University of Padova

Department of Department of Mathematics

Master Thesis in Data Science

HomologyRing: A Python Tool for Analyzing Intra- and Inter-Chain Contact Specificity in Protein Families

Supervisor
Damiano Piovesan
University of Padova

Master Candidate
Tanner Aaron Graves

Academic Year
2023–2024

Abstract

Physicial and chemical interactions, or contacts, between peptides are principally responsible for stabilizing a proteins structure and, consequently, determining its function. AlphaFold 2, developed by DeepMind, has proven capable of predicting protein structure with unprecedented accuracy. With predictions available for the majority of known proteins, AlphaFold has proven transformative for bioinformatics, resulting in an abundance of high-quality structural data. RING is a software that deterministically predicts non-covalent interactions in such structure files to create a residue interaction network (RIN). RINs created by RING capture high accuracy contact information with a wide verity of uses including: aiding in human understanding of protein structure or function, and training machine learning models. HomologyRing is a Python package that combines the functionality of RING with a homology search, enabling the analysis of contact conservation and variance across many of evolutionarily related proteins. Starting from a query structure or sequence, HomologyRing uses BLAST to perform a homology search against either UniProt or Protein Data Bank (PDB) databases, and subsequently using RING to collect contact information for corresponding AlphaFold or PDB structures. By mapping contacts to corresponding residues in a multiple sequence alignment (MSA), the pipeline at the core of HomologyRing synthesises a novel Homology enriched Residue Interaction Network (hRIN), and supplamentary tools included within the HomologyRing package aid in hRIN analysis. Using these tools, we demonstrate the utility of the resulting hRINS for characterizing how preservation and variance of contacts in homologs contributes to protein structure, function, and partner binding. HomologyRing compiles and visualizes detailed information on intra- and inter-chain contacts, and shows promise for a wide verity of potential applications, including: study of ligand, and partner binding specificity.

Contents

Listing of figures

Listing of tables

Listing of acronyms

RIN . . . . . Residue Interaction Network

hRIN . . . . . Homology enriched Residue Interaction Network

RING . . . . . Residue Interaction Netork Generator

MSA . . . . . Multiple Sequence Alignment

BLAST . . . . . Basic Local Alignment Search Tool

CIF . . . . . Crystallographic Information File

PDB . . . . . Protein Data Bank

VDW . . . . . Van Der Waals forces

NCBI . . . . . National Center for Biotechnology Information

JSON . . . . . JavaScript Object Notation

1 Introduction

Residue Interaction Networks (RINs) have received a lot of attention in Bioinformatics as useful means of capturing representing and analyzing many aspects of proteins in a simplified form. They represent proteins as a mathematical graph or network where nodes identify amino acid residues in a protein chain, and the edges between them represent various contacts or forces. This work describes an extension to conventional RINs created by synthesizing networks for a family of homologous proteins. We have termed this 'homology enriched Residue Interaction Network' (hRIN) which is created with a software pipeline, part of a greater HomologyRing python package that includes additional tools and visualizations of their analysis. We show how incorporating evolutionary information into RINs may be used to identify residues of particular significance for maintaining conformation, partner binding, or function.

However, RINs, like many aspects of bioinformatics, borrow heavily from many different disciplines, namely: Biology, Mathematics, and Computer Science. As such, a basic familiarity with concepts from each of these fields is required for understanding the abstractions with which RINs are built. This section serves to accommodate readers who may lack some requisite knowledge needed for discussion of RINs. This will include a brief introduction to central biological dogma and an account of relevant topics in structural bioinformatics. We then introduce some of the methods from mathematics and computer science that have proven useful for the study of biological information. Namely, biological sequence, or omics, processing; and basic graph theory, commonly used for the study of interaction networks.

Readers well acquainted with these topics may choose to skip to Section 1.3 later in this chap-

ter, describing RINs as they are used here or Chapter 2 detailing the implementation of the python pipeline responsible for the creation of homology enriched RINs and supplementary tools for their analysis. Chapter 3 demonstrates HomologyRing's utility, applying it Elongin C (ELOC) protein in the RNA polymerase II complex to create an hRIN. ELOC, as will be covered later, is particularly notable for its important function and interactions with other proteins associated with transcription and tumor suppression. We provide a discussion of many notable aspects of its structures and function in complex and demonstrate how hRINs created by HomologyRing capture these properties or how the supporting analyses in the greater package may aid researchers in quickly understanding critical aspects of a protein family such as this one.

1.1 Structural Biology

Proteins, organic polymers made of amino acids, are critical for all forms of life. They possess immense flexibility, assuming many different roles in all cells. Some protein roles may be primarily physical or mechanical. For instance, some proteins may be responsible for utilizing energy to contract muscle fibers, while others may form rigid aggregates like keratin to help maintain cell structure. Many proteins are enzymes, which, by way of their particular chemical and structural properties, are capable of facilitating one of the thousands of chemical reactions necessary for homeostasis and sustained life. Proteins possess remarkable specificity, binding with specific partners to form complexes or recognizing antigens in the immune system through their precise structure and chemical properties. The immense variety in protein structure and function is owed to their constituent parts and how they are constructed from linked amino acid units.

All proteins are transcribed from genes, which, put simply, are regions of DNA attributed to a specific product or function. The details of the transcription process are essential for this work. However, suffice it to say that protein-coding regions of a gene contain instructions for the creation of the protein. As polymers, proteins are constructed by joining many smaller molecular units to form a large macromolecule. These basic units are called amino acids and represent the fundamental building blocks of all proteins. Joining amino acid units, or residues, together end-to-end forms a protein chain. The various properties of these amino acids and their order in the sequence ultimately determine the properties of the protein. As will be discussed in later subsections, interactions between amino acid residues in a chain drive the process of protein folding, allowing a protein chain to assume its final, biologically functional form.

1.1.1 Amino Acid Properties

Amino acids, the building blocks of proteins, are discrete components with distinct physical and chemical properties. Their ability to be combined in a modular manner allows for the construction of a virtually infinite set of proteins from the same basic elements. There are 20 standard amino acid residues that vary in structure, but all possess a $-\mathrm{COOH}$ carboxyl group and a $-\mathrm{NH_2}$ amino group, to which they owe their name. In all amino acids, the carboxyl group and amino group are covalently bonded to a central carbon atom, referred to as the α-carbon. Together, the amino group, carboxyl group, central α-carbon, and a bound hydrogen atom constitute the 'backbone' of an amino acid, of all standard residues. Figure 1.1 depicts these basic components of amino acid structure.

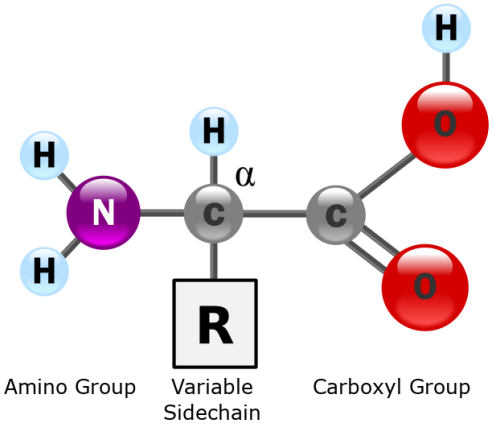

Figure 1.1: Cartoon representation of amino acid structure.

Modified from: GYassineMrabetTalk, Public domain, via Wikimedia Commons

Amino acids primarily vary by their distinct side chains, otherwise known as R-groups, bound to the α-carbon in the backbone. All R-groups, with the exception of glycine, bind to the α-carbon via a first β-carbon in the side chain. The R-groups differ in a few key aspects, which can be used to group the residues into sets with similar characteristics, as shown in Figure 1.2. These primary properties are charge and polarity.

Polarity significantly affects a compound's hydrophobicity, or its repulsion of water molecules. As the majority of proteins will fold and exist in an aqueous environment, the polarity of residues in a protein is a significant contributor to the protein's final structure. Nonpolar side chains, such as those in alanine, leucine, or phenylalanine, are hydrophobic due to their lack of

polar functional groups or charges. These hydrophobic amino acids tend to cluster together in the interior of proteins, away from water, stabilizing the protein's folded structure.

In contrast, polar side chains, like those in serine or glutamine, have functional groups capable of forming hydrogen bonds, making them hydrophilic. These residues are often found on the protein's surface, interacting with the aqueous environment or forming specific interactions within the protein structure such as maintaining a protein's *secondary structure*.

Charged side chains, which can be either positively charged (e.g., lysine, arginine) or negatively charged (e.g., aspartate, glutamate), play a critical role in ionic interactions and maintaining protein solubility. These charged residues often participate in salt bridges or interact with oppositely charged molecules, influencing the protein's stability and function.

The combination of these properties—hydrophobicity, hydrogen bonding capacity, and ionic interactions—determines how the amino acids influence protein folding, structure, and interaction with other molecules. The variability in R-group properties underpins the vast functional diversity of proteins.

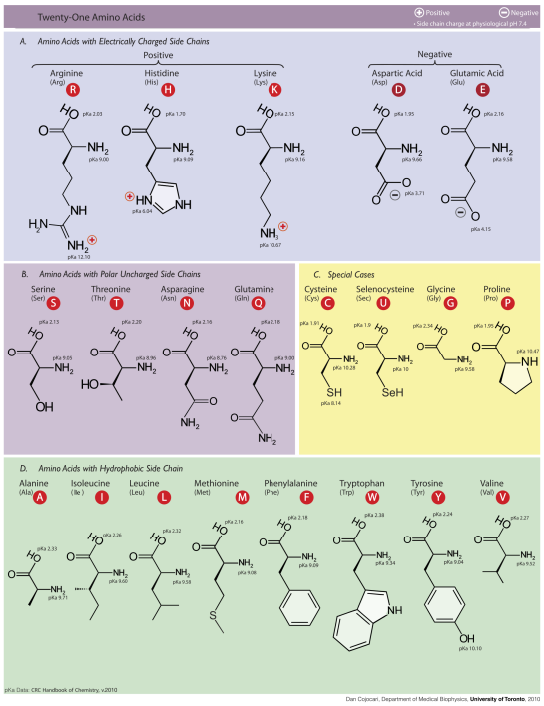

Figure 1.2: Chart displaying the chemical structures of the 20 standard amino acids and selenocysteine, grouped by properties.

Dan Cojocari, CC BY-SA 3.0, via Wikimedia Commons

In addition to the two properties used to form these groups, the amino acid residues vary in several other important aspects. Particularly relevant to this work are amino acids whose chemical structures enable specific interactions, such as: the aromatic carbon rings of phenylalanine, tryptophan, and Tyrosine that allow the formation of π interactions; or the sulfurs contained in the side chains of cysteine and methionine which permit the formation of di-sulfide bridges. The details of such interactions are discussed later in this chapter in Section 1.4.

1.1.2 Protein Sequences

In order to from a polymer, amino acids are bound together into a continuous protein chain by forming peptide bonds—a particular kind of covalent bond where the amino group of one residue will form a peptide bond with the carboxyl group of another. The sequence of the specific residues in a protein chain is referred to as the *primary structure* of a protein, and is the foundation for all higher-order structural features. The primary structure is specified by the sequence of codons in an organism's DNA, with each codon corresponding to a specific amino acid residue.

The order of amino acids in the primary structure dictates the chemical and physical properties of the protein, as well as its ultimate function. This is due to the diverse properties of the amino acid side chains, which govern how the chain will fold into secondary, tertiary, and quaternary structures. Thus, even minor changes in the sequence can lead to significant effects on protein behavior.

Proteomics in part aims to study proteins by representing protein chains as strings, or sequences where each amino acid is represented by a single-letter code (e.g., A for alanine, R for arginine, N for asparagine). This compact representation allows for efficient computational analysis of protein sequences, enabling tasks such as sequence alignment, motif discovery, and evolutionary comparisons. For example, the sequence "ACDEK" represents a protein segment, where A stands for alanine, C for cysteine, D for aspartic acid, E for glutamic acid, and K for lysine. Each letter corresponds to a specific amino acid in the protein's primary structure.

Many useful tasks are carried out on a sequence level, such as: inferring homology or evolutionary relationship between proteins or identifying motifs or sub-sequences associated with notable functions. Importantly, machine learning models the likes of RoseTTAFold, developed by the Baker Lab, or AlphaFold developed by Deepmind are capable of predicting the structure of protein chains from their amino acid sequence with incredible accuracy. This has been of great importance to structural bioinformatics and will be covered in more detail later

in this paper along with protein homology.

1.1.3 Protein Folding

The unique properties of the various amino acids joined together to form a chain alone is insufficient to allow proteins to give proteins their wide variety of functions. The functionality of a protein is intrinsically tied to its structure, or 3D conformation, which most, but not all proteins, assume through the process of folding. Protein folding is a notoriously complex process. However, this section will introduce the theory of protein folding, including: how protein flexibility allows for dynamic conformations that make folding possible; how frustration, entropy, and weak interactions between atoms lend stability to some conformations over others; and the countless variables that drive this process.

Firstly, the backbone of protein chains is far from rigid–the covalent bonds between an α-carbon–nitrogen of the amine section and the α-carbon– carboxyl group carbon are seen to rotate, permitting different dihedral torsion angles and allowing the protein to adapt different conformations. Measuring these dihedral angles, denoted φ and ψ, require observing that certain atoms in a polypeptide will be close to co-planar as a consequence of repulsive atomic forces. Specifically, a residue's α-carbon, amine group nitrogen, and adjacent residues carboxyl carbon define what we will call 'Plane 1'. The same Nitrogen and α-carbon and carboxyl carbon *of the same residue* define 'Plane 2'. The incidence, or angle between Plane 1 and Plane 2 defines the angle φ. Defining Plane 2 instead with a residues α-carbon, carboxyl carbon, and the nitrogen *of the next residue* and taking the angle between it and plane 1 gives the angle ψ. These planes and their corresponding torsion angles are depicted in Figure 1.3.

Visualizing the density of dihedral angles φ and ψ in proteins is a common practice and called a Ramachandran Plot. Figure 1.4 shows such an example of such a diagram showing around 100,000 pairs of dihedral angles observed in a set of protein crystal structures. This is immediately useful for visualizing several important aspects of protein folding. The areas of highest density can be attributed to secondary structure in proteins, or prolific patterns in local protein conformation, which will be covered in the following subsection. Of particular note here are the large regions of the plot where no or few torsion angles are observed. This is evidence of a major factor affecting protein conformation; namely, frustration which goes hand-in-hand with the energy required to adopt a particular local conformation. This drastically reduces the space of feasible protein conformations and explains why some conformations are more probable than others.

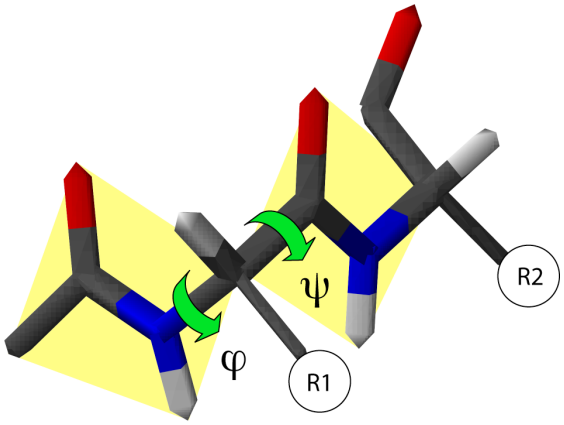

Figure 1.3: Depiction of planes defining torsion angles φ and ψ

Attribution: Frédéric Dardel, CC BY-SA 3.0, via Wikimedia Commons

The second primary factor influencing the conformation of proteins and the process of folding is of central importance to this work. Interchangeably referred to as contacts, interactions, or forces between the different residues in proximity to each other is essential for adapting and maintaining protein conformation. There are several types of interactions, some attractive and others repulsive. The most common interactions between residues are non-covalent interactions such as hydrogen bonding or Van Der Waals forces, and are relatively week when compared to the covalent peptide bonds that link residues together. The specifics of the different interaction types are detailed later in this chapter. However, here it is important that attractive forces between residues contribute to the stability of a conformation which places the residues in close proximity.

Several interaction types are considerably more specific than others. For instance, Van Der Waals forces are present between all atoms in close enough proximity, and the backbones of all amino acids are seen to participate in hydrogen bonding. In contrast, only charged amino acids readily form ionic bonds, only the few aromatic amino acids can act as π donors, and only the sulfur containing cystine and methionine may form sulfur bridges. These less common interactions act with notable specificity, which serves to promote folding consistency and specific folding pathways required to to assume protein conformations capable of their biological function.

The folding of a protein is a notoriously complicated process, dependent on many factors, such as: physiological temperature, pH, and the presence and concentration of other compounds in the environment. Furthermore, the folding process happens remarkably quickly;

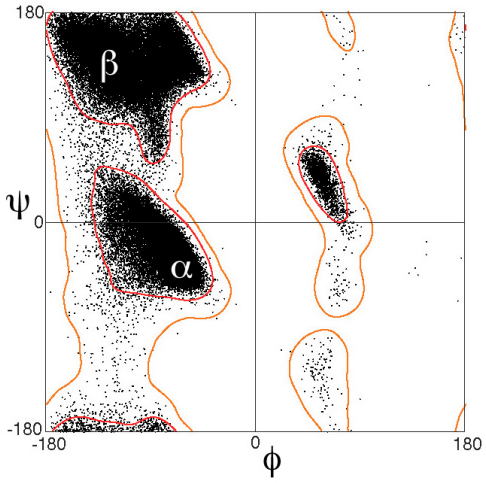

Figure 1.4: Example Ramachandran Plot

Attribution: Dcrjsr, CC BY 3.0, via Wikimedia Commons

With the forces driving it, like heat, changing incredibly quickly, it is often modeled as a stochastic process. Despite its complexity, protein folding must be reliable, as misfolded proteins can cause severe conditions like Parkinson's disease or Human Prion Disease. Both the consequence of protein mal-folding.

The result of the large amount of variables affecting folding, the amount of possible of intermediate conformations with countless possible interactions between residues, and the stochastic nature demanding impractical timescales, is a process that has eluded attempts to deterministically or methodically understand the process of protein folding given its sequence. Contemporary theories of protein folding describe an 'entropy landscape'. Where different conformations adapted by a protein chain are associated with an energy or disorder. The entropy of the protein, when considered across all possible conformations, reveals valleys and peaks that correspond to conformations of varying stability. Energy, often in the form of heat, is required to move the chain from one valley of marginal stability to another.

Given two interacting residues in the same protein chain, the number of residues between them is known as the sequence separation of the interaction; and interactions of high and low sequence separation are seen to have very different, but both critical, roles in contributing to the conformation of a protein. Low separation interactions frequently occur in patterns referred to as *secondary structure*, which serve as intermediate components of protein structure; where interactions with higher degrees of sequence separation contribute to stabilizing the proteins

tertiary structure or overall 3D conformation. Finally, interactions occur between amino acid residues not belonging to the same chain. In this case, it is seen to contribute to *quaternary structure*.

1.1.4 Secondary Structure

Secondary structure of a protein refers to the most common local conformational patterns observed in the polypeptide chains constituting proteins. These patterns promote hydrogen bonding between atoms in the backbone of residues, creating the beginnings of a structural scaffold. The most common of these patterns are α-helices and β-strands. α-helices are right-handed spirals stabilized by hydrogen bonds formed between the oxygen of a carboxyl group and the hydrogen of an amide of another residue – typically 3 to 4 residues away. Though, it is primarily the backbone of amino acids responsible for the hydrogen bonding most important for the formation of α-helices, residues alanine and leucine favor their formation due to their small, non-polar R group. α-helices can be interpreted to have immense importance to the structure and stability of a protein. Simple repeated motifs readily fold into a rigid structure. With the pattern being ubiquitous in proteins, α-helices, along with other patterns in secondary structure, like β-sheets, can be thought of as a sort of modular component in protein construction; as lumber is to houses.

β-strands refer to a series of linear segments of typically 5 to 10 amino acids lying side-by-side. The effect is a sheet like structure of variable width, with two or more β-strands needed to form a β-sheet. β-strands can run parallel or anti-parallel, with the latter being more stable. They frequently are seen forming substrates; regions on the surface of a protein facilitating an interaction: possibly either with another protein or chemical to facilitate a reaction as an enzyme. However, they also play a significant role as a major structural component, lending stability to a protein's *tertiary structure*, or overall conformation.

1.1.5 Tertiary Structure

Tertiary structure refers to the complete three-dimensional conformation of a single polypeptide chain, encompassing how its secondary structure elements—α-helices, β-sheets, and loops—are folded and packed into a functional protein. This level of organization represents the culmination of intramolecular interactions that define a protein's unique shape and function.

The folding of tertiary structure is driven by the interplay of various non-covalent forces. Hydrophobic interactions are a key factor. Nonpolar side chains aggregate in the protein's

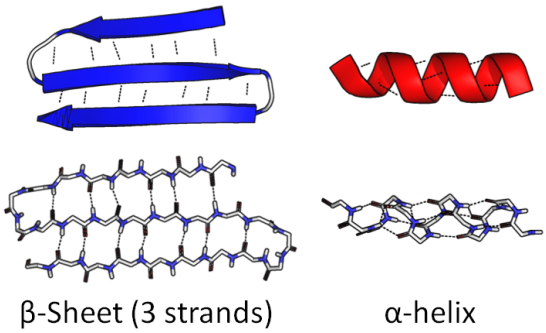

Figure 1.5: Example of the components of Secondary Protein structure: β-sheets and α-helicies.

Thomas Shafee, CC BY-SA 4.0, via Wikimedia Commons

interior to avoid water, forming a stable hydrophobic core.. Conversely, polar residues and charged side chains often reside on the protein's surface, where they can interact with the aqueous environment. Hydrogen bonds, van der Waals forces, and ionic bonds further stabilize the overall conformation. Additionally, covalent disulfide bridges, formed between cysteine residues, provide extra stability, particularly in extracellular proteins.

The tertiary structure is not merely an arrangement of secondary structures; it is the architecture that brings functional groups into precise spatial proximity. This enables active sites in enzymes, binding pockets in receptors, or structural scaffolds in globular proteins. A protein's tertiary structure is dynamic, enabling interactions with other molecules and critical conformational changes.

1.1.6 Quaternary structure & and Protein Complexes

Quaternary structure refers to the arrangement and interaction of multiple polypeptide chains (subunits) in a protein. Each subunit has its own tertiary structure, and their assembly forms a functional multimeric complex. These interactions are stabilized by non-covalent forces such as hydrogen bonds, ionic bonds, hydrophobic interactions, and, in some cases, covalent disulfide bonds.

Human Hemoglobin is an excellent example of quaternary protein structure; being a heterotetramer, biologically active hemoglobin is a complex of four protein chains: two α subunits encoded by either the human HBA1 or HBA2 gene and two β subunits encoded by the HBB gene. F Figure 1.6 shows the four subunits in an annular configuration. Orange and purple chains correspond to α subunits, where green and magenta are β subunits. Hydrogen bond-

ing and Van Der Waals forces are present along the interface of the distinct chains; however, conserved interactions between π systems in a select few aromatic residues in each chain are critical for the overall conformation of the complex. Namely, particular Phenylalanines and Tyrosines in each α subunit, and Histidines and Tryptophans in the β subunits are notable for their contributions to quaternary structure.

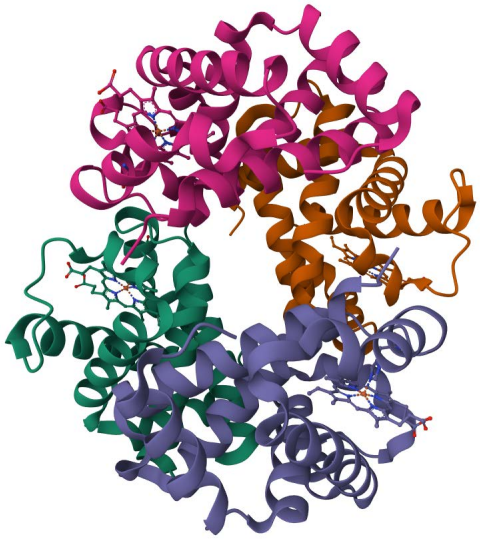

Figure 1.6: Example of Quaternary structure: Human Hemoglobin

From: RCSB PDB: 2DN2, visualized with Mol Viewer

Furthermore, proteins often form complexes with non-protein cofactors or ligands. In fact, hemoglobin owes its oxygen carrying capabilities to a heme groups, which are non-protein compounds held in complex with each hemoglobin chain. Similar to the contacts between protein chains in the hemoglobin complex, specific π interactions involving aromatic rings in only a small subset of residues are critical for the correspondence of the heme groups with their respective hemoglobin chains. The particulars of π interactions are detailed at the end of this chapter along with the other types of interactions formed between residues. These examples are highlighted here to illustrate the importance of some key interactions to protein complexes, something which HomologyRing aids in analyzing.

1.2 Homology & Molecular Evolution

It should be understood how several organisms may be evolutionarily related; from one common ancestor, mutations are accumulated that may affect an individual's fitness, or the ability to thrive for a given environment. These mutations are often neutral, having not have much if any observable effect; however, mutations that alter fitness are most likely to be negative, but some mutations may confer a slight benefit to the individual and become selected. Accumulation of these mutations in a sub-population can lead to notably different traits and eventually speciation. Generally, evolution can be thought of as a process driven by a series of incremental variations and adaptations over generations–admitting an evolutionary or phylogenetic tree tracking this process, which may also be used to assess the relatedness of two individuals.

While the evolutionary relationships between organisms can be traced through accumulated mutations and the eventual divergence of traits, similar principles apply at the molecular level. Just as populations adapt incrementally over generations, so too do proteins, undergoing changes that can affect their structure, function, and interactions. These molecular changes often mirror the broader evolutionary processes, giving rise to families of related proteins with shared ancestry and functional diversity. Molecular evolution describes the process by which proteins undergo small changes that may lead to functional variations or the formation of evolutionarily related protein families. Mutations in the DNA of a gene can alter the protein sequence, while structural mutations, such as insertions, deletions, and inversions, may lead to more significant changes. This process explains the various ways in which different proteins may be related.

Protein homology refers to the similarity between proteins that arises from shared ancestry. Homologous proteins often retain structural or functional characteristics, even as they diverge through evolutionary processes. These similarities can provide valuable insights into protein function and evolutionary relationships, making homology a key concept in understanding molecular evolution.

- Homologs: Proteins that share a common evolutionary origin. Homologs encompass both paralogs and orthologs and indicate a general evolutionary relationship. This broad category underscores the shared ancestry of proteins, which may have diverged significantly over time.

- Paralogs: Proteins that arise from gene duplication events within the same organism. These proteins may retain the original function of their ancestor or evolve new func-

tions, contributing to functional diversity within an organism. Paralogs often appear in protein families with specialized roles.

- Orthologs: Proteins in different species that diverged following a speciation event. These proteins typically retain similar functions across species and are used to infer evolutionary relationships and predict protein function. Orthologs are especially important in comparative genomics and evolutionary biology.

Direct evidence for two proteins being homologs is rarely available, as many intermediate mutations may be absent from a modern population. Instead, homology is generally inferred by analyzing sequence similarity, which can be assessed using computational tools. These tools compare amino acid sequences and identify conserved regions, which suggest shared evolutionary ancestry. In some cases, structural similarity further supports inferred homology, particularly when sequences are highly divergent but still maintain similar three-dimensional conformations. One such popular tool for finding other proteins with inferred homology to a given query is the Basic Local Alignment Search Tool (BLAST). It is made extensive use of in this paper and introduced in more detail in Subsection 1.2.4.

BLAST is just one of the many different methods with which proteins are grouped into families, or collections of proteins with similar functional, structural, or sequence characteristics. Studying these families is a fundamental aspect of bioinformatics and serves to identify conserved features, predict protein functions, and understand evolutionary relationships. In fact, it is the aim of HomologyRing presented later in this work to use homology information of a family of proteins to synthesize information about their structure, allowing for insight into which aspects may be critical for their function.

1.2.1 UniProt Database

The Universal Protein Resource (UniProt) is a comprehensive database providing high-quality protein sequence and functional information. It is an essential resource in bioinformatics, supporting research in areas such as comparative sequence analysis, functional annotation, and evolutionary studies.

UniProt's central component, the UniProt Knowledge-base (UniProtKB), is comprised of two member databases:

- Swiss-Prot: Contains over 560,000 manually curated and reviewed protein sequences, offering highly reliable annotations.

- TrEMBL: Includes approximately 245 million computationally analyzed protein sequences that have not yet been reviewed, providing broader coverage of known proteins.

Each entry in UniProtKB includes rich metadata, such as functional annotations, domain and motif descriptions, protein structure data, and post-transnational modifications. This detailed information helps researchers connect sequence data with functional and biological context, making UniProtKB a key resource for integrative studies.

One of the primary utilities of the UniProt database for the purposes of this work is the unique and consistent accessions assigned by UniProt for the identification of entries. This is indispensable for workflows in bioinformatics such as this one, allowing for integrating data from a variety of different online sources.

1.2.2 Protein Data Bank (PDB)

The Protein Data Bank (PDB) aims to be a complete catalog of published experimental structures of biological macromolecules, including proteins, nucleic acids, and complex assemblies. It provides researchers with access to detailed structural data, enabling the study and analysis of biomolecular conformations and interactions.

The PDB currently contains over 200,000 experimentally determined structures, primarily obtained using techniques such as X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and cryo-electron microscopy (cryo-EM). Each entry in the PDB includes detailed metadata, such as:

- Structural coordinates of macromolecules.
- Experimental methods and resolution of structure determination.
- Annotations on ligands, cofactors, and functional sites.
- Cross-references to other databases for sequence and functional information.

Experimental structures in the PDB provide critical evidence for understanding the conformations of biomolecules and their interactions with ligands and other proteins to form complexes. These data enable researchers to identify binding sites, predict interaction mechanisms, and study the structural basis of complex formation.

Protein structures are determined experimentally using methods such as X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, or cryo-electron microscopy (cryo-EM). X-ray crystallography requires a purified protein sample to be crystallized. Passing a high-power X-ray beam through the crystal results in the creation of a diffraction pattern, which is analyzed using Fourier transform techniques to calculate the electron density map, allowing the determination of the protein's 3-D structure. This results in high quality position information, with enough spatial resolution to accurately determine the position of individual molecules. However, the requirement for a highly purified crystal sample poses a significant limitation. Proteins with dynamic conformations, such as intrinsically disordered proteins (IDPs), or highly hydrophobic proteins, such as transmembrane proteins (TMPs) can be difficult to crystallize. This limitation introduces bias into the structural information obtainable by this method, as it excludes proteins that are difficult to crystallize[2].

NMR exploits the magnetic properties of atoms nuclei to resolve their position. NMR works by applying a magnetic field to a protein sample and detecting the energy released as the atomic nuclei return to their original states, which helps determine the distances between atoms. NMR has the benefit of being able to study proteins in solution, allowing for observation in conditions that may more similarly resemble their normal biological conditions. Furthermore, the technique is particularly useful for observing the conformal dynamics of flexible proteins. Cryo-EM involves rapidly freezing protein samples in a thin layer of ice to preserve their native structure, then using electron microscopy to capture high-resolution images of the sample. Thousands of these 2D images are collected and combined using computational techniques to reconstruct a 3D model of the protein. This method is particularly effective for studying large molecular complexes and proteins that are difficult to crystallize.

1.2.3 AlphaFold Database

The AlphaFold Database represents a revolutionary advancement in computational biology, addressing the longstanding challenge of predicting protein structures from sequences. Powered by the AlphaFold model developed by DeepMind, it provides structural predictions of unprecedented quality, rivaling experimental accuracy in many cases. This transformative resource has greatly expanded the availability of structural information, enabling researchers to access predicted structures for millions of proteins that previously lacked detailed structural data.

AlphaFold focuses on single protein chains and includes structural predictions for nearly all

known protein sequences in key reference databases, such as UniProt. Each entry includes a predicted 3D structure with confidence scores for individual regions, allowing researchers to assess the reliability of the predictions.

Despite its groundbreaking nature, the AlphaFold Database has certain limitations. It is restricted to single protein chains and does not account for complex interactions, such as those between proteins or with ligands. Additionally, the accuracy of predictions varies, particularly for disordered regions or proteins with highly dynamic conformations.

1.2.4 BLAST

The Basic Local Alignment Search Tool (BLAST) is used to quickly search biological sequence databases for entries with high pairwise similarity to a given query. Often, the pairwise similarity serves as evidence to infer some evolutionary relationship, or homology, between the query and results–making BLAST a useful tool for creating a collection or *family* of closely related sequences. Commonly containing DNA or amino acid sequences, such biological databases can be massive, with potentially millions of entries. BLAST can, in part, be seen to tackle a one-to-many alignment problem, and the task of optimally performing alignments with each sequence in a large database quickly becomes infeasible. Much of the utility of BLAST is owed to its speed and heuristic approach to identifying matches, balancing computational efficiency with biologically meaningful results.

1.2.5 Multiple Sequence Alignments

Multiple Sequence Alignment (MSA) is a computational technique used to align three or more biological sequences—most commonly RNA, DNA, or protein sequences, which we focus on here. In doing so, they are essential tools for identifying evolutionary relationships, or conserved motifs within a set of structures. Many software tools are capable of algorithmically creating MSAs; this work makes use of ClustalΩ, which is well regarded for its ability to quickly create high-quality alignments from large sets of sequences. However, they all aim to optimally insert gaps into sequences in the MSA such that residue identity or similarity is maximized in each column.

MSAs are frequently used to evaluate evolutionary relationships between sequences. Greater dissimilarity typically indicates a more distant common ancestor, while conserved regions in an MSA often correspond to functional motifs critical to protein function.

In this work, the utility of MSAs in grouping residues into related sets, as defined by the columns, is central to the synthesis of hRINs. These alignments will serve as a map, linking regions of different protein sequences to functionally corresponding structural elements within a protein family.

1.2.6 Structure Files

Structural data available in databases are standardized to enhance both machine and human readability. This work makes extensive use of the PDBx/mmCIF format (eXtended PDB / Macromolecular Crystallographic Information File) and will refer to the format simply as 'CIF'.

The CIF format employs a dictionary-based schema, offering extensibility for storing diverse data types. These files include a great deal of metadata, providing crucial biological context. Examples include information about the organism and genes from which protein chains were transcribed, mutation details, experimental methods, and authorship.

One of the most important sections of a CIF file details the position and identity of atoms within a structure. This information is stored in the `_atom_site` category, which uses a hierarchical structure for describing biological assemblies. In this structure, individual atoms are annotated with identifying information and additional data fields, including a numerical identifier, the atom's element symbol, 3D coordinates, composition ID, sequence ID, and chain identifier. These identifiers are used to assign atom membership to biologically relevant groups within the CIF data hierarchy, formalizing concepts for describing polymers like DNA and proteins at distinct levels: atom, residue, entity, chain (molecule), and structure (complex).

In addition to macromolecules—polymers like RNA, DNA, or proteins, which are the primary subjects of CIF files—some atoms belong to heterogenous groups referred to as heteroatoms or `HETATM`. These include metal ions, water molecules, and ligands that are frequently important for a macromolecules structure and function.

Use of these stucture files is important throughout this work, and the categories established by this hierarchy are used extensively to describe different interactions observed in structures. such as occurring within a same chain, across different chains, or involving a heteroatom. Additionally, metadata contained within these files is frequently utilized, so we take the time to discuss these files with some attention to technical detail.

Despite the immense utility of CIF files, there are some notable difficulties when working with them. Information about the geometry of the structure is given by the position of discrete atoms. This makes extracting biologically relevant information a somewhat complicated pro-

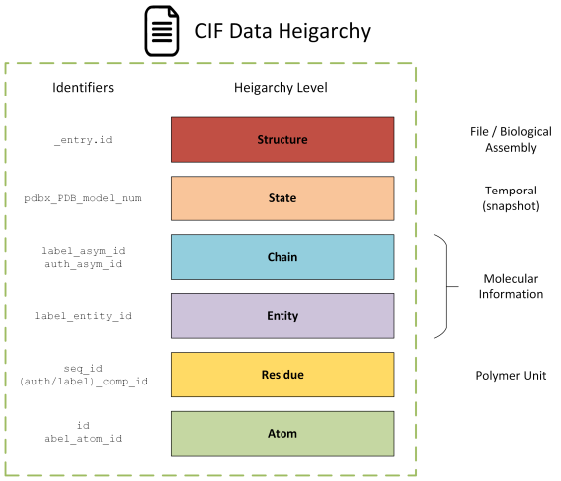

Figure 1.7: Diagram displaying the various levels of the data organization of structural information in the CIF format.

cess. CIF files are also frequently several megabytes in size. This is perfectly fine for just one file, but with the increasing abundance of structural information in databases, and recent advancements in machine learning, large-scale operations involving enormous amounts of CIF data are becoming very common, which unnecessary metadata bloating structure files can bog down. Methods of encoding these structures into a simpler format that retains biologically relevant information are becoming increasingly popular.

1.3 Residue Interaction Networks

To briefly summarize the earlier sections, protein chains are polymers of amino acid residues bound together by covalent peptide bonds. As a chain folds, it is weaker *non-covalent* interactions between residues that are principally responsible for the secondary and tertiary structure, or conformation of the protein. These non-covalent interactions could be a variety of different forces and are the result of the different chemical and physical properties of the 20 different base amino acids and their arrangement in the peptide sequence that constitutes the protein chain. Examples of some of the most common interactions observed between residues in proteins include Van Der Waals forces and hydrogen bonding. Other interactions are observed less frequently and exhibit high specificity with respect to amino acid participants, like Ionic bonds or $\pi - \pi$ interactions and several others. The structure of a protein is intrinsically linked to its function, and understanding these interactions provides critical insights into protein behavior in a biological context.

From knowledge of these interactions, researchers can infer structural and functional significance. However, the rapid growth of available structural data presents a significant challenge. Structural data from experimental sources like the Protein Data Bank (PDB) and high-accuracy predictions from AlphaFold2 are widely available in formats such as PDB and mmCIF. These files contain extensive information, including 3D atomic coordinates, chemical properties, and metadata. Parsing and extracting relevant functional information from these large datasets is a computationally intensive task.

One promising approach to address this challenge is the construction of Residue Interaction Networks (RINs). RINs simplify protein structure by representing residues and their interactions as a mathematical graph or network. In their most basic form, RINs may be denoted as $G = (N, E)$, where N is the set of nodes representing amino acid residues or compounds, and E is the set of non-peptide interactions observed in the structure. For example, an edge $(u, v) \in E$ may represent a hydrogen bond between two residues. Figure 1.8 provides a visual representation of a simple RIN.

RINs can be seen to provide a simplified means of capturing the topology of a protein structure, and has seen notable attention for its various applications in protein engineering, drug discovery[3]. Additionally, the well defined structure of a network readily lends itself to the application of computational or machine learning methods; and notions from graph theory, like *betweenness centrality*, to identify structurally important nodes.

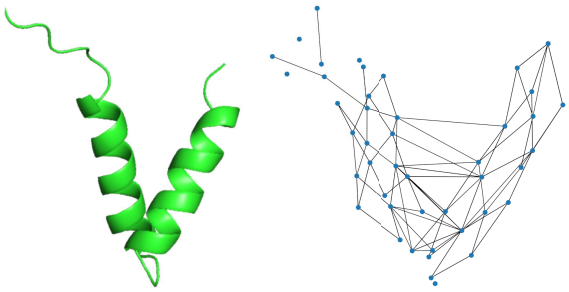

Figure 1.8: Example of a basic RIN for generated for a protein chain.

1.3.1 Residue Interaction Network Generator - RING

There are many tools available for generating RINs, each offering a slightly different interpretation of the concept. In this work, Residue Interaction Network Generator (RING) is used to create these networks. RING is notable for its ability to handle all ligands and modified residues

in the PDB and is highly regarded for its deterministic and accurate prediction of many types of interactions from structure files.

It is important to note that RING produces a multi-graph, which differs from the standard mathematical definition of a graph by allowing multiple edges between a pair of nodes (see Figure 1.9). Each edge in the multi-graph is associated with a specific interaction type, enabling the network to represent multiple interactions between residues. This design captures not only the presence of these interactions but also provides rich chemical information within the network structure.

The interaction types predicted by RING and their corresponding edge definitions are described in detail in Section 1.4 at the end of this chapter.

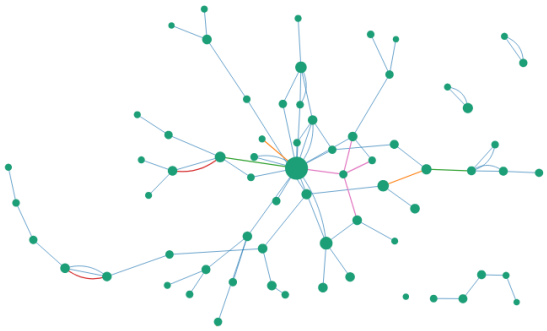

Figure 1.9: Example of a RIN multi-graph visualization created with RING.

As typical with RINs, nodes correspond to amino acid residues in a given protein structure. Edges correspond to non-peptide interactions or bonds between residues. Where edges of many interactions are types are permitted between a pair of nodes.

1.4 Predicted RING Interactions

RING predicts and classifies various types of molecular interactions based on deterministic factors such as geometric orientation, distance, and the chemical nature of the compounds involved. These interactions differ significantly in their chemical and physical properties, including strength and specificity, which in turn influence the likelihood of different amino acids forming specific bonds. Understanding these variations is essential for interpreting the structural consequences and functional roles of these interactions on the atomic and protein level as well as in the context of RINs and hRINs. This section will define and provide context for all interaction types classified by RING, enabling a better understanding of their role in protein structure and function.

For each interaction type, we will provide the names that biologists and other researchers will use to describe the phenomenon in addition to the interactions shorthand identifier which is used internally within RING and HomologyRing. Some of these short hand identifiers such as VDW for 'Van der Waals' are used throughout this text for brevity, and in the more technical discussions of HomolgyRing's implementation.

1.4.1 Van Der Waals Forces - VDW

Van der Waals forces are relatively weak non-covalent interactions between molecules or atoms, often considered among the weakest inter-atomic forces. Despite their individual weakness, their abundance and cumulative effects play a key role in stabilizing protein conformation. VDW forces arise from temporary shifts in electron density induced by the proximity of other molecules or atoms. These shifts occur because of fluctuations in the electron cloud, leading to temporary dipoles, or simply gradients in charge.

Van der Waals forces can be either attractive or repulsive. At very small distances, the repulsive forces dominate due to the proximity of electron clouds, while at larger distances, the attractive forces generated by induced dipoles become prominent. This balance creates a stable equilibrium distance, which plays a key role in maintaining weak but stable interactions. Given that VDW interactions can exhibit these different behaviors, they are generally grouped into different categories, distinguishing interactions based primarily on if dipoles involved are permanent, or temporarily induced by proximity to other atoms. However, it is not critically relevant to this work so the distinction will not be made.

Van der Waals (VDW) forces are ubiquitous in proteins, arising between atoms in close proximity. While individually weak, their cumulative effect can significantly stabilize protein tertiary and quaternary structures by contributing to the packing of residues in the hydrophobic core. However, their non-specific nature and dependence on atom proximity make them challenging to interpret in the context of functional or evolutionary studies.

In RINs, VDW forces can serve as indicators of residue proximity. However, their widespread presence often generates relatively dense networks, making it difficult to extract meaningful insights. Residues frequently interact with multiple neighbors through VDW forces, leading to high connectivity in RINs, which can obscure biologically relevant interactions.

Furthermore, VDW interactions are generally less conserved across protein families compared to covalent or hydrogen-bond interactions. This lack of conservation, coupled with their abundance, complicates the identification of functionally significant interactions. As a result,

analyses often filter out VDW interactions to focus on more specific and conserved forces, such as hydrogen bonds or salt bridges, that are more directly linked to protein function and stability.

1.4.2 Hydrogen Bonds - HBOND

Hydrogen bonds are second most prolific non-covalent interactions observed in protein structures, behind VDW forces. They are critical for the formation of α-helices and β-sheets, or the secondary structure of a protein [4]. Hydrogen bonds, when they occur in secondary structure, take pace between the backbones of interacting residues, typically carboxyl oxygen and amide hydrogen. These common patterns observed in hydrogen bonding in secondary structure is one of the basic components used to construct more complex structures.

Hydrogen bonds are additionally highly significant for protein conformation, beyond their contribution to secondary structure. They are frequently observed forming longer distance interactions, stabilizing the overall conformation of a chains tertiary and a complexes quaternary structure. In these cases it is common to see hydrogen bonds form between residues side chains, rather than their backbone as seen in secondary structure formation. As a result, hydrogen bonding is most prevalent in amino acids with polar side chains, where electronegative atoms like oxygen, nitrogen, and sulfur. Examples of such residues include: serine, threonine, asparagine, and glutamine.

In the context of RINs, information of hydrogen bonding in protein families provides is quite useful, despite their prolific nature. Within a protein family, it is commonly observed that the preservation of local conformation of functional sites will be critical, even in a family of distantly related proteins–which hydrogen bonding may often play a role in maintaining.

1.4.3 Ionic Bonds - IONIC

Ionic bonds occur at the atomic level when an electron is transferred from one atom to another. The resulting atoms, having transferred an electron, become ionized hand have and each have a net charge. The donor will exhibit a positive charge, while the recipient becomes negatively charged. This difference in charge, creates a strong electrostatic attraction, bonding the two atoms. In proteins, ionic bonds are often referred to as *salt bridges*, and their formation generally require the presence of amino acids with charged side chains. These include: lysine, arginine, and histidine with positive charge; and asperic and glutamic acid have negatively charged side chains. This creates a relationship between conservation of these residues in at a sequence-

level and the conservation of ionic bonds they may participate in in a structure. However, the ability of the charged amino acids to participate in ionic bonds is pH dependent. This allows some proteins to adapt variable conformations in response to cellular environment; this is one possible functional interpretation of an ionic bond with in a protein. Both factors affecting the conservation of ionic bonds in protein families.

1.4.4 $\pi - \pi$ Stacking - PIPISTACK

π-π stacks are specific interactions between pairs of aromatic rings; or stable, planar, ring-shaped molecular structures with alternating single and double bonds present in some amino acid side chains. In proteins, such amino acids with required for the interaction are: phenylalanine, tyrosine, histidine, and tryptophan. These aromatic rings have decentralized π-electrons, to which the interaction owes its name. The mobility of π electrons within aromatic rings give the structure its resonate properties—having electrons with distributions not centralized around a single atom, as typically the case, but rather the electrons may exist in a shared annular cloud around the aromatic ring. These peculiar electron distributions, or π systems, permit specialized interactions: either with other aromatic rings to form a $\pi - \pi$ interactions, or other compounds, as seen later.

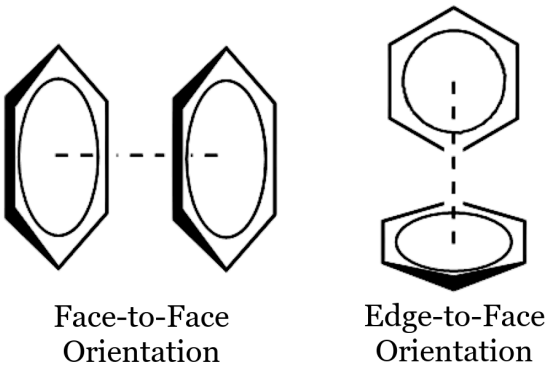

Figure 1.10: Depiction of the different π-π stack orientations.

Attribution: Emily ricq, CC BY-SA 3.0, via Wikimedia Commons

$\pi - \pi$ interactions may may be generally categorized into two configurations of the participating aromatic rings, depicted in Figure 1.10:

- **Face-to-face**: where the rings are stacked directly on top of each other with a slight offset.

- **Edge-to-face**: where the rings are oriented perpendicularly, with the edge of one ring pointing toward the center of the other.

In both cases, the interaction of the π-systems creates an attractive force that stabilizes the residues relative positions.

These interactions tend to be quite specific, as they require the presence of aromatic residues, and relatively well preserved. This specificity gives them a notable role in stabilizing tertiary and quaternary structure. The Hemoglobin complex discussed earlier in Figure 1.6 serves as a good example of $\pi - \pi$ interactions specificity aiding in maintaining a complex.

1.4.5 π-cation Bond - PICATION

Similar to π-π interactions, π-cation interactions require the presence of a π system, like that seen in an aromatic ring. However, instead of sharing decentralized electrons with another π system, they can be donated to a near by positively charged cation. In the context of proteins, amino acids with positively charged side chains such as: lysine, arginine, or histidine often fill this role of π-acceptor. The specific properties required of two residues to form π-cation bond, namely that one is aromatic and the other positively charged, contributes intra and inter-chain interactions that are generally well conserved within a family. However, charged positively charged ligands are frequently seen to act as π-acceptors, having notable roles in enzyme active sites—where aromatic residues can be used to locate cations and stabilize interactions.

1.4.6 π-hydrogen Bond - PIHBOND

Similar in action to both π-cation bonds and typical hydrogen bonding, π-hydrogen bonds are stabilizing interactions where a positively charged hydrogen bond donor, such as a polar $-$OH or $-$NH group, interacts with the electron rich π system of an aromatic ring. π hydrogen bonds are weaker and far less common than both π-cation and π-π stack interactions. As a result, they tend be less frequently important for the structure and function of a protein, and they are observed to often not constitute a significant part of residue interaction networks. However, there are notable examples where they perform a significant role, such as in DNA binding proteins such as transcription factors and aromatic rich enzyme active sites.

1.4.7 Halogen Bonds - HALOGEN

Halogen bonding is a form of non-covalent interaction involving a halogen atom (such as fluorine, chlorine, bromine, or iodine) and an electron-rich atom or group (such as oxygen, nitrogen, or sulfur) acting as an electron donor. Although halogens are typically considered electronegative, when they participate in covalent bonds, they exhibit a region of positive electrostatic potential called the σ-hole on the side opposite to the covalent bond. This σ-hole allows the halogen to act as an electron acceptor, facilitating the formation of a halogen bond with an electron donor.

Though halogens required for the formation of halogen bonds are not present in any of the standard amino acids, some residues may be post-transcriptionally modified to become halogenated; however, this is seldom observed[5]. Instead, it is more common to observe halogen binding with specific ligands, making their presence highly significant in some contexts.

1.4.8 Disulfide Bonds - SSBOND

Disulfide bonds, also called disulfide bridges, are notably different from previous interactions in this list, as they are covalent bonds; and the amino acid cystine is also unique for its sulfur containing $-SH$ thiol group in its side chain, needed for the formation of disulfide bonds between residues.

The formation of the sulfur bridge requires a post-translational modification of cystine residues. This involves an oxidation reaction where the hydrogen is displaced, and results in a $S-S$ interaction that is considerably stronger and more sable than many of the other inter-residue interactions within a protein. That said, they are reversible under reducing conditions, allowing for dynamic protein conformations that are responsive to certain cellular conditions. The strength of the covalent bond is durable in harsh extracellular conditions, and is often seen stabilizing structure in such environments[6].

These bonds are generally seen to have notable separation in a protein chain or present in typically well preserved inter-chain interactions. This gives cystine a clear role in contributing to tertiary and quaternary structure of proteins. Due to the unique nature of cystine and disulfide bridges, cystines are frequently well preserved at the sequence level within a protein family and so too are associated disulfide interactions.

1.4.9 Metal Coordination - METAL_ION

Like disulfide bridges, metal coordination interactions are notable for being a coordinate covalent bonds, where an atom will donate a pair of electrons to a metallic cation. The most common donors are amino acids containing nitrogen oxygen or sulfur groups. Since the donor contributes a pair of electrons, the most common acceptors for this type of interaction are Zn^{2+}, Fe^{2+}, and Cu^{2+}.

Coordinated metals in proteins are of great functional importance, and pay a role in may critical processes such as: electron transfer, facilitation of redox reactions, and maintaining the structure of zinc finger proteins responsible for facilitating RNA and DNA binding[7]. Given this importance, metal coordination interactions most often represent critical components of a proteins structure or function and consequently are frequently well preserved in families. However, experimental and procedural variation in structure creation in the PDB can make there presence inconsistent. Notably, AlphaFold2 does not predict the presence of metals, or other ligands, so will not be represented in predicted structures. Due to both of these factors, we may anticipate that that conservation of metal coordination interactions in families of RINs will be effected.

2
Methods

This chapter outlines the methodologies and technical implementations behind the RingHomology tool and its application. After providing an informal introduction to the hRIN object—the primary output of the tool—we present an overview of the pipeline aspect of HomologyRing responsible for their generation in Section 2.1. Section 2.3 introduces the auxiliary tools included in the HomologyRing package that support subsequent analysis of a constructed hRIN. Finally, Section 2.2 provides a formal definition of the hRIN object and a rigorous description of their construction, offering a theoretical foundation for future applications.

The HomologyRing package primarily constructs Homology-enriched Residue Interaction Networks (hRINs), which represent a family of proteins. Specifically, hRINs are weighted, undirected multigraphs. Seen as an extension of traditional RINs, such as those generated by RING, hRINs similarly capture information about non-peptide interactions as edges of a graph or network structure. However, instead of nodes representing individual amino acid residues, as in typical RINs, they represent a family of corresponding residues across multiple protein structures. Variation, or lack thereof, in the residues present or the interactions in which they participate, may hold significant structural or biological relevance. The aim of constructing hRINs is to capture this information.

One notable aspect of hRINs not seen in many implementations of RINs is assigning weights to edges in the network in order to capture the additional homology information. Weighting edges by the probability of the corresponding interaction's presence in any protein structure in a family encodes valuable information into the network, providing evidence to the impor-

tance of particular interactions. Critical interactions essential for the structure or function of proteins in a family, or flexible regions with conformal variability, can be quickly identified through interaction probabilities. Furthermore, integrating RING's ability to predict interactions with ligands or between residues in different protein chains with edge weights capturing homology information provides insights into the roles of particular interactions in complex formation.

To provide a practical understanding of the hRIN object and its contained information, we first describe the pipeline for their construction before discussing the analyses they enable.

2.1 Pipeline

HomologyRing is implemented as a Python package responsible for running the pipeline and providing tools for supporting analyses. The construction of an hRIN represents the primary functionality of the package and is handled by a pipeline that organizes the various software applications, internet databases, and Python tools required for hRIN creation. The pipeline exists in two variations: one performs a BLAST homology search to identify a family of proteins, and the other allows the user to provide a set of protein structure files and specify which chains are considered members of the protein family for analysis. At least one implementation of the pipeline must be executed after object initialization to perform any subsequent analysis on an hRIN.

First, we will describe the more common of the two pipelines, which performs the BLAST search to collect a set of homologous protein chains that define a family. Subsection 2.1.8 describes how the user may forego the homology search by providing a set of protein chains to define the family. An overview of this primary pipelines operation are depicted in Figure 2.1 and can be summarized follows:

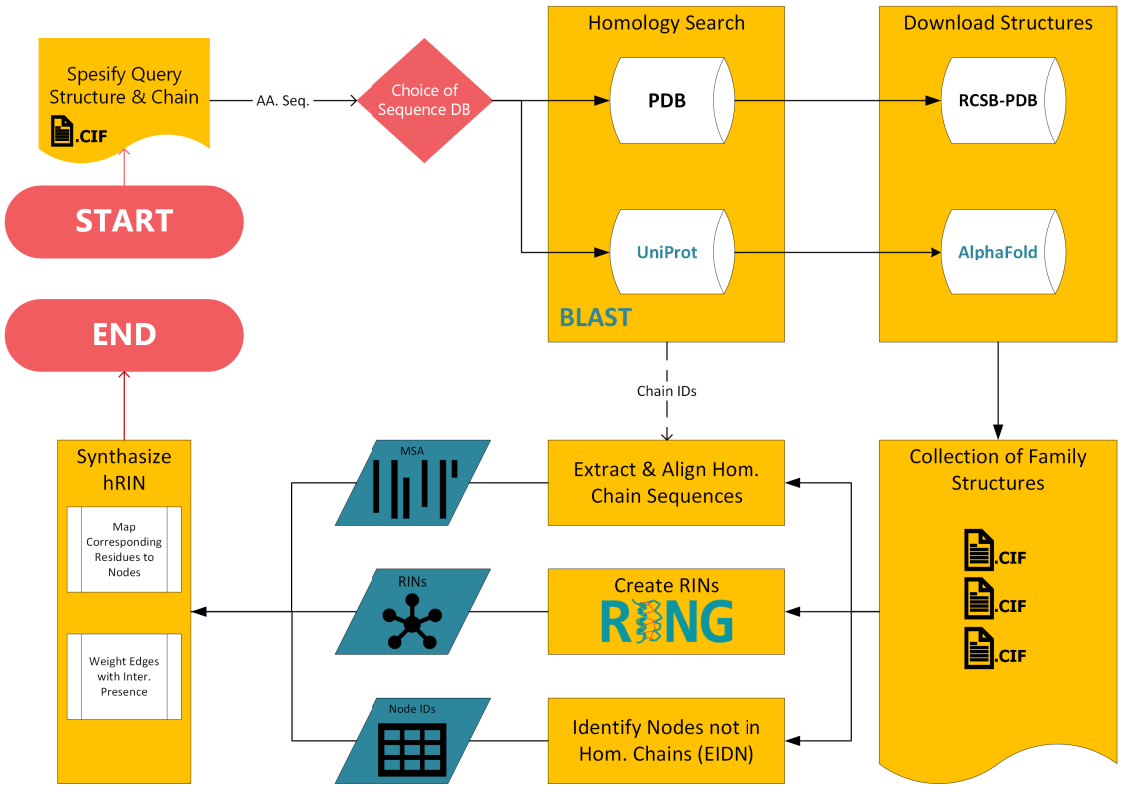

Figure 2.1: Overview of components and flow of primary HomologyRing pipeline operation.

1. **Initialization:** The pipeline is initiated for a query protein chain identified in a CIF file.
2. **Homology Search:** Canonical Amino Acid sequence is extracted from the query chain and used as a query for a BLAST homology search of either the AlphaFold or PDB database. Homologous protein chains identified in BLAST results define the protein 'family'.
3. **Download Structures:** Programmatically retrieve structures for family members in the form of CIF files from either AlphaFold or PDB databases.
4. **Multiple Sequence Alignment:** Extract AA sequence for chains in family and multiply align to create MSA defining the hRIN's 'residue nodes'.
5. **Non-Homolog Chain Processing:** Identify chains in structures indicated as not family members to define 'non-residue' nodes which participate in 'inter-chain' interactions.
6. **RING:** Predict non-covalent interactions for each structure with RING.
7. **Synthesize RINs** generated for each structure by mapping nodes according to information collected in steps 4 and 5 to create final hRIN.

In the subsequent subsections, we explain each pipeline component and its role in hRIN construction in greater detail.

2.1.1 Initiation

Initiating hRIN construction requires the user to identify a protein chain of interest to serve as a *query chain*. The primary structure, or amino acid sequence, is extracted from this query chain and used as the basis for the BLAST homology search. To begin, the user supplies a Crystallographic Information File (CIF) containing physical information about the query chain, and possible other chains present. From this file, the query polypeptide sequence is extracted in the form of the query chain's *canonical sequence*. Use of this field specifically serves as a data cleaning step, providing a simplified representation of the chain's primary structure by omitting non-standard or modified residues that are irrelevant to the homology search.

Upon initiation of the pipeline, the user will also supply additional parameters other than specifying a query chain. These parameters affect the behavior of subsequent steps in the pipeline, such as: maximum family size, which database should structures be retrieved from, among others. The parameters available are mentioned in the section they are most relevant.

2.1.2 Homology Search

With the query sequence extracted, it is used to perform a homology search using protein-BLAST (BLASTp). BLAST, introduced more thoroughly in Section 1.2.4 of the Introduction, is a powerful tool for quickly retrieving a set of proteins from a database that are inferred to have an evolutionary relationship to a given query protein.

Practically, BLAST can refer to an algorithm, an application, or a service that performs homology searches. In this case, HomologyRing uses either a local BLAST command-line application or remote BLAST hosted by NCBI servers[8] to perform the search. The user must specify one of two databases for the search: SwissProt or the PDB. The size, composition, and available metadata of these databases differ significantly, and the choice between them should be guided by the aims of the user's analysis. The practical differences between these databases are discussed in Subsection 2.1.5.

The proteins returned by the BLAST search are treated as members of a protein family for the purpose of constructing an hRIN. To ensure flexibility, many BLAST parameters can be adjusted when initiating the HomologyRing pipeline. Key parameters include:

1. E-value Threshold: This statistical measure indicates the quality of a match. Adjusting the threshold allows BLAST to be more permissive or restrictive when determining which entries are included in the results.

2. Maximum Results: This parameter limits the number of returned results to the top matches, helping to create a more unified and biologically relevant protein family.

By fine-tuning these parameters, users can optimize the construction of the hRIN to suit their specific research needs.

The BLAST output is provided in form of a JSON file, and parsed by HomologyRing to create and maintain a tabular record of sequences in the family. Importantly, this record will include identifying information such as database accessions, used in the following step to fetch results corresponding structural information from online databases.

HomologyRing comes naively packaged with the ability to search local SwissProt or PDB sequence databases. These smaller databases with 572,619 and 12,642 entries as of 2024 respectively[9, 10] are provided, as the much larger UniProt database requires too much memory for use on consumer machines, but remains available by performing the BLAST search remotely on NCBI servers. Choice of the database used for the Homology search will subsequently dictate which database the Structures will be downloaded from in the next subsection.

2.1.3 Download Structures

In the previous step, a family of proteins were created based solely on the sequence properties of the protein chains. The intrinsic connection between a proteins sequence and structure, would suggest that the assumption that the identified proteins will at least have comparable structure is well-founded. However, the pipeline, at this point, does not have access to any of the structural information about these protein chains required to construct RINs and eventually an hRIN.

To address this issue, HomologyRing needs to obtain structures for each protein identified in the previous step. These structures may be obtained from one of two different databases the PDB, or AlphaFold; and the choice of database used for the homology will inform this choice. Naturally, sequences listed in the PDB sequence database will correspond to experimental structures available form the PDB. However, experimental structures exist for only a subset of proteins with sequences in SwissProt or UniProt. For this reason, the structural predictions provided by the AlphaFold Database must be used in the case the homology search was performed over UniProtKB entries.

Once structure files for all family members are collected, some basic filtering is performed to ensure CIF structures are not malformed, as is sometimes the case in the PDB, preventing the running of RING due to missing secondary structure information. The peptide sequences of each homolog chain are extracted from the structure. This ensures parody with information contained in the hRIN and the component structures used for its construction, thus avoiding issues caused by out of data BLAST databases and updated PDB/UniProt entries can causing discrepancies.

In later sections, we regularly refer to a family of structures which to take to mean the collection of structures returned by this process that were not filtered.

A notable difference between stuctures obtained from the PDB compared to ones from the Alphafold Database is currently, structures downloaded from the AlphaFold Database only contain predictions for single protein chains, providing no information about ligands or protein complexes. This limitation is discussed further in Subsection 2.1.5. However, it should be noted that if the user intends to analyze portein complexes, PDB databases should be used, which permit structures consisting of many entities.

This may introduce some confusion to readers, as experimental structures from the PDB may contain many protein chains, but was only included in the family based on a single chains similarity to the query. We will often refer to such as 'homolog chains'. Additionally note that structures from the PDB may contain multiple homolog chains, like in the case of homomultimers, winch are complexes of multiple copies of the same protein. Lastly, will make particular note of entities, both protein chains and non-polymer compounds, which were not homolog chains, and will be straightforwardly referred to as 'non-homolog chains'.

2.1.4 Multiple Sequence Alignment

The aim of this pipeline is to create a homology-aware RIN. A critical source of homology information comes from the BLAST search, which assembles a family of related sequences. However, BLAST provides only pairwise alignments, where the query sequence is aligned individually to each result. This describes how the query relates to each sequence but does not produce a unified model of the entire family. To address this, the sequences are multiply aligned to create an MSA, as described in Introduction Subsection 1.2.5.

To ensure data consistency, the sequences from the BLAST results are not directly used for the MSA. Instead, sequences are extracted from each structure in the family. This approach avoids potential issues caused by discrepancies between the sequence databases used for the

BLAST search and the canonical sequences found in the structural data.

Once the canonical sequences for each homologous chain in the family are gathered, an MSA is created using ClustalΩ. The columns of the MSA identify groups of corresponding residues across the protein chains within the family. These groups of residues will eventually be represented as nodes in a network structure. The MSA serves as a reference, guiding the integration of information from various sources, including BLAST, structure files, and RINs generated by RING, and mapping them onto their respective features in the final hRIN.

2.1.5 Non-Homolog Chain Processing

As mentioned in Section 1.2.6 of the Introduction, the CIF schema allows files to include structures for multiple entities. This is particularly useful for studying biological complexes containing different protein chains, non-protein compounds, or ligands such as DNA or cofactors.

A key motivation behind the development of HomologyRing is to analyze how closely related proteins form complexes. In the context of RINs, this involves examining how inter-chain interactions are conserved or vary across a family of proteins. This capability enables studies on how residue modifications, mutations, or cofactors, such as phosphorylated adenosine molecules, influence protein interactions and conformations. However, the study of inter-chain interaction preservation across many homologous structures comes with some practical and technical challenges that must be addressed.

The first consideration is whether inter-chain contacts are relevant to the user's analysis. This determines which database should be used for the BLAST homology search described in Subsection 2.1.2. For example, the AlphaFold Database provides predictions only for single protein chains, making it unsuitable for studying inter-chain interactions. AlphaFold has the benefit of having structures available for nearly all entries in UniProt, allowing for families to be created from the over 214 million predictions available in AlphaFold [11]. This makes AlphaFold useful for analysis with the aim of observing how variation among closely related sequences effect residue interactions in a single chain.

The ability of the PDB to represent structures consisting of many protein chains and ligands is increadibly useful for the study of the interactions between protein chains, as many structures published in the PDB include proteins in their biologically active complex, or in a bound state with another protein. However, studying inter-chain interactions in experimental structures at scale comes with some practical problems. The variable presence of compounds in different structures makes it difficult to reliably create a map that can be used to robustly

and consistently identify corresponding entities across all structures. For example, consider a hemoglobin complex, as shown in Figure 1.6 in the Introduction. As discussed there, it is a hetero-multimer consisting of two HBA and two HBB subunits with four instances of heme ligands. Studying the conservation of a particular interaction between a residue of only one of the four protein chains and a consistent heme requires a means of consistently relating individual entities corresponding instances between structures in a family. Neither the CIF data schema or databases using the format provide a reliable way of doing this. A model for predicting corresponding instances of entities would be a needed, and has been considered for further work.

This problem is further complicated when considering inter-chain contacts between polymer entities like protein or nucleotide chains. Assuming, corresponding instances of polymer entities can successfully related across structures, there remains the issue of edges, or interactions, being defined as being between two residues, in this case, residues in different chains. This requires a means of consistently identifying residues in non-family protein chains. This is a non-issue if all corresponding polymer entities have 100% sequence identity; However, slight variations in corresponding entities primary structure would require a multiple alignment. This is achievable using the same methods discussed in the previous Subsection 2.1.4, but would require the aforementioned means of consistently identifying corresponding polymer structures.

To address this issue and allow HomologyRing to enable the study the conservation of inter-chain interactions, a system is implemented to simplify these two problems. To address the issue of consistently identifying instances of the same entity across structures, all individual entities–except for homologous protein chains within the family–are aggregated into a single entity.. For example, all instances of heme ligands in a Hemoglobin structure would be identified as a single entity, which in the hRIN is assigned its own node. In order to perform this aggregation, information contained in member CIFs about an entities identity are referenced. For non-polymer entities, they are identified by a `db_code` field which will provide the compounds name. All entities in structures that share a `db_code` will then be mapped to the same node in the process of synthesizing the final hRIN.

To address the second issue of needing to relate residues in non-homologous protein chains in structures, a similar logic is followed. CIFs obtained from the PDB will most often contain supplementary metadata mapping entities in a structure to the gene a protein was transcribed from. Like for non-polymer entities, this feild is use to aggregate non-homolog protein identities, collapsing proteins transcribed from the same gene to be identified by a single entity

identifier. Where This is then taken one step further to address potential variation in residues between the instances of proteins transcribed from the same chain. All residues belonging to non-family member protein chains are mapped to one node per gene.

This approach represents a compromise, allowing for a somewhat one-sided analysis of inter-chain interactions. An inter-chain edge in an hRIN is defined as being between two nodes, say (u, v). HomologyRing only considers interactions where atleast one node defining the identifies a residue. Without loss of generality, we may say that u is a residue node. Internally, these identifiers are numbers, in this case of u is the index of the column in the MSA used to define the node as described in Subsection 2.1.4. For inter-chain contacts, non-residue nodes (v) are assigned numeric identifiers. These identifiers are based on the chemical compound (if it is a ligand) or the gene from which the protein was transcribed (if it is a polymer).

2.1.6 RING

This subsection introduces the role of RING and the importance of RINs in the pipeline, followed by technical details on RING's integration. At this point in the process, the pipeline has gathered a great deal of data taking various forms. Information about the various proteins in a family exists at the sequence and structure-level. Additionally, homology information from the initial BLAST search and MSA provides the basis of comparing entries, or family members on only a sequence level. Structural information, such as that in a CIF file, poses significant computational challenges. These include handling 3D coordinate data for atomic positions, interpreting biologically relevant chemical information, and managing large file sizes that can create time and memory constraints at scale. One such problem is how can structural data be enriched with homology information, which here has been collected at the sequence-level. To address this issue, and some others associated with structural homology, we employ the use of RINs.

RINs simplify the representation of protein structures while capturing key characteristics. Firstly, RINs serve well to capture the topology of a protein chain into a network structure, where residues in close proximity are more likely to participate in an interaction, represented in the network as an edge. Similarly, a smaller degree of separation between nodes in a RIN corresponds to closer spatial proximity in the protein structure. Encoding biological structures as an RINs enables trading spacial-based analysis of structures, which may often be intensive or imprecise with network-based analyses, for instance: the application computational methods like Graph Neural Networks (GNN).

Secondly, a network-based structure easily accommodates biologically significant chemical information like the amino acid identity of a residue associated with a given node, or the type of interaction identified by an edge like those described in Subsection 1.4 of the introduction. Additional benefits of using RINs include their smaller file sizes compared to CIF or PDB formats and their ease of interpretation when well-annotated..

Lastly, a critical benefit of representing structures as a RIN for or use case is it easily enables incorporation of the sequence-level homology information. The details of this process are detailed in Subsection 2.1.7.

To conclude this subsection, we describe the technical details of RING's integration into the pipeline. RINs are generated by running command-line RING on each structure file in the family. These processes are executed in parallel, significantly reducing pipeline runtime. After sucessful execution of all process, we consider the resulting networks to be a family of RINs. However, each is likely to use different identifiers for entities and residues present in their corresponding structures, and will need to be synthesized, as described in the next subsection, into a single hRIN.

It is worth noting that RINs generated by RING are represented as edge lists, defined by two components: a node list and an edge list. The node list which enumerates nodes in the network, providing their identifiers and additional that may be associated with any given node like: its numerical index in a protein sequence, which entity or chain in the structure it belongs to, or what particular type of amino acid residue is identified by the node (e.g. Arg, Lys...). Secondly, the edge list will consist of columns that identify the two nodes by their respective IDs that the edge connects. Similarly, the edge list contains columns providing additional information about an interaction like the type of interaction, or indicate if it is an intra-chain or inter-chain interaction. Each of these files is parsed and combined into master edge and master node tables, which will be processed to synthesize the final hRIN.

2.1.7 hRIN Synthesis

The final step in the HomologyRing pipeline is synthesizing the hRIN from previously collected data. This involves defining the nodes and edges that will comprise the network and enriching them with useful information about the protein family.

Firstly, the process of defining the set of nodes that will be in the final hRIN must be completed before the edge data that will connect them can be processed. Nodes in the hRIN are categorized into three types, which are defined through different processes. These categories

are as follows and will be explained in detail:

1. **Residue Nodes** represent families of amino acid residues, who each belonging to a *homolog chain*.

2. **Chain Nodes** identify entire *non-homolog chains* present in the family of structures. They additionally may represent multiple instances of identical protein chains in one structure as a single node.

3. **Ligand Nodes** identify unique nucleotides and non-polymer entities as single nodes.

The first and most important category is residue nodes, which represent the homolog chains described in Subsection 2.1.3. That is to say they represent the residue information of the protein chains indicated as homologous by the BLAST search. Unlike typical RINs, residue nodes in hRINs do not represent individual residues. Instead, they represent families of residues defined by the columns of the MSA created in Subsection 2.1.4. It is this crucial distinction that enables hRINs to incorporate homology information and come to describe a family of protein structures. The index of each column of the MSA is used to assign a numerical identifier to each *residue node* in the hRIN. Relating nodes in the original RINs to nodes in the final hRIN can be conceptualized as a mathematical mapping between their respective node sets. It is described more rigorously in the following section. However, it can be easily understood as the map that takes nodes, corresponding to residues in a sequence, to their respective columns defined by the MSA.

The remaining two node categories, chain nodes and ligand nodes, represent cases where a node in an RIN is not associated with a homolog chain. These nodes are handled using the logic described in Subsection 2.1.5. In short, the system assigns each entity which is not a *homolog chain* across all structures a numeric identifier that is unique to either the gene a protein was transcribed from in the case of *chain nodes* or the identity of a chemical compound in the case of *ligand nodes*. It is important to note, that though DNA or RNA strands present in structure files are technically polymers, they are classified by RING as ligands, and infrequently are annotated with information making it possible to group instances of nucleotides by their sequence or gene. s a result, these entities are treated as ligands within the RingHomology pipeline. The numeric identified assigned to either *chain nodes* or *ligand nodes* by this system will be equal or greater to the number of columns in the MSA. This creates a node identification system where IDs in the range $[0, m)$ correspond to residue nodes, and IDs $\geq m$ correspond to chain or ligand nodes, with m being the number of columns in the MSA.

Once the node set for an hRIN has been defined, constructing the edge set becomes more straightforward. This process relies on the previously established mappings between nodes in individual RINs and their corresponding nodes in the hRIN. In any graph, edges are defined by the nodes they connect. For instance, an edge between two nodes u and v in a simple graph can be represented as an ordered pair: (u, v). RINs, however, are more complex, as their edges include additional attributes, such as the type of interaction they represent. Mapping edges from RINs to the hRIN is accomplished by applying the node mappings to each node that defines an edge.

This process is described more formally in Section 2.2, but the general concept is straightforward: an edge in an hRIN corresponds to multiple edges from the individual RINs. For example, interactions between corresponding residues in different family structures are aggregated into a single edge in the hRIN. The information from these interactions is summarized as weights associated with each edge. These weights capture homology information by reflecting the number of structures in the family that include the interaction. The exact method for calculating weights will be covered in detail later, but they generally quantify the conservation of interactions across the family.

2.1.8 Pipeline Variant: User-Provided Families

This section describes a feature of the HomologyRing tool that allows users to create an hRIN for their own pre-defined families consisting of a set of homolog chains in local structure files, forgoing the need of performing a homology search of fetching structures from an external database. This is done with a variation of the main HomologyRing pipeline, where upon initialization, the user will specify the typical query structure and chain, but also provide a file giving a description of the family. The details of initiating the pipeline in this way are provided in Section 5.2 of the appendix, but this file must provide the following information:

1. An identifier unique each structure in the family. Typically, PDB IDs or UniProt accessions.
2. A path to a local structure file.
3. The chain identifier (`auth_asym_id` or `label_asym_id`) containing a homolog chain in that structure.

Note that in the case that a structure file contain multiple homolog chains, the same identifier may be provided for each, given that they have different chain IDs, which will be used in their distinction.

Beyond this initialization step, the remainder of the pipeline functions identically to the primary HomologyRing workflow. This feature provides users with significant flexibility in defining and controlling the protein family of study. Additionally, HomologyRing offers an option to export a family file in this format for any previously constructed hRIN. This functionality allows users to save details of families constructed using the primary pipeline, enabling the network to be recreated later.

2.2 Homology Residue Interaction Network - hRIN

An hRIN can be broadly described as a homology-aware extension of a RIN, designed to characterize non-peptide interactions formed by residues across a family of proteins. While the previous discussion of the pipeline focused on the steps for constructing an hRIN, it aimed to highlight the type and extent of homology information encoded in the network. However, the precise methods for calculating and encoding contact conservation within an hRIN are yet to be detailed, as they can best be described in terms of graph based abstractions of hRINs.

In practice, a RINs and hRINs contain a considerable amount of metadata which is utilized throughout the HomologyRing pipeline and supporting analysis. However, the details of which are not needed for a functional understanding of the hRIN object. Abstracting hRINs as networks provides several advantages. The network representation simplifies complex structural and sequence-based homology information into a form that is more amenable to computational and graph-based analyses.

We will begin by providing a definition for the RIN. How RINs are used to synthesized to form an hRIN will be described in a semi-formal manner, as a several aspects abstracted without rigorous definition. Then, this notation developed in this section will be used deliver a concise definition of contact conservation.

2.2.1 Formal Definition

We use the variable x to represent a homolog chain and its corresponding structure information. There is notoriously a great deal of information contained within the CIF file, so a rigorous definition of this object could get quite tedious. Thankfully, here primarily use the variable to identify homologs within the family for which the hRIN is built, which we call X. For example, a homolog search executed on the PDB will return a family of structures such as $X = \{\texttt{8bb5\_B}, \texttt{8bb4\_O}, \texttt{8je2\_C}, \ldots\}$.

We may then abstraction our notion of an RIN created RING for single structure as a multi-graph:

$$H_x = \{V_x, E_x, \varphi, c\}$$

Where V_x is the set of nodes and E_x is the multiset of edges in the RIN created for a single structure x. φ is the incidence map, and c is the edge typing function, which are explained in turn. The incidence map, $\varphi : E \to V \times V$, is used to relate a given edge to the pair of vertices, or nodes, which it connects. For example: $\varphi(e) = (v_1, v_2)$. This is used because in an hRIN a lot of information may be associated with an edge, like its interaction type and its associated weight. All of this information is required to define an edge, but in many contexts its incident nodes are most relevant.

The edge typing function c notation borrows from a common formalism of multi-edge graphs where edges may be colored, where the map c is used to identify the *color* associated with a given edge. Here, $c : E \to T$ maps a given edge to the type of interaction it corresponds to. We denote arbitrary interaction types $\xi \in T$ where T is the set of possible interaction types predicted by RING such as: `HBOND`, `IONIC`, `PIPISTACK`.... A complete list of interactions in T and their properties is provided in Subsection 1.4 of the Introduction.

Importantly, such a multi graph H_x is created for each homolog in the family $x \in X$. We denote the set of all RINs in a family H. It is worth noting that H is simply a collection multi-graphs, and does not contain any of the homology information at the residue level necessary to create the hRIN. Transforming this information into an hRIN requires a series of maps, which in the implementation of HomologyRing, is handled by the MSA and the system identifying nodes associated with entities other than non-homolog chains described in Subsection 2.1.7 of this chapter. Formalizing the exact implementation of this system would be neither useful nor appropriate here, so we leave this process as a black-box, and represent it as a collection of maps $f_x : V_x \to V$ that for a given structure x, sends nodes in its RIN H_x, to the set of homology-aware nodes in the hRIN which we denote V, and is the union of the image f_x over $x \in X$, or more precisely:

$$V = \bigcup_{x \in X} f_x(V_x) \tag{2.1}$$

Observe that f_x is a graph-homomorphism — Formally, a mapping between (multi)graphs G_1 and G_2 such that vertices in G_1 are mapped to vertices in G_2 and edges in G_1 are mapped to edges in G_2. This is frequently denoted as function of the vertices of the two graphs, $f : V_1 \to V_2$ such that if $(u, v) \in \varphi(E_1)$, then $(f(u), f(v)) \in \varphi(E_2)$. Additionally, observe

that taking the family of such maps, given by the family of structures, admits a family of graph-homomorphisms onto a single graph, namely the hRIN of the family. This family is at the core of the definition of the hRIN for a family of structures:

$$F = \{f_x : V_x \to V \mid x \in H\} \tag{2.2}$$

Similar to the definition of V, we use this family of homomorphisms to define the edgeset of the RIN:

$$E = \{(f_x(u), f_x(v), c(e)) \mid e \in E_x, x \in X, (u, v) = \varphi(e)\} \tag{2.3}$$

This is then sufficient to give a definition of the hRIN object:

$$\mathcal{H} = (V, E, \varphi, \rho, c) \tag{2.4}$$

Where V is the set of nodes and E is the set of edges as defined in equations (2.1) and (2.3) respectively. $\varphi : E \to V \times V$ is the incidence map, which given an edge e returns the pair of nodes the edge connects. ρ is the *edge weight function*, returning the conservation scores associated with each edge in the hRIN. The details of its formulation are given in the following subsection. Finally, $c : E \to T$ is the *edge type map* and is invariant under homomorphism, meaning the type is the same as the edges in the pre-image: $c(e) = c(f_x^{-1}(e)) \forall x : e \in H_x$.

As noted earlier, the individual RINs represented by H_x do not contain any homology information, as they were created for a single structure x which is likely incompatible identifiers for its nodes. Later in this section we make use of a homology-aware analog of H_x defined as follows:

$$\mathcal{H}_x = (f_x(V_x), \{(u, v) \mid (u, v) = \varphi(e), e \in E_x\}, \varphi, \rho, c) \tag{2.5}$$

Where edge weights ρ are recalculated on the subset of edges E_x. This can also be seen as the subgraph of $\mathcal{H}$ admitted by a particular structure of interest x which will only contain nodes corresponding to non-gapped residues in columns of the MSA or entities present in the structure file for x.This is used to define a dissimilarity measure in Section 2.3.5, which as the practical use of grouping structures based how similar their interactions are.

2.2.2 Contact Conservation

On of the key aspects of homology information encoded to hRINs are edges weighted with the probability or conservation score. This value captures the propensity of an interaction represented by an edge being present in any given member chain of the protein family. This required developing different notions of *contact conservation* that can be interchanged depending on the needs of the user. The first, more basic, definition simply normalizes the number of observed interactions of a given type for a given pair of nodes by the number of homolog structures in the family. That is, for residue nodes $u, v \in N$ and interaction type $\xi \in \{\texttt{VDW}, \texttt{HBOND}, ...\}$ * (see Introduction Section 1.4) , we define:

$$p_\xi(u, v) = \frac{1}{|H|} \sum_{x \in H} I_\xi(u, v, x) \tag{2.6}$$

where N is the number of homolog structures in the family and I_ξ is simply the indicator function: returning 1 if nodes u and v are connected by an edge of type ξ, else 0. This does well to normalize interaction counts to probabilities, but has the effect of diminishing some interactions conservation score when it may, in fact, not even be possible in every structure. This is because some nodes may represent families of families of residues given by columns of the MSA with low occupancy, meaning they have gaps or no corresponding residue for some structures. Given the variation seen in the PDB for both indels in sequences, and entities present in complexes, it may be undesirable to penalize such situations for some analyses. To address this, we provide an alternative formulation of conservation:

$$p_\xi(u, v) = \sum_{x \in H} \frac{I_\xi(u, v, x)}{I_{\text{non-gap}}(u, x) \cdot I_{\text{non-gap}}(v, x)} \tag{2.7}$$

2.3 Analysis Tools

Beyond simply creating hRINs, the HomologyRing package offers several tools to aid in their visualization and interpretation. These support analysis on a sequence, network, and structural level. As an example for the purposes of demonstrating various visualizations, we will create a

*The HomologyRing tool supports calculating conservation scores for multiple interaction types at once to support analysis requiring queries selecting multiple interaction types of interest. It only requires that conservation scores also be normalized by the number of selected interactions.

small hRIN to represent only 32 chains–using a Hemoglobin chain from the PDB:2DN2 used in Subsection 1.1.6 of the introduction and shown in Figure 1.6.

2.3.1 Analysis queries

Before introducing the visulizations, it is worth mentioning some considerations users may have when performing their analysis. Even though RINs in many ways represent a large simplification of structural data, large protein structures may still contain 1000 of interacting residues, only a small subset of which may be of interest to a user. To address this issue, many of the visualizations provided with HomologyRing support a custom query system, allowing users to quickly filter hRINs for relevant interactions, which in turn greatly simplifies visualizations.

When creating visualizations specifying query parameters effectively limit the visualization to a sub-graph of the hRIN. This sub-hRIN is defined by 3 primary parameters: *interaction type*, *interaction class*, and *region*. Filters on interaction type allows for more focused analyses by considering only edges in the network having relevant interaction types. The interaction types predicted by RING are listed and described in the Subsection 1.4 of the Introduction. A common use of this parameter is to remove the more prolific, and less specific interactions from an analysis. For example: RINs constructed from all possible interaction types tend to be quite dense, primarily as a consequence of prolific nature of interactions like `VDW` and `HBOND`. Temporally discarding information about these interactions can result in a sub-hRIN that highlights aspects of more novel or specific interactions like `IONIC` or `PIPISTACKS`.

Specification of `interaction class` allows the user to construct a graph out of a subset of edges conditioned on the entities the edges are between. Valid choices for this parameter are 'intra', 'inter', 'LIG', or 'all' for considering only intra-chain, inter-chain, chain-ligand contacts, or all interactions respectively.

Intra-chain contacts include interactions and corresponding edges between residues in the same protein chain. Consideration of exclusively these edges enable analysis of secondary and tertiary structure variance and conservation within a family.

Inter-chain contacts refer exclusively interactions with participating residues in different protein chains. At a glance, conservation of contacts can deliver insight into which contacts are key to the binding of protein chains to form ligands. This enables an analysis of binding domains to gain insight into which residues are evolutionary important for interactions or what adaptations in specific sequences may contribute to specificity in partner binding.

Ligand contacts focus on interactions between residues in protein chains and ligand molecules,

including small molecules, cofactors, or nucleotides. Conservation of these contacts in a family can reveal key residues involved in ligand binding, highlighting their functional importance.

The last parameter that may be specified in a query is the *region*. Unlike the previous parameters–interaction type and interaction class, which filter edges in an hRIN–the region parameter filters nodes to define the resulting sub-hRIN. By default, no filtering is applied, and all nodes are included.

The region parameter allows users to specify regions of interest in a structure by providing a range node IDs, creating a sub-hRIN that focuses on interactions involving at least one node that falls within that region. For example, generating an hRIN for an entire structure may include irrelevant information when analyzing inter-chain binding substrate conservation within a family. Specifying the set of continuous node IDs that correspond to the substrate results in much more focused analyses. Another common issue is some sequences included in a family may be abnormally long, creating regions of low-occupancy at the beginning or end of the MSA, which results in many superfluous nodes in the network. To address this, users may apply an occupancy ratio filter, selecting residue nodes corresponding to MSA columns with a non-gap content above a specified threshold. This enables quick exclusion of poorly conserved regions, such as unrepresentative tails of the MSA.

Note that region filters are not strict. If an edge connects to at least one node in the selected region, both the edge and its connected nodes will be included in the resulting hRIN.

2.3.2 Network Visualization

Visualizing the network representation of an hRIN is a good starting point for interpreting its structure and relationships. The left-most network in Figure 2.2 shows the full hRIN created by HomologyRing for the Hemoglobin family, alongside two sub-hRINs generated using HomologyRing's query feature. As previously mentioned, these networks can be dense and challenging to interpret visually. However, they still provide valuable insights into the family. For instance, interaction conservation scores are encoded in the line weights, where thicker lines represent more conserved interactions within the family. Additionally, edge colors correspond to different interaction types, as indicated in the legend in the upper-right corner. Node colors convey specific node types: gray nodes represent Residue Nodes, green nodes correspond to Chain Nodes, and red nodes indicate Ligand Nodes. These node types are detailed in Subsection 2.1.7.

Network (b) in Figure 2.2 represents a sub-hRIN filtered to include only intra-chain edges,

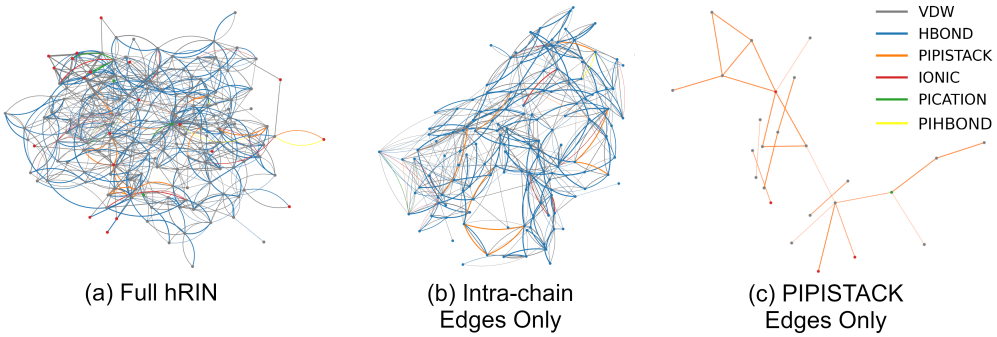

Figure 2.2: Example of hRIN visualizations created with HomologyRing.

meaning interactions connecting two Residue Nodes within homolog chains of the structures. Applying this edge filter provides a clearer view of key conserved interactions, offering insights into the structural characteristics of the family. One noticeable feature is the prevalence of Van der Waals interactions, which, despite their abundance, are relatively poorly conserved. In contrast, hydrogen bonds show a high degree of conservation, reflecting the critical role of secondary structural elements such as α-helices, which are foundational components of protein structure.

Another observation is the conservation of π-π stacking interactions, which are essential in the Hemoglobin family. As discussed in Subsection 2.1.7, aromatic residues in Hemoglobin chains play a crucial role in forming these interactions. These well-conserved aromatic residues are vital for Hemoglobin's tertiary structure, the assembly of the Hemoglobin tetramer, and the coordination of Heme ligands. In less well-characterized protein families, identifying conserved and specific interactions like these can be aid in uncovering interactions critical to a protein's structure and function.

2.3.3 Contat Mask Plots

While conserved interactions can be identified by visualizing the hRIN, this approach offers limited insight at the sequence level. Contact mask plots address this gap by enabling the quick identification of conserved or variable interactions formed by specific regions of a protein family. These plots are built on the MSA used to create the hRIN, leveraging the MSA's ability to highlight conserved sequence features. Instead of displaying the residue symbols typically

shown in an MSA, residues in the contact mask are represented using CLUSTAL Colors.[†] This coloring scheme groups residues with similar properties (e.g., charge, polarity, aromaticity) into distinct colors.

To overlay contact information, a binary mask is applied to the MSA colors. If a residue participates in an interaction of interest, its color remains bold; otherwise, it is faded to indicate the absence of such interactions.

Figure 2.3 shows a contact mask created for our Hemoglobin A family, indicating which residues specifically participate in π-π interactions. This can provide a different perspective on how interactions are conserved in a family than what is communicate by the edge weights in the previous network visualization.

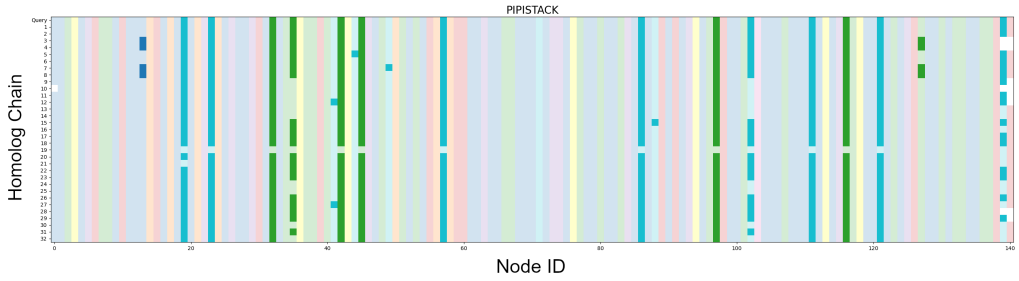

Figure 2.3: Example of *contact mask plot*. Created query PDB: 2DN2_A, HBA_HUMAN

Unlike other analyses provided by HomologyRing, conservation of contacts in a contact mask plot is seen at the sequence level as conserved columns in the MSA, rather than as edge weights or conservation scores in the hRIN. However, a notable limitation of this approach is that it only indicates whether a residue participates in an interaction; it does not specify the interaction partner. To address this limitation, we later introduce a contact similarity analysis as an extension of this visualization. Despite this, contact mask plots remain a valuable tool for identifying regions of a protein family that participate in specific or important interactions.

2.3.4 Interaction Conservation Plots

Interaction conservation plots are a useful way to see information about specifically the edges and their conservation scores in a network, in a visually clear manner, but can also give a surprising amount of insight to a structure of a protein family.

[†]A description of the Clustal Coloring system and the properties of the different groups can be found at: jalview.org/help/html/colourSchemes/clustal.html

They are created by creating a matrix where each cell represents a potential edge in an hRIN, identified by the row and column indices which correspond to the node IDs that the edge connects. The value of each cell will be a function of the conservation score associated with that edge, or 0 if no such edge exists. Figure 2.4 shows conservation plots for three interaction types seen in the Hemoglobin family, in order: Hydrogen Bonds, Van Der Waals, and π-π stacking.

Immediately, the most conserved interactions in the family can be identified: the notably very specific π-π stacking, and more prolific hydrogen bonds. However, where it was relatively difficult to interpret VDW information in the hRIN network visualizations, some rather striking patterns emerge here. Which highlight a nice secondary use of these plots for getting information about the conformation of structures in the family. Highly conserved interactions are indicative of regions of structures that are consistently in close proximity, but the poorly conserved Van Der Waals forces would indicate some variability, or flexibility, of other parts of the structure.

This is a very nice example about how critical aspects of a proteins function are captured by hRINs incorporation of homology information. In this case, variability of specific VDW interactions may be attributed to a slight conformal changes to the hemoglobin complex in its oxygenated and de-oxygenated states[12], which are both contained in the family.

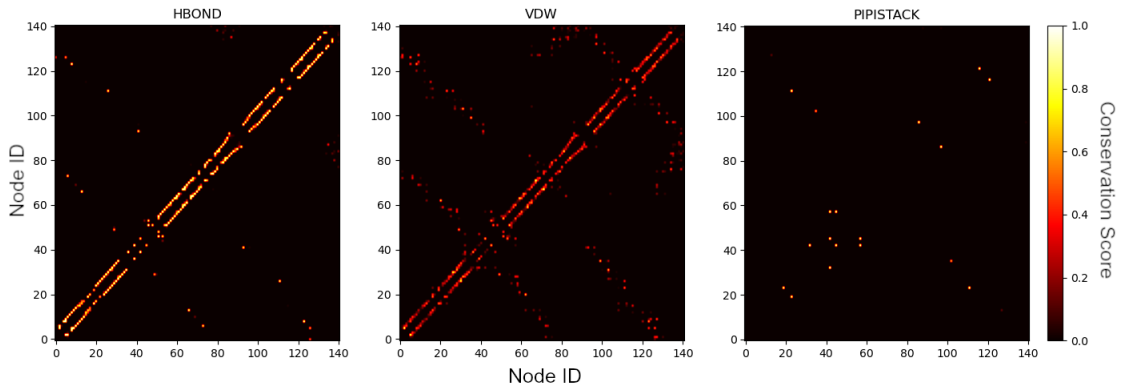

Figure 2.4: Examples of *contact conservation plots* considering different interaction types.

These conservation plots bear great resemblance to visualizations of distance matrices commonly used in structural bio informatics, where row and column indicies both correspond to amino acids in a protein chain, but each cell in the matrix will represent the distance between the two residues. The are commonly employed in machine learning applications for their effectiveness at encoding information about a proteins conformation. Some of the basic elements of secondary structure can be easily identified as patterns in a contact matrix: α-helices appear

as parallel lines slightly offset from the diagonal, and β-sheets can appear as lines either parallel or perpendicular to the diagonal.

Interaction Conservation plots are similarly useful. These patterns are readily identifiable in the HBOND and VDW conservation plots, and in fact, contain additional useful chemical information. For example, it is easy to see from the plots the significance hydrogen bonding has for the formation of α-helices compared to other interactions.

Beyond just being useful for intrepreting hRINs, interaction conservation plots serve well to demonstrate how hRINs effectively capture important structural, chemical, and functional aspects of a protein family.

2.3.5 Contact Similarity Analysis

With the importance of non-covalent interactions for conformal and functional properties of a protein, a reasonable analysis task is to compare and group structures included in a family based on the similarities in the contacts present in each structure. To achieve this, we provide a means means of classifying structures into a pseudo-taxonomy, grouping structures based on homology-aware RIN similarity.

Once a means of comparing family members has been established, they will be used to create an extension of the contact mask visualization seen previously. A notable draw back of contact masks for visualizations, was presence of many interactions in a column would seem to indicate a well conserved interaction for those nodes. However, it remains possible that though the same node is seen to participate in many interactions, that the partners may be different, which may constitute a significant dissimilarity between the structures.

We will after providing the details of the methods implementation, we will use it to group structures in the example Hemoglobin family by the interactions present.

Typically, evolutionary relationships can be inferred by sequence similarity: using disparities between multiple residue sequences to infer the elapsed time since a common ancestor. Further structure is given to these relationships by organizing individuals into a tree to map their relationships. We utilize this concept to compare and group structures based not on sequence similarity, but rather the similarity of their respective RINs.

At the core of this approach is a means of quantifying the dissimilarity of RINs based on their graph definition and the node maps integral to hRINs. Given an hRIN $\mathcal{H} = (V, E, \varphi, \rho, c)$, for homolog protein structure x we denote the edge set of $\mathcal{H}_x$ as $E_x \subseteq E$, or the set of edges in an hRIN that have corresponding edges in structure x. We continue to define the edge

difference set, $D_{x,y} \subseteq E$ for two homolog structures x and y to be the set of edges present in either x or y exclusively. This can otherwise be thought of as the symmetric difference of the subgraph edgesets:

$$D_{x,y} = (E_x \backslash E_y) \cup (E_y \backslash E_x) = E_x \oplus E_y \tag{2.8}$$

Visualization of this operator is shown in Figure 2.5. Observe, edges are considered to be shared between two graphs if their incident nodes have the same labels. We are able adapt this notion thanks to the node maps created with the MSA, and may avoid using more complex and computationally expensive notions of graph similarity. Additionally, the result may have node set of the result will be the union of the two graphs.

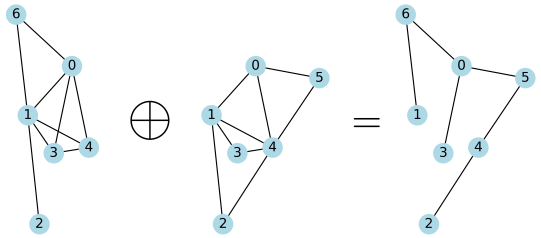

Figure 2.5: Example of the graph *Symmetric Difference* (XOR) operator on graphs.

With this we define both a unweighted and weighted pseudo-distance[‡] to quantify the dissimilarity of homolog chains based on the interactions the admit.

$$d(x,y) = |D_{x,y}| \quad \text{or} \quad d(x,y) = \sum_{e \in D_{x,y}} p(e) \tag{2.9}$$

This is also referred to as Hamming distance, though it is worth noting that the weighted version in this context differs from most uses of the Weighted Hamming Distance, where the difference in edge weights between the two different graphs are summed. Here the edges, if they are shared between the two subgraphs, will have the same weights, and will not contribute to the distance.

The ladder weighted variant will penalize more discrepancies of highly conserved edges - considering homologs with edges , where the unweighted variant considers only the number of unshared edges significant. The type of analysis being performed should inform the choice between the two. Studies of contact conservation will desire that structures that admit highly

[‡]This cannot be called a true distance metric, as the triangle inequality does not necessarily hold. However, does not represent a significant issue for practical issue.

conserved edges will be grouped closer together in the resulting taxonomy, while studies of specificity will value shared novel or uncommon edges more important when grouping structures.

With the distance metric defined, the sequences organized into a rooted tree, or deprogram, using the popular unweighted pair group method with arithmetic mean (UPGMA) method. This is an iterative method, employed here to create a hierarchical clustering of structures based on non-covalent interactions of interest. At each iteration, the method defines a distance between groups of examples based on the average inter-group distance:

$$d(X, Y) = \frac{1}{|X||Y|} \sum_{x \in X} \sum_{y \in Y} d(x, y) \tag{2.10}$$

This is referred to a pseudo-taxonomy because it does not aim to capture inferred evolutionary relationships, rather group structures based on the similarity of the non-covalent interactions the admit. This is useful for finding patters of contact conservation or specificity.

We may use these methods to create a visualization to investigate some notably variable interactions within the Hemoglobin A hRIN. Specifically, we will consider inter-chain VDW interactions in the tail region of the family. Specifically, nodes $100 - 140$, which correspond to the C-terminal region of the HBA and consist of an α-helix and a short disordered tail. Figure 2.6 shows the contact similarity plot for this region.

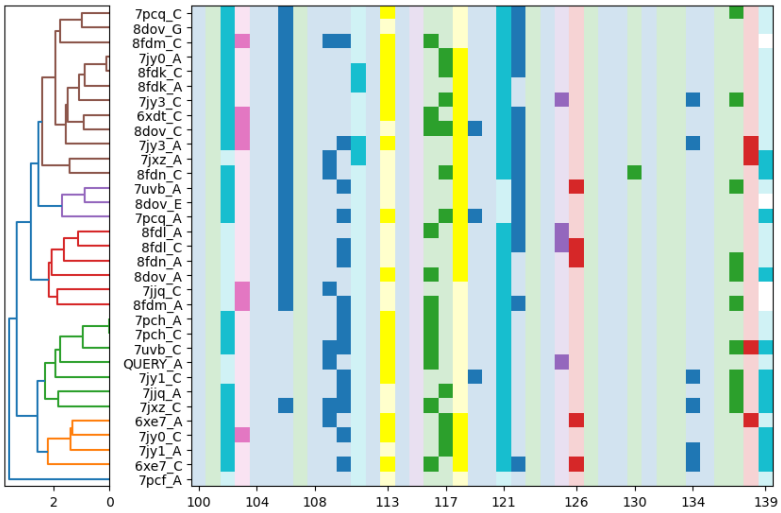

Figure 2.6: Example of *contact conservation plot*

The plot highlights several well-conserved interactions in this region, most notably involving

residue 127, which interacts with an HBB chain in the complex. Other interactions, however, show greater variability. For instance, pairs of columns such as 109 and 110 or 116 and 117 exhibit inconsistencies in which nodes form interactions. These regions correspond to turns in the secondary structure, which exhibit some flexibility, allowing them to form transient interactions.

The disordered tail, corresponding to the last three columns, is seen to variably interact with both HBA and HBB chains in the complex. Although these interactions occur in similar regions across the family, differences in interaction patterns contribute to the structural variability among the members. This variability is captured by the graph-based methods underlying the dissimilarity measure, enabling the identification of subtle differences between structures.

The combined effect of these factors results in the grouping of structures, as seen in the dendrogram on the left, which accounts for these nuanced variations.

3

Example Application

3.1 Elongin C - pVHL Interaction

This section demonstrates the utility of HomologyRing for providing structural insights into a specific protein family and its complexes. Specifically, we analyze inter-chain interactions within an hRIN constructed using Elongin C (ELOC), a component of the Elongin complex, as a query. By leveraging HomologyRing's analysis tools, we aim to explore the complex, focusing on ELOC's interactions and conserved binding interfaces.

Transcription Factor B, also known as SIII or simply Elongin, is a protein complex responsible for promoting RNA polymerase II elongation by suppressing the function of arresting sites within a transcription window [13]. It has been identified as an elongation factor that regulates transcription by RNA polymerase II. While transcription is commonly thought to be regulated by initiation factors, regulation during elongation is increasingly recognized as critical. The Elongin complex is dynamic and interacts with several other factors that influence elongation behavior, and one should expect to find considerable variation in the inter-chain found in experimental structures.

Transcription Factor B, also known as SIII or simply Elongin, is a protein complex that promotes RNA polymerase II elongation by suppressing arresting sites during transcription [13]. While transcription regulation is often associated with initiation factors, elongation regulation is increasingly recognized as critical. The Elongin complex is dynamic, interacting with several

factors that influence elongation behavior, contributing to its structural variability.

The Elongin complex comprises three subunits:Elongins A, B, and C. Respectively denoted as ELOA, ELOB, and ELOC genes, respectively. ELOB and ELOC often form a subcomplex independent of ELOA [13]. A notable interaction involves Elongins B and C binding with the Von Hippel–Lindau tumor suppressor protein (pVHL). This interaction inhibits transcription elongation under hypoxic conditions. The pVHL protein plays a critical role in regulating tissue responses to low oxygen levels, primarily through ubiquitination of hypoxia-inducible factors (HIFs), which mediate cellular responses to hypoxia. Dysregulation of this pathway is strongly linked to tumor development, particularly in vascularized tissues, as observed in Von Hippel–Lindau syndrome.

Figure 3.1 shows the complex formed by ELOB, ELOC, and pVHL in association with a fragment of HIF.

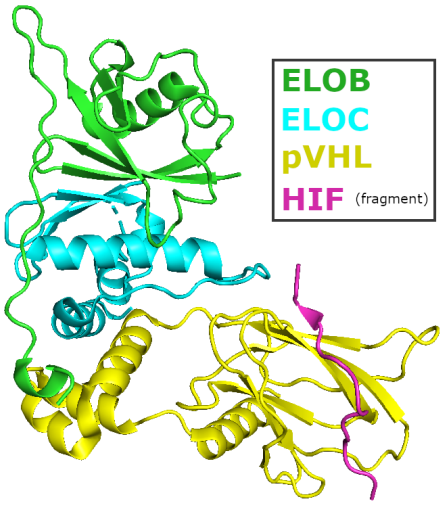

Figure 3.1: PDB: 1LM8; Complex of Elongin B, C with pVHL and HIF-region [1].

In this analysis, we focus on ELOC, using it as a query to construct an hRIN with HomologyRing. This will allow us to observe inter-chain interactions within Transcription Factor B (SIII) and related complexes. By comparing structural insights gained through HomologyRing with findings from the literature, we aim demonstrate the utility of the tool and the information contained in the resulting hRINs.

The pipeline begins by performing a homology search of the PDB to retrieve structures containing ELOC or its homologs. While AlphaFold provides high-quality predictions, it is

limited to single-chain structures, preventing the study of inter-chain interactions. The PDB, by contrast, includes structures reflecting diverse biological contexts, such as variations in pH, temperature, or interaction partners.

ELOC is taken as the query structure from chain `AUTH C` in PDB entry `1LM8`, an X-ray crystal structure of the ELOB-ELOC dimer in complex with a pVHL fragment [1]. The sequence used comprises 88 residues, corresponding to UniProt accession `Q15369-2`, an isoform differing from the canonical sequence by the omission of 16 N-terminal residues.

A homology search retrieves 128 structures from the PDB, which collectively contain over 500 MB of data. To simplify visualization, the search is restricted to sequences with high pairwise similarity (E-value threshold: 5×10^{-3}). The pipeline quickly constructs an hRIN, capturing information about conserved and variable interactions across the family.

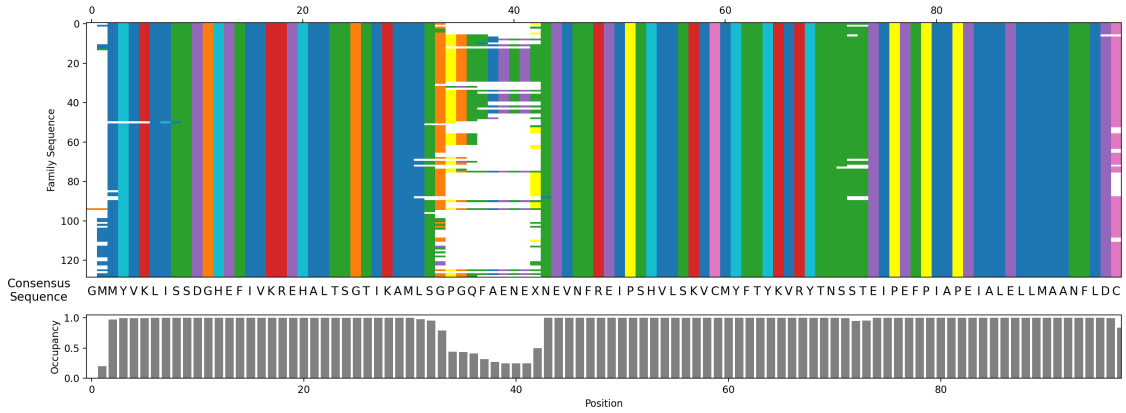

Figure 3.2: Representation of a multiple sequence alignment for the ELOC family. Created as part of the HomologyRing pipeline with ClustalΩ

As stated, a primary aim of this analysis is structural insight into inter-chain interactions. To begin, we will summarize which protein chains are present in the families structures. This gives indication of which proteins are commonly observed in complex with homologs of our query in the PDB and which interactions are captured by the network. Table 3.1 provides an overview of the presence of protein chains in the structure collected to form the family. HomologyRing exposes this information originally collected from CIF metadata and includes additional information on ligand presence and are provide reference for detailed information associated with nodes in the network.

ELOC_HUMAN, despite being used as the query sequence, is only listed as present in 98.2% of structures. This is due to close homologs like ELOC_MOUSE comprising a small handful of the homolog chains included in the family. We also observe that the distribution of

protein chain counts has a long tail. Truncated from Table 3.1 was 30 different proteins that appeared in no more than 2 structures. Some are the close homologs, as previously mentioned, while others are lesser studied proteins associated with Elongin or RNA Poly II.

Gene	Chains	Structures with Chain	Structures with Chain (%)
ELOB_HUMAN	158	63	98.4%
ELOC_HUMAN	124	59	92.2%
VHL_HUMAN	78	39	60.9%
CUL2_HUMAN	38	18	28.1%
RBX1_HUMAN	24	14	21.9%
FEM1B_HUMAN	19	10	15.6%
APBP2_HUMAN	16	5	7.8%
PEBB_HUMAN	15	3	4.7%
BRD4_HUMAN	15	8	12.5%
CUL5_HUMAN	14	2	3.1%
RASK_HUMAN	6	3	4.7%
HIF1A_HUMAN	6	3	4.7%
OTHER	77	34	53.1%

Table 3.1: Number of occurrences of each the most common protein chains in family of Elongin complexes.

This gives us information the most common interactions with ELOC in the PDB, but lacks information about what regions in ELOC are forming them. In order to get a sense of where this might be happening we may use an interaction conservation plot to examine which residues in ELOC form the most VDW interactions, indication the presence of a binding substrate.

Patterns indicating common interaction substrates are visible in Figure 3.3. HomologyRing allows you to access the maticies produced by the contact conservation method. Filtering this information further allows for a better understanding of the regions and proteins involved.

This plot clearly indicates 3 primary regions constituting ELOCs interface with ELOB. And referring to the structure depicted in Figure 3.1, we can indeed see that these correspond with a continuous β-sheet, small region of contact in a central hairpin, and a disordered C-terminal region at which they make contact.

We can additionally see how the ELOC-pVHL interface is primarily dependent on a conserved region at the C terminal of ELOC, wich is known to be a region particularly sensitive to mutations preventing their binding[16].

Looking for conserved regions of VDW interactions or hydrogen bonding is a useful approach for quickly finding important binding sites in proteins. And now that they have been identified, we may examine these regions for more specific interactions. We will consider the

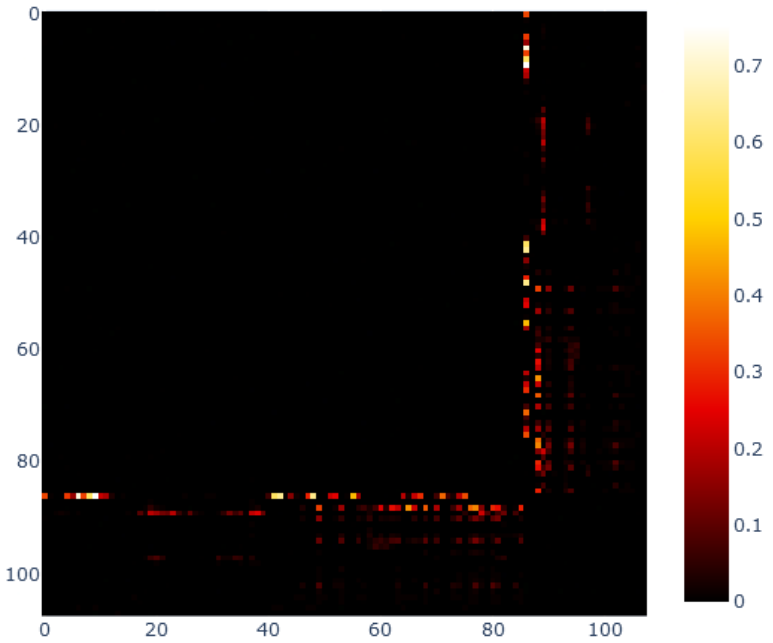

Figure 3.3: Contact conservation plot for inter-chain VDW interactions formed by ELOC family.

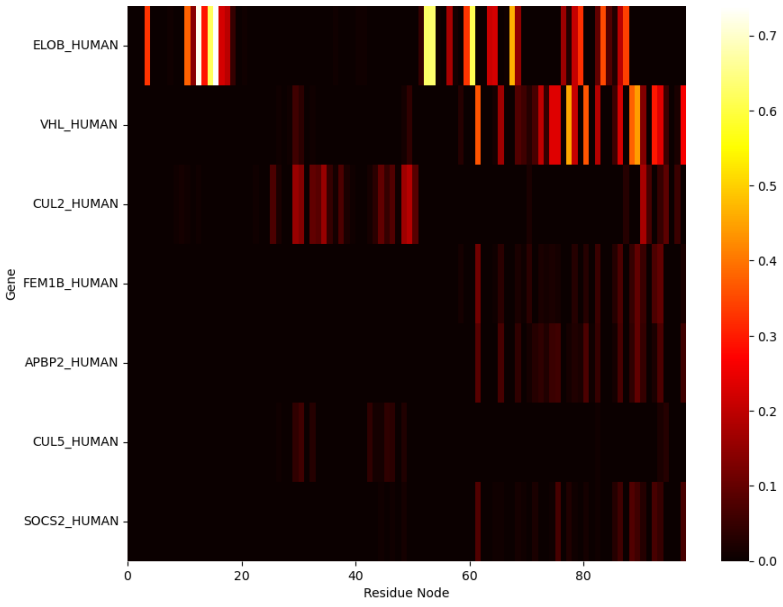

Figure 3.4: Filtered Contact conservation data displaying reigions of ELOC family that most frequently participate in inter-chain interactions.

highly conserved region of ELOC-ELOB interaction at nodes 11-17. These correspond to residues 26 - 32 in ELOC_HUMAN, UniProt: Q15369. Figure 3.5

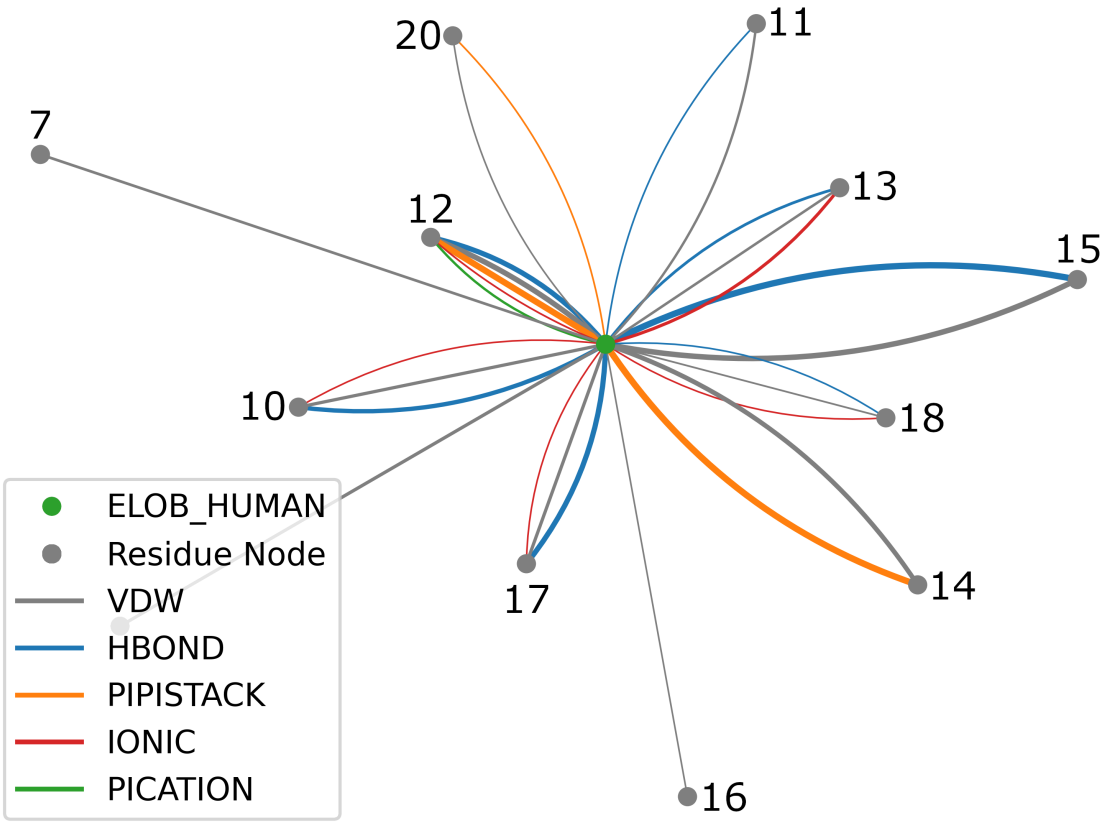

Figure 3.5: Sub-network of hRIN created for studying a region of ELOC-ELOB binding.

The figure conveys at a glance information from over 64 structures and 128 distinct ELOC structures. it indicates highly conserved π-π stack, and hydrogen bonds comprising the two proteins interface.

4

Conclusion

This work introduces an extension of Residue Interaction Networks (RINs), incorporating homology information to encode key biological aspects of protein families. By multiply aligning sets of homologous protein sequences, we defined nodes to represent families of corresponding residues across structures, identified by the columns of the MSA, rather than single residues as they do in typical RINs. In doing so we defined a map between nodes that also be used to relate corresponding edges, which are aggregated to create our definition of interaction conservation scores. Weighting edges with these scores allows of the identification of critical interactions, whose high conservation is evidence to evolutionary importance, attributed to its role in contributing to a proteins structure of function.

Termed *Homology enriched Residue Interaction Networks*, or hRINs, they combine powerful analysis enabled by homology information with the many attractive features of RINs. The simplified representation of protein information enables the extension of sequence-level homology information from BLAST searches and MSAs to enhance the structural information captured by RINs. In doing so, tapping abundance of information available from structure databases like the PDB and AlphaFold. Network representation is seen to retain useful information of structures topology while not having to process atomic coordinate data. Furthermore, the different interaction types predicted by RING serve provide a significant amount of chemical information to the object.

Novel applications of graph methods allow for quantitative analysis of dissimilarity between proteins, based on non-covalent interactions which capture many biologically relevant aspects

of proteins. This was created after observing that individual structures in a family will only contain a subset of the edges seen in an hRIN for that family, and different structures, then were represented as sub-graphs of the hRIN. The different sub-graphs are then compared with the hamming distance on graphs, which allowed the application of aggregating methods to assign structures accordingly to groups.

In addition to conceptualizing of the hRIN, it required the development of the HomologyRing Python Package so they could be realized and analyzed. Consisting, which has come to posses a sizable feature set to accommodate many different uses and users.

Firstly, we implemented a pipeline for the creation of hRINs. Which coordinated all files, applications, services, databases, and artifacts need. From Python, a CIF file is parsed to extract a query sequence that becomes the basis of a blast search, handled by BLAST command-line, from there structure files are programmatically downloaded from either AlphaFold or PDB databases. The resulting structures are parsed, and their sequences aligned using command-line ClustalΩ. Additionally the collection of structures are processed by RING to produce normal RINs which HomologyRing must then synthesize into a final hRIN.

To include in the package, we additionally developed several visualizations and tools to aid users in accessing information encoded within hRINs, and promoting HomologyRings use as a research tool. Some of these methods included visualization of the hRIN as a network, interaction conservation plots, and interaction similarity plots. The package includes visualization and analysis tools to facilitate user access to the information encoded in hRINs, such as network visualizations, interaction conservation plots, and interaction similarity plots. Each of which were useful and provided interesting insight into different aspects of hRINs.

We aimed to create a flexible tool, allowing users to work with in very different ways to accommodate different users preferences, experience, and use cases. The HomologyRing package naturally contains methods for its use in python directly, but also supports use from the command line interface, allowing users to easily export hRINs as files for use in other network visualization software. The support for users to define their own protein family for analysis offers great flexibility, and positions the tool for many more possible uses.

Implementation of a web application which allows for dynamic and reactive visualizations of hRINs has been very useful for exploratory analysis, and shows promise to increase the tools accessibility due to its relative ease of use.

hRIN visualizations, such as contact conservation plots, are effective for interpreting functional aspects of proteins by identifying key interactions whose conservation indicates evolutionary importance, attributed to its role in contributing to a proteins structure of function,

but can also signal which regions in a family constitute a binding substrate by using the plots to study inter-chain contacts in families of experimental structures.

We provided a semi-formal description of the hRIN as a multi-graph, borrowing a few notions graph theory. In doing so we abstracted some of the more technical details of the pipeline to create a more robust theory of an hRIN, the information it contains, and how it was created. All of this with the aim that a more mathematically rigorous description of the hRIN would lend itself to future applications of computational and more theoretical methods.

Finally, we presented two examples showcasing the insights derivable from hRINs. Looking first at a family of human Hemoglobin A, we observed the role of highly specific and well conserved interactions in the complex π-π stacks play in maintaining the tertiary and quaternary structure of Hemoglobin. Additionally we observed how conformal variations in a family of structures can manifest in the structure of an hRIN.

After that, we showd how HomologyRing may be used to quickly gain familiarity with binding sites in highly variable complexes such as Elongin.

5 Appendix

5.1 Installation and Dependencies

The code repository can is available from the BiocomputingUP GitHub:
https://github.com/BioComputingUP/ring-homology-pipeline
Execution of the pipeline has a set of commandline programs be installed and correctly added to PATH.

BLAST Command Line	`https://www.ncbi.nlm.nih.gov/books/NBK279690/`
ClustalΩ	`http://www.clustal.org/omega/`
RING standalone	`https://ring.biocomputingup.it/`

Note that ring4.0 command line is needed for use of `use_label_asym_id` argument when building a hRIN. Additionally, commandline BLAST requires copies of databases be built and installed locally inorder to perform a homology search. The two that we display in this paper and are likely to be of particular interest to a user is '`swissprot`' and '`pdb`' which may be downloaded from the BLAST FTP server. Files, typically in FASTA format must be built into blast databases using the BLAST commandline tool and placed into the `blast_db` directory. Specifying that the homology search should be performed remotely with the `remote_BLAST` argument can avoid the need for a local blast installation. However, it should be noted that the search usually takes much longer on NCBI servers.

5.2 Usage

Though implemented as a python package, HomologyRings functionality may be acces in many different ways, with the aim to accommodate users of a verity of different backgrounds. We will introduce some of te ways HomologyRings functionality can be accessed. A more through guide to using HomologyRing can be found in the 'ExampleUsage' notebook within the repository.

5.2.1 Python

Running the pipeline and supporting analyses witin one Python Environment is the most flexible and powerful way to use this tool. As HomolgyRing exposes several data attributes, and methods that describe and can modify an hRIN. These include tabular records of Node and Edge metadata, and reference to an instance of useful artifacts, such as the family MSA, or BLAST results. Additionally, network representations of the hRIN object are provided and can easily be incorporated into other python workflows.

5.2.2 Command Line Interface (CLI)

Creation of hRINs, either from a user-defined family or by homology search given a query, may be executed by invoking HomologyRing through the CLI, and detailed information about arguments is accessible by invoking the help argument with '`python3 homologyRing --help`'. Use of HomologyRing via the CLI works well for integration of the output network files with other applications as discussed in the following Subsection.

5.2.3 File Output

HomologyRing supports outputting a hRIN – either from an entire network or a user query – to files which can be directly imported into other applications popular in bioinformatics workflows. Cytoscape is frequently used for the visualization of network information arising from biological contexts, including: protein-protein interaction networks (PPI), metabolic pathways, and more recently Residue Interaction Networks. HomologyRing allows exporting hRIN as a pair of tsv files which allows them to be visualized and analyzed in cytoscape among other software. The hRIN is defined in this context with a node file and an edge file, both of which contain identifying and metadata information about the entity. In cytoscape, the

user may synthesize this information with data from other sources, or perform analysis on the network.

5.2.4 Interactive Application

We have aimed to illustrate thus far how HomologyRing as a tool can be used to gain useful information in a wide range of applications. We do acknowledge, however, that knowledge of python and learning the functionality of the methods required to preform analysis represents significance obstacles in its use. To address this, we have an interactive web application that allows a user to access the core functionality of the software in an rich interactive manner. Because this represents an easy way to begin using HomologyRing, we provide a detailed description of its use and its interface.

The web app is ran locally from the users machine in implemented in plotlys Dash — a powerful python framework to build data driven applications in a reactive paradigm where both the front and back end are handled in Python. To lauch the app, simply execute the app_test.py script included in the repository. After it loads, it will output a local web address. Following this address in a web browser will open the application.

The use of this software allows users to explore interactions in a protein family in a dynamic and intuitive manner. As a user builds a query specifying the interaction types, classes, and regions to be included in the hRIN – The supplementing visualizations are updated.

An overview of these visualizations is shown in Figure 5.1, and include: A 3D display of the network, a contact conservation score plot, and the contact plot. One of the primary appeals of Plotly is the ability to enrich visualizations with additional data such as hover-text to plots. Here, this aids use usability of contact plots and conservation plots, as hovering over a cell, provides the user with information pertaining to a particular, node, edge, or record corresponding to a structure. Furthermore, there is the addition of click events, which are utilized by the interactive HomologyRing app to allow the user to create dynamic selections. Should the user click on a cell in the contact conservation plot, n set of multi edges associated with that cell will become selected; and clicking a cell in the 3D network plot results in a node being selected. A summary of information pertaining to the users current selection is displayed to the right of the 3D Network visualization. Summary content varies depending on the type of object selected. Should a edge be selected, brief information identifying the nodes that edges are between are provided. Additionally interaction conservation score is displayed and information of interaction observations in structures captured by the edge is displayed in a tabular

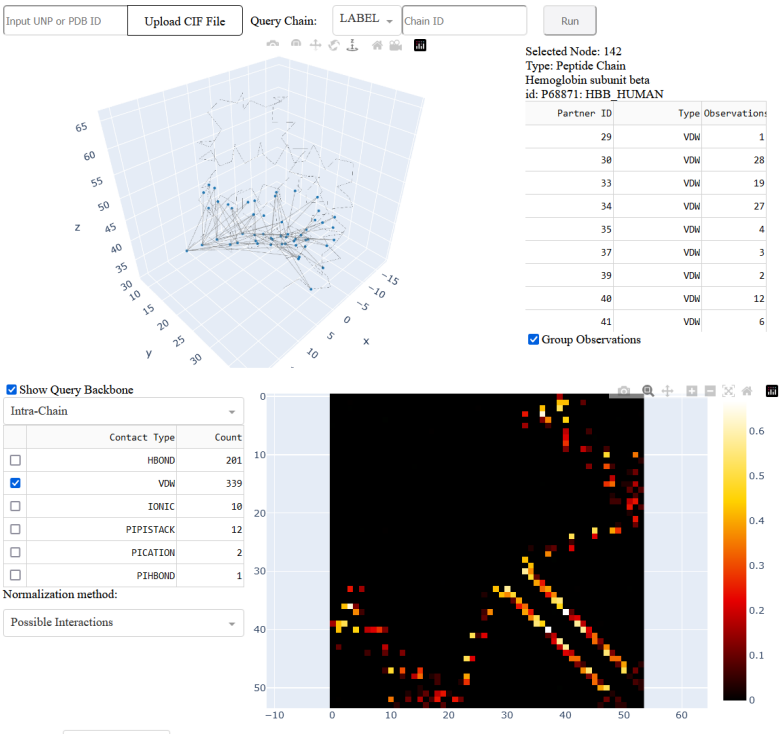

Figure 5.1: Sample view of the HomolgyRing web app.

format below. Observations are grouped by interaction type, and the user has the option of either displaying a count of observations of interactions of that type is provided, or a list of structures said interactions are observed in. Alternatively, should a node be selected, more detailed information about node identifying information is provided. For the different types of nodes this information will vary. All nodes will display their numeric id used to identify the node within the hRIN. For residue type nodes, this id corresponds to the column of the MSA the node represents. Additionally, an amino acid residue from the consensus sequence is given, giving the user indication as to the properties of the residues in that family. Nodes representing protein chains, the gene the protein is transcribed from along with the structure identifier.

The 3D visualization allows users to see nodes and how they generally relate to each other spicily. Interactions specified in the user query will be displayed in the graph. It is important to note that only nodes connected by an edge returned by the user query will be displayed in the plot. This reduces visual clutter, and also reflects the internal representation of the custom hRIN defined by the user that may contain information of particular interest to their use case. The 3D representation of the network takes position information from the Query chain the user specified when the pipeline was initiated. It is often the case that the MSA will open

gaps in the query sequence, and as a result, there will be no position information associated with some nodes for use in the visualization. To solve this issue, the positions of nodes that do not have a member residue in the query chain are linearly interpolated or extrapolated from other nodes position information. Finally, the user can choose to display the 'backbone' of the polypeptide chain. this is displayed as a polyline connecting the positions of the alpha carbons in each residue in the chain, allowing the user to see the conformation of the query chain and infer family structure when only a sparse subset of nodes may be displayed.

The query builder functionality is divided into two spaces Edge filtering options are placed to the left of the contact conservation plot and node filtering options are placed above the contact mask plot which provides a visualization of the MSA that defines the residue type nodes.

Filters on the edges include allowing to users to specify the interaction class (intra-chain, inter-chain, ligand, or all), and the interaction types. These filers may be used to easily create focused analyses. The interaction class dropdown allows users to focus their analysis on all possible interactions to form edges, or limit results to intra-chain, inter-chain, or chain-ligand interactions. Just below is a table containing basic information on the types of interactions present in the protein family. This display the number of interactions of each given type and the ability to toggle inclusion of the interactions in the network.

Above the *contact mask plot* are controls for region filtering. The user may either choose to include all nodes, specify a range in which residue nodes will be included, or an occupancy ratio may be given. Use of this filters allows for greatly simplifying visuals and focusing on regions of interest.

References

[1] J.-H. Min, H. Yang, M. Ivan, F. Gertler, W. G. J. Kaelin, and N. P. Pavletich, "Structure of an hif-1alpha -pvhl complex: hydroxyproline recognition in signaling," *Science*, vol. 296, no. 5574, pp. 1886–1889, 2002, epub 2002 May 9.

[2] E. P. Carpenter, K. Beis, A. D. Cameron, and S. Iwata, "Overcoming the challenges of membrane protein crystallography," *Current Opinion in Structural Biology*, vol. 18, no. 5, pp. 581–586, Oct 2008, epub 2008 Aug 11.

[3] D. Yehorova, R. M. Crean, P. M. Kasson, and S. C. L. Kamerlin, "Key interaction networks: Identifying evolutionarily conserved non-covalent interaction networks across protein families," *Protein Science*, vol. 33, no. 3, p. e4911, Mar 2024.

[4] K. A. Dill and J. L. MacCallum, "The protein folding problem, 50 years on," *Science*, vol. 338, no. 6110, pp. 1042–1046, 2012.

[5] M. R. Scholfield, C. M. Zanden, M. Carter, and P. S. Ho, "Halogen bonding (x-bonding): a biological perspective," *Protein Science*, vol. 22, no. 2, pp. 139–152, Feb 2013, epub 2012 Dec 29.

[6] M. J. Feige, I. Braakman, and L. M. Hendershot, "Disulfide bonds in protein folding and stability," in *Oxidative Folding of Proteins: Basic Principles, Cellular Regulation and Engineering*. The Royal Society of Chemistry, 07 2018. [Online]. Available: https://doi.org/10.1039/9781788013253-00001

[7] E. A. Permyakov, "Metal binding proteins," *Encyclopedia*, vol. 1, no. 1, pp. 261–292, 2021. [Online]. Available: https://www.mdpi.com/2673-8392/1/1/24

[8] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, October 1990. [Online]. Available: https://doi.org/10.1016/S0022-2836(05)80360-2

[9] U. Consortium, "Uniprotkb/swiss-prot: The manually annotated and reviewed section of the uniprot knowledgebase," 2024, accessed: November 27, 2024. [Online]. Available: https://web.expasy.org/docs/relnotes/relstat.html

[10] P. D. Bank, "Protein data bank statistics: Sequence clustering and unique protein sequences," 2024, accessed: November 27, 2024. [Online]. Available: https://www.rcsb.org/stats/nr/cluster-ids-all

[11] M. Varadi, S. Anyango, M. Deshpande, S. Nair, C. Natassia, M. Schmitz, J. Shi, and J. Charchan, "Alphafold protein structure database in 2024: Providing structure coverage for over 214 million protein sequences," *Nucleic Acids Research*, vol. gkad1011, 2023. [Online]. Available: https://doi.org/10.1093/nar/gkad1011

[12] M. R. Mihailescu and I. M. Russu, "A signature of the t → r transition in human hemoglobin," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 7, pp. 3773–3777, Mar 2001, epub 2001 Mar 20.

[13] Y. Chen, G. Kokic, C. Dienemann, O. Dybkov, H. Urlaub, and P. Cramer, "Structure of the transcribing rna polymerase ii–elongin complex," *Nature Structural & Molecular Biology*, vol. 30, no. 12, pp. 1925–1935, dec 2023. [Online]. Available: https://doi.org/10.1038/s41594-023-01138-w

[14] T. Aso, W. S. Lane, J. W. Conaway, and R. C. Conaway, "Elongin (siii): a multisubunit regulator of elongation by rna polymerase ii," *Science*, vol. 269, no. 5229, pp. 1439–1443, Sep 1995.

[15] P. Carmeliet, Y. Dor, J. M. Herbert, D. Fukumura, K. Brusselmans, M. Dewerchin, M. Neeman, F. Bono, R. Abramovitch, P. Maxwell, C. J. Koch, P. Ratcliffe, L. Moons, R. K. Jain, D. Collen, and E. Keshert, "Role of hif-1alpha in hypoxia-mediated apoptosis, cell proliferation and tumour angiogenesis," *Nature*, vol. 394, no. 6692, pp. 485–490, 1998.

[16] M. Ohh, Y. Takagi, T. Aso, C. E. Stebbins, N. P. Pavletich, B. Zbar, R. C. Conaway, J. W. Conaway, and W. G. J. Kaelin, "Synthetic peptides define critical contacts between elongin C, elongin B, and the von Hippel-Lindau protein," *J Clin Invest*, vol. 104, no. 11, pp. 1583–1591, 1999.