements of PPIs were detected using a graph-based approach with atoms representing nodes and edges represented as non-covalent interactions. The distance criteria and the physicochemical properties of atoms were used to label the nodes and edges. The process leading to the detection of conserved structure includes modeling of data acquisition, clustering analysis, and conserved substructure mining through the use of Frequent Subgraph Mining (FSM) to obtain subgraph isomorphism. This Subgraph isomorphism depicts the substructures that are embedded within the whole PPIs dataset. Biologically, this conserved substructure is the binding strength of the PPIs of trypsin and trypsin-like proteins from serine protease dataset.

One of the sources of data for the protein dataset being explored in bioinformatics is the Protein Data Bank (PDB). PDB is a database of biological data that contains over 171,916 entries of protein structure and structures of nucleic acids such as RNA, DNA, and protein-nucleic acid (Sharma & Yadav, 2022). If the already established conserved substructures can be used to search for the similar substructures at Protein Data Bank (PDB), then is there the possibility that these conserved substructures can be used to predict proteins/peptides to interact with a protein target of interest? Therefore, considering the fact that the graph modeling of PPI is acclaimed to be effective and less labor intensive compared to experimental methods and other computational methods such as text mining, there are established conserved structural arrangements of trypsin and trypsin-like proteins from the serine protease and BCL-2 datasets. Hence, this research proposes pGReMLIN, a computational strategy that

calculates similar structures in protein-protein complexes and then searches PDB for similar substructures. We hypothesize that such substructures captured can support us on detecting ligands for targets of interest.

The modelling of biological data such as molecular data using graphs is a basic task in chemoinformatics that has been on for decades. The graph representation of molecular data forms the basic and theoretical foundation of using computer aided processing of molecular data. This allows graphs to easily solve the substructure searching problem through the use of graph isomorphism techniques (Ehrlich & Rarey, 2012). Application of graph searching in solving matching problems is an early attempt and still popular method widely used by researchers (Song et al., 2022). Some of the algorithms explored in solving graph or subgraph matching are Ullmann algorithm (Gouda et al., 2022), VF2 algorithm (Kusari et al., 2022), exact and error-tolerant graph matching algorithms (Dwivedi, 2019) and probabilistic relaxation algorithms (Zhang et al., 2018) with each algorithm having their strengths and weaknesses. To find substructures in molecular networks, some of the algorithms that solve the subgraph isomorphism problem include the Ullmann and VF2 algorithms (Ehrlich & Rarey, 2012). Researchers mostly investigate subgraph isomorphism using the VF2 algorithm (Jüttner & Madarasi, 2018). Based on VF2 algorithm comparison with other algorithms, it saves time, convenient in terms of one-to-one matching and mostly used for exact graph matching (Cordella et al., 2004). While the Ullmann algorithm is a backtracking procedure that employs a relaxation-based refinement step to reduce the search space (Cibej & Mihelic, 2015), the VF2 algorithm iteratively extends a partial solution using a set of feasibility criteria to decide whether to extend or backtrack (Kusari et al., 2022). However, the VF2 algorithm outperforms the Ullman in all test cases when supplied with a favorable substructure formulation and seems to be more robust in terms of run time outliers (J. Lee et al., 2012). Also, VF2 remains the most suitable algorithm for large dataset (Dwivedi, 2020) which is also reported in the (Jüttner & Madarasi, 2018) study being the commonly used algorithm for graph subgraph isomorphism. Thus, the research question is, can these conserved substructures be used to predict proteins/peptides do interact with a protein target of interest?

1.2. Statement of the Problem

Proteins are macromolecules that play an important biological role in plant and animal cells. Biochemists are motivated by the functional roles of proteins to conduct

various in vitro laboratory research projects to improve protein functionalities and control pests and diseases in plants and animals. However, the funding required to carry out wet laboratory experiments to enhance protein functionality and reduce protein inhibitors is expensive. In this work we propose pGReMLIN, a computational strategy that calculates similar structures in protein-ligand complexes and then searches PDB for similar substructures. We hypothesize that such substructures, also called patterns, can support us in detecting ligands for targets of interest.

1.3. Justification of the Research

This proposed computational strategy will be useful in conducting *in silico* experiments for prospecting similar conserved substructures within large protein datasets, which can potentially support us on detecting ligands for targets of interest. Detecting ligands for targets of interest is relevant for laboratory experiments, for predicting drugs for animal and plant diseases, among others. To illustrate, in a recent work of our group (de Souza Gomes et al., 2022), a set of potential ligands for Covid-19 main protease (which was the target protein) was predicted. In this work, we will use 2 datasets, one dataset of proteases similar to the ones of a caterpillar wich attacks soybean plant, causing considerable damage to this crop, of which Brazil is the largest producer in the world (Botelho & Junior, 2022) (Toloi et al., 2021). The other dataset is composed of proteins similar to the structure of Mcl-1, an anti-apoptotic human protein of the BCL-2 family an anti-apoptotic human protein of the BCL-2 family. BCL-2 is composed of key regulators of programmed cell death (Krajewski et al., 1993); (Cai et al., 1998). Actually, apoptosis has been established as a critical tumor suppression mechanism (Delbridge et al., 2012). These datasets provide a real and relevant scenario which will be used to evaluate our computational strategy.

1.4. Scope of the Study

This research work is limited to search the PDB for similar conserved structures of protein-protein and protein-peptides using serine protease and BCL-2 as query patterns. More specifically, we will use a strategy developed by our group to calculate conserved substructures (graphs) on protein-protein interfaces from serine protease and BCL-2 (Queiroz et al., 2020). Then we will use the mentioned substructures as queries to search the whole PDB for patterns similar to these queries. This search will be performed by a computational strategy developed for this work.

1.5. Objectives of the Research

The research will leverage on the result of conserved structural arrangements of trypsin and trypsin-like proteins from serine protease and BCL-2 datasets established in the previous research (Queiroz et al., 2020) to find similar structure in the large dataset of protein-protein and protein-peptide complexes. The specific objectives are to;

i. Download all PDB entries and filter out those that represent protein-protein or protein-peptide complexes.
ii. Compute and model the interaction between protein-protein complexes and protein-ligand complexes as graphs.
iii. Use the established conserved structure to search PDB for the similar conserved patterns.
iv. Visualize the computed similar patterns

1.6. Research Methods

i. The PDB file protein-protein complexes were filtered, and their chains were selected and downloaded from the Protein Data Bank, while the Bio Python parser was used to download, read, and write the protein structures of each PDB id and their chains for the purpose of computations.

ii. The protein-protein interactions were computed to derive PPI graphs using a cutoff-dependent strategy, such that, at the atomic level, the interactions between atoms depends on the types of atom and distance criteria. Graph modeling was used to model protein-protein interactions as graphs with atoms as nodes and interactions between atoms as edges using atom types, interaction types, and distance criteria.

iii. Searching for similar conserved structure on the graph dataset using a set of candidate pairs based on state space representation and feasibility rules that embed equality of graph structure and attributes.

iv. The visualization of the result was achieved using D3, CSS and HTML

1.7. Expected Contribution to Knowledge

The study will bring about a user-friendly web-based tool that implements a computational strategy that searches for similar structural patterns of protein-protein

or protein-peptide in a large dataset of complexes, using already established conserved substructure.

Chapter 2

Overview of the Previous Works

Proteins are biological biomolecules and macromolecules that are widely responsible for various biological functions in the cell (Rasheed et al., 2020). Proteins are one of the most vital molecules in which their structure and combinational behavior are vital to many functions in the living organisms. This protein molecule plays important roles in plants to carry out many processes that occur within the cell of the organisms (Silva et al., 2020). Some of those functions include structural, enzymatic and functional roles such as in biosynthesis, immunity, photosynthesis and transportation. In seedlings development, proteins similarly act as medium for storage in order to meet the nutritional demands and growth. These functions are performed in exact structural forms and their composition through folding ranging from well-ordered and compact to intrinsically disordered and unfolded (Rasheed et al., 2020).

However, the protein hardly acts alone, but rather forms a physical network interaction with other proteins. Protein can also form other relationship types such as signaling cascades or regulatory and metabolic relationship (Garcia-Garcia et al., 2012). Protein consists of unfolding and long chains of amino acids linked by peptide bonds. It has different sequences of natural amino acids which can subsequently fold into various degrees of geometric similarity of tertiary structures (Jaiswal et al., 2020). Two or more groups of associated proteins are otherwise called protein complex. Precisely, protein complexes are referred to as aggregation of protein molecules commonly connected together through several interactions among proteins. Many functionalities of proteins are achieved after protein-protein interactions that lead to formation of protein complex (Grbic et al., 2020).

2.1. Protein Structures

Protein usually folds in 3D unique structures. Naturally, the shape to which protein fold into is called native conformation. Similarly, while many proteins can fold unaided through amino acid chemical properties, others require aid for the folding through molecular chaperones into their native states. According to biochemists, the four distinct protein structures are (Murray et al., 2016);

i. *Primary structure:* It is the sequence of amino acids in which its protein is called polyamides

ii. *Secondary structure:* It is an ordered linear array of amino acids that confer local regular conformational forms which constitute the secondary structure of the protein. The main elements of the secondary structures are α-helix, β-sheet, and random structures called loops or coils. Because secondary structures are local, many regions of different secondary structures can be present in the same protein molecule.

iii. *Tertiary structure:* The third level of protein structure is the tertiary structure which refers to the overall three-dimensional arrangement of the various secondary structure elements (Ahmed & Gomaa, 2011). It is the overall shape of a single protein molecule; the spatial relationship of the secondary structures to one another. This tertiary structure is generally stabilized by nonlocal interactions, most commonly the formation of a hydrophobic core, but also through salt bridges, hydrogen bonds, disulfide bonds, and even posttranslational modifications. The term "tertiary structure" is often used as synonymous with the term *fold*. The tertiary structure is what controls the basic function of the protein.

iv. *Quaternary structure*: the structure formed by several protein molecules (polypeptide chains), usually called *protein subunits* in this context, which function as a single protein complex.

v. *Quinary structure*: the signatures of protein surface that organize the crowded cellular interior. Quinary structure is dependent on transient, yet essential, macromolecular interactions that occur inside living cells. In contrast to the other first four types of protein structures, which are closely connected to isolated proteins in dilute conditions, the emergence of quinary structures is due to the crowdedness of the cellular context, in which transient encounters among macromolecules are constantly occurring (Danielsson & Oliveberg, 2017). The quinary structure performs its functions by finding a certain partner that it can bind with in a relatively long encounter. Therefore, the adaptation of the protein interface that proteins use to navigate the complexity of the cellular environment led to the formation of the quinary structure (Mika & Poolman, 2011).

Furthermore, protein has the possibility of shifting between many associated structures during the process of performing its functions because it is not a totally rigid molecule. The tertiary and quaternary structure are called conformations in this functional rearrangement context while the transitions that occur in between the structure are referred to as conformational changes. These changes are commonly induced through substrate molecule binding to an active site of enzymes or the protein physical region that takes part in the chemical catalysis. Also, in solution, proteins experience structure variation over thermal vibration and other molecules collision (Christopher et al., 1996). The protein 3D structure provides information that plays a vital role in numerous fields such as biomaterials, health, biotechnology, bioinformatics to mention but a few. The information derived from proteins' 3D structure assists to understand the possibilities of interactions that might occur with other molecules (Kurniawan et al., 2020). This information also plays an important role in determining protein function and its interaction with enzymes, DNA and RNA. Its conformation information is essential in providing information that can be used in drug design and protein engineering (Pan, 2007).

2.2. Protein Data Bank (PDB)

The technological advancement has led to the introduction of different techniques that avail the determination of protein structure. One of the avenues provided by advanced technology is a repository of protein structures called Protein Data Bank (PDB) which allows sharing various data regarding protein structures (Tariq et al., 2020). The PDB currently contains thousands of protein structures and new structures are continuously being released. The PDB is a very important resource to the community of structural bioinformatics for software development that can use, mine, classify, label and analyze data. Figure 1 presents the number of protein structures in PDB over the years.

2.3. Protein–Protein Interactions

Protein–Protein Interactions **(**PPIs) are caused by various interactions such as hydrogen bonding, electrostatic force and hydrophobic effect which result in biochemical events that lead to the establishment of physical contacts between protein molecules with high specificity. Many of these interactions are association of molecular contacts between chains that arise in a living organism or cell in a particular biomolecular context. Proteins hardly act alone as their roles have the tendency to be

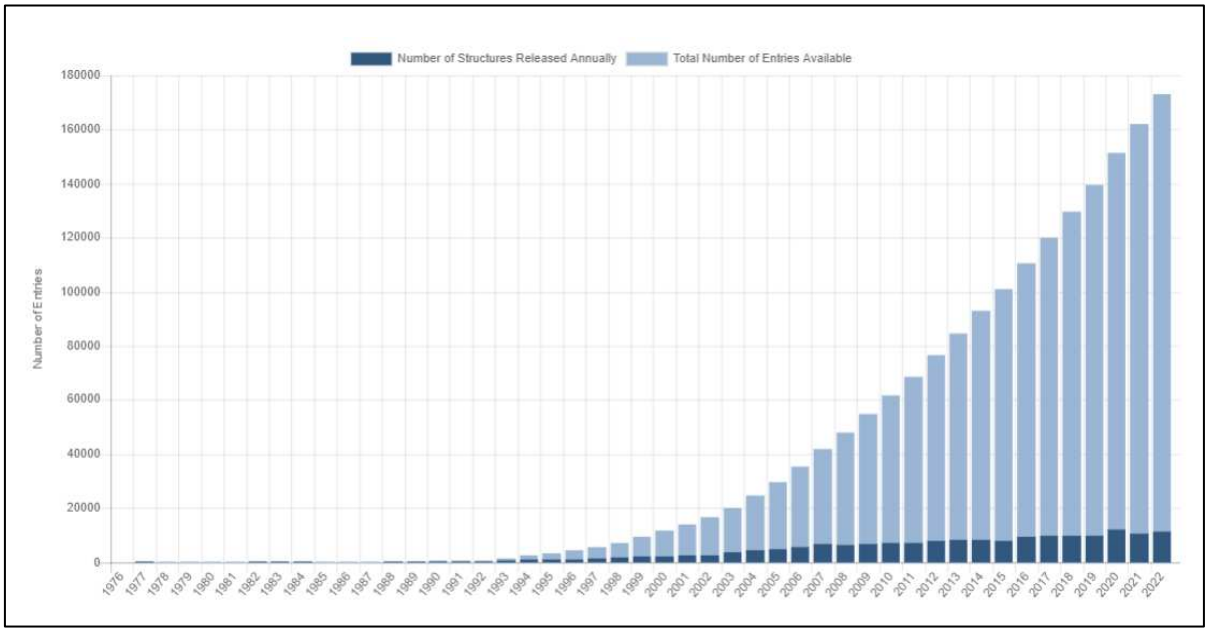


Figure 1 – Growth of protein structure released per year (RCSB 2022). The horizontal line represents the year protein structures were released, and the vertical line represents the entries and accumulated number of protein structures released in each year.

regulated. Various processes of molecules that happen in the cell are performed by molecular machines built from several components of protein arranged by their PPIs. The organism interactions are as a result of physiological interactions while unusual PPIs are the background of related aggregation of multiple diseases (Gonzalez & Kann, 2012) (Kuzmanov & Emili, 2013). Many methods have been deployed for the study of PPI using different perspectives such as quantum chemistry, biochemistry, signal transduction, molecular dynamics to mention but a few. The information derived from these methods enables the formation of large networks of protein interaction similar to genetic network or metabolic network that enhance the discovery of current knowledge on molecular etiology of disease and biochemical cascades. It also assists in the unraveling of the supposed protein target of interest. PPIs are described based on the possibility of protein interaction in a transient or stable way to form complexes that subsequently turn to molecular machines within the living organisms. The assembly of protein complexes can lead to the formation of hetero-oligomeric or homo-oligomeric complexes.

PPIs can be detected using experimental or computational methods. The most common and conventional high-throughput experimental methods used are affinity purification joined to mass spectrometry (Wodak et al., 2013) and yeast two-hybrid

screening (Terentiev et al., 2009). Others are protein microarrays, nucleic acid programmable protein array, intragenic complementation (Bertolini et al., 2021), co-immunoprecipitation, analytical ultracentrifugation to mention but a few. Each of these methods has their own weaknesses and strengths in relation to the specificity and sensitivity of the methods (Titeca et al., 2018). The experimental detection and characterization of PPIs is labor-intensive and time-consuming. In computational methods, PPIs can be also predicted usually by using experimental data as a starting point. However, many methods have also been developed that allow the prediction of PPI de novo, that is, without prior evidence for these interactions. Examples of such methods include Genomic Context Methods (Raman, 2010)**,** Text mining methods (Badal et al., 2015), Machine learning methods (Sarkar & Saha, 2019) and graph theory or statistical methods (Yang et al., 2020).

2.3.1. Types of Protein–Protein Interactions (PPIs)

In this work, we focus on tertiary structural data of protein families to discuss the functional and structural diversity of protein-protein interactions (PPIs). Biologically, the roles of PPI are diverse, and this difference is based on the affinity, composition, and type of association, which can either be transient or permanent (Phizicky, 2018). In vivo, the interaction between protomers can be impacted by the protomer's local surroundings, concentration, and localization, which are important to the oligomeric condition of protein complexes and composition control. Transient PPIs are significant biological regulators because a change in quaternary state is frequently accompanied by a change in biological activity or function. The structural features of various PPI types are described together with their physiological function, evolution, and specificity (Nooren & Thornton, 2003). The types of protein interactions are classified into homo- and hetero-oligomeric complexes, transient and permanent complexes, and non-obligate and obligate complexes (Acuner Ozbabacan et al., 2011).

i. *Homo- and Hetero-oligomeric Complexes:* Based on the composition resulting from the physical interaction between two or more proteins, groups of complexes can be classified into homo-oligomeric and hetero-oligomeric complexes. A homo-oligomer is formed when a PPI occurs between identical chains, whereas a hetero-oligomer is formed when a PPI occurs between non-identical chains. Protein oligomers that include identical or homologous protein units can be arranged symmetrically in isologous or heterologous ways. A 2-fold symmetry

axis connects the same surface on both monomers in an isologous relationship. Contrary to isologous associations, which can only oligomerize further using a different interface, heterologous assemblies utilize several interfaces that, in the absence of a closed (cyclic) symmetry, can result in limitless aggregation (Phizicky, 2018).

ii. *Transient and permanent complexes:* PPIs can also be identified based on the complex's lifespan. A transient interaction associates and dissociates in vivo as opposed to a permanent interaction, which is typically quite stable and only exists in its complexed form. A weak or strong association can exist in a transient interaction. Strong transient associations need a molecular trigger to change the oligomeric equilibrium, whereas weak transient associations have interactions that continuously break and form oligomeric equilibrium in solution. While non-obligatory interactions might be temporary or permanent, structurally or functionally obligate interactions are often permanent. In addition, many PPIs do not fit neatly into distinct types but rather fluctuate between obligate and non-obligate interactions, and the environment and physiological factors have a significant impact on the stability of all complexes. In vivo, an interaction might be primarily temporary, but under specific cellular circumstances, it might become permanent. The biologically significant form of interaction will frequently be suggested by the protein's function. Examples of this include intracellular signaling interactions, which are predicted to be transient because they are necessary for the functions of ready association and dissociation. For instance, (Lee et al., 2014) developed stabilized alpha-helices of BCL-2 domains (SAHBs) to dissect and target protein interactions of the BCL-2 family, which is a critical network that regulates the apoptotic pathway. These SAHBs are α-helical surrogates that bind both stable and transient physiologic interactors and have successfully uncovered new BCL-2 family protein interaction sites.

iii. *Non-obligate and obligate complexes:* In addition to composition, two distinct types of complexes can be identified based on whether they are obligate or non-obligate. In an obligate PPI, the protomers are not observed in vivo as stable structures on their own. These complexes typically also have functional obligations; the Arc repressor dimer, for instance, is required for DNA binding. Numerous hetero-oligomeric structures in the Protein Data Bank feature non-

obligate interactions of independent protomers, such as intracellular signaling complexes, antibody-antigen, receptor-ligand, and enzyme-inhibitor complexes (Bartholow et al., 2021). Since the components of these protein-protein complexes frequently do not initially co-localize, each one must be stable on its own. However, some co-localized homo-oligomers, which are by definition, can form non-obligate assemblies. In summary, all obligate PPIs are permanent, while not all permanent interactions are obligate. Non-obligate interactions are transient, but some non-obligate interactions are permanent, like some enzyme-inhibitor interactions. The strong transient category includes protein interactions that shift from an unbound or weakly bound state to a strongly bound state, which is generally triggered by an effector molecule.

The protein-protein interaction types in relation to affinity and stability is shown in Figure 2.

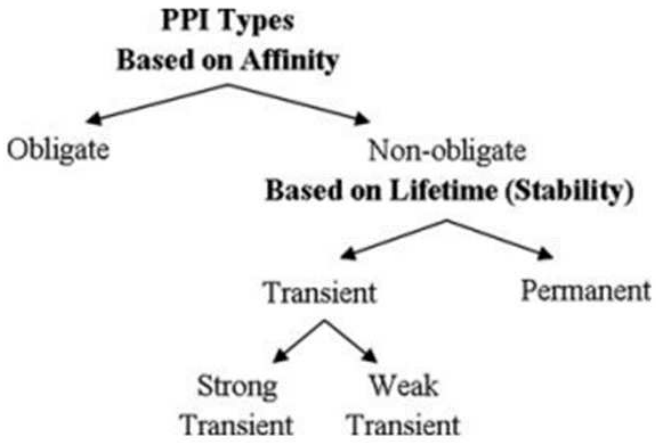


Figure 2 – Relation of affinity and stability-based protein–protein interaction types. Non-obligate interactions are not stable but there are some examples of stable non-obligate interactions (Acuner Ozbabacan et al., 2011).

2.3.2. Characteristics of Protein-protein Interactions Based on Structure

The nature of the structural interfaces involved can be investigated in order to determine the kind of PPI based on knowledge of the structure and the many forms of functional interactions at protein-protein interfaces. For different protein or protomer complexes, structural information is available, and the structure of the interfaces can be evaluated using a variety of criteria, including the size of the contact area, flatness,

protrusion, and polarity of the interface. Protein-protein interfaces, for example, bury 1200-5000 Å2 of surface at their most fundamental level, with direct interactions accounting for approximately 70% of the interface. Water molecules take up the remaining surface area of the interface (Keskin et al., 2016) (Reichmann et al., 2008). The interacting proteins appear to have a high degree of form complementarity (although not perfect) based on the basic interface design. Additionally, a high degree of chemical complementarity between the amino acids creating inter-protein contacts is necessary for them to form hydrophobic, polar, or charged interactions, which are necessary for the proteins to bind (Cohen et al., 2008).

2.3.3. The Possible Interactions Between Proteins

Because proteins are incredibly heterogeneous molecules, they may carry out a wide range of biological functions. Proteins can play structural roles, catalyze chemical reactions (enzymes), and are components of more intricate, large-scale molecular machines. Most proteins have the ability to interact with other molecules. Among the interaction partners that proteins can interact with are other proteins, small molecules, nucleic acids, peptides, fatty acid chains, carbohydrates, and more. Proteins carry out their diverse functions in the congested cellular environment, where they are able to remain efficient and specific (Schreiber, 2020).

2.3.4. Impact of Protein-Protein Interactions on Biological Processes

Numerous biological processes, such as cell-to-cell interactions, signal transduction, cell cycle progression, and metabolic pathways, depend on protein-protein interactions. The following are some of the crucial characteristics of PPIs, according to (Phizicky, 2018):

i. PPIs can change an enzyme's kinetic characteristics, which may result in minor adjustments to substrate binding or allosteric effects.
ii. By transferring a substrate between domains or subunits, PPIs can function as a generic mechanism to enable substrate channeling.
iii. PPIs are able to develop a brand-new binding location for small effector molecules.
iv. PPI has the power to inhibit or inactivate proteins.
v. PPI interacts with several binding partners to alter the specificity of a protein for its substrate.

vi. PPIs can play a regulating role in an upstream or downstream event.

2.4. Graph Subgragh Isomorphism

Some of the algorithms explored in solving graph or subgraph matching are Ullmann algorithm (Gouda et al., 2022), VF2 algorithm (Kusari et al., 2022), exact and error-tolerant graph matching algorithms (Dwivedi, 2019) and probabilistic relaxation algorithms (Zhang et al., 2018). Probability relaxation techniques shows to be an effective method for graph matching with relational attributes (Zhang et al., 2018). The basis and methodology design of the relaxation process are subsequently based on heuristics. This technique is so amazing as to lessen the complexity from exponential to polynomial in most cases. It is however nondeterministic and not ensured to track down an accurate and optimal solution. Also, a nauty algorithm (McKay & Piperno, 2014) which changes the matching graph to a generally acceptable form prior to its examination for isomorphism is considered among the fastest available algorithms for graph isomorphism. But it has been established that it makes use of exponential time for some classes of graphs and cannot be utilized in solving problem related isomorphism in graph and subgraph (McKay, 1981). In the same vein, (Bunke & Messmer, 1995) endeavors to diminish the generally speaking computational expense using exact and error-tolerant graph matching algorithms while matching an example graph against a huge set of models. This algorithm brings about a quadratic time concerning graph size, yet with an exponential memory necessity and preprocessing time. The primary improvement presented in the VF algorithm is that the data structures utilized are coordinated in the course of exploration of the search space in such a manner to essentially lessen memory requirement. Along these lines, the algorithm is appropriate for graph matching with number nodes and branches (Cordella et al., 1999). The Ullmann algorithm is a backtracking procedure that employs a relaxation-based refinement step to reduce the search space (Ullmann, 1976). In the Ullmann algorithm, the detection of graph matching is possible for both graph and subgraph isomorphism. A short description of the manner in which data is structured with the view to enhance graph matching algorithm known as VF2 was unraveled by (Cordella et al., 2004). In his review, the description of the aftereffects of the hypothetical examination of its general effectiveness was introduced in relation to spatial and computational complexity. The most effective technique for handling enormous datasets is still VF2 (Dwivedi, 2020), and it is also being utilized as the basis

for a specialized subgraph matching algorithm in other recent research (Adams et al., 2022; Guo et al., 2022).

Chapter 3

Materials and Methods

This work leverage on the result of conserved structural arrangements of trypsin and trypsin-like proteins and BCL-2 from serine protease dataset established in the previous research (Queiroz et al. 2020) to find similar structure in the large dataset of protein-protein complexes or protein-peptide complexes. To achieve this, we downloaded PDB entries and filter out those that represent protein-protein complexes or protein-ligand complexes in order to compute and model the interaction between protein-protein complexes and protein-ligand complexes as graphs. Figure 2 explains the steps performed to detect similar structures (conserved subgraphs) in the large dataset of PDB entries modeled as graphs. In Figure 2A, the graph modeling used protein-protein complexes retrieved from the PDB to compute the Protein-Protein Interactions (PPI) at the atomic level, utilizing physicochemical properties and distance criteria to generate PPI graphs. To search for subgraph isomorphism as presented in Figure 1B, the established subgraphs, which are the conserved substructure of the PPIs of trypsin and trypsin-like proteins from the serine protease dataset, were used to search for a similar subgraph in the large dataset of graphs from PDB from PDB using VF2 algorithm.

3.1. Data Acquisition

The data acquisition aims to get all the PDB entries that represent protein-protein complexes or protein-peptide complexes with their chains from the Protein Data Bank. Considering the large number of protein complexes available in the Protein Data Bank, the promising approach is to find the similar structure using a substantial number of sample size of protein-protein complexes. The PDB ids of protein-protein complexes were filtered and downloaded from Protein Data Bank (RSCB) while the Bio Python parser (Chang et al., 2021) was used to download and read the PDB files of each PDB id in order to automatically extract the chains contained in each PDB file. These PDB chains were extracted based on information in the PDB file manual by parsing such files into the local system and subsequently saved in the predefined format (PDBID, chain 1, chain 2 … chain N).

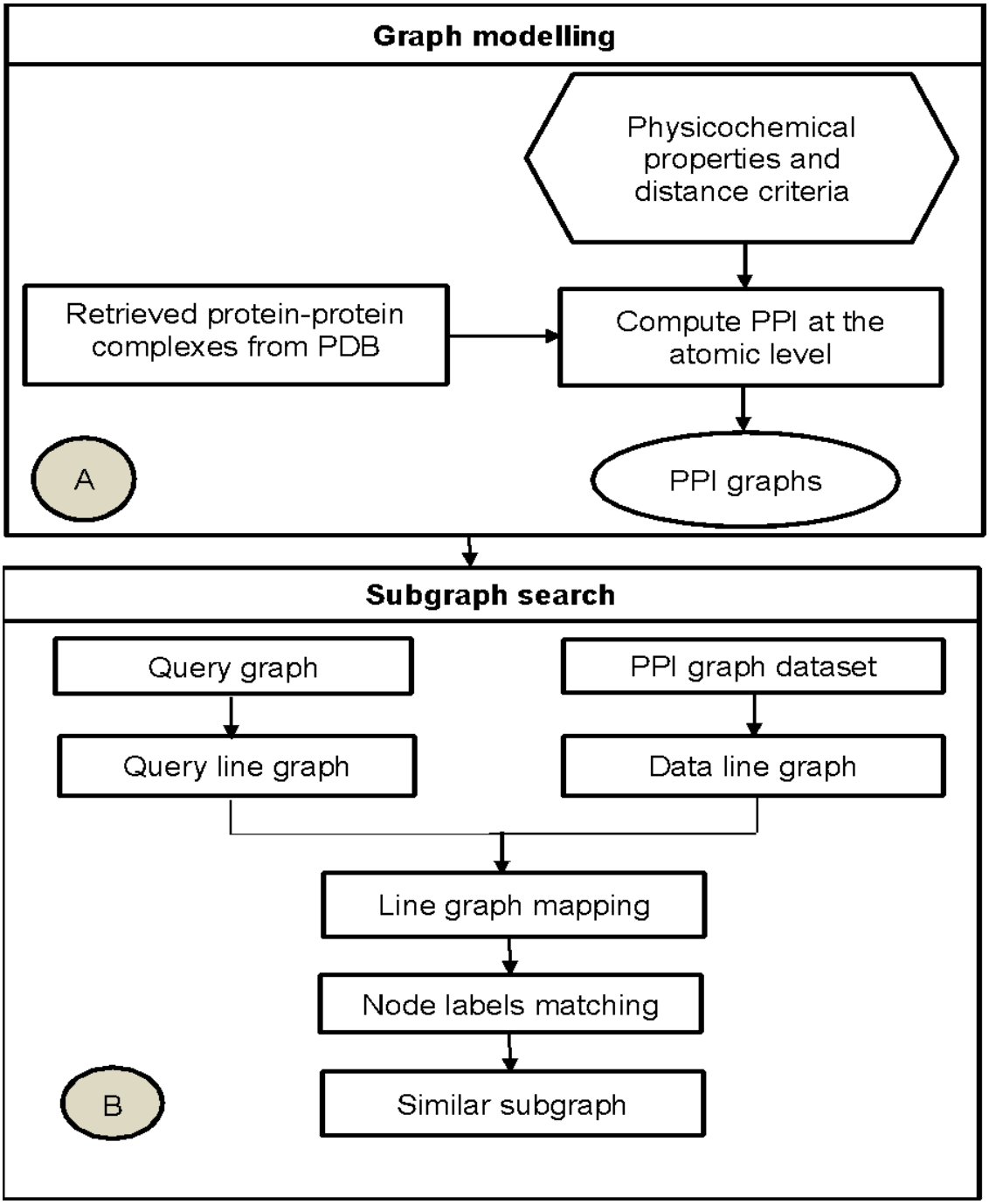


Figure 3 – pGReMLIN workflow. The workflow is composed of two blocks: Graph modeling and Searching strategy. Rectangles indicate processing steps; ellipsoids denote output files; and hexagons represent input files or parameters.

3.2. Physicochemical Properties of Atoms and Distance Criteria

One of the goals of pGReMLIN is to use graphs to model protein-protein interactions with atoms as nodes and interactions between atoms as edges. As a result, atom types, interaction types, and distance criteria are required. Therefore, this work employed the physicochemical properties of atoms and the standard distance criteria established in previous research (Fassio et al., 2020). Table 1 presents the list

of physicochemical properties of atoms and distance criteria to compute interactions among atoms. These atoms' physicochemical properties and standard criteria for distance between the pair of atoms in each interaction type are expected to produce bipartite graphs as an output (Queiroz et al., 2020); (Ribeiro et al., 2020).

Table 1 – Physicochemical properties of atoms and distance criteria to compute interactions

Interactions type	**Atom types**	**Distance**	
		Min	**Max**
Aromatic stacking	2 aromatic atoms	1.5	3.5
Hydrogen bond	1 Acceptor and 1 donor atom	2.0	3.0
Hydrophobic	2 hydrophobic atoms	2.0	3.8
Repulsive	2 atoms with the same charge	2.0	6.0
Salt bridges	2 atoms with opposite charge	2.0	6.0

3.3. Computation and Modeling of Protein-Protein Interactions (PPI)

The protein-protein interaction, which is a mathematical representation of the physical contacts between proteins, was computed to derive PPI graphs using a cutoff-dependent strategy. This was computed based on the physicochemical properties and the distance criteria between pairs of atoms. At the atomic level, there exists contact between atoms i and j provided that the Euclidean distance that separates them is less than or equal to the cutoff (C. H. Da Silveira et al., 2009), (Pires et al., 2011). Therefore, in the pGReMLIN modelling, only the contacts that connect atoms from different chains of protein were considered, such that they can be named chains X and Y. This means that a bipartite graph G(A, B, E) can be achieved with vertices segmented into 2 disjoint sets, A (nodes from X) and B (nodes from Y) such that every edge in E connects a vertex in A to a vertex in B. The illustration of how bipartite graph modelling was achieved is shown in Figure 3. The nodes are categorized and labeled based on their corresponding physicochemical properties, while the edges are categorized and labeled according to their atoms' physicochemical properties and distance criteria.

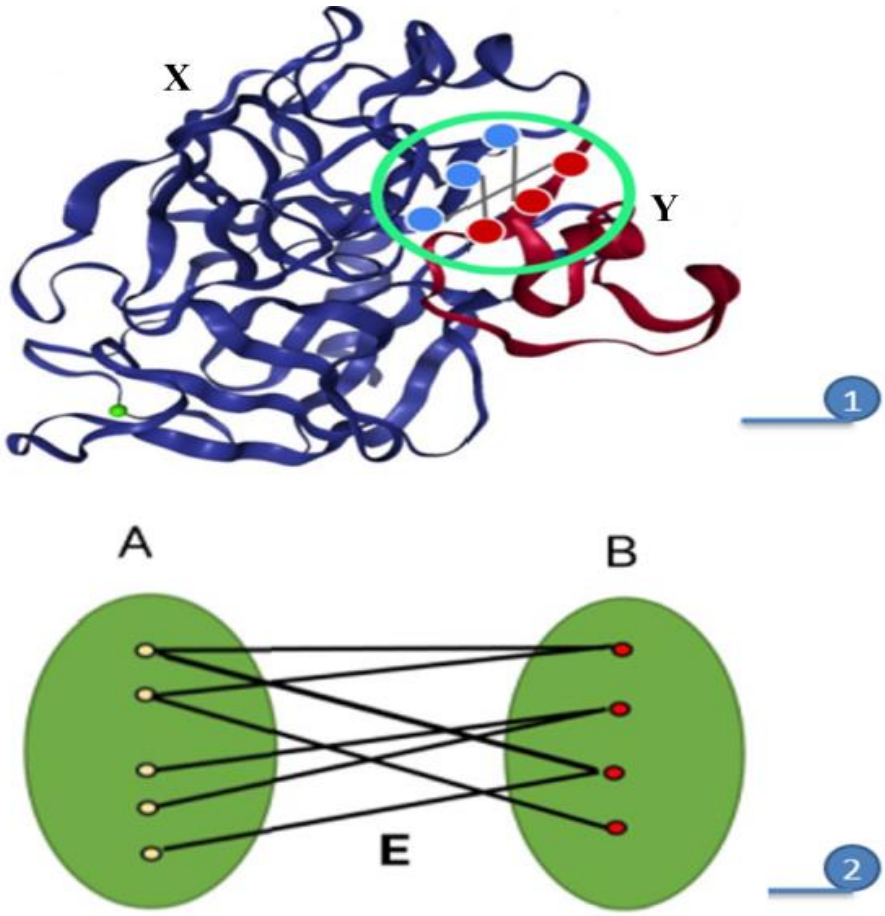


Figure 4 – pGReMLIN as bipartite graph. 1 represents the interactions that connect atoms from the chains of protein residues such that it can be named chains X and Y; 2 represents the model as bipartite graph G(A, B, E) with vertices segmented into 2 disjoint sets, A(nodes from X) and B(nodes from Y).

As shown in the pGReMLIN graph modelling in Figure 4, the attributes (physicochemical properties) used for labelling nodes are aromatic (RAM), acceptor (ACP), hydrophobic (HPB), positive (POS), donor (DON) or negative (NEG) (Santana et al., 2016), (Gonçalves-Almeida et al., 2012), (S. A. Silveira et al., 2014) (Fassio et al., 2018) while the attributes (atoms' physicochemical properties and distance criteria) used for labeling edges are hydrogen bond, aromatic, hydrophobic, salt bridge and repulsive (Queiroz et al., 2020). Because of the interaction tendencies of node labels, multiple edges having different labels can possibly be associated with an individual pair of nodes. In our effort to lay hold of all the physicochemical properties of protein interactions, the interaction graphs are modelled as multigraphs. Subsequently, the final labels for the various nodes in the graph are considered based on the node labels that make up their interaction with other nodes. This is because not every node label is associated with the node. Therefore, connected components are computed for each

protein-protein complex (PDB id and chains), which are eventually used as dataset input to the graph subgraph searching.

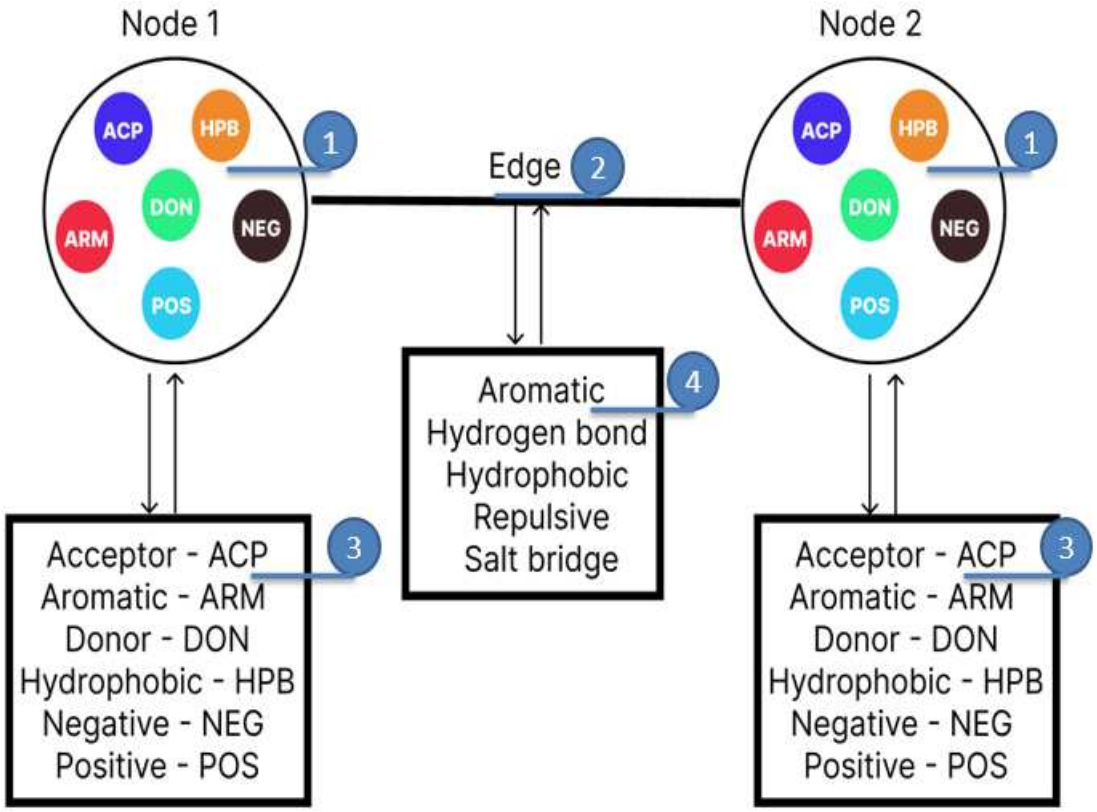


Figure 5 – pGReMLIN graph modeling. 1 depicts how nodes are used to represent atoms; 2 depicts how edges are used to represent interactions between atoms; 3 represent nodes labeled according to their physicochemical properties; 4 represents edges labeled based on the interaction types and distance criteria.

3.4. Searching Strategy

In this section, we used the already established conserved structures in (Queiroz et al., 2020) to search for the same conserved patterns in our PPI graph dataset. This literally means searching for the presence of graph Q (query graph) in another graph D (data graph), otherwise known as a subgraph, as well as subgraph isomorphism (Nabti & Seba, 2016). Therefore, searching for Q(G) (query graph) in D(G) (data graph) for subgraph isomorphism means finding all subgraphs of D that are isomorphic to Q, that is, finding all the embeddings Q(G) in D(G). It should be noted, however, that both Q and D are products of the same PPIs computation and graph modelling approach. Figure 5 explains the processes and steps employed in searching for the similar structures (conserved subgraphs) in the large dataset of PDB entries modelled as data graphs.

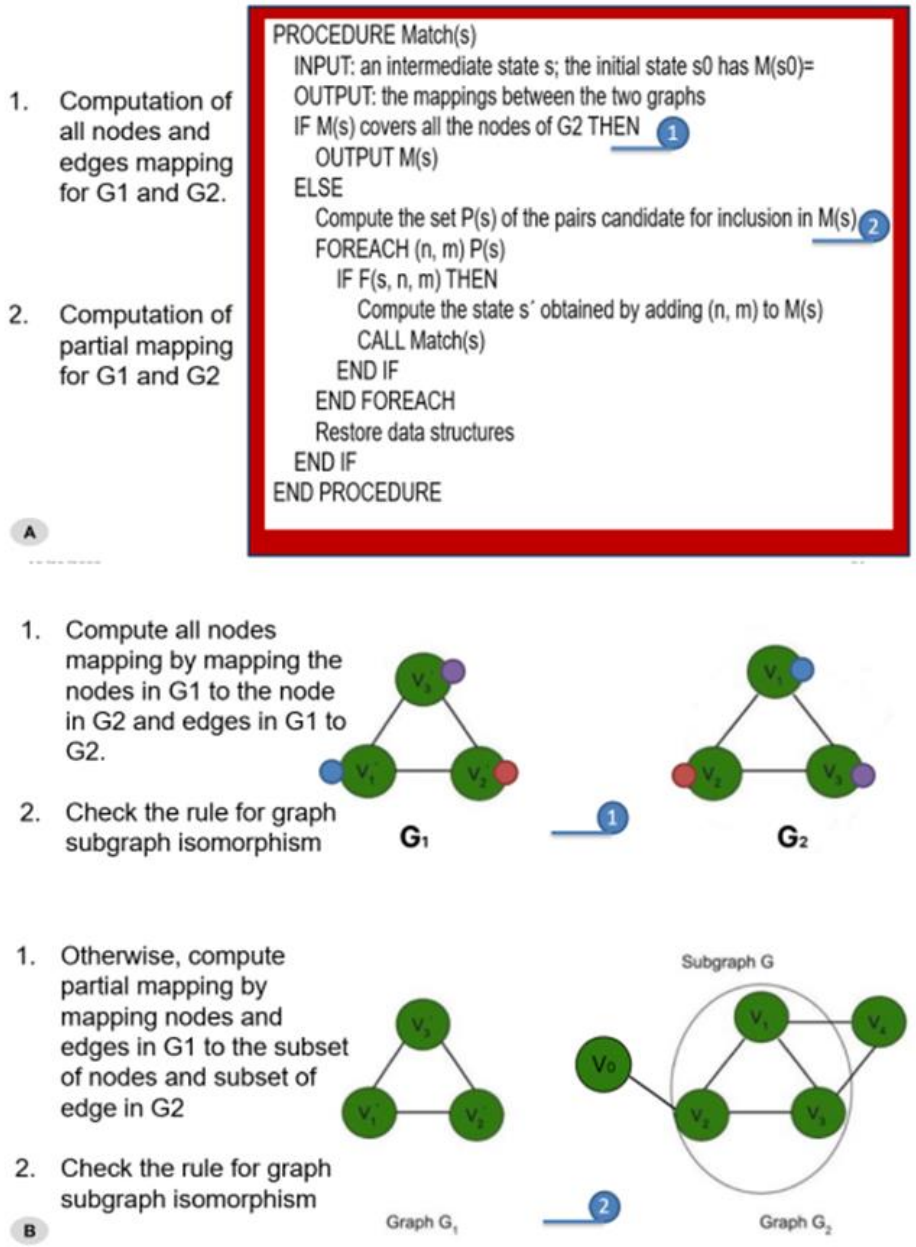


Figure 6 – Pattern searching strategy. A depicts the high-level abstraction of the description of the algorithm for the computation of all nodes and edges mapping in G1 and G2; and the computation of partial mapping of G1 and G2. B depicts samples of the conditions for graph isomorphism and graph-subgraph isomorphism.

The high-level abstraction of the description of the algorithm is shown in Figure 5A, while the conditions for graph isomorphism and graph-subgraph isomorphism is shown in Figure 5B. The two operations embedded in the searching process are the line graph mapping and node matching functions. In the first operation, the line graphs of the data graph and query graph are parsed, and only those line graphs that map each other are retained. The second operation matches the labels in the nodes of the query graphs

with the labels in the nodes of the data graphs and retains only those that match each other.

In our search strategy, we used subgraph isomorphism algorithm (VF2), which takes as input the Q(G) and D(G) and subsequently gives a report of the embeddings. The computation of subgraph isomorphism in the data graph was achieved using a set of candidate pairs of data graph D(G) and query graph Q(G) based on state space representation and feasibility rule that embed equality of structure (syntactic feasibility) and attributes (semantic feasibility). Lastly, the subgraph isomorphism of both query graphs Q(G) and data graphs D(G) was performed based on node match functions. The labels in the query graphs Q(G) nodes that are a subset of the data graphs D(G) nodes are considered and estimated as a match in this process.

3.4.1. Candidate Pairs and Feasibility Rules

The graph-subgraph isomorphism is a computation of the set of candidate pairs based on state space representation and feasibility rule that embed equality of structure (syntactic feasibility) and attributes (semantic feasibility). The process of matching between the query graph and data graph, that is, $Q(G) = (U_q, E_q)$ and $D(G) = (V_d, E_d)$ are incorporate in the determination of a mapping M which associates D(G) vertices with Q(G) vertices and vice versa, according to the predefined constraints. Generally, mapping M is defined as the set of pairs (x,y) where $x \in D(G)$ and $y \in Q(G)$ with each representing the mapping of a vertex *x* of D(G) with a vertex y of Q(G). The mapping $M \subset V_d \mathbf{X} V_q$ is regarded as isomorphism if M is a bijective function that preserves the branch structures of the two graphs. A mapping $M \subset V_d \mathbf{X} V_q$ is said to be a graph-subgraph isomorphism if M is an isomorphism between Q(G) and a subgraph of D(G).

This finding of mapping function process is based on State Space Representation in which each states of the matching process is associated to a partial mapping solution M(s) that comprises only a subset of M. M(s) specifically pinpoint two subgraphs of D(G) and Q(G) that is $D(G_s)$ and $Q(G_s)$, which are obtained through the selection of only the vertices of D(G) and Q(G) included in the M(s), and the edges connecting them. The transition of mapping will be denoted by $M_d(s)$, $M_q(s)$, $E_d(s)$, and $B_q(s)$ which are the sets of vertices of $D(G_s)$ and $Q(G_s)$ and the associated edges. The processes employ syntactic feasibility and semantic feasibility to ensure verification of consistency in order to generate consistent results. The syntactic feasibility uses line

mapping functions while semantic feasibility uses node matching functions and recursively adds the satisfied vertices to the current state until all the nodes in the graph are verified using breadth first strategies. This feasibility rule is denoted as shown in Equation 1.

$F(s,x,y) = F_{syn}(s,x,y) \wedge F_{sem}(s,x,y)$

(1)

Where;

F_{syn} *= syntatic feasibility depending only on the graph structure*

F_{sem} *= semantic feasibility depends on the attributes.*

3.4.2. Query Dataset

The dataset used for the query graph pattern consists of established, conserved structural arrangements that represent highly conserved interactions at the specificity binding pocket of trypsin proteins from Serine protease and BCL-2 datasets. The serine protease is composed of trypsin and trypsin-like proteins together with some peptide inhibitors, while BCL-2 is composed of protein complexes belonging to the BCL-2 family.

Soybean is one of the most widely grown farm products because of its importance as a source of protein for human and farm animal consumption (Xu & Zhang, 2022). It contains significant quantities of dietary minerals, phytic acid, and B vitamins, as well as vegetable oil utilised in various industrial and food applications (Valerie et al., 2020). It is also a source of income for many countries around the world, with Brazil, the United States, Paraguay, and Argentina being the top exporters and China being the top importer. However, one of the major problems affecting the yield of soybean in Brazil is the presence of Anticarsia Gemmatalis (AG), one of the various caterpillars that strip off the leaves of soybean prematurely. Therefore, this serine protease is a protease sequence that originated from the digestive system of a caterpillar, which was subsequently explored further to detect the highly conserved substructures in (Queiroz et al., 2020). Finding the ligands that interact with our target in our similar conserved structure predictions can go a long way in reducing the menace of caterpillars affecting soybean production. It will also reduce the toxicity to human health and the environment caused by the application of agrochemicals, which have a great potential

to cause the emergence of new parasites with high resistance that will further require high dosages or more powerful products.

BCL-2 is a set of protein complexes that share homology with the BCL-2 domains (1-4), comprising pro-apoptotic proteins and anti-apoptotic proteins, whose balance is the ultimate determinant of cell status. The anti-apoptotic protein originated from BH domains 1–4, in which its functionality is to preserve the integrity of the outer mitochondrial membrane through the inhibition of its pro-apoptotic counterparts. while the pro-apoptotic proteins have the possibility of having a single BH3 domain or numerous BH domains. The expression of these proteins can be modulated to point toward equilibrium in the direction of death or survival through cell signalling (Opferman & Kothari, 2018). As the apoptosis evasion capability of cells, together with oncogenic mutations that cause deregulation of cell cycling and cell growth, significantly increase tumorigenesis, apoptosis has been established as a critical mechanism for the suppression of tumors (Delbridge et al., 2012) which otherwise have a great potential to circumvent cancer (Delbridge & Strasser, 2015); (Hanahan & Weinberg, 2011). As a result, BCL-2 proteins can be used to regulate the release of cytochrome c from the mitochondria for the regulation of programmed cell death (Krajewski et al., 1993); (Kroemer, 2003). The already established conserved substructures in this BCL-2 are the second dataset we explored as the query pattern in our implementation.

3.5. Visualization of the Similar Patterns

The result of the visualization consisted of two group menus linked to the serine protease and BCL-2. Each of the menus has submenus consisting of the graph dataset detail, the graph pattern table, and the graphical analysis. The graph dataset detail shows the result of a similar pattern in the implementation based on pattern sizes, with the first column representing a group of patterns and the second column showing the PDB id(s) represented in the group pattern. The graph pattern table displays the pattern's size and the number of occurrences in each dataset, with many filtering options for minimum pattern size and minimum pattern occurrence. Also, the graphical analysis shows the composition of atomic type and interaction type based on the pattern size in each dataset. The link to the visualization of the implementation can be found in Appendix 2 and some of the sample of the visualization can be found in Appendix 3.

Chapter 4

Result and Discussion

In the implementation process, a large dataset of graphs was modeled from the PDB id and chains downloaded from the PDB. For the query patterns, two datasets used for the query graphs are serine protease and BCL-2. These query graphs are the conserved structures for which we intend to search for similar patterns in the data graph. In order to make sense of the result of our implementation, we segmented each subgraph derived from the isomorphism based on the number of pattern sizes. This is with a view to making it possible and easy to select subgraphs of interest among those with large pattern sizes in our visualization. As our interest is in those patterns with large sizes, we considered pattern size with a minimum threshold of 15 in our selection. Some of the most interesting patterns among the selected patterns in the two datasets are subsequently identified for the purposes of analysis and discourse.

4.1. Serine Protease Dataset

In the serine protease dataset, the minimum threshold of 0.6 was used for the gSpan, and a total of 99,709 subgraph isomorphisms were found in our data graphs. Among the subgraphs generated in our implementation, some subgraphs occur only in serine protease while some are also found in BCL-2 subgraph isomorphism (e.g., PDB id 5XAC, chain B, graph id 100721, size 62). Those that are only found in serine protease in this context are referred to as "unique patterns". Also, as the pattern size decreases up to the smallest pattern size (2 nodes), the number of occurrences of pattern size in this group increases to 66,268. The graph that shows the detail of the pattern's size and occurrence is shown in Figure 6. The sample from the first 15 groups of patterns identified in the serine protease dataset, starting from the highest pattern size, is shown in Figure 7. In this figure, the first unique pattern in the serine protease dataset is pattern size 28 at graph id 37,233 PDB id 3HU3 in chain A, followed by graph id 37,159 PDB id 3HUI in chain C, graph id 37,148 PDB id 3HUI in chain A, graph id 15,024 PDB id 2BWE in chain A, and graph id 37,185 PDB id 3HU2 in chain A at pattern sizes 24, 23, 22, and 21 respectively.

This highest unique pattern in serine protease is composed of acceptor/negative, positive, donor/postive, acceptor, and acceptor/aromatic/donor/postive atoms connected by hydrogen bond and/or salt bridge interactions in glutamic, lysine, leucine,

arginine, and histidine amino acid. Lysine, leucine, and histidine belong to the essential group of amino acids while glutamic and arginine belong to non-essential group of amino acids.

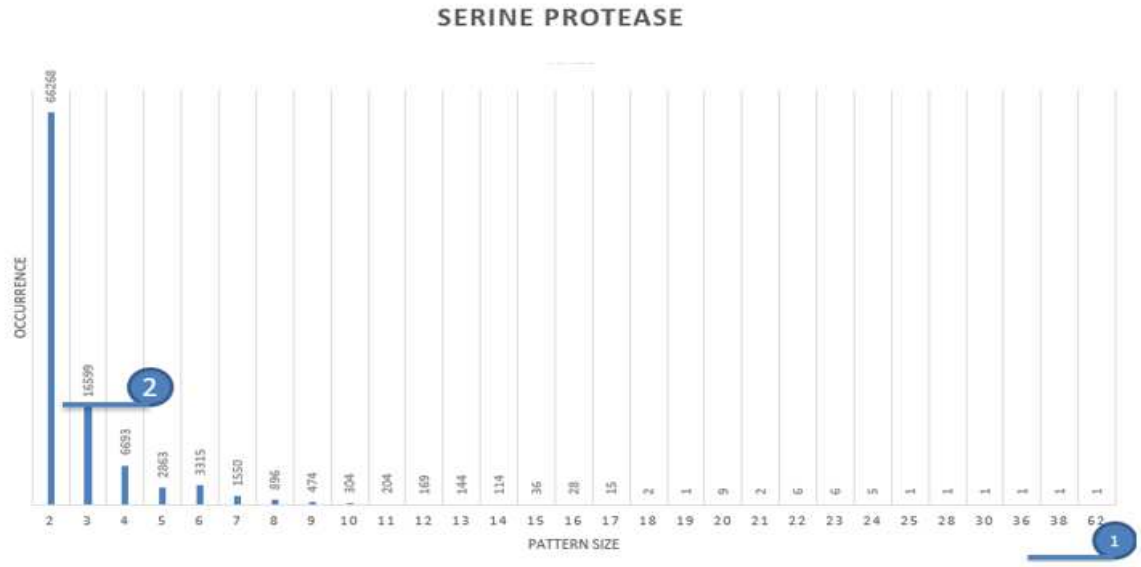


Figure 7 – Pattern size and occurrence in serine protease. 1 represent the unique size of the pattern in serine protease. 2 represent the total number of times the pattern occurs in each pattern size.

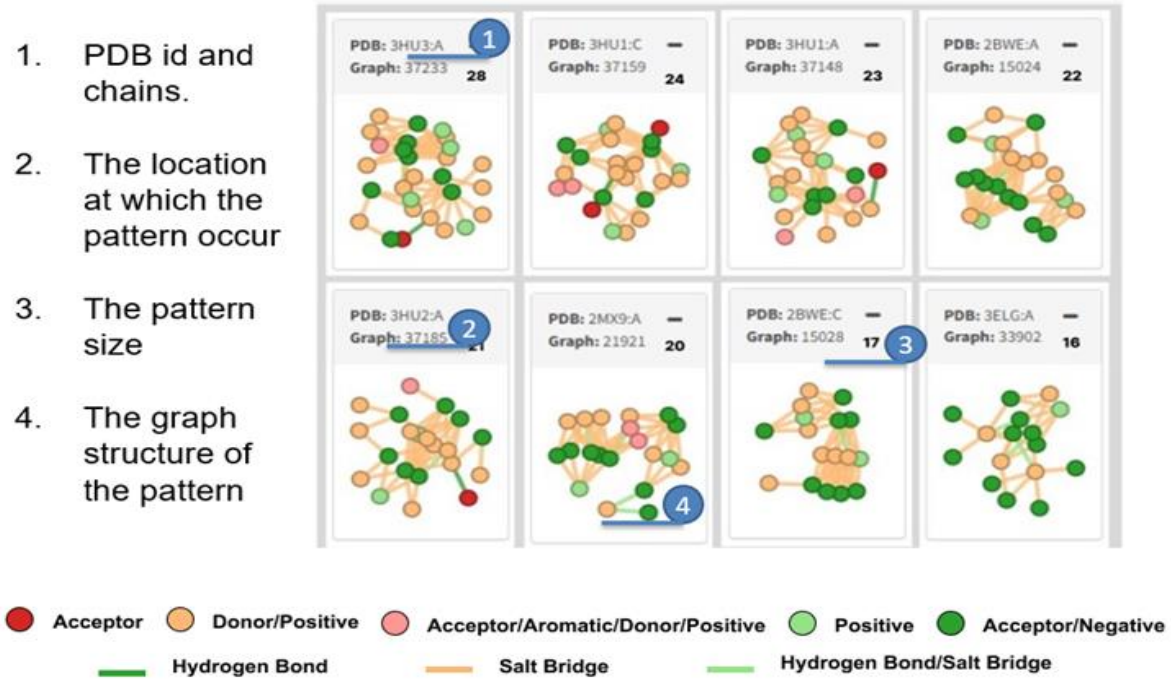


Figure 8 – Unique patterns in serine protease. Some selected patterns similar to conserved patterns in serine protease. Graphs legend displays the colors used to depict atoms and edges.

4.2. BCL-2 Dataset

In the BCL-2 dataset, the minimum threshold of 0.7 was used for the gSpan, and a total of 108,972 subgraph isomorphisms were found in our data graphs. Also, as the pattern size is reduced up to the smallest pattern size (2 nodes), the number of occurrences of pattern size in this group increases to 68,651. Figure 8. depicts the details of the pattern size and occurrence along the group of graph structures. The sample from the first 15 groups of patterns uniquely identified in the BCL-2 dataset, starting from the highest pattern size, is shown in Figure 9. In this figure, the first unique pattern in the BCL-2 dataset is pattern size 38 at graph id 101,575 PDB id 5Y40 in chain C, followed by graph id 43,365 PDB id 3NHQ in chain G, graph id 18,101 PDB id 2GSZ in chain E, graph id 6,550 PDB id 6DJL in chain B, and graph id 63942 PDB id 40IY in chain A at pattern sizes 35, 33, 32, and 31 respectively. This highest unique pattern in BCL-2 is composed of acceptor, donor/positive, acceptor/aromatic/donor/positive, positive, and acceptor/donor atoms connected by hydrogen bond, repulsive, and salt bridge interactions in histidine, glutamic, arginine and serine amino acid.

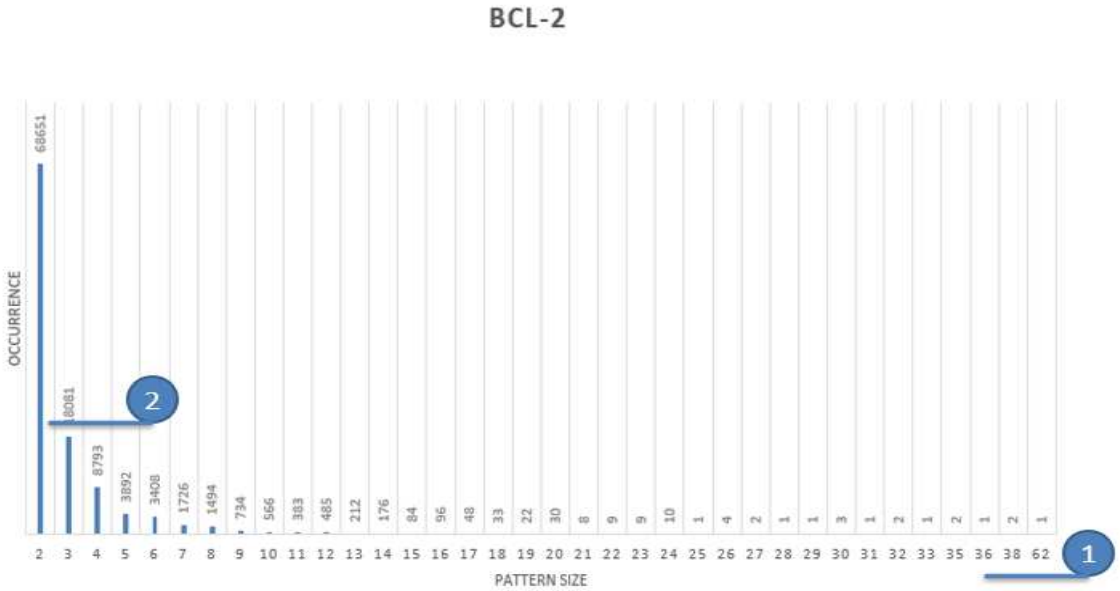


Figure 9 – Pattern size and occurrence in BCL-2. 1 represents the unique size of the pattern in BCL-2. 2 represent the total number of times the pattern occurs in each pattern size.

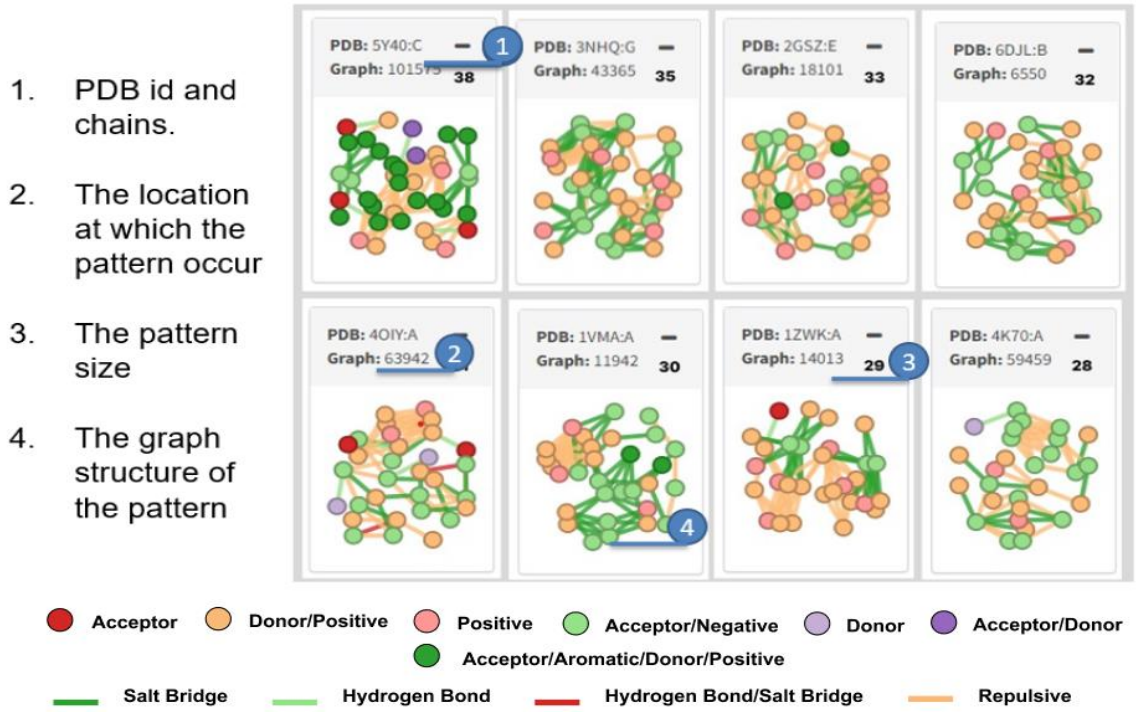


Figure 10 – Unique pattern in BCL-2. Samples similar to the conserved patterns in BCL-2 selected from the first 15 patterns starting from the highest pattern size. Graphs legend displays the colors used to depict atoms and edges.

4.3. Comparison of Serine Protease and BCL-2

We can infer that it is biologically possible to identify some patterns in our isomorphism that are common and those that are unique to serine proteases and BCL-2 datasets within the same pattern group. In order to arrive at this, the result of the isomorphism in both serine protease and BCL-2 was matched based on pattern size and graph id to derive the same occurrence of pattern and those that are unique to serine protease and BCL-2. The table in Appendix 2 shows the details of the pattern size and form of occurrence in both serine protease and BCL-2 up to the defined threshold.

The sample from the first 15 patterns identified as the same in both the serine protease and the BCL-2 datasets, starting from the highest pattern size, is shown in Figure 10. In this figure, it is observed that the first three highest pattern sizes in the isomorphism result of both serine protease and BCL-2 are the same. These patterns occurred in graph id 100,721 PDB id 5XAC at chain B, graph id 21,313 PDB id 2LZL at chain A, and graph id 101,947 PDB id 5YGH at chain A in pattern sizes 62, 38, and 36, respectively. Furthermore, the highest pattern identified in the entire dataset, which is graph id 100,721 PDB id 5XAC at chain B, is common to both serine protease and

BCL-2. This highest common pattern is composed of hydrophobic and aromatic atoms connected by hydrophobic and/or aromatic interactions in tyrosine, isoleucine and phenylalanine amino acids. The details of this pattern and other smaller patterns identified in the same PDB id are shown in Figure 11. It is also worthy of note that after pattern size 36, BCL-2 is the only dataset that has a representation of pattern sizes 35, 33, 32, 31, 29, 27, and 26, while serine protease is the only dataset that has a representation of pattern size 25.

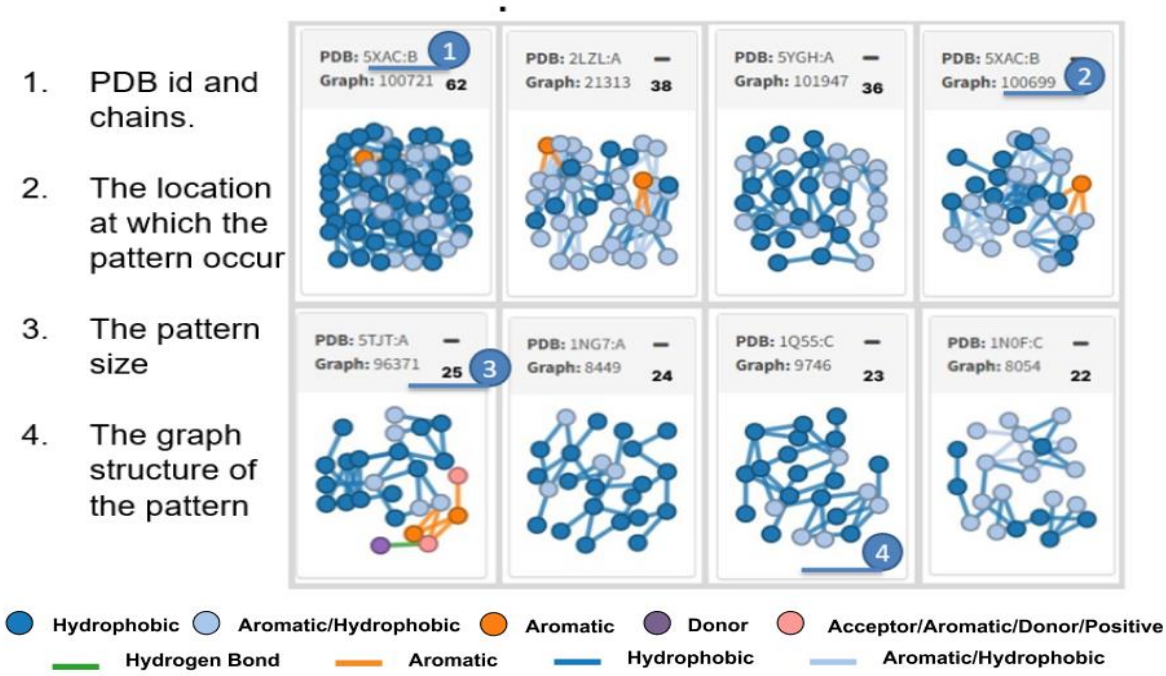


Figure 11 – Common pattern in serine protease and BCL-2. Some selected patterns similar to both conserved patterns in serine protease and BCL-2. Graphs legend displays the colors used to depict atoms and edges.

1. The largest pattern size

2. Other smaller pattern size from the same PDB id

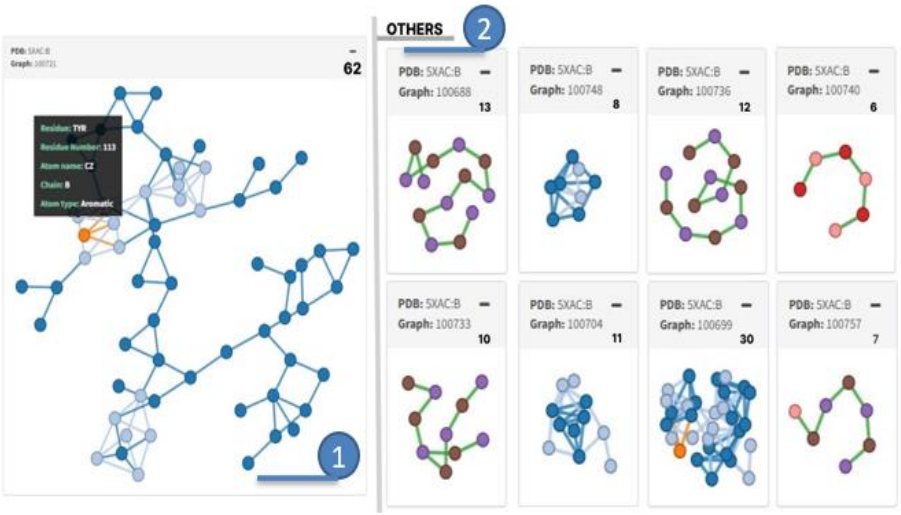


Figure 12 – Subgraph isomorphism with the highest pattern. 1 represent the pattern with highest size; 2 represent other patterns with the same PDB id.

Conclusion

In this work, a graph-based approach named pGReMLIN is proposed, which predicts similar structures in protein-protein or protein-ligand complexes using the established conserved structural arrangements of trypsin and trypsin-like proteins from the Serine protease. In the graph modeling strategy, the interface between proteins and/or ligands at the atomic level is modeled as a graph, with atoms as nodes and the non-covalent interactions between atoms as edges. The atoms are labeled based on their physicochemical properties, while the interactions are labeled based on their distance criteria. The VF2 algorithm was used to search for a similar pattern of structural arrangement in the data graph using graph line mapping and the node matching function in the isomorphism process.

In the serine protease and BCL-2 datasets, a total of 99,709 and 108,972 similar structural arrangement patterns were found in the data graphs, with both datasets having a single common occurrence of the largest pattern size of 62. Also, as the pattern size decreases up to the smallest pattern size 2, the number of occurrences in this group increases to 66,268 and 68,651 in the serine protease and BCL-2 datasets, respectively.

As similar patterns with large pattern sizes have a higher chance of containing protein targets of interest, this work grouped the similar structural arrangement patterns found in the entire dataset based on pattern size, which has pattern sizes of 62, 38, and 36 and has the first three highest patterns common to both serine proteases and BCL-2. In the same vein, serine protease has the first three unique highest pattern sizes of 28, 24, and 23, while BCL-2 has the first three unique highest pattern sizes of 38, 35, and 33. These predicted similar structures discovered in protein-protein or protein-peptide complexes can be subjected to additional analysis such as docking and molecular dynamics to determine their potency.

Biologically, having patterns in conserved structures similar to the patterns in our dataset indicates the possibility that both patterns share a genetic background or the presence of convergent evolution. Proteins having structural similarities with other proteins means they can possibly share a common evolutionary origin (Andreeva et al., 2022). This evolutionary relationship generally means that pairwise residue identities between the conserved structure and similar pattern found in our protein-protein dataset are 30% and greater (Masso, 2022). Furthermore, protein similarity

relationships can be based on structural or sequence similarity, and protein energetics can help us understand these relationships. High structural but low sequence similarity is due to unfavorable energetics, while high sequence but low structural similarity is well-represented by proteins that can accommodate large structural changes (Gan et al., 2002). In addition, the biochemical activity of a protein can sometimes be predicted by searching for known proteins that are similar in their amino acid sequences (Alberts et al., 2002). Hence, the presence of similar structures in our protein-protein dataset indicates the possibility that patterns found in our dataset have similar functions to the conserved structures.

References

Acuner Ozbabacan, S. E., Engin, H. B., Gursoy, A., & Keskin, O. (2011). Transient protein protein interactions. *Protein Engineering, Design and Selection*, *24*(9), 635–648. https://doi.org/10.1093/protein/gzr025

Adams, J. N., Schuster, D., Schmitz, S., Schuh, G., & Van Der Aalst, W. M. P. (2022). Defining Cases and Variants for Object-Centric Event Data. *International Conference on Process Mining, ICPM*, *4*, 128–135. https://doi.org/10.1109/ICPM57379.2022.9980730

Ahmed, W. F., & Gomaa, W. (2011). Approaches to prediction of protein structure. *Proceedings of IEEE/ACS International Conference on Computer Systems and Applications, AICCSA*, 284. https://doi.org/10.1109/AICCSA.2011.6126616

Alberts, B., Johnson, A., Julian Lewis, Raff, M., Roberts, K., & Walter, P. (2002). Analyzing Protein Structure and Function. In *Molecular Biology of the Cell*. https://www.ncbi.nlm.nih.gov/books/NBK26820/

Andreeva, A., Kulesha, E., Gough, J., & Murzin, A. (2022). *Structural Classification of Proteins Learn*. SCOP2 Prototype: A New Approach to Protein Structure Mining. http://scop.mrc-lmb.cam.ac.uk/

Badal, V. D., Kundrotas, P. J., & Vakser, I. A. (2015). Text Mining for Protein Docking. *PLoS Computational Biology*, *11*(12), 1–21. https://doi.org/10.1371/journal.pcbi.1004630

Barabasi, A., & Oltvai, Z. (2004). Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, *5*(February), 101–113. https://doi.org/10.1038/nrg1272

Bartholow, T. G., Sztain, T., Patel, A., Lee, D. J., Young, M. A., Abagyan, R., & Burkart, M. D. (2021). Elucidation of transient protein-protein interactions within carrier protein-dependent biosynthesis. *Communications Biology*, *4*(1), 1–10. https://doi.org/10.1038/s42003-021-01838-3

Bertolini, M., Fenzl, K., Kats, I., Wruck, F., Tippmann, F., Schmitt, J., Auburger, J. J., Tans, S., Bukau, B., & Kramer, G. (2021). Interactions between nascent proteins translated by adjacent ribosomes drive homomer assembly. *Science*, *371*(6524). https://doi.org/10.1126/science.abc7151

Botelho, V. B., & Junior, J. Y. S. (2022). THE SOYBEAN PRODUCTION AND

SUPPLY CHAIN IN BRAZIL. *MUUA Grand Challenges Scholars Program*, *3*(2), 58–66. http://www.tjyybjb.ac.cn/CN/article/downloadArticleFile.do?attachType=PDF&id=9987

Bunke, H., & Messmer, B. T. (1995). Efficient Attributed Graph Matching and its Application to Image Analysis. In V. G. Braccini C., DeFloriani L. (Ed.), *International Conference on Image Analysis and Processing* (Vol. 974, pp. 44–55). Springer. https://doi.org/DOI:10.1007/3-540-60298-4_235

Cai, J., Yang, J., & Jones, D. P. (1998). Mitochondrial control of apoptosis: The role of cytochrome c. *Biochimica et Biophysica Acta - Bioenergetics*, *1366*(1–2), 139–149. https://doi.org/10.1016/S0005-2728(98)00109-1

Chang, J., Chapman, B., Friedberg, I., Hamelryck, T., Hoon, M. De, Cock, P., Antao, T., Talevich, E., & Wilczy, B. (2021). *Biopython Tutorial and Cookbook (Biopython 1.79)* (Vol. 2021). http://biopython.org/DIST/docs/tutorial/Tutorial.pdf

Christopher, K., Holde, V., & Edward, K. E. K. (1996). *Biochemistry*. Menlo Park, Calif. : Benjamin/Cummings Pub. Co.

Cibej, U., & Mihelic, J. (2015). Improvements to Ullmann's Algorithm for the Subgraph Isomorphism Problem. *International Journal of Pattern Recognition and Artificial Intelligence*, *29*(7), 1–26. https://doi.org/10.1142/S0218001415500251

Cohen, M., Reichmann, D., Neuvirth, H., & Schreiber, G. (2008). Similar chemistry, but different bond preferences in inter versus intra-protein interactions. *Proteins: Structure, Function and Genetics*, *72*(2), 741–753. https://doi.org/10.1002/prot.21960

Cordella, L. P., Foggia, P., Sansone, C., & Vento, M. (1999). Performance evaluation of the VF graph matching algorithm. *Proceedings - International Conference on Image Analysis and Processing, ICIAP 1999*, *February 1999*, 1172–1177. https://doi.org/DOI:10.1109/ICIAP.1999.797762

Cordella, L. P., Foggia, P., Sansone, C., & Vento, M. (2004). A (sub)graph isomorphism algorithm for matching large graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26*(10), 1367–1372. https://doi.org/10.1109/TPAMI.2004.75

Da Silveira, C. H., Pires, D. E. V., Minardi, R. C., Ribeiro, C., Veloso, C. J. M., Lopes,

J. C. D., Meira, W., Neshich, G., Ramos, C. H. I., Habesch, R., & Santoro, M. M. (2009). Protein cutoff scanning: A comparative analysis of cutoff dependent and cutoff free methods for prospecting contacts in proteins. *Proteins: Structure, Function and Bioinformatics*, *74*(3), 727–743. https://doi.org/10.1002/prot.22187

Day, P. R. (2009). The biology of plant proteins. *Food Science and Nutrition*, *36*(S), 39–47. https://doi.org/https://doi.org/10.1080/10408399609527758

de Souza Gomes, I., Santana, C. A., Marcolino, L. S., de Lima, L. H. F., de Melo-Minardi, R. C., Dias, R. S., de Paula, S. O., & de Azevedo Silveira, S. (2022). Computational prediction of potential inhibitors for SARS-COV-2 main protease based on machine learning, docking, MM-PBSA calculations, and metadynamics. *PLoS ONE*, *17*(4 April), 1–20. https://doi.org/10.1371/journal.pone.0267471

Delbridge, A. R. D., & Strasser, A. (2015). The BCL-2 protein family, BH3-mimetics and cancer therapy. *Cell Death and Differentiation*, *22*(7), 1071–1080. https://doi.org/10.1038/cdd.2015.50

Delbridge, A. R. D., Valente, L. J., & Strasser, A. (2012). The role of the apoptotic machinery in tumor suppression. *Cold Spring Harbor Perspectives in Biology*, *4*(11), 1–15. https://doi.org/10.1101/cshperspect.a008789

Ding, Z., & Kihara, D. (2019). Computational identification of protein-protein interactions in model plant proteomes. *Scientific Reports*, *9*(1), 1–13. https://doi.org/10.1038/s41598-019-45072-8

Dwivedi, S. P. (2019). Some Algorithms on Exact , Approximate and Error-Tolerant Graph Matching [Banaras Hindu University]. In *PhD Thesis*. https://arxiv.org/pdf/2012.15279.pdf

Edwards, A. M., Kus, B., Jansen, R., Greenbaum, D., Greenblatt, J., & Gerstein, M. (2002). Bridging structural biology and genomics: Assessing protein interaction data with known complexes. *Trends in Genetics*, *18*(10), 529–536. https://doi.org/10.1016/S0168-9525(02)02763-4

Ehrlich, H. C., & Rarey, M. (2012). Systematic benchmark of substructure search in molecular graphs - From Ullmann to VF2. *Journal of Cheminformatics*, *4*(7), 1–17. https://doi.org/10.1186/1758-2946-4-13

Fassio, A. V., Santana, C. A., Cerqueira, F. R., da Silveira, C. H., Romanelli, J. P. R., de Melo-Minardi, R. C., & Silveira, S. de A. (2018). An interactive strategy to visualize common subgraphs in protein-ligand interaction. *Lecture Notes in*

Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), *10813 LNBI*, 383–394. https://doi.org/10.1007/978-3-319-78723-7_33

Fassio, A. V., Santos, L. H., Silveira, S. A., Ferreira, R. S., & De Melo-Minardi, R. C. (2020). NAPOLI: A Graph-Based Strategy to Detect and Visualize Conserved Protein-Ligand Interactions in Large-Scale. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *17*(4), 1317–1328. https://doi.org/10.1109/TCBB.2019.2892099

Gan, H. H., Perlow, R. A., Roy, S., Ko, J., Wu, M., Huang, J., Yan, S., Nicoletta, A., Vafai, J., Sun, D., Wang, L., Noah, J. E., Pasquali, S., & Schlick, T. (2002). Analysis of protein sequence/structure similarity relationships. *Biophysical Journal*, *83*(5), 2781–2791. https://doi.org/10.1016/s0006-3495(02)75287-9

Garcia-Garcia, J., Bonet, J., Guney, E., Fornes, O., Planas, J., & Oliva, B. (2012). Networks of protein-protein interactions: From uncertainty to molecular details. *Molecular Informatics*, *31*(5), 342–362. https://doi.org/10.1002/minf.201200005

Gonçalves-Almeida, V. M., Pires, D. E. V., De melo-minardi, R. C., Da silveira, C. H., Meira, W., & Santoro, M. M. (2012). Hydropace: Understanding and predicting cross-inhibition in serine proteases through hydrophobic patch centroids. *Bioinformatics*, *28*(3), 342–349. https://doi.org/10.1093/bioinformatics/btr680

Gonzalez, M. W., & Kann, M. G. (2012). Chapter 4: Protein Interactions and Disease. *PLoS Computational Biology*, *8*(12). https://doi.org/10.1371/journal.pcbi.1002819

Gouda, K., Bujdosó, G., & Hassaan, M. (2022). Scaling Subgraph Matching by Improving Ullmann Algorithm. *Computing and Informatics*, *41*(4), 1002–1024. https://doi.org/10.31577/cai_2022_4_1002

Grbic, M., Crnogorac, V., Predojevic, M., Kartelj, A., & Matic, D. (2020). How well are known protein complexes supported in PPI networks? *INISTA 2020 - 2020 International Conference on INnovations in Intelligent SysTems and Applications, Proceedings*. https://doi.org/10.1109/INISTA49547.2020.9194663

Guo, Z. X., Luo, L. M., & Zhang, S. D. (2022). Composite Reuse Method Based On Graph Database. *International Conference on Advances in Computer Technology, Information Science and Communications*, *4*. https://doi.org/10.1109/CTISC54888.2022.9849796

Hanahan, D., & Weinberg, R. (2011). Hallmarks of Cancer: Supplement. In *Cell*

Press.

Jaiswal, M., Saleem, S., Kweon, Y., Draizen, E. J., Veretnik, S., Mura, C., & Bourne, P. E. (2020). Deep Learning of Protein Structural Classes: Any Evidence for an “Urfold”? *2020 Systems and Information Engineering Design Symposium, SIEDS 2020*, 1–6. https://doi.org/10.1109/SIEDS49339.2020.9106642

Keskin, O., Tuncbag, N., & Gursoy, A. (2016). Predicting Protein-Protein Interactions from the Molecular to the Proteome Level. *Chemical Reviews*, *116*(8), 4884–4909. https://doi.org/10.1021/acs.chemrev.5b00683

Koutrouli, M., Karatzas, E., Paez-Espino, D., & Pavlopoulos, G. A. (2020). A Guide to Conquer the Biological Network Era Using Graph Theory. *Frontiers in Bioengineering and Biotechnology*, *8*(January), 1–26. https://doi.org/10.3389/fbioe.2020.00034

Kroemer, G. (2003). Mitochondrial control of apoptosis: An introduction. *Biochemical and Biophysical Research Communications*, *304*(3), 433–435. https://doi.org/10.1016/S0006-291X(03)00614-4

Kurniawan, A., Jatmiko, W., Hertadi, R., & Habibie, N. (2020). Prediction of protein tertiary structure using pre-Trained self-supervised learning based on transformer. *2020 International Workshop on Big Data and Information Security, IWBIS 2020*, 73–78. https://doi.org/10.1109/IWBIS50925.2020.9255624

Kusari, A., Sun, W., & Arbor, A. (2022). *Uncertainty-aware Efficient Subgraph Isomorphism using Graph Topology* (pp. 1–20). https://arxiv.org/pdf/2209.09090.pdf

Kuzmanov, U., & Emili, A. (2013). Protein-protein interaction networks: probing disease mechanisms using model systems. *Genome Medicine*, *5*(37), 1–12.

Laskowski, R. A., & Thornton, J. M. (2015). Proteins: Interaction at a distance. *IUCrJ*, *2*(2015), 609–610. https://doi.org/10.1107/S2052252515020217

Lee, S., Braun, C. R., Bird, G. H., & Walensky, L. D. (2014). Photoreactive stapled peptides to identify and characterize bcl-2 family interaction sites by mass spectrometry. In *Methods in Enzymology* (1st ed., Vol. 544). Elsevier Inc. https://doi.org/10.1016/B978-0-12-417158-9.00002-9

Masso, M. (2022). Protein Structure Analysis. In *Lecture Note* (p. 49). http://www.binf.gmu.edu/jafri/binf630/Lecture10.pdf

McKay, B. D. (1981). Practical graph isomorphism. In *Congressus Numerantium* (Vol.

30, pp. 45–87). http://users.cecs.anu.edu.au/~bdm/papers/pgi.pdf

McKay, B. D., & Piperno, A. (2014). Practical graph isomorphism, II. *Journal of Symbolic Computation*, *60*, 94–112. https://doi.org/10.1016/j.jsc.2013.09.003

Medlineplus. (2021). *What are proteins and what do they do ?* National Library of Medicine. https://medlineplus.gov/genetics/understanding/howgeneswork/protein/

Murray, R. K., Granner, D. K., Mayes, P. A., & Rodwell, V. W. (2016). Overview of Metabolism. In *Harper's Illustrated Biochemistry*. New York: Lange Medical Books/McGraw-Hill. https://doi.org/10.5005/jp/books/13014_10

Nabti, C., & Seba, H. (2016). Subgraph isomorphism search in massive graph databases. *IoTBD 2016 - Proceedings of the International Conference on Internet of Things and Big Data*, 204–213. https://hal.archives-ouvertes.fr/hal-01313922

Nooren, I. M. A., & Thornton, J. M. (2003). Diversity of protein-protein interactions. *EMBO Journal*, *22*(14), 3486–3492. https://doi.org/10.1093/emboj/cdg359

Opferman, J. T., & Kothari, A. (2018). Anti-apoptotic BCL-2 family members in development. *Cell Death and Differentiation*, *25*(1), 37–45. https://doi.org/10.1038/cdd.2017.170

Pan, Y. (2007). Protein Structure Prediction and Its Understanding Based on Machine Learning Methods. *International Symposium on BioInformatics and BioEngineering*. https://doi.org/10.1109/bibe.2007.4375533

Phizicky. (2018). Creative Proteomics Blog. In *BRIEF INTRODUCTION OF PROTEIN-PROTEIN INTERACTION (PPI)*. https://www.creative-proteomics.com/blog/index.php/brief-introduction-of-protein-protein-interaction-ppi/

Pires, D. E. V., de Melo-Minardi, R. C., dos Santos, M. A., da Silveira, C. H., Santoro, M. M., & Meira, W. (2011). Cutoff Scanning Matrix (CSM): Structural classification and function prediction by protein inter-residue distance patterns. *BMC Genomics*, *12*(SUPPL. 4). https://doi.org/10.1186/1471-2164-12-S4-S12

Queiroz, F. C., Vargas, A. M. P., Oliveira, M. G. A., Comarela, G. V., & Silveira, S. A. (2020). PpiGReMLIN: A graph mining based detection of conserved structural arrangements in protein-protein interfaces. *BMC Bioinformatics*, *21*(1), 1–25.

https://doi.org/10.1186/s12859-020-3474-1

Qvit, N., & Crapster, J. A. (2014). Peptides that target protein-protein interactions as an anti-parasite strategy. *Chimica Oggi/Chemistry Today*, *32*(6), 31–36.

Raman, K. (2010). Construction and analysis of protein – protein interaction networks. *Automated Experimentation*, *2*(1), 1–11. https://doi.org/doi:10.1186/1759-4499-2-2

Rao, V. S., Srinivas, K., Sujini, G. N., & Kumar1, G. N. S. (2014). Protein-Protein Interaction Detection: Methods and Analysis. *International Journal of Proteomics*, 341–365. https://doi.org/http://dx.doi.org/10.1155/2014/147648

Rasheed, F., Markgren, J., Hedenqvist, M., & Johansson, E. (2020). Modeling to Understand Plant Protein Structure-Function Relationships—Implications for Seed Storage Proteins. *Molecules*, *25*(4), 873–889.

Reed, J. C., Tanaka, S., Takayama, S., Schibler, M. J., & Fenton, W. (1993). Investigation of the Subcellular Distribution of the bcl-2 Oncoprotein: Residence in the Nuclear Envelope, Endoplasmic Reticulum, and Outer Mitochondrial Membranes. *Cancer Research*, *53*(19), 4701–4714.

Rehman, Z. ur, Idris, A., & Khan, A. (2018). Multi-Dimensional Scaling based grouping of known complexes and intelligent protein complex detection. *Computational Biology and Chemistry*, *74*(June), 149–156. https://doi.org/https://doi.org/10.1016/j.compbiolchem.2018.03.023

Reichmann, D., Phillip, Y., Carmi, A., & Schreiber, G. (2008). On the contribution of water-mediated interactions to protein-complex stability. *Biochemistry*, *47*(3), 1051–1060. https://doi.org/10.1021/bi7019639

Ribeiro, V. S., Santana, C. A., Fassio, A. V., Cerqueira, F. R., Da Silveira, C. H., Romanelli, J. P. R., Patarroyo-Vargas, A., Oliveira, M. G. A., Gonçalves-Almeida, V., Izidoro, S. C., De Melo-Minardi, R. C., & Silveira, S. D. A. (2020). VisGReMLIN: Graph mining-based detection and visualization of conserved motifs at 3D protein-ligand interface at the atomic level. *BMC Bioinformatics*, *21*(Suppl 2), 1–12. https://doi.org/10.1186/s12859-020-3347-7

Santana, C. A., Cerqueira, F. R., De Silveira, C. H., Fassio, A. V., De Melo-Minardi, R. C., & Silveira, S. D. A. (2016). GReMLIN: A Graph Mining Strategy to Infer Protein-Ligand Interaction Patterns. *Proceedings - 2016 IEEE 16th International Conference on Bioinformatics and Bioengineering, BIBE 2016*, 28–35.

https://doi.org/10.1109/BIBE.2016.48

Sarkar, D., & Saha, S. (2019). Machine-learning techniques for the prediction of protein–protein interactions. *Journal of Biosciences*, *44*(4). https://doi.org/10.1007/s12038-019-9909-z

Schreiber, G. (2020). CHAPTER 1: Protein-Protein Interaction Interfaces and their Functional Implications. In *RSC Drug Discovery Series* (Vols. 2021-Janua, Issue 78). https://doi.org/10.1039/9781788016544-00001

Sharma, P. K., & Yadav, I. S. (2022). Biological databases and their application. In *Bioinformatics Methods and Applications* (pp. 17–31). Elsevier Inc. https://doi.org/https://doi.org/10.1016/B978-0-323-89775-4.00021-3

Silva, L., Pereira, C., & Arrais, J. P. (2020). Using a Novel Unbiased Dataset and Deep Learning Architectures to Predict Protein-Protein Interactions. *Proceedings - 2020 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2020*, *DI*, 213–216. https://doi.org/10.1109/BIBM49941.2020.9313393

Silveira, S. A., Fassio, A. V., Gonçalves-Almeida, V. M., De Lima, E. B., Barcelos, Y. T., Aburjaile, F. F., Rodrigues, L. M., Meira, W., & De Melo-Minardi, R. C. (2014). VERMONT: Visualizing mutations and their effects on protein physicochemical and topological property conservation. *BMC Proceedings*, *8*(Suppl 2), 1–10. https://doi.org/10.1186/1753-6561-8-S2-S4

Song, H., Dai, Z., Xu, P., & Ren, L. (2022). Interactive Visual Pattern Search on Graph Data via Graph Representation Learning. *IEEE Transactions on Visualization and Computer Graphics*, *28*(1), 335–345. https://doi.org/10.1109/TVCG.2021.3114857

Tariq, T., Frezund, J., Farhan, M., Latif, R. M. A., & Mehmood, A. (2020). Structure Analysis of Protein Data Bank Using Python Libraries. *Proceedings of 2020 17th International Bhurban Conference on Applied Sciences and Technology, IBCAST 2020*, 201–209. https://doi.org/10.1109/IBCAST47879.2020.9044525

Terentiev, A. A., Moldogazieva, N. T., & Shaitan, K. V. (2009). Dynamic proteomics in modeling of the living cell. Protein-protein interactions. *Biochemistry (Moscow)*, *74*(13), 1586–1607. https://doi.org/10.1134/S0006297909130112

Titeca, K., Lemmens, I., Tavernier, J., & Eyckerman, S. (2018). Discovering cellular protein-protein interactions: Technological strategies and opportunities. *Mass Spectrometry Reviews*, *38*(1), 79–111. https://doi.org/10.1002/mas.21574

Toloi, M. N. V., Bonilla, S. H., Toloi, R. C., Silva, H. R. O., & Nääs, I. de A. (2021). Development indicators and soybean production in Brazil. *Agriculture (Switzerland)*, *11*(11), 1–15. https://doi.org/10.3390/agriculture11111164

Ullmann, J. R. (1976). An Algorithm for Subgraph Isomorphism. *Journal of the ACM (JACM)*, *23*(1), 31–42. https://doi.org/10.1145/321921.321925

Valerie, H., Tran, G., & Kaushik, S. (2020). Soybean meal. In *Feedipedia, a programme by INRAE, CIRAD, AFZ and FAO.: Vol. March*. https://www.feedipedia.org/node/674 Last

Vianna, U. R., Pratissoli, D., Zanuncio, J. C., Alencar, J. R. C. C. de, & Zinger, F. D. (2011). Espécies E/Ou Linhagens De Trichogramma Spp. (Hymenoptera: Trichogrammatidae) Para O Controle De Anticarsia Gemmatalis (Lepidoptera: Noctuidae). *Arquivos Do Instituto Biológico*, *78*(1), 81–87. https://doi.org/10.1590/1808-1657v78p0812011

Wodak, S. J., Vlasblom, J., Turinsky, A. L., & Pu, S. (2013). Protein-protein interaction networks: The puzzling riches. *Current Opinion in Structural Biology*, *23*(6), 941–953. https://doi.org/10.1016/j.sbi.2013.08.002

Xu, X., & Zhang, Y. (2022). Commodity price forecasting via neural networks for coffee, corn, cotton, oats, soybeans, soybean oil, sugar, and wheat. *Intelligent Systems in Accounting, Finance and Management*, *29*(3), 169–181. https://doi.org/10.1002/isaf.1519

Yan, Y., Zhang, S., & Wu, F. X. (2011). Applications of graph theory in protein structure identification. *Proteome Science*, *9*(SUPPL. 1), 1–10. https://doi.org/10.1186/1477-5956-9-S1-S17

Yang, F., Fan, K., Song, D., & Lin, H. (2020). Graph-based prediction of Protein-protein interactions with attributed signed graph embedding. *BMC Bioinformatics*, *21*(1), 1–16. https://doi.org/10.1186/s12859-020-03646-8

Yu, Y., & Kong, D. (2022). Protein complexes detection based on node local properties and gene expression in PPI weighted networks. *BMC Bioinformatics*, *23*(1), 1–15. https://doi.org/10.1186/s12859-021-04543-4

Zhang, J., Wang, Y., & Zhao, W. (2018). An improved probabilistic relaxation method for matching multi-scale road networks. *International Journal of Digital Earth*, *11*(6), 635–655. https://doi.org/10.1080/17538947.2017.1341557

Appendix 1: Link to the result of visualization

https://ecventures.com.ng/research

Appendix 2: Pattern size and occurrence in both serine protease and BCL-2

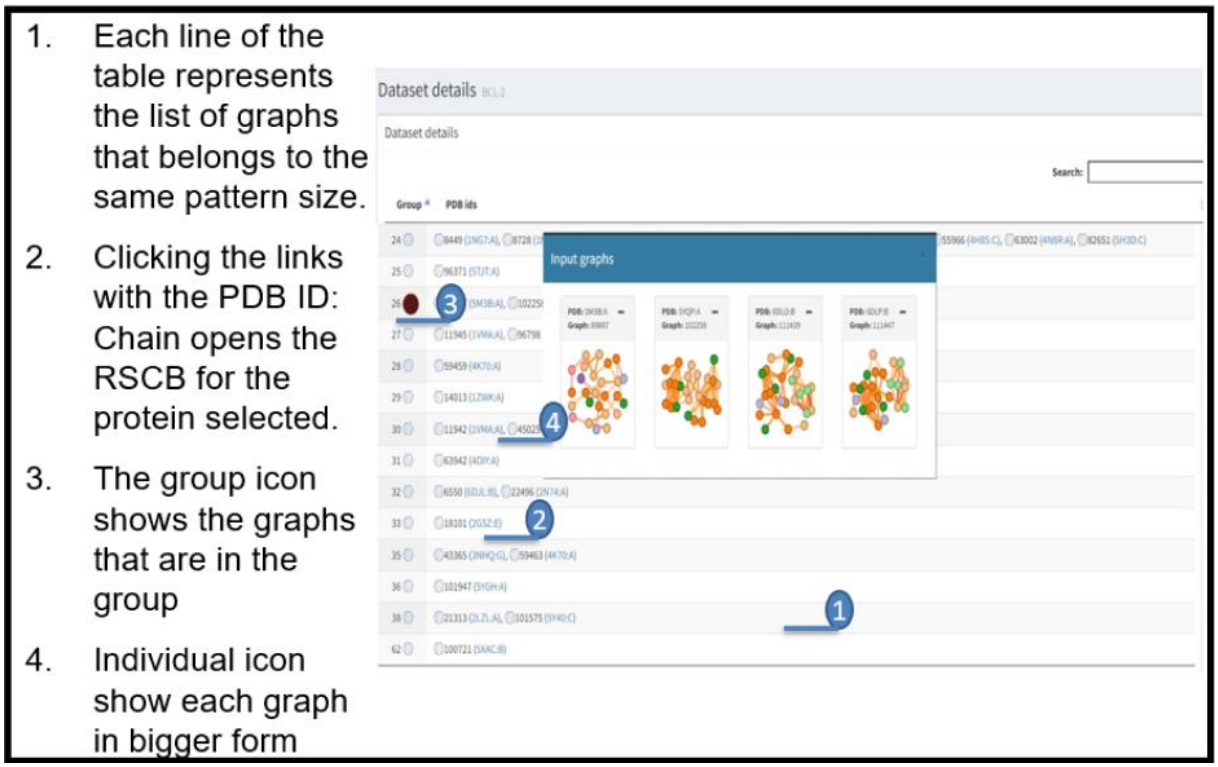
Pattern Size	Same	Difference	
		Serine	BCL-2
62	100721 (5XAC:B)		
38	21313 (2LZL:A),		101575 (5Y40:C)
36	101947 (5YGH:A)		
35			43365 (3NHQ:G), 59463 (4K70:A)
33			18101 (2GSZ:E)
32			6550 (6DJL:B), 22496 (2N74:A)
31			63942 (4OIY:A)
30	100699 (5XAC:B)		11942 (1VMA:A), 45025 (3PJZ:A),
29			14013 (1ZWK:A)
28		37233 (3HU3:A)	59459 (4K70:A)
27			11945 (1VMA:A), 96798 (5U9J:A)
26			89887 (5M3B:A), 102258 (5YQP:A), 111439 (6DLO:B), 111447 (6DLP:B)
25	96371 (5TJT:A)		
24	8449 (1NG7:A), 10558 (1SFK:E), 22638 (2NAO:E),	37159 (3HU1:C), 37171 (3HU1:E)	8728 (1NS1:A), 19165 (2IBM:A), 34679 (3FHN:A), 49889 (3V64:A), 55966 (4H8S:C), 63002 (4N8R:A), 82651 (5H3D:C)
23	974 6 (1Q55:C), 22604 (2NAO:A),	37148 (3HU1:A), 37199 (3HU2:C), 37212 (3HU2:E), 43984 (3O78:A)	7662 (1MN8:C), 8399 (1NCE:A), 8412 (1NCE:A), 12600 (1XEQ:A), 33895 (3EIJ:A), 43330 (3NHQ:C), 88596 (5L9S:A)
22	8054 (1N0F:C),	15024 (2BWE:A), 15034 (2BWE:E), 15039 (2BWE:G), 15048 (2BWE:O), 21923 (2MX9:A)	250 (2DH3:A), 8614 (1NO7:A), 16878 (2F36:A), 63003 (4N8R:A), 73573 (4XH9:B), 91964 (5NUG:A), 96255 (5TEB:A), 97353 (5UWA:A)
21	7193 (1L0O:A),	37185 (3HU2:A)	11483 (1U4Q:A), 24592 (2QDI:A), 31834 (3C2W:C), 31855 (3C2W:E), 33130 (3DED:C), 43981 (3O66:A), 109294 (6CLX:A)
20	2965 (1DXX:A), 2987 (1DXX:C), 26099 (2VEE:E), 52706 (4APW:E), 52731 (4APW:G), 52771 (4APW:I), 52796 (4APW:K),	21921 (2MX9:A),	1282 (4EO2:A), 1301 (4EO2:C), 1320 (4EO2:E), 1342 (4EP0:A), 1384 (4EP0:E), 9033 (1OOF:A), 9040 (1OOG:A), 11149 (1T72:F), 27129 (2WHU:A), 32845 (3CVF:C), 33124 (3DED:A), 33335 (3DOB:A), 36494 (3H51:A), 71620 (4W75:A), 82844 (5H89:C), 91121 (5NBD:A), 103155 (5ZFU:A), 105598 (6AMC:A), 105631 (6AMH:A), 105666 (6AMI:A), 105682 (6AMI:C), 113017 (6E27:C)

	89784 (5M1G:A),		
19	13278 (1Z2W:A),		14545 (2AWO:C), 18633 (2HMV:A), 31748 (3BXJ:A), 33135 (3DED:E), 33890 (3EIJ:A), 34067 (3ES2:A), 48361 (3TFZ:E), 48363 (3TFZ:E), 52528 (4AE5:A), 63734 (4O6C:E), 66491 (4Q7T:A), 71404 (4V07:A), 71444 (4V2E:A), 71610 (4W72:A), 84599 (5IWN:C), 84714 (5IXZ:A), 98418 (5VRF:B), 102358 (5YT1:A), 105564 (6AM7:A), 105616 (6AMC:C), 105942 (6AXG:J)
18	7521 (1LYN:A), 91150 (5NBD:A)		321 (2FU3:A), 9905 (1QDV:A), 11143 (1T72:D), 18731 (2HYD:A), 25300 (2RA6:C), 28401 (2Y3V:C), 32241 (3CF0:E), 32249 (3CF0:G), 32263 (3CF0:I), 42036 (3M95:A), 42216 (3MLH:A), 43837 (3O1J:A), 46065 (3RA5:A), 47287 (3SAQ:A), 56320 (4HJ3:A), 58273 (4J77:A), 59276 (4JUP:A), 71423 (4V08:A), 73754 (4XSD:C), 74345 (4YJM:A), 74351 (4YJX:A), 74366 (4YKA:A), 80692 (5F4Y:A), 84290 (5INE:A), 89848 (5M2X:C), 98405 (5VRF:B), 103632 (5ZSU:A), 103655 (5ZSU:C), 103677 (5ZSU:E), 108533 (6CCB:E), 114073 (6EK4:A)
17	10540 (1SFK:C), 95344 (5T2P:E), 96858 (5UF5:A),	15028 (2BWE:C), 15043 (2BWE:K), 15051 (2BWE:Q), 38193 (3JA7:A), 38237 (3JA7:C), 38281 (3JA7:E), 38325 (3JA7:G), 38369 (3JA7:I), 38413 (3JA7:K), 39608 (3KGL:E), 78516 (5BXB:A), 78530 (5BXB:G),	11835 (1VB6:A), 14858 (2BB6:A), 15076 (2C04:A), 19988 (2JA3:A), 24357 (2Q0A:A), 27067 (2WFN:A), 30448 (3ADR:A), 31984 (3C9I:A), 33962 (3EPM:A), 33970 (3EPN:A), 33996 (3EPO:A), 37806 (3IKY:A), 37813 (3IKY:C), 37820 (3IKY:E), 37827 (3IKY:G), 37834 (3IKY:I), 37841 (3IKY:K), 42903 (3N0W:A), 43244 (3NF5:A), 46079 (3RBX:A), 46090 (3RBX:A), 46117 (3RBX:C), 46128 (3RBX:C), 46166 (3RBX:E), 48444 (3TMR:A), 48987 (3U94:A), 49873 (3UYX:A), 56965 (4HZI:A), 57141 (4I63:C), 66322 (4Q28:A), 73735 (4XSD:A), 81440 (5FVI:A), 81448 (5FVI:C), 82355 (5H1Q:A), 82374 (5H1Q:C), 82393 (5H1Q:E), 82412 (5H1Q:G), 84683 (5IXV:A), 97662 (5V85:A), 99637 (5WJ5:A), 99674 (5WJ5:C), 102799 (5ZAB:O), 105430 (6AJE:C), 108231 (6C3R:A), 114082 (6EK4:C)
16	3542 (1F3C:A), 6715 (6OL1:C), 9731 (1Q55:A),	33902 (3ELG:A), 39575 (3KGL:A), 40853 (3LJ5:A),	130 (1T3E:A), 4809 (1H4R:A), 5178 (1I3C:A),

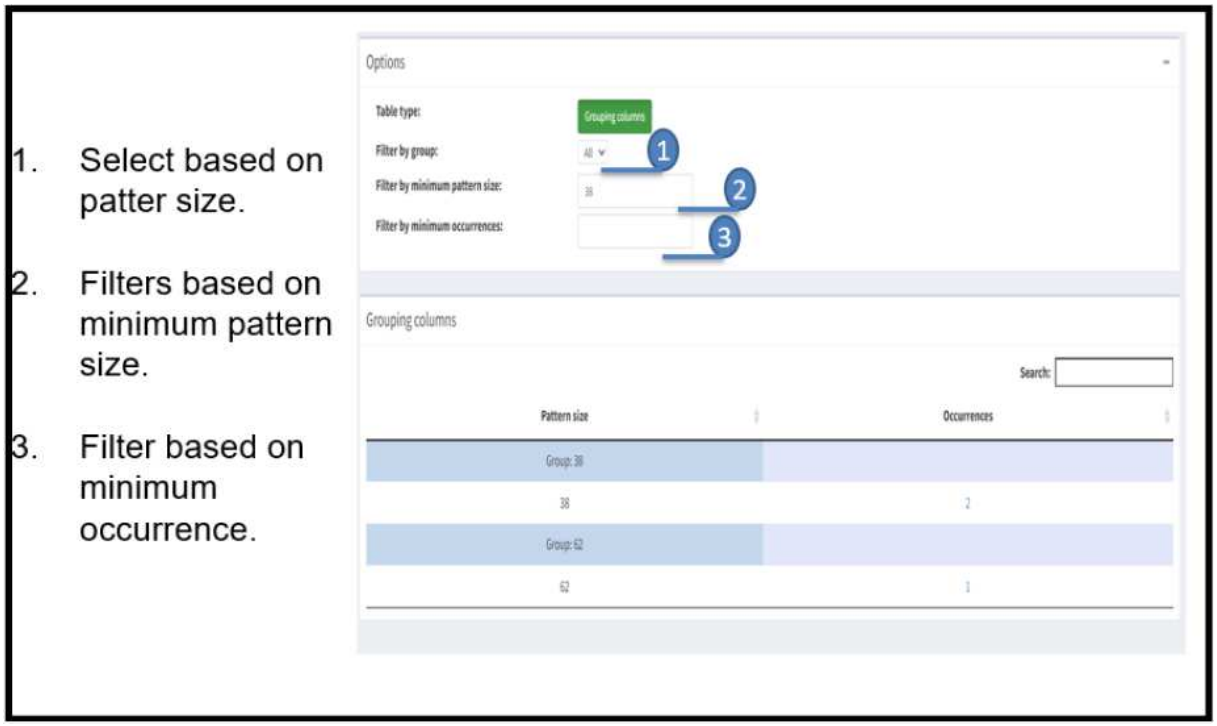
	9770 (1Q5C:A), 9789 (1Q5C:C), 13292 (1Z2X:A), 14330 (2ABM:G), 23889 (2P4M:A), 23941 (2P4M:G), 25680 (2RH7:A), 33058 (3D7P:A), 40949 (3LJ5:E), 41092 (3LJ5:K),), 41651 (3LVU:A), 53559 (4BW5:A), 57378 (4IK7:A), 58355 (4J7C:K), 59017 (4JQ5:A), 59107 (4JQ5:K), 68023 (4RPT:A), 73296 (4XDK:A), 87471 (5KHN:B),	40904 (3LJ5:C), 40997 (3LJ5:G), 41047 (3LJ5:I),	8753 (1NYJ:A), 8757 (1NYJ:C), 11818 (1V9Z:A), 13205 (1YZW:A), 13225 (1YZW:C), 19604 (2IUB:I), 24355 (2Q0A:A), 24481 (2Q7F:A), 24525 (2Q9B:A), 25824 (2V4E:C), 25838 (2V4E:E), 25850 (2V4E:G), 27012 (2WET:A), 28062 (2XUL:E), 31678 (3BXB:E), 32234 (3CF0:A), 32273 (3CF0:K), 32283 (3CF0:M), 32992 (3D31:C), 33991 (3EPN:A), 34013 (3EPO:A), 35760 (3GL4:A), 41380 (3LQM:A) 44802 (3PDI:B), 47933 (3T3A:A), 48479 (3TMT:C), 48599 (3TTN:A), 50398 (3W9F:C), 53165 (4B90:A), 53180 (4B91:A), 53224 (4BBD:E), 53256 (4BJM:A), 59558 (4KGE:A), 59574 (4KGF:A), 63130 (4NHF:A), 63138 (4NHF:C), 63147 (4NHF:E), 63683 (4O6C:A), 63707 (4O6C:C), 63712 (4O6C:C), 63782 (4O9U:E), 63784 (4O9U:E), 64742 (4P4H:I), 65025 (4PD3:A), 69342 (4TXV:B), 71521 (4W6B:A), 71649 (4W7R:A), 71652 (4W7R:C), 72310 (4WYK:B), 73886 (4XVN:E), 74386 (4YKJ:A), 79250 (5CM6:A), 84623 (5IWO:A), 85662 (5JAE:A), 85912 (5JJ1:G), 87074 (5K5S:A), 88268 (5L22:B), 93295 (5OMC:C), 96579 (5TZ0:A), 97542 (5V6B:A), 101021 (5XJA:A), 102369 (5YT1:C), 103190 (5ZFU:E), 103872 (5ZVS:D), 104494 (6A6F:A), 108863 (6CJT:A), 108889 (6CJT:C), 110836 (6DEJ:A), 110862 (6DEJ:C), 114266 (6EU7:C), 114269 (6EU7:E)
15	2274 (5L1B:A), 2397 (1BFM:A), 9788 (1Q5C:C), 21705 (2MN6:B), 23906 (2P4M:C), 23924 (2P4M:E), 24000 (2PE9:A), 24648 (2QEL:C), 39948 (3KXS:A), 43890 (3O4X:E), 52459 (4AC4:A), 53453	32778 (3CPI:G), 66641 (4Q91:A), 66644 (4Q9G:A), 66676 (4QAL:A), 66679 (4QBC:A), 66682 (4QBV:A), 66693 (4QC4:A), 66796 (4QO3:A), 67669 (4RG6:A), 67743 (4RG9:A), 96420 (5TJT:C), 101839 (5YBC:C),	2337 (1AN7:A), 7926 (1MWS:A), 9049 (1OOH:A), 11046 (1T14:A), 14522 (2AWN:C), 16283 (2E4W:A), 16288 (2E4X:A), 19570 (2IUB:A), 19587 (2IUB:G), 23081 (2NZ0:B), 23725 (2OTE:A), 24052 (2PL2:A), 27006 (2WES:A), 27290 (2WRB:A), 27315 (2WRC:A), 28042 (2XUL:A), 28586 (2YMN:C), 30428 (3ADF:A), 31212 (3B7A:A), 33377 (3DPU:A), 33945 (3EPM:A), 39557 (3KG5:A), 45249 (3PXG:C), 45258 (3PXG:E), 46154 (3RBX:E), 47687 (3SVR:C), 48358 (3TFZ:C), 48469 (3TMT:A), 50782 (3WVQ:A), 50789 (3WVQ:A), 50801 (3WVQ:C), 59174 (4JTX:B), 59219 (4JU0:J), 63679 (4O6C:A), 65222 (4PKO:Y), 67243 (4QYN:A), 68014

	(4BMG:C), 53592 (4BW5:A), 54621 (4CZ8:A), 55070 (4DES:A), 57295 (4IIZ:A), 57298 (4IIZ:A), 57382 (4IK7:A), 57420 (4IKJ:A), 59077 (4JQ5:G), 71910 (4WD9:A), 73383 (4XDL:A), 82330 (5H0X:B), 110785 (6DD9:D),		(4RPD:A), 71636 (4W7A:C), 76894 (5AKP:A), 77680 (5B3D:A), 79361 (5CNJ:A), 80695 (5F4Y:A), 81094 (5FM5:O), 83660 (5HL8:A), 84028 (5I27:A), 84944 (5J1H:A), 84990 (5J4E:C), 85801 (5JJ1:A), 86328 (5JJ5:A), 86990 (5K3M:A), 89810 (5M29:A), 97389 (5UYR:A), 97514 (5V6B:A), 100624 (5X5Y:B), 101396 (5XT6:A), 102957 (5ZDQ:C), 103467 (5ZRX:A), 110254 (6CVL:A), 114065 (6EJP:C), 114264 (6EU7:A)

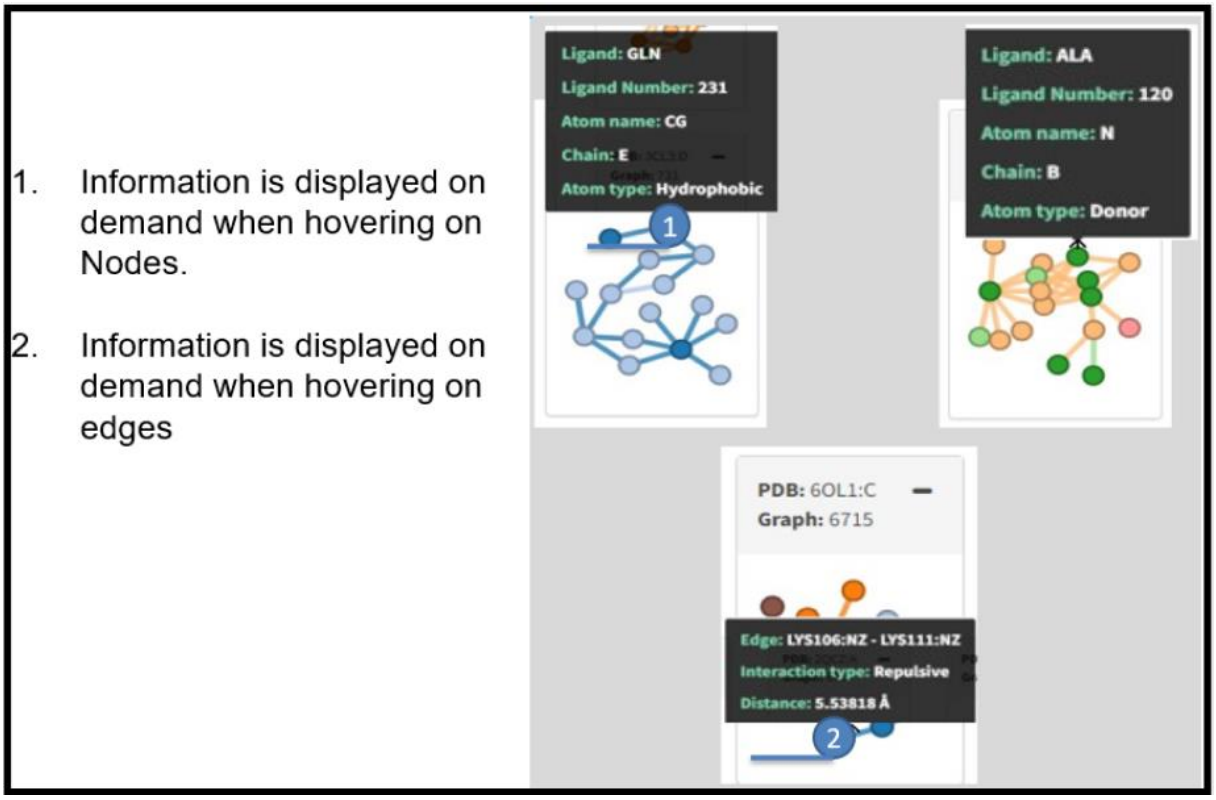
Appendix 3: Sample for the visualization of similar pattern size



Group of patterns based on size

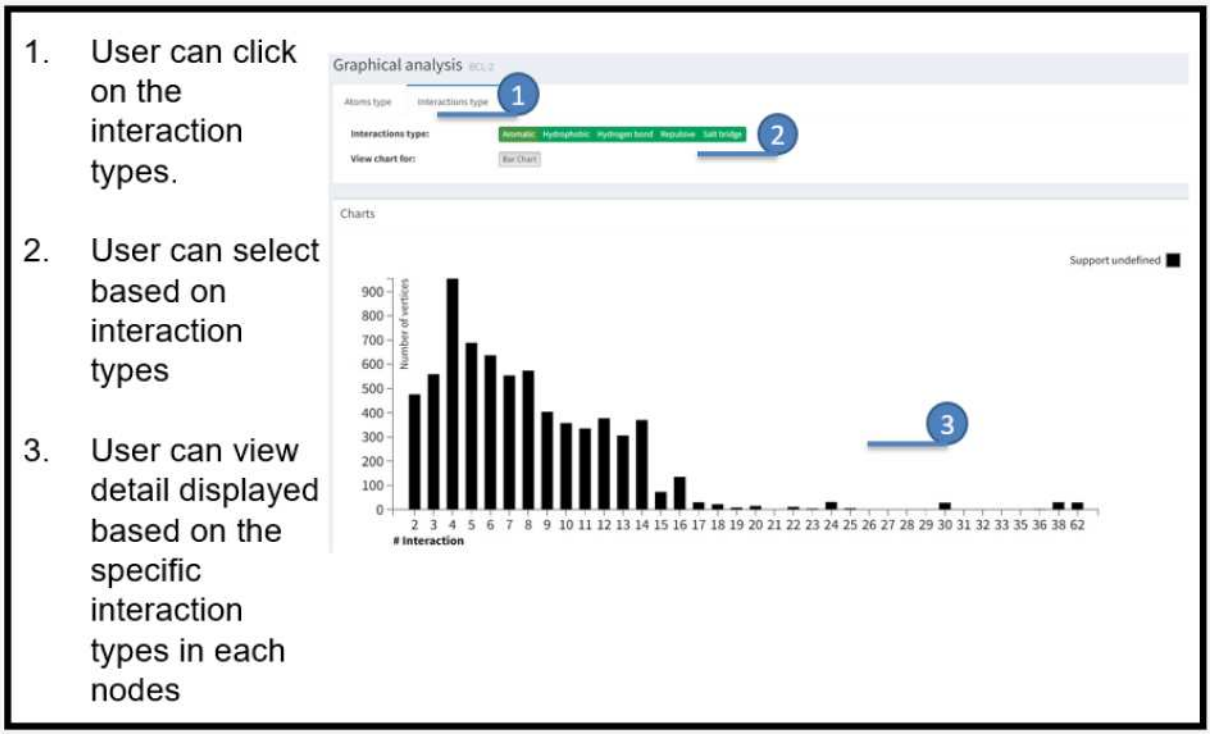


Filtering options to visualize result based on patterns size and occurrence



Graph patterns and detail for user interaction

Graphical analysis based on atom types



Graphical analysis based on interaction types