People's Democratic Republic of Algeria

The Ministry of Higher Education and Scientific Research

Faculty of Natural and Life Sciences
Department of Agronomy & Nutrition Sciences
University of Saida - Dr. Moulay Taher

Memory

Presented for obtaining the Master's degree in: Biology

Specialty: Plants Biotechnology

TITLE:

Discovery and Characterization of a set of Ligand Binding Structural Motifs in Pathogenesis-Related Proteins; A Structural Bioinformatics Study Towards Unravelling Essentials in Protein Structure-Function Relationships.

Prepared and presented by: Belkhiri Kheira

Examined by a jury commission, composed by:

❖ President: Dr. Saidi Abdelmoumen	MCA	Univ. of Saida - Dr. Moulay Taher.
❖ Examiner: Dr. Adeli Djallal Eddine	Prof.	Univ. of Saida - Dr. Moulay Taher.
❖ Supervisor: Dr. Abdelkrim Rachedi	MCA	Univ. of Saida - Dr. Moulay Taher

2024/2025

People's Democratic Republic of Algeria

The Ministry of Higher Education and Scientific Research

Faculty of Natural and Life Sciences
Department of Agronomy & Nutrition Sciences
University of Saida - Dr. Moulay Taher

Memory

Presented for obtaining the Master's degree in: Biology

Specialty: Plants Biotechnology

TITLE:

Discovery and Characterization of a set of Ligand Binding Structural Motifs in Pathogenesis-Related Proteins; A Structural Bioinformatics Study Towards Unravelling Essentials in Protein Structure-Function Relationships.

Prepared and presented by: Belkhiri Kheira

Examined by a jury commission, composed by:

❖ President: Dr. Saidi Abdelmoumen	MCA	Univ. of Saida - Dr. Moulay Taher.
❖ Examiner: Dr. Adeli Djallal Eddine	Prof.	Univ. of Saida - Dr. Moulay Taher.
❖ Supervisor: Dr. Abdelkrim Rachedi	MCA	Univ. of Saida - Dr. Moulay Taher

2024/2025

Dedications

First thanks to ALLAH who gave me the strength to follow the right path in doing such work and may peace and blessings be upon Mohammad who I felt always beside me.

This research is sincerely dedicated to the source of happiness in my whole life: my dear parents those who blessed me with their Love, patience, understanding and support. To my grandmother, may God have mercy on her, I will never forget you.

I also extend extend my most sincere to my sisters specially Chaimaa and my only brother.

Furthermore, I also dedicate this work to my nephews and nieces.

Without forgetting the distinguished professors "Dr Haji. B" "Dr Boudaa. N" who provided me with a helping hand and support, my God protect them.

And finally, to myself, the one who faces challenges with courage, Thank you.

BELKHIRI KHEIRA

Acknowledgments

Firstly, I would like to thank the Almighty God, the most powerful and merciful, Who gave me strength and patience to complete this modest work.

I would like to extend my gratitude to my supervisor Dr. Abdelkrim Rachedi, for his invaluable advice, continuous support, and patience and his guidance; I am really thankful to have acquired them.

My thanks are also to the members of the jury; examiner and the president, for accepting to examine and evaluate this work

BELKHIRI KHEIRA

Abstract

This project seeks to explore more the basis behind structure-function relationship in the biological context of macromolecules; the proteins related to plant biology as in the case of this study. Furthermore, the study draws attention to results that would touch upon distant evolutionary relations across species.

The understanding of the structure-function relationship is important for deeper studies of the biological function of proteins, and macromolecules in general, in both health and disease situations. One way to undertake this kind of study is to analyse the protein structures involved in selected biological functions and examine their ligand-binding environments.

In this project, a set of 3D structures of Pathogenesis-Related proteins (PR-proteins), representing 115 full chains from 62 Protein Databank (PDB) entries from different source organisms or species, has been studied in terms of their binding site environments and a significant number of bound ligands. The study implemented structural bioinformatics tools developed at the Department of Biology, University of Saida.

Recognized public databases relevant to protein data like the UniProt and the Protein Databank (PDB) have been used as main sources of the data. Methods including data mining, data parsing and extraction, molecular graphical software (RasMol), and a locally developed bioinformatics software, namely the Structure, Sequence and Function Server (SSFS), have been used to carry out the study in this project.

The study resulted in the identification and construction of a set of PR-protein structural and functional binding motifs that have been classified into seven (7) structural classes, including: loops only, α-helix only, β-strand only, α & β only, loops & α-helices, loops & β-strands, and loops & α-helices & β-strands. These motifs have been characterised and analysed based on their binding site residue content and properties. They are associated with a number of vital biological functions that fall under the theme of plant immunity.
The data and analysis pertaining to this project have been stored in a flat-file database to be made accessible for querying and exploration by scientists in the field.

Keywords: Pathogenesis-Related proteins, PR-proteins, Plant Immunity, Structural Bioinformatics, Structural & Functional Motifs, Ligand Binding Environment, Amino Acids, Residues, Databases, SSFS.

الملخص

يهدف هذا المشروع إلى استكشاف الأسس التي تقوم عليها العلاقة بين البنية والوظيفة في السياق البيولوجي للجزيئات الضخمة، وتحديدًا البروتينات المرتبطة بعلم النبات كما هو الحال في هذه الدراسة. بالإضافة إلى ذلك، تسلط الدراسة الضوء على نتائج قد تمس علاقات تطورية بعيدة بين الأنواع المختلفة.

تُعدّ معرفة العلاقة بين البنية والوظيفة أمرًا مهمًا لإجراء دراسات أعمق حول الوظيفة البيولوجية للبروتينات، وللمركبات الحيوية عمومًا، سواء في حالات الصحة أو المرض .ومن الطرق الممكنة لتنفيذ مثل هذه الدراسة تحليل بُنى البروتينات المرتبطة بوظائف بيولوجية مختارة، ودراسة بيئات ارتباطها بالليغاندات.

في هذا المشروع، تم دراسة مجموعة من البُنى الثلاثية الأبعاد لبروتينات مرتبطة بالممراضية (بروتينات(PR ، تمثل 115 سلسلة كاملة من 62 إدخالًا في بنك بيانات البروتين(PDB) ، تعود لمصادر أو أنواع كائنات مختلفة، من حيث بيئة مواقع الارتباط وعدد كبير من الليغاندات المرتبطة بها. وقد تم تنفيذ الدراسة باستخدام أدوات البيوانفورماتيك البنيوية المطوّرة في قسم البيولوجيا بجامعة سعيدة.

وقد استُخدمت قواعد بيانات عامة ومعترف بها في مجال البروتينات مثل UniProt وبنك بيانات البروتين (PDB) كمصادر رئيسية للبيانات. وشملت المنهجية تقنيات تنقيب البيانات، واستخراج وتحليل الملفات، واستخدام برامج رسومية جزيئية مثل RasMol، إلى جانب استخدام برنامج محلي التطوير في البيوانفورماتيك يُعرف بـ "خادم البنية والتسلسل والوظيفة.(SSFS) "

وقد أسفرت الدراسة عن تحديد وبناء مجموعة من "أنماط الارتباط البنيوية والوظيفية" الخاصة ببروتيناتPR ، والتي تم تصنيفها إلى سبع (7) فئات بنيوية: الحلقات فقط، الحلزونات α فقط، الشرائح β فقط، αو β فقط، حلقات وحلزوناتα ، حلقات وشرائحβ ، حلقات وحلزونات α وشرائح β معًا. وقد تم توصيف هذه الأنماط وتحليل محتوى وخصائص بقايا الأحماض الأمينية في مواقع الارتباط الخاصة بها. وترتبط هذه الأنماط بعدد من الوظائف البيولوجية الحيوية التي تندرج تحت موضوع المناعة النباتية.

تم تخزين البيانات والتحليلات المتعلقة بهذا المشروع في قاعدة بيانات بصيغة ملفات مسطحة، لتكون متاحة للعلماء المختصين للاستعلام والاكتشاف في المستقبل.

الكلمات المفتاحية :بروتينات مرتبطة بالممراضية، بروتيناتPR ، المناعة النباتية، المغلوماتية_الحيوية البنيوية، أنماط بنيوية ووظيفية، بيئة ارتباط الليغاند، أحماض أمينية، بقايا بروتينية، قواعد بيانات، .SSFS

Résumé

Ce projet vise à approfondir les fondements de la relation structure-fonction dans le contexte biologique des macromolécules, en particulier des protéines liées à la biologie végétale, comme c'est le cas dans cette étude. Par ailleurs, l'étude attire l'attention sur des résultats pouvant révéler des relations évolutives lointaines entre différentes espèces.
Comprendre la relation entre la structure et la fonction est essentiel pour mener des recherches plus approfondies sur la fonction biologique des protéines, et des macromolécules en général, tant en conditions de santé que de maladie. Une approche consiste à analyser les structures des protéines impliquées dans certaines fonctions biologiques sélectionnées et à examiner leurs environnements de liaison aux ligands.

Dans ce projet, un ensemble de structures tridimensionnelles de protéines liées à la pathogénèse (protéines PR), représentant 115 chaînes complètes issues de 62 entrées dans la banque de données sur les protéines (PDB), provenant de divers organismes ou espèces, a été étudié, notamment en ce qui concerne leurs environnements de sites de liaison et un nombre significatif de ligands associés. L'étude a utilisé des outils de bioinformatique structurale développés au Département de Biologie de l'Université de Saïda.

Des bases de données publiques reconnues, telles que UniProt et le Protein Databank (PDB), ont été utilisées comme principales sources de données. Les méthodes appliquées comprennent l'exploration et l'extraction de données, le traitement de fichiers structurés, l'utilisation de logiciels graphiques moléculaires (RasMol), ainsi qu'un logiciel localement développé de bioinformatique nommé Structure, Sequence and Function Server (SSFS), utilisé pour mener à bien l'étude.

L'étude a conduit à l'identification et à la construction d'un ensemble de motifs structuraux et fonctionnels de liaison des protéines PR. Ces motifs ont été classés en sept (7) catégories structurales : boucles uniquement, hélices α uniquement, brins β uniquement, hélices α et brins β uniquement, boucles & hélices α, boucles & brins β, et boucles & hélices α & brins β. Ces motifs ont été caractérisés et analysés en fonction du contenu et des propriétés de leurs résidus au niveau des sites de liaison. Ils sont associés à un certain nombre de fonctions biologiques essentielles relevant du domaine de l'immunité végétale.
Les données et analyses issues de ce projet ont été stockées dans une base de données au format fichier plat, afin d'être rendues accessibles pour les chercheurs du domaine, pour des requêtes et des explorations futures.

Mots-clés: Protéines liées à la pathogénèse, protéines PR, immunité végétale, bioinformatique structurale, motifs structuraux et fonctionnels, environnement de liaison aux ligands, acides aminés, résidus, bases de données, SSFS.

Table of Contents

Abstract

Chapitre 2: Materials and methods

Chapitre 3 : Results and Discussion

List of Abbreviations

2AN:	**8**-ANILINO-1-NAPHTHALENE SULFONATE
3D:	**T**hree-dimensional
A8I:	**3**-BENZOYLBENZOIC ACID
AA:	**A**mino acid
ACT:	ACETATE ION
ACE:	**A**CETYL GROUP
ACY:	**A**CETIC ACID
ALV:	(2S)-2-AMINOPROPANE-1,1-DIOL
BHO:	**B**ENZHYDROXAMIC ACID
CA :	**C**ALCIUM ION
CAC:	**C**ACODYLATE ION
CAPE1:	**C**AP-derived peptide 1
CK:	**C**reatine kinase
CL:	**C**HLORIDE ION
Cryo-EM:	**M**odern cryo-electron microscopy
CTPP:	**C**-terminal prepropeptide
DED:	**D**irect electron detector
DIO:	1,4-DIETHYLENE DIOXIDE
DXC:	(3ALPHA,5BETA,12ALPHA)-3,12-DIHYDROXYCHOLAN -24-OIC ACID
ECDs :	**E**xtracellular Domains
EDO :	1,2-ETHANEDIOL
EMU:	**N**-BENZYL-9H-PURIN-6-AMINE
EPE:	4-(2-HYDROXYETHYL)-1-PIPERAZINE ETHANESULFONIC ACID
ET:	**E**thylene
ER:	**E**ndoplasmic reticulum

FER:	3-(4-HYDROXY-3-METHOXYPHENYL)-2-PROPENOIC ACID
FMT:	**F**ORMIC ACID
FUC:	**A**LPHA-L-FUCOPYRANOSE
GOL:	**G**LYCEROL
GSH:	**G**LUTATHIONE
H35:	**N**-(FURAN-2-YLMETHYL)-7H-PURIN-6-AMINE
HEM:	**P**ROTOPORPHYRIN IX CONTAINING FE
HR:	**H**ypersensitive Response
HyPep:	**p**roteinase inhibitor peptide
IHP:	**I**NOSITOL HEXAKISPHOSPHATE
IMD:	**I**MIDAZOLE
JA:	**j**asmonic acid
KXN:	(2R,3S)-2-(3,4-DIHYDROXYPHENYL)-3,4-DIHYDRO -2H-CHROMENE-3,5,7-TRIOL
LDH:	**L**actate dehydrogenase
LRR:	**l**eucine rich repeat
MES:	2-(N-MORPHOLINO)-ETHANESULFONIC ACID
MG:	**M**AGNESIUM ION
MPD:	(4S)-2-METHYL-2,4-PENTANEDIOL
MSE:	**S**ELENOMETHIONINE
MLI:	**M**ALONATE ION
NAG:	2-ACETAMIDO-2-DEOXY-BETA-D-GLUCOPYRANOSE
NB:	**N**ucleotide binding
NLRs:	**N**ucleotide-binding Leucine-rich repeat immune Receptors
NMR:	**N**uclear magnetic resonance
O:	**O**XYGEN ATOM
OCS:	**C**YSTEINESULFONIC ACID
OXY:	**O**XYGEN MOLECULE
0QE:	**C**HLOROMETHANE
P4C:	**O**-ACETALDEHYDYL-HEXAETHYLENE GLYCOL
PAMPs:	**P**athogen-associated molecular patterns
PEG:	**D**I(HYDROXYETHYL)ETHER
PR:	**P**athogenesis-Related Proteins

PRRs:	**P**attern recognition receptors
PIs:	**P**roteinase inhibitors
PTI:	**P**AMP-triggered immunity
PDB:	**P**rotein Data Bank
PPI:	**P**rotein-Protein Interaction
PEO:	**H**YDROGEN PEROXIDE
ROS:	**R**eactive Oxygen Species
SAR:	**S**ystemic Acquired Resistance
SA:	**S**alicylic Acid
SSFS:	**S**equence Structure and Function Server
SIFTS:	**S**tructure Integration with Function, Taxonomy and Sequence
SO4:	**S**ULFATE ION
SIN:	**S**UCCINIC ACID
TRS:	**2**-AMINO-2-HYDROXYMETHYL-PROPANE-1,3-DIOL
TQG:	**N**-[(3-CARBOXYPHENOXY)ACETYL]-L-GLUTAMIC ACID
TLA:	**L**(+)-TARTARIC ACID
UniProt:	**U**niversal Protein Resource
UNL:	**U**NKNOWN LIGAND
ZN:	**Z**INC ION
ZIP:	**N**-(3-METHYLBUT-2-EN-1-YL)-9H-PURIN-6-AMINE
ZEA:	(2E)-2-METHYL-4-(9H-PURIN-6-YLAMINO)BUT-2-EN-1-OL

List of Figures

Chapter 1

Chapter 2

Chapter 3

List of Tables

Chapter 1

Chapter 2

Chapter 3

General Introduction

The biochemical function of a protein is generally dictated by the 3D structure of the polypeptide chain. Structural motifs refer to specific 3D arrangements or patterns of atoms, functional groups, or molecular fragments within a molecule. These motifs play a significant role in determining the physicochemical properties, biological activity, and pharmacokinetic behavior of a medicament.

The biochemical function of protein structure is governed by many factors including structure motifs which are specific regions in proteins that are associated, amongst a number of roles, such as binding substrates, cofactors and other ligands necessary for relevant biological functions. One example of importantly functional structural motifs is **helix-turn-helix** motif (Figure 1) which can bind DNA and help in gene transcription regulation and is difficult to identify from the amino acid sequence alone (EMBL-EBI, Voet and Voet, 2004).

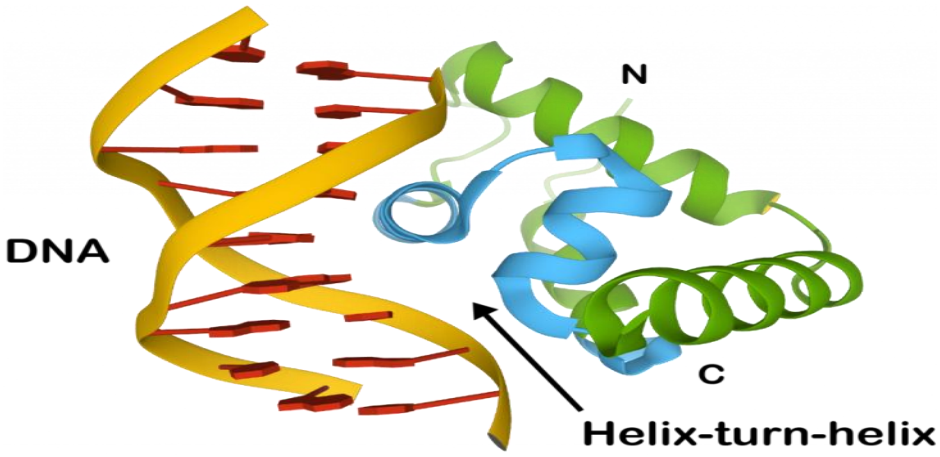


Figure1.1. The Structural motif **helix-turn-helix** (blue) as it binds to DNA (left) in the transcription factor SATB1 (right: green and blue).

The list of structural motifs identification is growing as research in the field of structural biology continues. The study undertaken in this project, looks at whether structural motifs play role in binding ligands in the different classes of Pathogenesis-Related proteins which are essential for defence processes relevant to plant immunity during pathogenic attacks. Characterising binding motifs that may be linked to the functionality of these proteins would

play major contribution into the understanding better their function and their defence mechanisms that can help in novel-drug discoveries and protection of plants including crops important as food sources.

The study undertaken in this project used 62 PR-proteins experimentally determined structures found in the Protein Databank (PDB) including 115 full protein chains. This resulted in the characterisation of a number of structural motifs that has been classified into seven classes of motifs based on the types of secondary structure elements; Loops only, α_helix only, β_strand only, α & β only, Loops & α_helices, Loops & β_strands, and Loops & α_helices, Loops & β_strands all of which binding a variety of ligands. Amino-acids residues responsible for the binding have also been arranged in non contiguous sequences (of 3D-space context) that exhibit different chemical functions giving uniqueness to each and every motif.

It should be noted that this project implements techniques that fall under the field of Structural Bioinformatics as this latter uses informatics science to study and analyze biological structures such as proteins structure motifs and their biological functions.

Recognized public databases relevant to protein data like the UniProt and the Protein Databank (PDB) have been used as main sources of the data. Methods including data mining, data parsing and extraction, molecular graphical software (RasMol), and a locally developed bioinformatics software, namely the Structure, Sequence and Function Server (SSFS), have been used to carry out the study in this project.

This project is distributed as follows:

First chapter: A literature review through generality on the protein and Structural Motifs. The chapter covers protein possible motifs, Structural motifs, Amino Acids, Porphyrins definition, Porphyrins classes and the PDB database (Protein Data Bank).

Second chapter: Describes the Materials and Methods used to study and analysis of the data pertaining to the themes of the project.

Third chapter: Results and Discussion, contains presentation of the results obtained from the structural data analysis followed by discussion of what the results may mean and indicate to. This chapter is ended with a general conclusion around the benefits of the study and future orientations.

Chapter 1

Literature review

Literature review

I. Introduction:

In general, plants are infected by diverse organisms such as fungi, oomycetes and bacteria, and by viruses. When a plant is susceptible to a virulent pathogen race, disease ensues. Plants have evolved multiple layers of defence mechanisms to confer resistance against pathogens and pathogens have evolved to supress plant immune responses.

II. Strategies of plant pathogens

Plant pathogens use diverse life strategies. Pathogenic bacteria proliferate in intercellular spaces (the apoplast) after entering through gas or water pores (stomata and hydathodes, respectively), or gain access via wounds. Nematodes and aphids feed by inserting a stylet directly into a plant cell. Fungi can directly enter plant epidermal cells, or extend hyphae on top of, between, or through plant cells. Pathogenic and symbiotic fungi and oomycetes can invaginate feeding structures (haustoria), into the host cell plasma membrane. Haustorial plasma membranes, the extracellular matrix, and host plasma membranes form an intimate interface at which the outcome of the interaction is determined. These diverse pathogen classes all deliver effector molecules (virulence factors) into the plant cell to enhance microbial fitness **[1]**.

III. Definition of an immune system

An immune system is a complex network of different cellular actions and signals, allowing an organism to defend itself against a pathogen. Simply put, as this is the case with most biological systems, the immune system is based on an exchange of input and output. In order for an immune response to occur, there must be an input. However, the immune system does this in a very unique way. As know, Plants have three layers of immunity that do partially overlap: (1) preformed barriers, **[2]** a PAMP-triggered immune response and **[3]** the effector-triggered immunity. Unlike humans, plants lack an adaptive immune system but rely solely on an innate immune system whose components are present in every cell of the plant **[2].**

III.1 How Plant Defence against Pathogens?

Plants have developed an innate immunity comprising several structural, chemical, and protein-based defenses designed to detect and stop invading organisms (microbes, pests and herbivores) **[3].**

III.1.1 Structural defence mechanisms

Is often found on the plant surface, are generally of the categories that present physical barriers to pathogen entry. The physical barriers not only protect the plant from invasion, they also give the plant strength and rigidity and exist as integral component physiological structures throughout the lifespan of the plant **[3].**

- **Cell walls:** The primary cell wall of plant cells is composed mainly of cellulose, hemicellulose, and pectin. It provides structural support and acts as a physical barrier against pathogens trying to invade the plant cells. The rigidity of the cell wall can deter pathogens that attempt to breach it.
- **Cuticle:** The cuticle is a waxy, hydrophobic layer covering the outer surface of the epidermis of leaves, stems, and other aerial plant parts. It consists of cutin and waxes and serves multiple functions:

  Water Regulation: It helps prevent excessive water loss by reducing transpiration.

  Protection: The cuticle acts as a barrier against physical damage, UV radiation, and pathogens. It makes it difficult for pathogens to penetrate the plant surface and enter plant tissues.

- **Trichomes:** Trichomes are hair-like structures found on the surface of leaves, stems, and other plant parts. They can be glandular (producing substances) or non-glandular. Trichomes serve several protective functions:

  Mechanical Defense: They physically deter herbivores by causing irritation or difficulty in movement.

  Reduced Water Loss: Some trichomes trap a layer of moisture around the plant surface, reducing transpiration.

Chemical Defense: Glandular trichomes can produce and store toxic compounds that deter herbivores or inhibit pathogen growth.

- **Bark:** In woody plants, the bark consists of the outermost layers of stems and roots. It includes the cork cambium (producing cork cells) and the outer cork layer itself. Bark provides protection against physical damage, pathogens, and environmental stresses.

Physical Barrier: It forms a tough outer layer that prevents pathogens from entering deeper tissues.

Chemical Protection: Bark may contain secondary metabolites like phenolics and terpenes that deter herbivores and pathogens.

- **Thorns and Prickles:** Thorns are modified stems while prickles are modified epidermal outgrowths. Both are sharp structures that deter herbivores and protect plants from damage.
- **Lignification:** Lignification involves the deposition of lignin, a complex polymer, in cell walls. This process occurs in response to mechanical damage or pathogen attack and contributes to:

Reinforced Cell Walls: Lignin strengthens cell walls, making them more resistant to physical damage and penetration by pathogens.

Barrier Formation: It can create physical barriers that limit the spread of pathogens within plant tissues.

- **Tyloses:** Tyloses are outgrowths of parenchyma cells that extend into xylem vessels. They can block or plug the vessel lumen in response to infection or injury, preventing the spread of pathogens through the xylem.
- **Rapid Closure Response:** Some plants exhibit rapid movements or changes in response to physical damage or pathogen attack.

III-1-2- Biochemical defence mechanisms:

- **Secondary Metabolites:** Plants produce a vast array of secondary metabolites such as phenolics (including flavonoids and tannins), terpenoids (including essential oils), alkaloids, and glucosinolates. These compounds can have antimicrobial properties, deter herbivores, and even signal defense responses.
- **Phytoalexins:** Phytoalexins are low molecular weight compounds synthesized de novo by plants in response to pathogen attack. They have antimicrobial properties and can inhibit the

growth of fungi, bacteria, and viruses. Examples include resveratrol in grapes and pisatin in peas.

- **Hypersensitive Response (HR):** HR is a rapid, localized cell death response at the site of pathogen invasion. It restricts the spread of the pathogen by sealing off infected tissues and can involve the production of reactive oxygen species (ROS) and activation of defense genes.
- **Systemic Acquired Resistance (SAR):** SAR is a defense response that occurs in uninfected parts of the plant following localized infection. It involves the systemic induction of defense-related genes and accumulation of PR proteins and phytohormones like salicylic acid (SA), which enhance resistance to subsequent infections.
- **Phytohormones:** Various phytohormones play critical roles in regulating plant immunity. Salicylic acid (SA) is particularly associated with defense against biotrophic pathogens (which require living tissue to survive), while jasmonic acid (JA) and ethylene (ET) are more involved in defense against necrotrophic pathogens (which kill host cells to obtain nutrients).
- **Cell Wall Reinforcements:** In addition to providing structural support, cell walls can be reinforced in response to pathogen attack. This reinforcement can involve the deposition of callose, lignin, and other compounds that strengthen cell walls and restrict pathogen spread.
- **Signal Transduction Pathways:** Plants have complex signaling networks that detect pathogen signals (pathogen-associated molecular patterns or PAMPs) and activate defense responses. These pathways involve receptor proteins, kinases, transcription factors, and other signaling components.
- **Pathogenesis-Related (PR) Proteins:** PR proteins are a group of proteins synthesized by plants in response to pathogen attack. They include enzymes like chitinases, glucanases, peroxidases, and protease inhibitors that degrade fungal cell walls, inhibit pathogen growth, or protect plant proteins from degradation.

III. Notion of proteins

Molecular biology is the part of biology that studies biological processes at the molecular level. It concerns itself with the processes within a cell and with interactions between cells.

The central dogma of molecular biology describes the relationship between three important classes of molecules within a cell. These three are: DNA, RNA and proteins. Fig1.2 provides a simplified view of the relationship between these three molecules.

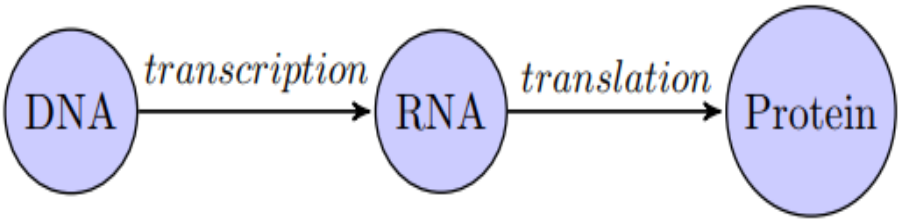


Figure1.2. The flow of information within a cell.

DNA can be seen as containing the blueprint of an organism. Proteins effectuate the realization of this blueprint; they are used as the building blocks of cells, used for signaling, used in the metabolism, and much more. RNA functions as a carrier of information: first, DNA is transcribed to RNA in the nucleus. Then, the RNA travels outside of the nucleus to the ribosomes. There, it is translated into proteins **[5].**

IV.1 What are proteins?

The word "protein" comes from the Greek word "prota", meaning "of primary importance" **[4]** ; are vital components of the biological world. They interact with other molecules through their interfaces and participate in crucial cellular processes **[6]**.

Proteins are linear polymers of amino acids that have a great variety of functional, structural and regulatory roles in organisms. Most proteins are assemblages of 20 possible standard amino acids that are directly coded for by DNA codons and some nonstandard amino acids in some organisms **[4]**.

VI.1.1 Amino acids:

Are the basic building blocks of proteins, and they serve as the nitrogenous backbones for compounds like neurotransmitters and hormones. In chemistry, an amino acid is an organic compound that contains both an amino (-NH2) and carboxylic acid (-COOH) functional group, hence the name amino acid, Fig1.3.

Figure 1.3. Amino Acid Formula Structure

Natural proteins are composed of 20 different amino acids that belong to distinctive chemical groups, see Figure 3.

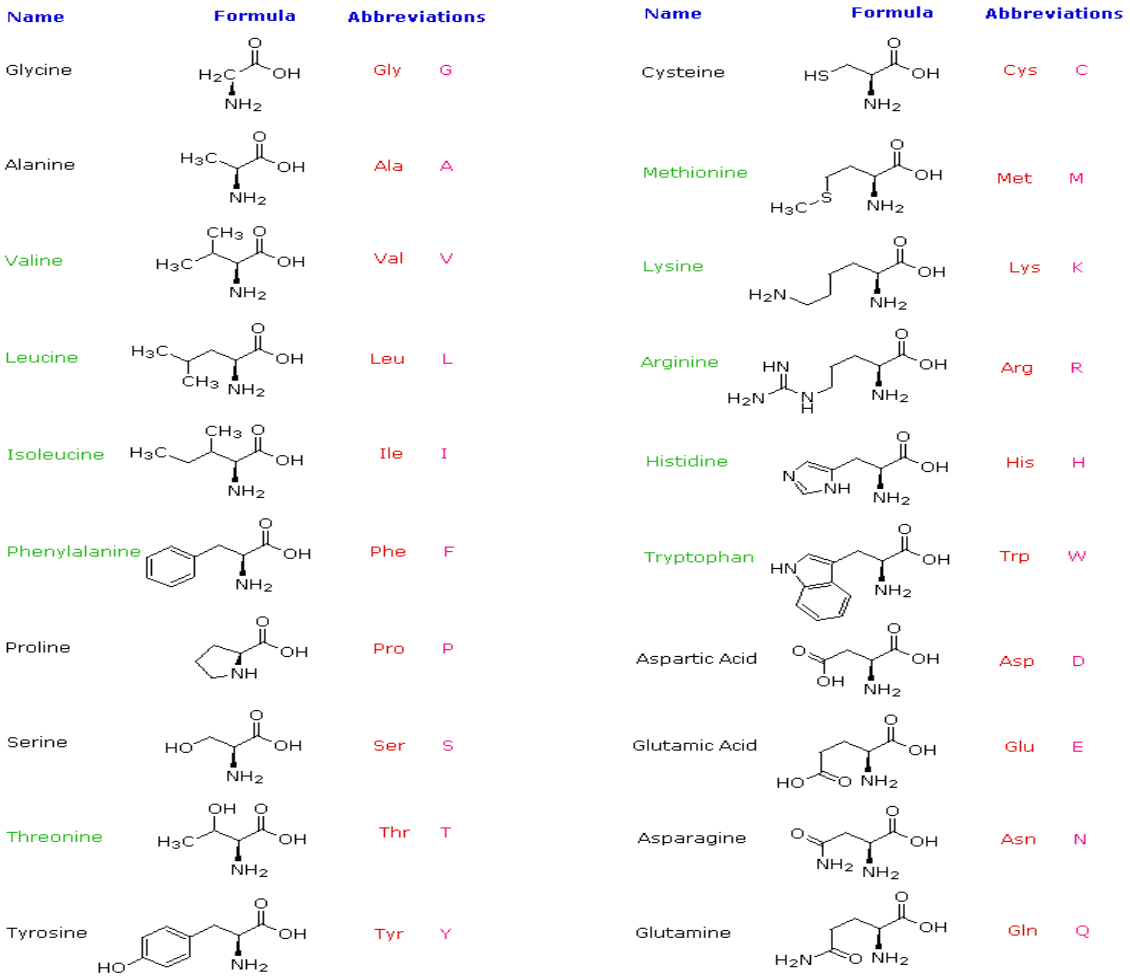


Figure 1.4. List of the 20 Amino Acids Usually Associated with Proteins Composition.

IV.2. Structural and Functional Properties of Proteins

IV.2.1. Structural Properties of Proteins

Protein structure is determined by the basic structure of the peptide bond, the sequence of its amino acids, folding of the primary chain and interactions, as well as intra-protein and inter-protein cross-links among its constituent amino acid side chains **[7]**. Proteins have different levels of structural organization; primary, secondary, tertiary and quaternary.

IV.2.1.1. Primary Structure

The primary structure is the arrangement of amino acids into a linear polypeptide chain. Here, peptide bonds and disulfide bonds only join the amino acids. The sequence of amino acids in a polypeptide chain is responsible for the functions of the proteins, which is decided by decoding the genetic codes or genes corresponding to the proteins, Fig1.5.

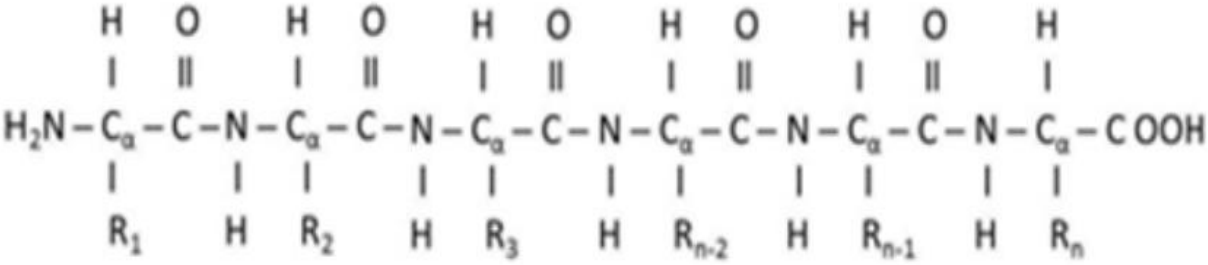


Figure 1.5. The Arrangement of Amino Acids into A Linear Polypeptide Chain

Furthermore, the partial double bond nature of the peptide bond, which is the result of the resonance structure created by the delocalization of the lone pair of electrons on the nitrogen, creates a slight positive charge on the amino group (N – Hδ+) and restricts the rotation of the CO–NH bond to a maximum torsion angle (ω) of 6°. This rigidity of the peptide bond keeps its six atoms in a single plane with mostly (1000:1) trans configurations, which reduces steric interactions of the α-carbons, except for X-proline peptide groups, which have 3:1 (*trans:cis*) ratios as shown in the figure 5. Rotation about the other bonds in the primary chain, namely, between N–Cα (Φ) and Cα–CO (Ψ), occurs freely but is affected by interactions among R groups

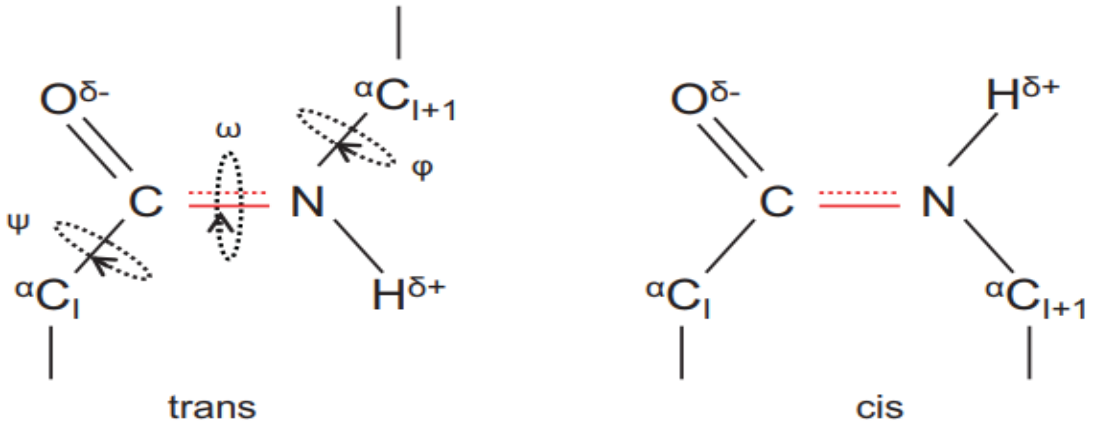


Figure 1.6. Schematic Representation of Tran-cis Orientation of Amino Acids in Polypeptide Chain

IV.2.1.2. Secondary Structure

The term **secondary structure** refers to the local conformation of some part of a polypeptide **[8]**, are formed by hydrogen bonding **[9]**. The majority of proteins fold into α-helices and β-sheets, which make up their secondary structures **[4]**.

- **The alpha helix:** It is the most common and stable conformation for a polypeptide chain. It is a spiral structure, Figure 6. The polypeptide bonds form the back-bone and the side chains of amino acids extend outward. The structure is stabilized by hydrogen bonds between NH and C=O groups of the main chain.

  Each turn is formed by 3.6 residues. The distance between each amino acid residue (translation) is 1.5 Å. The alpha helix is generally right handed. Left handed alpha helix is rare, because amino acids found in proteins are of L-variety, which exclude left handedness. Proline and hydroxyproline will not allow the formation of alpha helix [10]

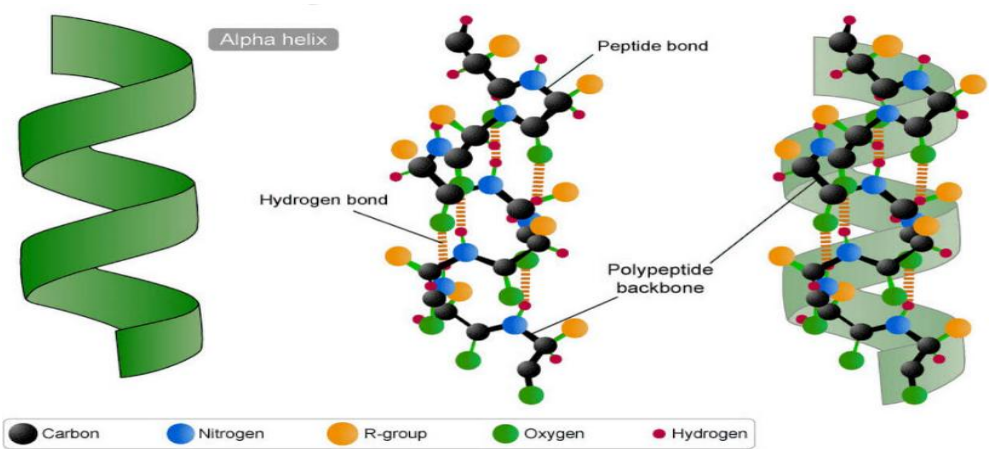


Figure 1.7. The alpha helix in secondary structure

- **Beta-Pleated Sheet:** The polypeptide chains in beta-pleated sheet is almost fully extended. The distance between adjacent amino acids is 3.5Å. It is stabilized by hydrogen bonds between NH and C=O groups of neighboring polypeptide segments. Adjacent strands in a sheet can run in the same direction with regard to the amino and carboxy terminal ends of the polypeptide chain (parallel) or in opposite direction (anti-parallel beta sheet), Figure 7. Beta- pleated sheet is the major structural motif in proteins like silk Fibroin (anti-parallel), Flavodoxin (parallel) and Carbonic anhydrase (both). Beta bends may be formed in many proteins by the abrupt U-turn folding of the chain. Intrachain disulfide bridges stabilize these bends [10].

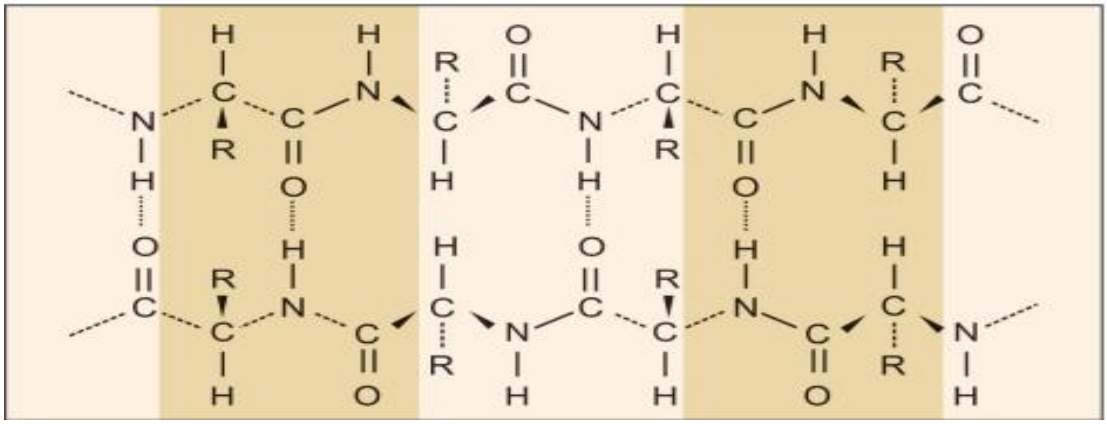


Figure 1.8. Structure of Beta-Pleated Sheet

IV.2.1.3. Tertiary Structure

Secondary structure denotes the configurational relation ship between residues which are about 3–4 amino acids apart; or secondary level defines the organization at immediate vicinity of amino acids. The tertiary structure denotes three dimensional structure of the whole protein Fig1.9. The tertiary structure defines the steric relationship of amino acids which are far apart from each other in the linear sequence, but are close in the three-dimensional aspect.

The tertiary structure is maintained by **noncovalent** interactions such as hydrophobic bonds, electrostatic bonds and van der Waals forces. The tertiary structure acquired by native protein is always thermodynamically most stable **[10]**.

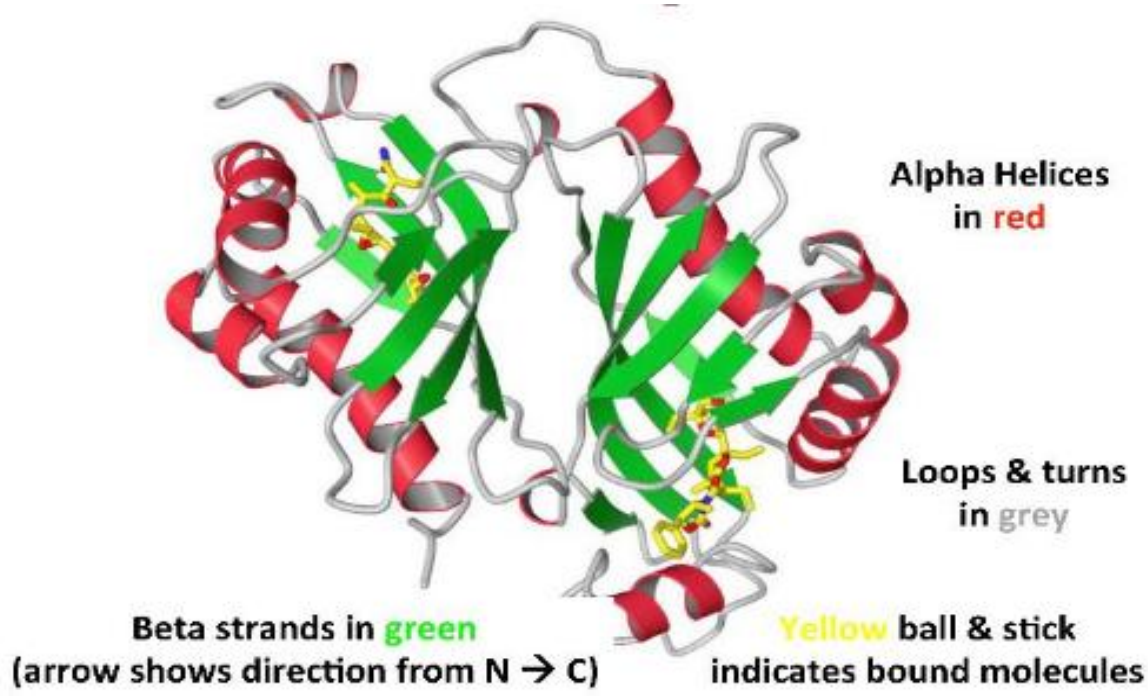


Figure 1.9. Tertiary Structure of Proteins. It is the final folding sate of a large set of Protein and Enzymes.

IV.2.1.4. Quaternary Structure

Certain polypeptides will aggregate to form one functional protein, Fig1.10. This is referred to as the quaternary structure. The protein will lose its function when the subunits are dissociated. The forces that keep the quaternary structure are hydrogen bonds, electrostatic bonds, hydrophobic bonds and van der Waals forces.

Depending on the number of polypeptide chain, the protein may be termed as monomer (1 chain), dimer (2 chains), tetramer (4 chains) and so on. Each polypeptide chain is termed as **subunit** or **monomer. Homodimer** contains two copies of the same polypeptide chain. **Heterodimer** contains two different types of polypeptides as a functional unit. For example, 2 alpha-chains and 2 beta-chains form the **hemoglobin** molecule. Similarly, 2 heavy chains and 2 light chains form one molecule of **immunoglobulin** G. Creatine kinase (CK) is a dimer. Lactate dehydrogenase (LDH) is a tetramer **[10]**.

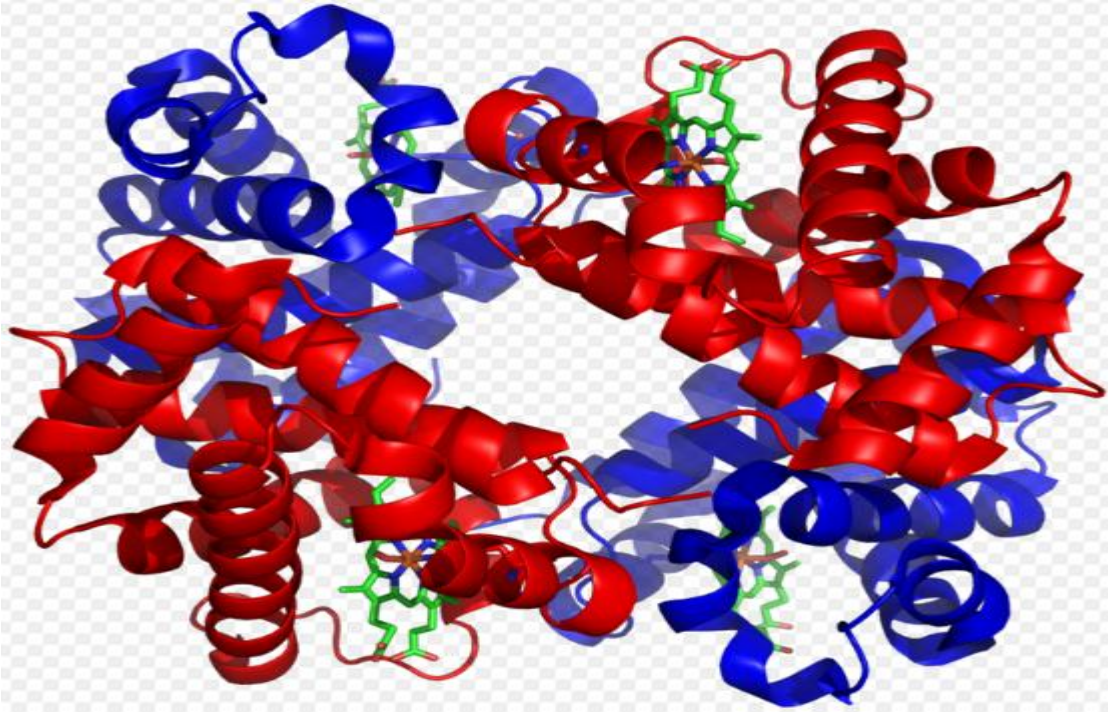


Figure 1.10. Quaternary structure of the protein of hemoglobin.

VI. The relationship between the levels of structure in proteins

The primary structure of a protein is defined as a linear sequence of amino acids and the location of disulfide (-S-S-) bonds if any. A peptide group can make two dihederal angles namely ψ (psi) and φ (phi). A polypeptide segment can acquire a regular secondary structure such as different helical structure, β-pleated sheet, β-turn, coils or loop, tertiary structure, quaternary structure **[12]**, Fig1.11.

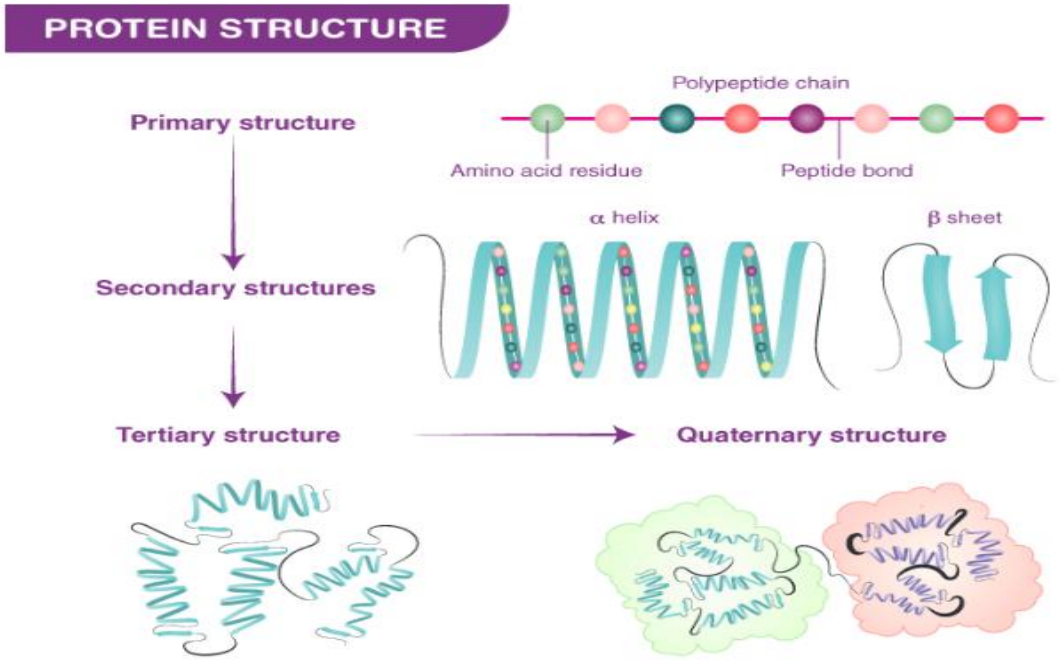


Figure 1.1 The Protein Levels of Structure

VI.1. What is the functional properties of a protein?

Proteins are the most versatile macromolecules in living systems and play important functions in essentially all biological processes. They function as catalysts, they transport and store other molecules such as oxygen, they provide mechanical support and immune protection, they generate movement, they transmit nerve impulses, and they control growth and differentiation **[12]**. Some important biological functions are outline in Table1.1, but it is to be expected that this rudimentary list of properties will expand each year as new proteins are characterized **[11]**.

Tableau1.1. The selective list of some functional roles for proteins within cells.

Function	Examples
-Enzymes or catalytic proteins	-trypsin, DNA polymerase and ligases
-Contractil proteins	-actin, myosin, tubulin, dynein
-Structural or cytoskeletal proteins	-Tropocollagen, keratin
-Transport proteins	-Hemoglobin, myoglobin, serumalbumin, ceruloplasmin, transthyretin
-Effector proteins	-insulin, epidermal grouth factor,thyroid stimulating hormone
-Deffence proteins	-Ricin, immunoglobins, venoms a,d toxins, thrombin
-Electron transfer proteind	-cytochrome oxidase, bacterial photosynthetic, reaction centre, plastocyanin, ferredoxin
-Receptors	-CD4, acetycholine receptor
-Repressor proteins	-Jun, Fos, Cro
-Chaperones(accessory folding proteins)	-GroEL, DnaK
-Storage proteins	-Ferritin, gliadin

VI.2. Structure-function Relationship

The structural and functional properties of proteins make them useful for various industrial applications. For example, proteins can be made into elastic materials that can survive repeated stress-strain cycles, and they can aggregate to form films that provide barriers to gases, moisture and bacteria. Films produced from protein could be used in food packaging, in paper coatings and in bandaging materials. Proteinbased materials have become a research focus because of their high performance, low cost and environmentally friendly characteristics **[4]**.

VII. An overview of plant immunity

From a conceptual point of view plant immunity can be divided into cell-surface immunity and intracellular immunity. The cell-surface immunity can be seen as a first line of defense. If a

pathogen is in close proximity of the plant cell and a signature is recognized, an immune response will be elicited. If the pathogen overcomes this first-line of defense and manages to penetrate a cell, the intracellular immunity comes to play.

While the cell-surface immunity is achieved via cell-surface receptors, whereby pathogen signatures are recognized by the extracellular domains (ECDs), activation of the intracellular immunity is triggered, if pathogen signatures are recognised inside of the cell. Hence, cell-surface immunity is mediated through pathogen-recognition receptors, referred to as PRRs. The cell's intrinsic immunity is mediated via NLRs (nucleotide-binding leucine-rich repeat immune receptors), Fig1.12. In other words, intracellular immunity receptors called NLRs detect signatures of better adapted pathogens. These signatures are translocated proteins, aimed at modulating the metabolism or the physiological response of the host, maximizing the pathogenicity **[13]**.

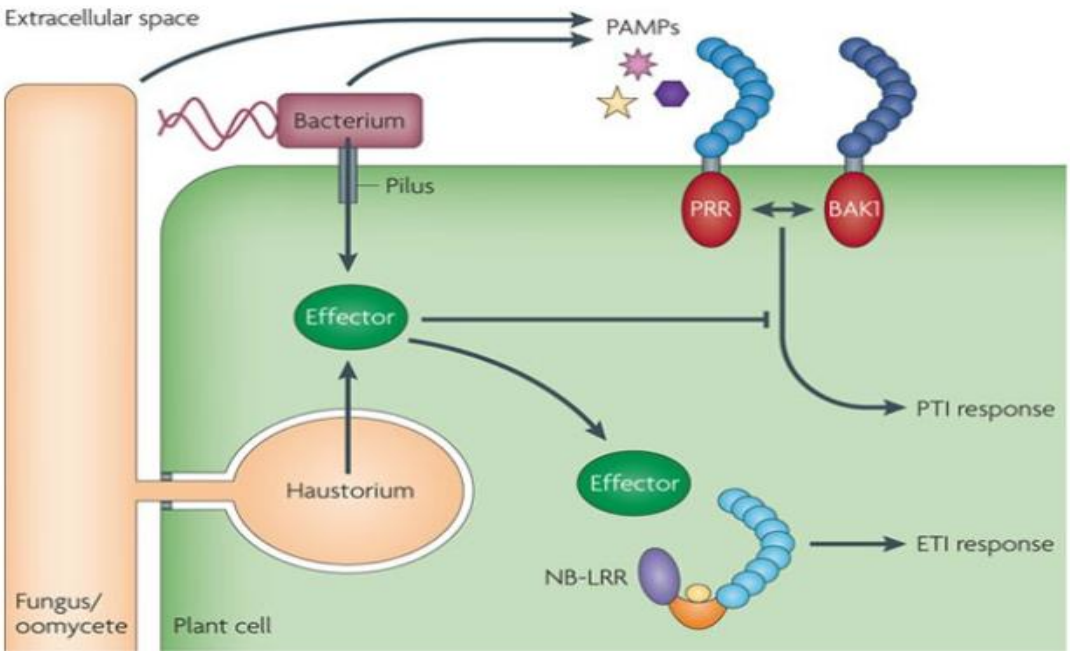


Figure 1.13 The Principals of Plant Immunity.

VII.1. What is Pathogenesis-Related (PR) Proteins?

In the last 30 years, the definition of pathogenesis-related proteins has been constantly reviewed . Today, the authors have reached a consensus defining PRs **[15]** Pathogenesis related proteins are defined as proteins encoded by the host

plant induced under pathological or related conditions. Pathological conditions refer to all types of pathogen stresses and not just to resistant hypersensitive responses in which the PR's are more common. Apart from other pathogens they also include parasitic attack by nematodes, insects and herbivores. In addition, necrosis caused by abiotic stress and toxin related chlorosis also induce PR Proteins **[14]**.

Plant reactions to these factors are very complex, and involve the activation of set of genes, encoding different proteins. how These stresses can induce biochemical and physiological changes in plants. Among these, production and accumulation of pathogenesis related proteins in plants in response to invading pathogen and/or stress situation is very important. Phytoalexins are mainly produced by healthy cells adjacent to localized damaged and necrotic cells, but PR proteins accumulate locally in the infected and surrounding tissues, and also in remote uninfected tissues. Production of PR proteins in the uninfected parts of plants can prevent the affected plants from further infection PR protein in the plants was first discovered and reported in tobacco plants infected by tobacco mosaic virus Later, these proteins were found in many plants. Most PR proteins in the plant species are acid-soluble, low molecular weight, and protease-resistant proteins. PR proteins depending on their isoelectric points may be acidic or basic proteins but they have similar functions. Most acidic PR proteins are located in the intercellular spaces, whereas, basic PR proteins are predominantly located in the vacuole, Currently PR-proteins were categorized into 17 families according to their properties and functions (Table1.2) **[16]**.

Tableau1. 2. Classification of pathogenesis related proteins.

Family	Domain classification	Proteins	Functions
PR-1	*IPR034111* *IPR001283*	PR-1 a, PR-1 b, and PR-1 c	Antifungal (CAP)
PR-2	(GH17)	β-1,3-Glucanases	Cleaves β-1,3-glucans
PR-3	*IPR016283*	Chitinase types I, II, IV, V, VI, and VII	Endochitinase
PR-4	*IPR001153*	Barwin domain chitinase I/II	Antifungal and chitinase
PR-5	*IPR001938*	Thaumatin-like	Antifungal
PR-6	*IPR000864*	Potato protease I	Proteinase inhibitor
PR-7	(Subtilisin-like)	Tomato endoproteinase P69 (O82007)	Endoproteinase
PR-8	(GH18)	Cucumber chitinase	Chitinase III
PR-9	(Haem peroxidase III)	Tobacco lignin-forming peroxidase (P11965)	Peroxidase
PR-10	*IPR024949* *IPR000916*	Parsley "PR-1"	Ribonuclease-like
PR-11	(GH18)	Tobacco chitinase V (Q43576)	Chitinase
PR-12	*IPR008176*	Radish Rs-AFP3 (O24332)	Plant Defensin
PR-13	*IPR001010*	Arabidopsis THI2.1 (Q42596)	Thionin
PR-14	*IPR000528*	Lipid transfer proteins	Shuttling of phospholipids and fatty acids
PR-15	*IPR001929*	Barley OxOa (P45850)	germin; Oxalate oxidase
PR-16	*IPR001929*	Barley OxOLP (O49871)	germin-like
PR-17	*IPR007541*	Tobacco NtPRp27 (Q9XIY9)	late blight resistance

VII.2. Classification of Pathogenesis Related Proteins

The PR's have been classified into various families based on the shared sequence homology. In addition, the isoelectric points of the PR's are considered as an important parameter for sub group of classification. PR's can also be grouped into different classes based on the migration in the native PAGE, reaction with specific antisera and mRNA probes. PR's have also been

classified based on the biological activity of the induced defense proteins. Seventeen different groups of PR's have been identified so far based on the above-mentioned properties **[14]**.

VII.3. Functional characterization and mode of action

Many PR-proteins have been shown to possess antifungal, antibacterial, antiviral and antinematode properties **[13]**. Different PR proteins have a distinct mode of action against the pathogen depending upon the type of pathogen and the activities of the majority of these protein families are known or can be inferred. PR-1 protein, one of the dominant groups of PRs induced by the pathogen, inhibits pathogen growth by binding and sequestration of sterols from the pathogen. Moreover, the programmed cell death is also inhibited by PR1 upon pathogen infection by releasing a defense signal peptide CAPE1 (CAP-derived peptide 1). Some PR proteins function as hydrolytic enzymes, *viz.* the PR-2 (endo-β-1,3-glucanases) and PR-3, -4, -8 and - 11 (endochitinases.

They function as antifungal proteins by catalyzing hydrolytic cleavage of major components of fungal and oomycete cell wall, i.e. β-1,3-glucan (by the breakdown of β-1,3-glucosidic linkages) or chitin (by the breakdown of internal β-1,4-glycoside bonds) respectively, resulting in the breakdown of the fungal cell wall. Different isoforms of glucanases and chitinases are produced depending upon the plant-pathogen interaction.

Thaumatin-like proteins or Osmotin-like proteins such as PR5 inhibit hyphal growth and spore germination by producing transmembrane pores leading to fungal cell leakiness and blocking the function of plasma membrane receptors molecules involved in cAMP/RAS2 signaling pathways. Also, antifungal action has been demonstrated in some family members, predominantly against oomycetes. PR-5 was also demonstrated to exhibit potato cell's defense against *Phytophthora infestans* by forming a cytoplasmic aggregation through an actin-binding complex. Proteinase inhibitors (PIs) such as trypsin inhibitors and serine inhibitors) belonging to PR6 family proteins, implicated in broad-spectrum defense activity, including suppressing pathogenic nematodes, insects and other herbivores, fungi and bacteria. PIs can provide defense against pathogens, decreasing the lyase activity essential for fungal pathogenicity, inhibiting the viral replication cycle and restricting the digestive enzyme activity of nematodes and insects, limiting amino acid release]. In addition, HyPep (proteinase inhibitor peptide) also causes cell

aggregation and pseudo-mycelia development by inhibiting amylase and serine proteinases. Also, PIs can block chitin synthesis in fungal cell walls by inhibiting endogenous trypsin that is essential for chitin synthase, thus inhibiting fungal growth and development.

PR-7 is a major protein that has only been examined in tomatoes as an endoproteinase. It is an antifungal auxiliary protein that aids in destroying fungal cell wall proteins, chitinases, and glucanases. The PR-9 family of peroxidases is believed to have a role in plant cell wall strengthening by facilitating lignin deposition in response to microbial invasion. In susceptible wheat varieties, the transcription level of PR9 is considerably reduced after infestation with the aphid-transmitted fusarium virus and hessian flies. This showed that PR9 catalyzes lignin deposition to protect susceptible cultivars from BPH.

The members of PR10 protein families exhibit ribonuclease activity required to inhibit the growth of pathogenic fungi. The antifungal activity of ribonucleases develops due to penetration of the pathogen and the destruction of cellular RNAs due to phosphorylation of PR10. It further leads to plant cell death at the inoculation site, causing apoptosis and the hypersensitivity reaction **[18]**. These intracellular PRs may be active against viruses due to their ribonuclease activity, although their ability to cleave viral RNA has yet to be shown.

The PR-12 type defensins, PR-13 type thionins, and PR-14 type lipid transfer proteins show antifungal and antibacterial activity, interacting with the target microorganism's biological membrane, leading to altered membrane permeability. Plant defensins are divided into two groups based on the structure of their precursor proteins: class I and class II. Class I defensins have endoplasmic reticulum (ER) signaling sequences along with defensin domains. In contrast, class II defensins contain an additional domain of 27–33 amino acid residues called C-terminal prepropeptide (CTPP) **[15]**. Due to a lack of signal sequences, class I defensins do not undergo post-translational modification or subcellular targeting. They accumulate in the cell wall and extracellular space directly upon synthesis through the secretory pathway. However, class II defensins undergo proteolysis in the vesicles due to CTPP signal peptides targeting vesicles and releasing mature short peptides. Mature defensins consist of five segments of non-conserved loops, linking α-helices and β-strands to form high-level structures. Differences in the loop

sequences confer different functions, including inhibition of protein synthesis, antimicrobial activity, heavy metal tolerance, plant development, and blocking of ion channels.

Oxalate oxidases (PR-15 family) and oxalate-oxidase-like proteins (PR-16 family) play an important role in plant defense. These are essential enzymes to produce reactive oxygen species (ROS) during apoplastic oxidative burst **[19]**. ROS are produced in the apoplast by an enzyme that produces H2O2 and CO2 when it reacts with oxalic acid. Proteolytic enzymes of the PR17 family play an important role in defense against fungi and viruses.

VII.4. PR-proteins activation as a defense response

Plant cells have evolved to activate and recruit the cellular machinery in response to various stresses to optimally utilize resources and sustain life. Accordingly, plants modulate genes' expression, activating a wide range of plant protectants and defense genes. The pathogenesis-related (PR) protein activation and production are crucial in response to an invading pathogen **[17]**.

There are, in essence, two branches of the plant immune system. One uses transmembrane pattern recognition receptors (PRRs) that respond to slowly evolving microbial- or pathogen-associated molecular patterns (MAMPS or PAMPs), such as flagellin proteins. The second acts largely inside the cell, using the polymorphic NB-LRR protein products encoded by most Rgenes. They are named after their characteristic nucleotide binding (NB) and leucine rich repeat (LRR) domains. NB-LRR proteins are broadly related to animal CATERPILLER/NOD/NLR proteins7and STAND ATPases8. Pathogen effectors from diverse kingdoms are recognized by NB-LRR proteins, and activate similar defence responses. NBLRR-mediated disease resistance is effective against pathogens that can grow only on living host tissue (obligate biotrophs), or hemibiotrophic pathogens, but not against pathogens that kill host tissue during colonization (necrotrophs)

Our current view of the plant immune system can be represented as a four phased 'zigzag' model, Fig 1.13, in which we introduce severa important abbreviations. In phase 1, PAMPs (or MAMPs) are recognized by PRRs, resulting in PAMP-triggered immunity (PTI) that can halt further colonization.

In phase 2, successful pathogens deploy effectors that contribute to pathogen virulence. Effectors can interfere with PTI. This results in effector-triggered susceptibility (ETS). In phase 3, a given effector is 'specifically recognized' by one of the NB-LRR proteins, resulting in effector-triggered immunity (ETI). Recognition is either indirect, or through direct NB-LRR recognition of an effector. ETI is an accelerated and amplified PTI response, resulting in disease resistance and, usually, a hypersensitive cell death response (HR) at the infection site. In phase 4, natural selection drives pathogens to avoid ETI either by shedding or diversifying the recognized effector gene, or by acquiring additional effectors that suppress ETI. Natural selection results in new R specificities so that ETI can be triggered again. Below, we review each phase in turn, we update the experimental validation of the 'guard hypothesis', and we consider future challenges in understanding and manipulating the plant immune system. We will not discuss the small RNA-based plant immune system active against viruses10or the active response of plants to herbivores **[1]**.

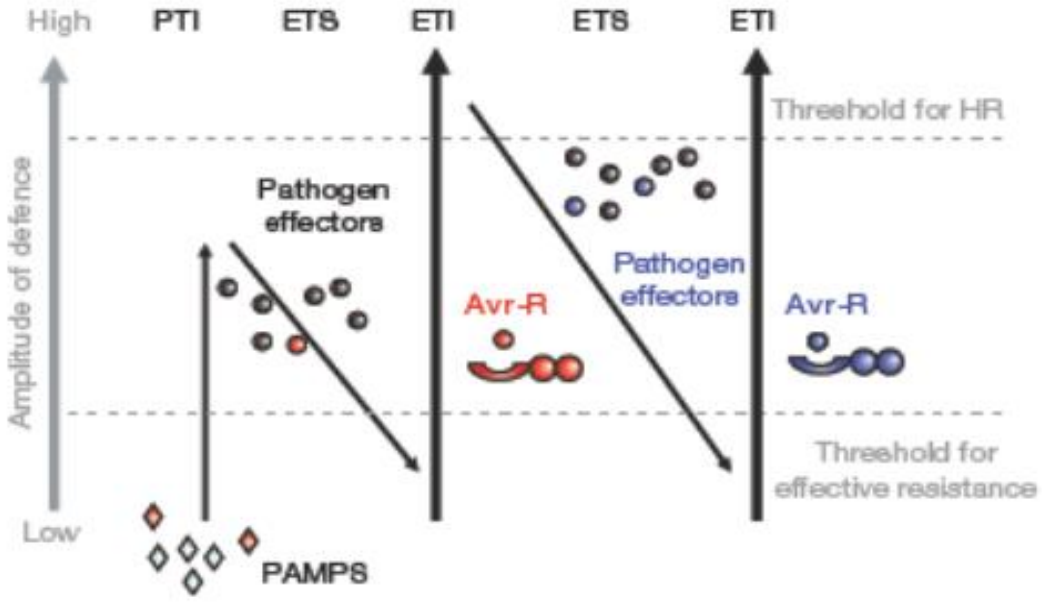


Figure 1.13. A zigzag Model Illustrates the Quantitative Output of Plant Immune System

VII.5. Biochemical characteristics. Cellular and tissue localization.

PR-proteins are distinguished by specific biochemical properties. They are low-molecular proteins (6-43 kDa), extractable and stable at low pH (< 3), thermostable, and highly resistant to proteases.

PRs have dual cellular localization – vacuolar and apoplastic, the apoplast being the main site of their accumulation. Apart from being present in the primary and secondary cell walls of infected plants, PR-proteins are also found in cell wall appositions (papillae) deposited at the inner side of cell wall in response to fungal attack. Interestingly, they are detected in the cell walls of invading fungal pathogens and in the space formed between cell walls and invaginated plasma membrane of fungi.

Acidic and basic PR-proteins are identified, each of these counterparts having both apoplastic and vacuolar localisation. Earlier data show that acidic tobacco PR-1 are localized in the apoplast, whereas basic tobacco PR-1 accumulate in the vacuole. This may be valid for one PR family (PR-1) in a host plant, such as tobacco, but cannot be generalized as a differential localization feature of acidic and basic proteins in plants.

Presently, PRs are established in all plant organs – leaves, stems, roots, flowers, being particularly abundant in the leaves, where they can amount to 5-10% of total leaf proteins. Thus, the original claiming that PRs occurrence is limited to photosynthetizing tissues has been abolished. The application of sensitive immunological techniques allowed the detection of PR-proteins in roots of tobacco and tomato plants inoculated with the fungal pathogens *Chalara elegans* and *Fusarium oxysporum*, respectively, as well as in lupine and birch roots exposed to abiotic stress.

In the leaves PR-proteins are present in mesophyll and epidermal tissues. They are also localized in the abscission zone of leaves and inflorescence, abscission zone at the stem-petiole junction, and vascular tissue of stems and petioles. In inflorescences PR-proteins are detected in sepals, pedicels, anthers, pistils, stigmata and ovaries. In seeds of maize, sorghum, oat, barley, and wheat a group of PR-proteins is established,

commonly named permatins, characterized as PR-5 thaumatin-like proteins. Linusitin from flax seeds is referred to the same group. Noteworthy, specific cell types, such as cultured plant cells, are highly active in PRs expression.

VIII. Proteins and Structural Motifs

Proteins are linear chains of covalently connected molecules called amino acids. Their sequences are encoded in DNA segments called genes. Protein is a complex organic molecule that plays numerous critical roles in living organisms. They are composed of one or more chains of amino acids linked together by peptide bonds. Proteins are involved in various functions such as structural support, transport of molecules, catalyzing biochemical reactions, and regulation of gene expression, among others.(Alberts B, Johnson A, Lewis J, Raff, M., Roberts, K., & Walter, P.2014).

VIII.1. Protein Sequence Motifs

Protein motifs refer to small, conserved amino acid sequences or patterns within a larger protein sequence that have specific functions. These motifs can be structural, functional, or both, and they play a crucial role in protein-protein interactions, protein localization, and enzymatic activity (Eddy SR .1998).

Some common examples of protein motifs include the the helix-turn-helix motif, zinc finger motif, and the leucine zipper motif. These motifs are often found in DNA-binding proteins involved in DNA transcription regulation through the interaction with specific sequences (Keskin O, Nussinov R. 2007).

VIII.1.1. Helix-turn-helix motif:

This motif consists of two alpha helices separated by a turn composed by few amino acids. The motif is commonly found in DNA-binding proteins, Fig 1.14, where the two alpha helices fit into the major groove of DNA to form a protein-DNA complex. Examples of proteins containing the helix-turn-helix motif include the Lac Repressor protein and the homeodomain proteins (Aravind L, Anantharaman V, Balaji S, Babu MM, Iyer LM .2005).

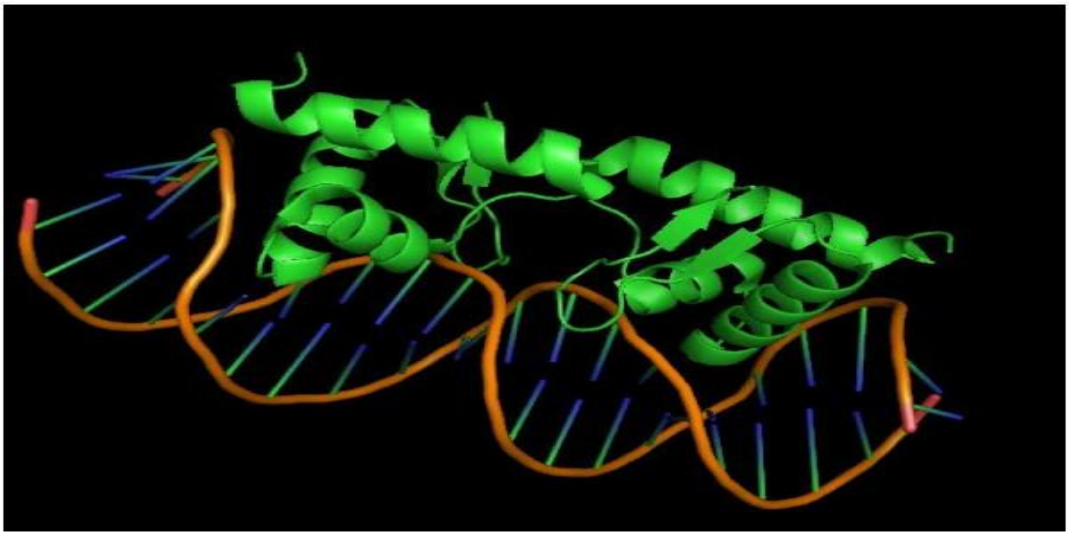


Figure 1.14. A typical winged Helix-turn-helix motif (PDB: 3JSO)

VIII.1.2. Zinc finger motif:

This motif involves the coordination of a zinc ion by several cysteine and histidine residues, forming a finger-like structure. Zinc finger motifs, Fig 1.15, are found in a variety of DNA-binding proteins, RNA-binding proteins, and enzymes. Examples of proteins containing zinc finger motifs include the steroid hormone receptor and the transcription factor TFIIIA (Laity JH, Lee BM, Wright PE .2001).

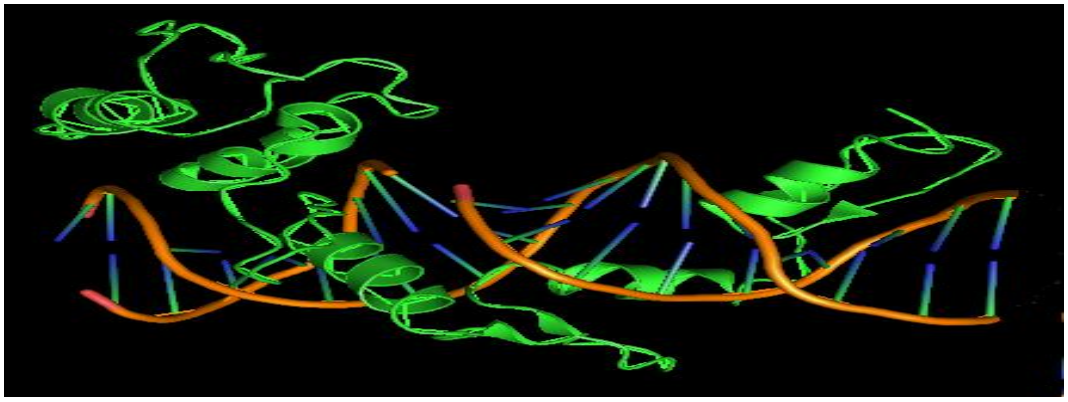


Figure 1.15. A typical Zinc finger motif

VIII.1.3 Coiled-coil motif:

This motif is formed by two or more alpha helices winding around each other to form a supercoil. The coiled-coil motif, Figure 15,. is commonly found in proteins involved in cell signaling and structural proteins such as myosin. Examples of proteins containing the coiled-coil motif include the transcription factor c-Jun and the protein keratin (Lupas AN, Gruber M 2005).

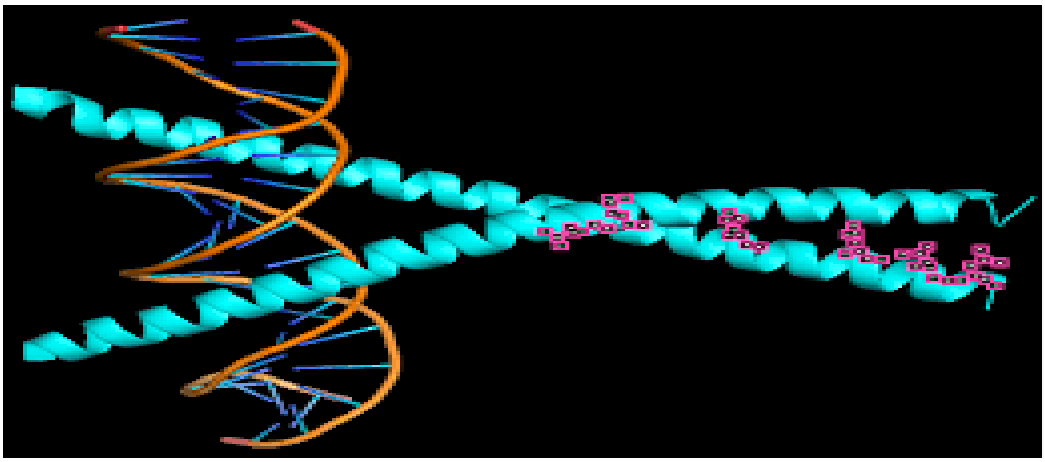


Figure 1.16. A typical Coiled-coil motif.

VIII.2. Protein Structural Motifs:

Structural motifs are recurring three-dimensional arrangements of amino acid residues or nucleotides that play important roles in the folding, stability, and function of proteins and nucleic acids.

There are three types of structural motifs:

VIII.2.1. α-Helices: α-helices are right-handed coils of amino acid residues stabilized by intramolecular hydrogen bonds, Fig 1.17. They are one of the most common structural motifs found usually associated with the secondary-structure level of proteins and are important in protein-protein interactions, DNA binding, and membrane-spanning domains (Richardson, J. S. 1981).

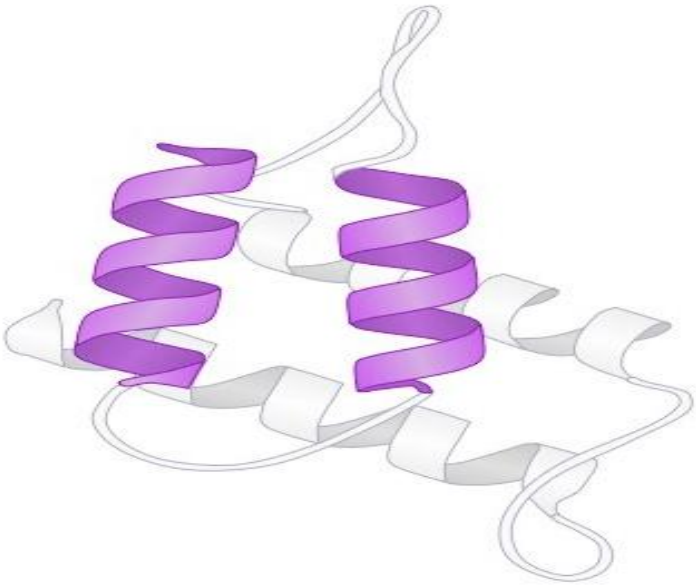


Figure 1.17. Secondary Structure of Alpha-Helix.

VIII.2.2. β-Sheets: β-sheets are planar arrangements of amino acid residues stabilized by hydrogen bonds between adjacent strands, Fig1.18. They are important in protein-protein interactions, enzymatic catalysis, and as part of the core of many proteins (Janin, J., 1997).

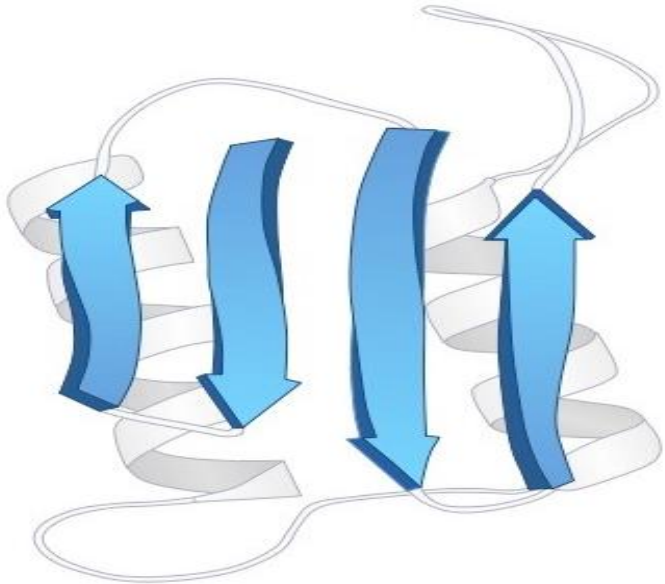


Figure1.18. Secondary Structure of β-Sheets

VIII.2.3. Loops:

Loops are flexible regions of a protein that connect the secondary structural elements (helices and sheets), Fig1.19. They often form functional sites, including active sites for enzymatic catalysis and ligand-binding sites (Chothia, C. & Lesk, A. M., 1986).

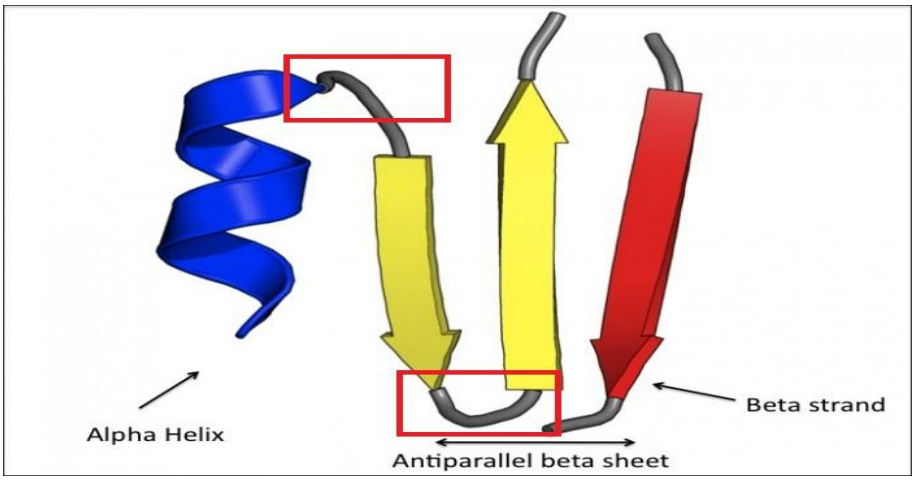


Figure 1.19. Secondary Structure of Loops

IX. Methods for experimentally determining protein structures

This field includes many techniques, and it can be determined by either experimental methods or computational methods, each method is characterized by certain controls, in this part, the various methods will be described briefly.

X-ray crystallography and NMR spectroscopy and cryo-electron microscopy are currently the three major experimental techniques for protein structure determination all of them are, however, time- and manpower-consuming, the most used technique is X-ray crystallography with over 89% of structures deposited in Protein Data Bank, followed by NMR over 8%, and Cryo-EM about 2,5%, and the remaining 0,5 structures were solved using other methods (Mutharasappan N et al., 2020)

IX.1. X-Ray Crystallography

X-Ray crystallography is currently the most favoured technique for structure determination of proteins and biological macromolecules. To perform protein crystallography, a reliable source of protein must be available, together with a purification/concentration protocol that will yield high quality, homogeneous, soluble material.

Principles

The aim of *x*-ray crystallography is to obtain a three dimensional molecular structure from a crystal. A purified sample at high concentration is crystallized and the crystals are exposed to an *x*-ray beam. The resulting diffraction patterns can then be processed, initially to yield information about the crystal packing symmetry and the size of the repeating unit that forms the crystal. This is obtained from the pattern of the diffraction spots. The intensities of the spots can be used to determine the "structure factors" from which a map of the electron density can be calculated. Various methods can be used to improve the quality of this map until it is of sufficient clarity to permit the building of the molecular structure using the protein sequence. The resulting structure is then refined to fit the map more accurately and to adopt a thermodynamically favoured conformation **[19]**.

IX.2. Nuclear Magnetic Resonance Spectroscopy of Proteins

Nuclear magnetic resonance (NMR) spectroscopy enables the determination of three-dimensional protein structures at atomic resolution under near-physiological conditions in solution. In structural biology, NMR complements X-ray crystallography, which provides similar information on proteins in single crystals.

Principles

A NMR structure determination involves special techniques of sample preparation, NMR measurements, assignment of the NMR lines to individual atoms in the polypeptide chain, collection of conformational constraints, and structure calculation and refinement. In present practice the sequence of steps usually corresponds to the standard protocol of Fig 1.20, which

has over the years made use of ever more refined techniques for the individual steps of the procedure **[20]**.

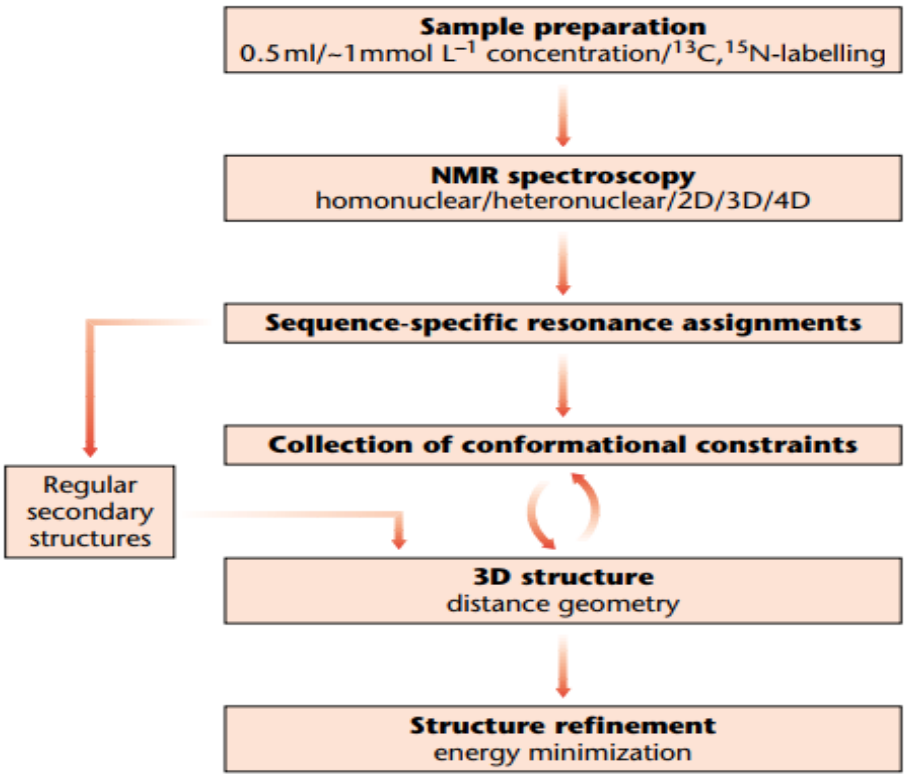


Figure 1.2 Diagram outlining the course of a protein structure determination by NMR in solution.

IX.3. Cryo-electron microscopy

Modern cryo-electron microscopy (cryo-EM) began with the introduction of a unique specimen preparation method by Dubochet and coworkers in the 1980s, resulting in the preservation of biological specimens at near native condition within a thin amorphous ice film, which allowed direct observation on a low dose transmission electron microscope operating at liquid nitrogen temperature or below. It has become one of the most powerful tools to solve the structure of biomolecules at near-atomic resolution, accompanied by X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy. With the cryo-EM method, which does not require three-dimensional (3D) crystals, organic macromolecules can be observed directly in multiple conformations in their native environment.

Recent advances of cryo-EM are provided by two major innovations. One is the employment of direct electron detector (DED) for electron microscopy. DED can detect electrons directly and read them at high frame rate without a mechanical shutter. Their higher performance is due to greatly improved quantum efficiency as compared to previous generations of detectors. Motion correction has become the standard to compensate for the blurring effect of stage drift and beam-induced movement. The other is advancement of image processing methods and the constant increase of microprocessor performance, which allow accurate classification of hundreds of thousands of EM images with computationally expensive algorithms. These two technologies have led to a "resolution revolution" with atomic structures no longer being the exclusive prerogative of x-ray crystallography or NMR spectroscopy, and cryo-EM becomes a tool to be able to analyze the structure of dynamical biomolecules **[21]**.

X. Use of Bioinformatics Tools in Different Spheres of Life Sciences

X.1. What is bioinformatics?

Bioinformatics is an interdisciplinary science, emerged by the combination of various other disciplines like biology, mathematics, computer science, and statistics, to develop methods for storage, retrieval and analyses of biological data. Paulien Hogeweg, a Dutch system-biologist, was the first person who used the term "Bioinformatics" in 1970, referring to the use of information technology for studying biological systems. The launch of user friendly interactive automated modeling along with the creation of SWISS-MODEL server around 18 years ago resulted in massive growth of this discipline. Since then, it has become an essential part of biological sciences to process biological data at a much faster rate with the databases and informatics working at the backend **[22]**.

X.2. DNA and Protein Sequence Databases

As we have mentioned, the role of databases in bioinformatics is of major importance. Nowadays, there exist three most prominent databases of sequences, American GenBank (Benson et al., 2009), European EMBL (European Molecular Biology Laboratory Data) (Stoesser *et al.*, 1998) and Japanese DDBJ (DNA Data Bank of Japan) (Tateno *et al.*, 2002), which are not moderated (meaning that anybody well-founded can add a sequence into it).

Besides these databases, there exist several other (mostly protein) databases. In our experiments we used the Swiss-Prot (Bairoch et al., 2004), a moderated database of proteins. The Swiss-Prot together with TrEMBL (Translated EMBL) and PIR (Protein Information Resource) constitute the UniProt database (Wu et al., 2006), which serves as a central repository of protein sequences and their functions **[25]**.

X.3. Databases used in this work

X.3.1. UniProt

The UniProt databases exist to support biological and biomedical research by providing a complete compendium of all known protein sequence data linked to a summary of the experimentally verified, or computationally predicted, functional information about that protein. The UniProt Knowledgebase (UniProtKB) combines reviewed UniProtKB/Swiss-Prot entries, to which data have been added by our expert biocuration team, with the unreviewed UniProtKB/TrEMBL entries that are annotated by automated systems. The UniRef databases cluster sequence sets at various levels of sequence identity and the UniProt Archive (UniParc) delivers a complete set of known sequences, including historical obsolete sequences. UniProt additionally integrates, interprets, and standardizes data from multiple selected resources to add biological knowledge and associated metadata to protein records and acts as a central hub from which users can link out to 180 other resources.

Therefore, the aim of the UniProt Knowledgebase is to provide users with a comprehensive, high-quality and freely accessible set of protein sequences annotated with functional information **[23]**.

X.3.2. Protein Data Bank (PDB)

It is the largest protein database that contains only experimentally resolved structures and is submitted by biologists and biochemists from around the world. used by various persons like scientists, bioinformatics, biologists and biochemists who use structure or sequence databases. Here we will learn protein databases (PDB), which have all known 3D structures of proteins. To find the PDB on the web search online by exploiting the webserver: https://www.rcsb.org/

One can download the protein structure data (i.e, the PDB files) you need in your studies to your computer. On the PDB homepage type the keyword/name of your protein in the search bar, and in the left you can select the organism to filter out the proteins associated. Each PDB file has a special name and coordinates of all the atoms in the protein (x;y;z). Every structure in the PDB is given a four-character alphanumeric identification known as PDB ID (the PDB identifier), e.g. “1MVS”.

X.3.3. PDB Members

The PDB is overseen by an organization called the Worldwide Protein Data Bank, wwPDB, and it has four members RCSB PDB, PDBe (Europe), PDBj (Japan), and BMRB (USA), to ensure the PDB archive is global and uniform. The RCSB PDB presently acts as the "archive keeper", The archive is updated once a week and it is distributed by wwPDB sites via FTP.

X.3.3.1. Protein Data Bank (RCSB)

Provides freely and publicly searching available to the global community for macromolecular structural data, Ligands, sequence-structure comparisons, 3D shapes of proteins, and nucleic acids. And also provides a renewed view of the molecule of the month, a feature for the phone called RCSB PDB Mobile, other educational resources at PDB-101, and more.

X.3.3.2, Protein Data Bank Europe (PDBe)

The Protein Data Bank Europe (PDBe http://www.ebi.ac.uk/pdbe/), is an important resource for high-quality molecular structures and related data and is also rich in information about all entries, It has many activities as an initiative "Structure Integration with Function, Taxonomy and Sequence" (SIFTS), and provides advanced visualization such Atlas pages, and validation of NMR and EM structures tools for bioinformaticians, and different advanced services such as PDBe PISA, PDBe Fold, PDBe Motif.

X.3.3.3. Protein Data Bank Japan (PDBj)

The Protein Data Bank Japan (PDBj, http://pdbj.org), provides a comprehensive range of services and tools for evaluating protein structures and functions, accepts and executes PDB entries mostly from Asia and Oceania, and browses in multiple languages, including Japanese, Chinese, and Korean. And more **[24]**.

Chapter 2

Materials and methods

Materials and methods

Introduction:

In order to realize the structural study of this project, the binding structural motifs that maybe found in the Pathogenesis-related (PR) proteins, structural bioinformatics methods employed required the steps explained in the following:

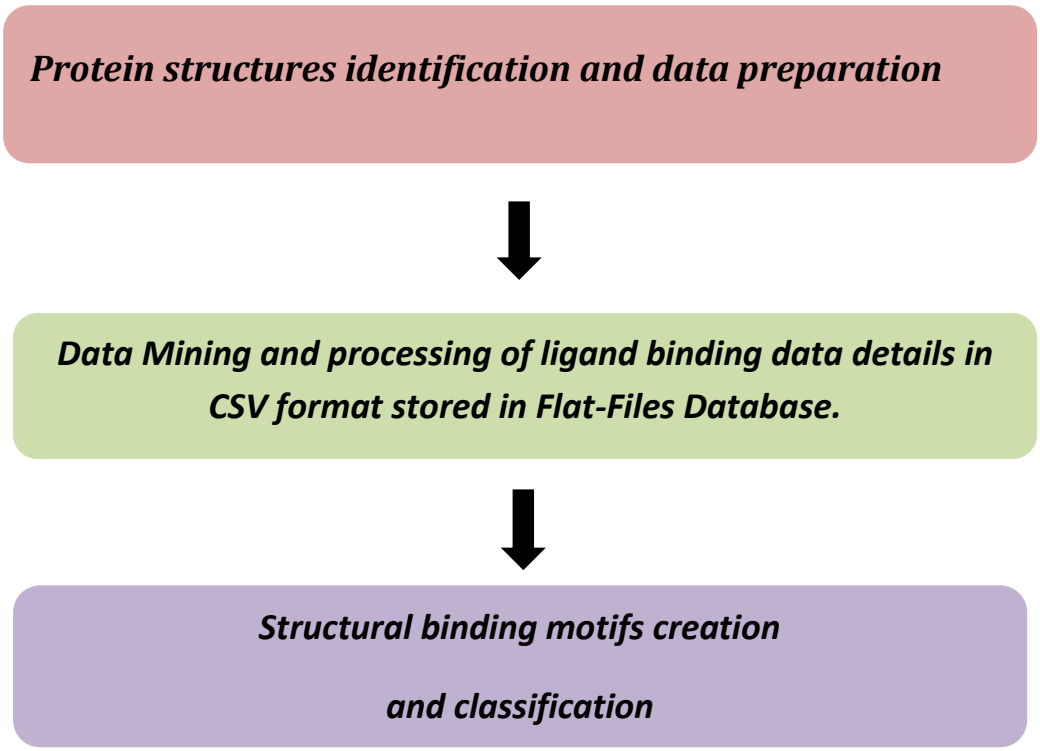


Figure 2.1. Steps of protein structures identification and Data Preparation.

As we explained in the previous chapter, the PDB is the database which provides Structural data for the selected PR proteins. The PDB assigns a unique identifier to each entry, called a PDB ID, which is used to reference the structure in scientific literature and other databases. The PDB ID consists of four characters, and it is assigned when a new structure is deposited in the database.

Data preparation for the proteins motifs structures may involve extracting relevant information such as the protein name, PDB ID, amino acid sequence, and PR protein coordination geometry. This data is extracted and binding motifs calculated using the SSFS tools (see previous chapter).

I.1.Protein structures (PDB entries):

The PDB entries used in this project amount to 62 structures each with its own PDB id all of which are X-ray Crystallography determined and of high resolution.

The X-ray determined structures are relatively less complicated in format and easily accessible for researchers to parse and extract relevant data from. The Table 1 below represents the list of PR protein classes and their protein name, PDB id.

Tableau 2.1 : The classes of PR Proteins, their PDB entries and the Ligands maybe bound to.

PR	Name	Uniprot	PDB	Method
PR-1	PR-1 a, PR-1 b, and PR-1 c *IPR001283*	Q9H4G4 · GAPR1_HUMAN	1SMB 4AIW 5VHG	X-ray X-ray X-ray
		Q9LL85 · WHY1_SOLTU	1L3A	X-ray
		P12670 · NP24_SOLLC	2I0W	X-ray
		O04298 · DAU1_DAUCA	2WQL	X-ray
		Q9FH97 · EPS1_ARATH	6WAO 6WCS	X-ray X-ray
		P48060 · GLIP1_HUMAN	3Q2R 3Q2U	X-ray X-ray
		P26987 · SAM22_SOYBN	2K7H	NMR
		P47032 · PRY1_YEAST	5JYS	X-ray
		P04284 · PR06_SOLLC	1CFE	NMR
PR-2	β-1,3-Glucanases (GH17)	P15737 · E13B_HORVU	1GHS	X-ray
		Q7DNA1 · CHI2_ORYSJ	2DKV 3IWR	X-ray X-ray
		P29022 · CHIA_MAIZE	4MCK	X-ray

PR-3	Chitinase types I, II, IV, V, VI, and VII *IPR016283*	Q5NTA4 · CHI4_CRYJA	5H7T	X-ray
		Q9FRV0 · CHIC_SECCE	4DWX 4DYG 4J0L	X-ray X-ray X-ray
		Q949H3 · CHI1_HEVBR	4MST	X-ray
		P85084 · CHIT_CARPA	3CQL	X-ray
		P23951 · CHI2_HORVU	1CNS 2BAA	X-ray X-ray
		Q8GUD7 · CHIL_HEVBR	4MPI	X-ray
PR-4	Barwin domain chitinase I/II *IPR001153*	P28814 · BARW_HORVU	1BW3 1BW4	NMR NMR
PR-5	Thaumatin-like *IPR001938*	P81370 · TLP_ACTDE	4BCT	X-ray
		P12670 · NP24_SOLLC	2I0W	X-ray
		P25871 · OLPA_TOBAC	1AUN	X-ray
		P33679 · ZEAM_MAIZE	1DU5	X-ray
		P14170 · OSMO_TOBAC	1PCV	X-ray
		Q9FSG7 · TP1A_MALDO	3ZS3	X-ray
		P81295 · PRR3_JUNAS	7P20	X-ray
PR-6	Potato protease I *IPR000864*	P09229 · CYT1_ORYSJ	1EQK	NMR
		P04517 · POLG_TEV	1LVB 1LVM 1Q31 6SUQ	X-ray X-ray X-ray X-ray
		P37842 · CYTM_SOLTU	2W9P 2W9Q 4LZI	X-ray X-ray X-ray
PR-7	Tomato endoproteinase P69 (O82007) (Subtilisin-like)	O82777 · SBT3_SOLLC	3I6S 3I74	X-ray X-ray
		Q8L5C6 · XIP1_WHEAT	1OM0 1TA3 1TE1	X-ray X-ray X-ray

PR-8	Cucumber chitinase (GH18)	P48827 · CHI42_TRIHA	6EPB 6YLJ 6YN4 7AKQ	X-ray X-ray X-ray X-ray
		Q12713 · CHI33_TRIHA	7ZY9 7ZYA	X-ray X-ray
PR-9	(Haem peroxidase III)	Q39034 · PER59_ARATH	1QGJ	X-ray
		P00433 · PER1A_ARMRU	1ATJ 1GW2 1GWO 1GWT 1GWU 1GX2 1H55 1H57	
			1H58 1H5A 1H5C 1H5D 1H5E 1H5F 1H5G 1H5H 1H5I 1H5J 1H5K 1H5L 1H5M 1HCH 1KZM	
PR-10	IPR024949	O04298 · DAU1_DAUCA	2WQL	X-ray (PR-1)
		P43211 · MAL11_MALDO	5MMU	NMR
		P93330 · NOD13_MEDTR	4GY9 4JHG 4JHH 4JHI 7QB6	X-ray X-ray X-ray X-ray X-ray
		Q256S2 · FRA1E_FRAAN	4C9C 4C9I	X-ray X-ray
		P52778 · L18A_LUPLU	1ICX 4RYV 4Y31 5C9Y	X-ray X-ray X-ray X-ray

		P43185 · BEV1L_BETPN	1FM4	X-ray
		Q9LLQ2 · P102B_LUPLU	2QIM 3E85 5MXB 5MX W	X-ray X-ray X-ray X-ray
		D0E0C7 · FRA13_FRAAN	4C94	X-ray
		P15494 · BEV1A_BETPN	1BV1 1FSK 1LLT 1QMR	
			4A80 4A81 4A83 4A84 4A85 4A86 4A87 4A88 4A8G 4B9R	
			4BK6 4BK7 4BKC 4BKD 4BTZ 4MNS 4QIP 4Z3L 7N0U 7N0V	
		D0E0C6 · FRA12_FRAAN	5AM W 6ST8 6ST9 6STA 6STB	
		P26987 · SAM22_SOYBN	2K7H	NMR
		Q9LLQ3 · P102A_LUPLU	1XDF	X-ray
		P43183 · BEV1J_BETPN	4A8U 4A8V	X-ray X-ray
		P49372 · ALL1_APIGR	2BK0	X-ray
		C0HK49 · DEF_NICSU	5KK4 5VYP	X-ray X-ray
		P30230 · DEF2_RAPSA	2N2R	NMR

PR-12	*IPR008176*	Q8GTM0 · DEF_NICAL	4AAZ 4AB0 4CQK 6B55	X-ray X-ray X-ray X-ray
		P69241 · DEF1_RAPSA	1AYJ	NMR
		P20158 · DEF1_WHEAT	1GPS	NMR
		P32026 · DEF_TOBAC	1MR4	NMR
		P81930 · DEF2_PEA	6NOM	NMR
		A4L7R7 · DEF1_PINSY	5NCE	NMR
		B3F051 · DEF_LENCC	2LJ7	NMR
		P81929 · DEF1_PEA	1JKZ	NMR
		A4L7R8 · DEF2_PINSY	7LNS	NMR
		P0C8Y5 · DEF1_HEUSA	2N2Q	NMR
		P20230 · DEF1_HORVU	1GPT	NMR

II. Extraction and process data in a CSV format and data Minning:

In order to study the porphyrin structural motifs and their relationship with their function in the 63 selected PR proteins, the bioinformatics tool SSFS (Sequence Structure and Function Server) was used to extract the bound Ligands existing in every one of these proteins.

The SSFS is a bioinformatics tool developed here at the Department of Biology by Dr. Abdelkrim Rachedi (Golovin et. al., 2005). It has been used to carry out the calculation of the binding motifs and ligand motifs environment details, the url address to access and use the tool is: https://bioinformatics.univ-saida.dz/ssfs/.

Figure 2.2 . Capture of the interface of the PPI tool.

II.1.Binding Details and download data in a CSV file:

For the extracting and processing hemeprotein PDB ID data in CSV format involves several steps. Here is a general outline of the process, Fig .2.2.

First step we insert the PDB ID in the interface of the site click Go in this case for a example insert the PDB ID of the cytochrome P450 “2CPP”, After that we click on ligand binding that

lead us to a new page we gonna choose and click on explore environment of the protein and click on it and download the CSV file

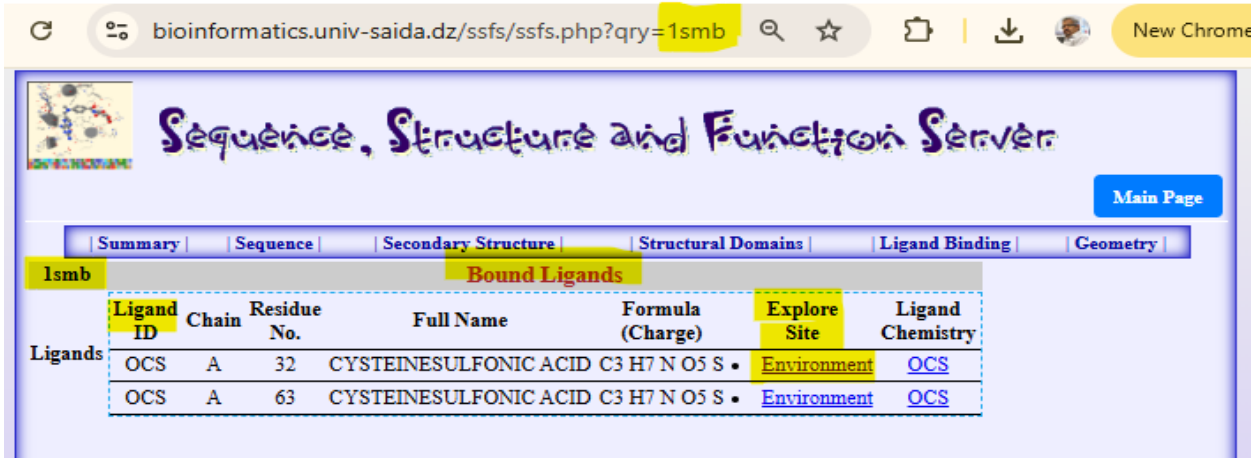


Figure 2.3: Capture of the following steps for the data extraction.

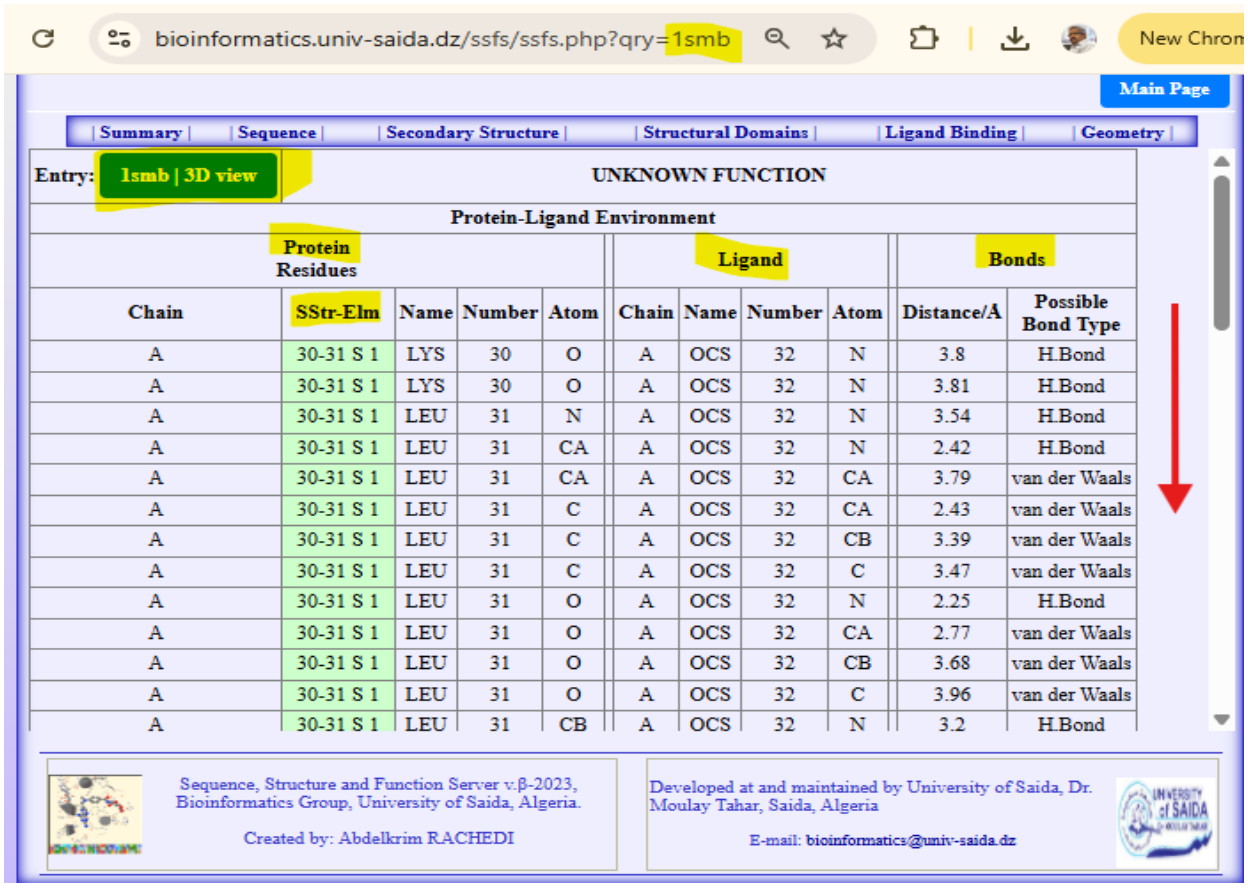


Figure 2.4: Capture of the table of the PR Proteins classes, their PDB entries, and the Ligands maybe bound to.

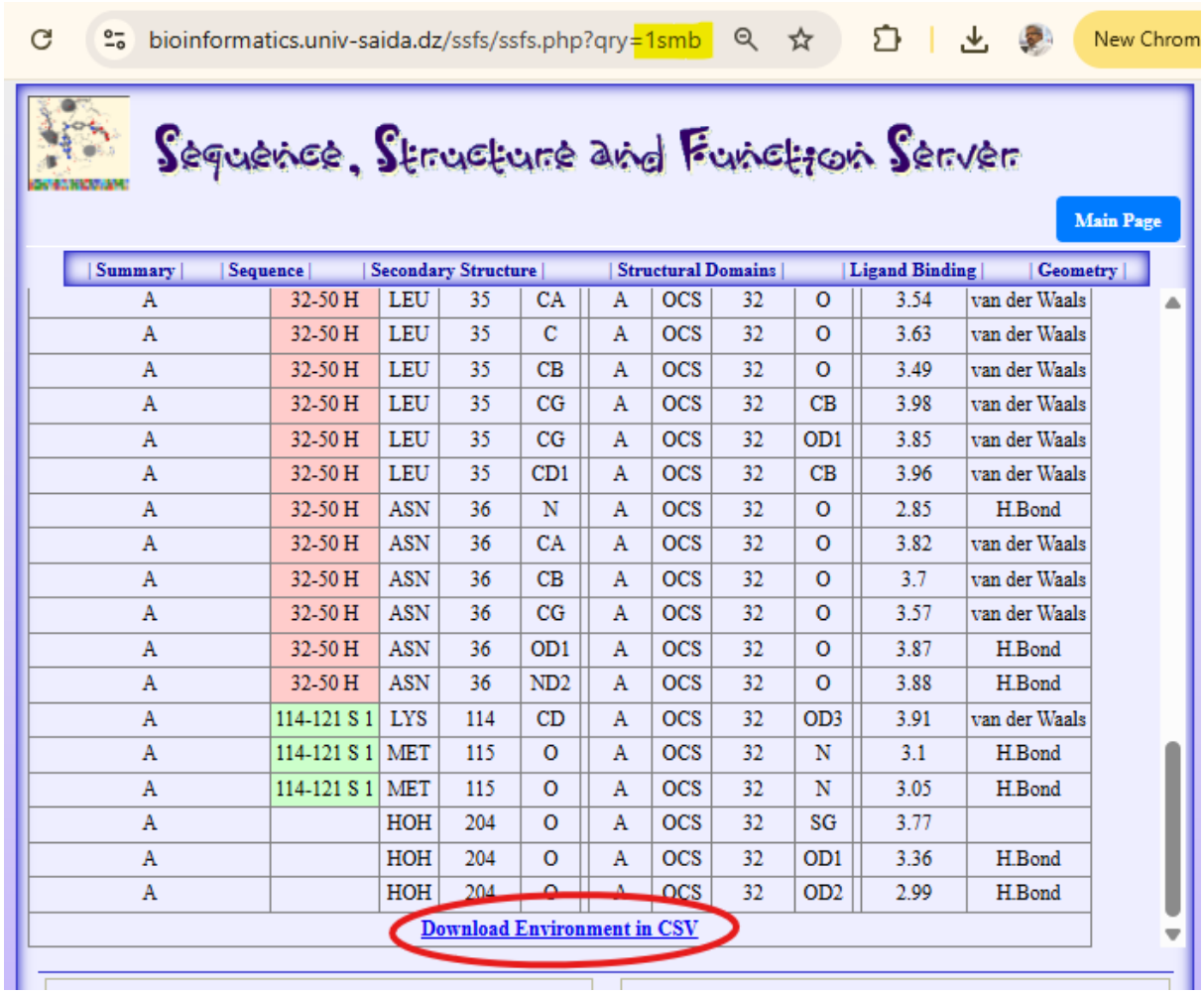
The ligand binding details shown in the above Fig.2.4 is organized in the following columns:

- The columns under the title ”Protein residues”: These columns show the atoms of the enzyme residues (AA) that bind with the ligand. The residues are also denoted in terms of what secondary elements (**α-helix, β-sheet or loop**) they may belong to.
- The columns under the title “Ligand”: These columns show the atoms of the ligand HEM, its number and the ligand id.
- The columns under the title “Ligand”: These columns show the distance between atoms (**Å: Angstroms**) and the possible bonds which can for example be a Hydrogen or Van der Waals bonds.. etc

Using the SSFS system, the binding environment details of all ligand associated with the 63 PDB PR proteins have been calculated, Parse the PDB file using a scripting language like php to extract the relevant data.

II.2. Data of the ligand binding details:

Once the SSFS find the binding environment of the ligands, it provides a link that enables downloading of the binding data, Fig.2.5. The downloaded files, in csv format, are automatically given suitable name that allows for storing them in a Flat-Files database, refer to the database schema show further below. The text editor, the Notepad++, is used to explore and verify the accuracy of the data.



Summary	Sequence	Secondary Structure			Structural Domains				Ligand Binding	Geometry
A	32-50 H	LEU	35	CA	A	OCS	32	O	3.54	van der Waals
A	32-50 H	LEU	35	C	A	OCS	32	O	3.63	van der Waals
A	32-50 H	LEU	35	CB	A	OCS	32	O	3.49	van der Waals
A	32-50 H	LEU	35	CG	A	OCS	32	CB	3.98	van der Waals
A	32-50 H	LEU	35	CG	A	OCS	32	OD1	3.85	van der Waals
A	32-50 H	LEU	35	CD1	A	OCS	32	CB	3.96	van der Waals
A	32-50 H	ASN	36	N	A	OCS	32	O	2.85	H.Bond
A	32-50 H	ASN	36	CA	A	OCS	32	O	3.82	van der Waals
A	32-50 H	ASN	36	CB	A	OCS	32	O	3.7	van der Waals
A	32-50 H	ASN	36	CG	A	OCS	32	O	3.57	van der Waals
A	32-50 H	ASN	36	OD1	A	OCS	32	O	3.87	H.Bond
A	32-50 H	ASN	36	ND2	A	OCS	32	O	3.88	H.Bond
A	114-121 S 1	LYS	114	CD	A	OCS	32	OD3	3.91	van der Waals
A	114-121 S 1	MET	115	O	A	OCS	32	N	3.1	H.Bond
A	114-121 S 1	MET	115	O	A	OCS	32	N	3.05	H.Bond
A		HOH	204	O	A	OCS	32	SG	3.77	
A		HOH	204	O	A	OCS	32	OD1	3.36	H.Bond
A		HOH	204	O	A	OCS	32	OD2	2.99	H.Bond

Download Environment in CSV

Figure 2.5: Binding data output. Highlighted in red the link for downloading the data. In this case, the link would download the csv file, named 1smb_OCSA32.csv

which contain the binding details of the protein residues with the **OCS** ligand and PDB id: 1SMB, chain A.

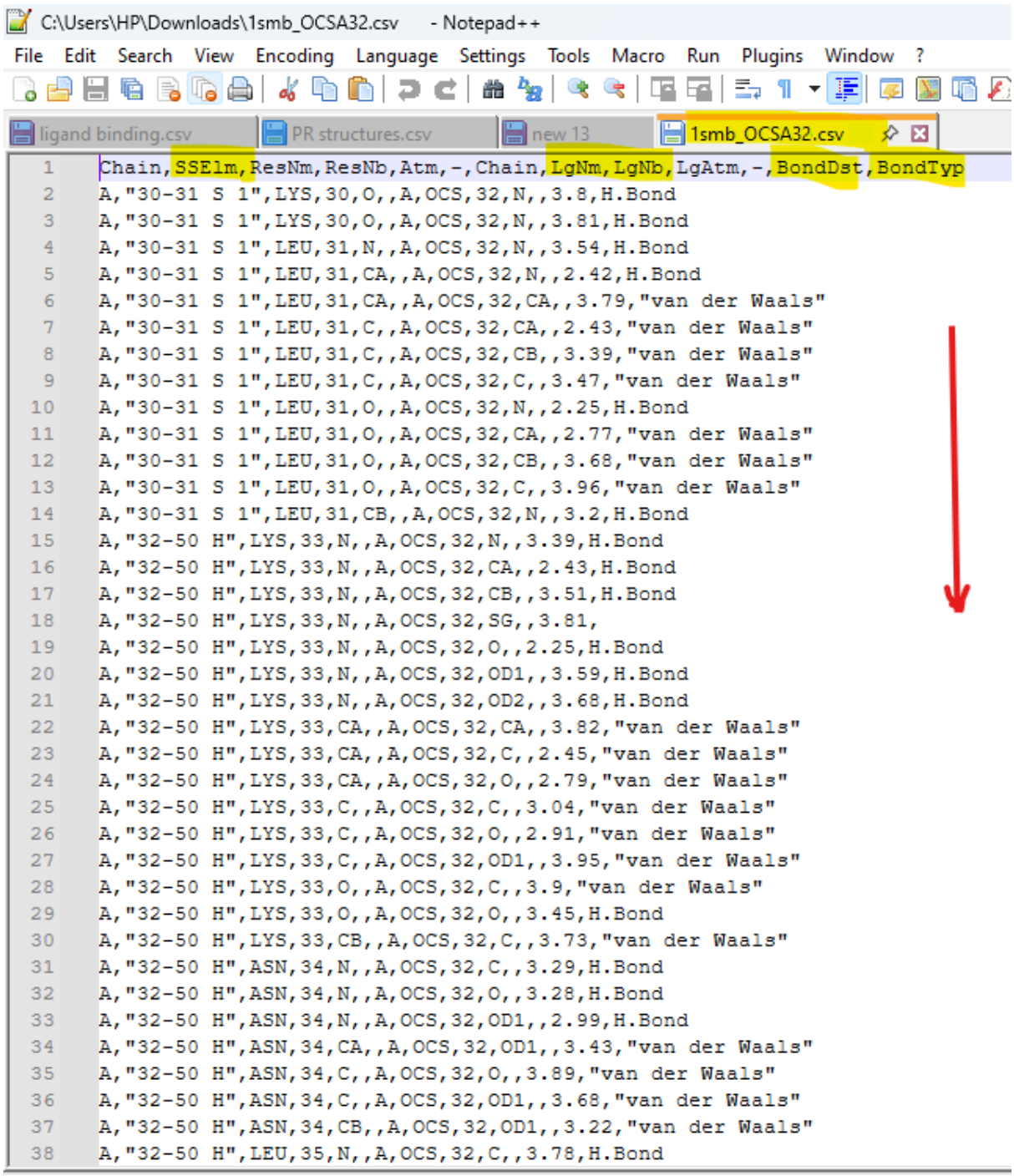


Figure 2.6. Notepad ++ example of a csv file, named 1smb_OCSA32.csv

which contain the binding details of the protein residues with the **OCS** ligand and PDB id: 1SMB, chain A.

II.3. Binding Motifs Construction and Representation:

In the binding details presented earlier, the residues that interact with the ligand are associated with certain secondary structure elements. Fig2.6, found above, shows the association of the protein binding residues with specific regions that represent secondary structure elements. The annotation for this is illustrated by the following example, which pertains to the ligand OCS and its association with the PDB ID 1SMB, chain A:

- **The protein region labeled "32-50 H"** corresponds to the secondary structure alpha-helix, designated by the symbol H.
- **The protein region labeled "114-121 S:1"** corresponds to the secondary structure beta-strand, designated by the symbol S.
- **In other outputs, a protein region labeled as "No SSE"** indicates the absence of secondary structure, implying that the binding residues belong to a loop region and are denoted by the symbol (**.**).

Thus, in our case of the PR structure 1SMB, the pattern representing the binding site of the ligand OCS 63 is: **HH..SS.**

These patterns have been observed for all ligand binding sites and appear to be associated with specific functions that recur consistently (see chapter 3). They can be further annotated as structural and functional motifs, also known as binding motifs.

II.3. Graphical Represntation of the Motifs :

In order to do a graphical representation of the motifs, the Rasmol molecular graphics program has been used to generate graphical representations of macromolecules such as proteins, DNA, and RNA.

Rasmol scripts were written to generate graphical representation scenes for the motifs in complex with their bound ligands, see Fig 2.6 as an example. These graphical representations can be used to analyze the structure of these molecules and to better understand their function

Rasmol can display various types of secondary structure motifs, such as alpha-helices, beta-strands, and loop regions, in a variety of graphical representations.

Screen-captures (images) where saved in PNG or JPEG for selected scenes which where also saved in images files with relevant names, for each and every motifs, and stored in the Flat-Database.

Alpha-helix representation(H): The alpha-helix is a common structural motif in proteins, consisting of a helical arrangement of amino acids. In Rasmol, alpha-helices can be represented as cylinders or ribbons. The cylinder representation shows the helix as a tube-like structure, while the ribbon representation shows the helix as a twisted ribbon. To display an alpha-helix in Rasmol, the user can select the helix region and use the "cartoon" command to choose between the cylinder or ribbon representation. For example, the command "cartoon cylinder" will display the helix as a cylinder show as **Red ribbons**.

Beta-strand representation(S): Beta-strands are another common structural motif in proteins, consisting of a sheet-like arrangement of amino acids. In Rasmol, beta-strands can be represented as arrows or ribbons. The arrow representation shows the strand as an arrow-like structure, while the ribbon representation shows the strand as a twisted ribbon. To display a beta-strand in Rasmol, the user can select the strand region and use the "cartoon" command to choose between the arrow or ribbon representation. For example, the command "cartoon ribbon" will display the strand as a ribbon show as **Yellow ribbons**.

Loop region representation (.): Loop regions are non-repetitive sections of protein structure that connect alpha-helices and beta-strands. In Rasmol, loop regions can be represented as coils or lines. The coil representation shows the loop as a coiled structure, while the line representation shows the loop as a straight line. To display a loop region in Rasmol, the user can select the loop region and use the "cartoon" command to choose between the coil or line representation. For example, the command "cartoon coil" will display the loop as a coil show as a **Light Grey strips.**

There is three types of Rasmol images were produced for each ligand binding:

- binding details general view representation.

- Motif+Ligand+Binding Residues(dots form) representation.
- Motif+ Ligand+ Binding Residues(cpk form) representation.
- Stereo graphic representation.

To generate the visual representation of the patterns in the binding sites, the Rasmol program employs a scripting language. This scripting language instructs the program on how to depict the molecular data in a graphical format. The resulting graphical representation can be observed in the figure. Essentially, the script serves as a set of commands that specifies how the program should generate the images, including the size, shape, color, and position of the different molecular components.

These commands may include instructions to highlight specific atoms, bonds, or residues that are of interest, or to manipulate the perspective and lighting conditions to optimize the visual clarity of the final image.

For the rasmol graphical representation we using the PPI for the download of the 4 rasmol scene before the edition and modification of this scene using notepad ++ .

II.3.1. binding details general view representation :

Blow is a Rasmol script, that produces a graphical representation of the **"..HH..SS."** motif without displaying the ligand "OCS", see Fig2.7**.**

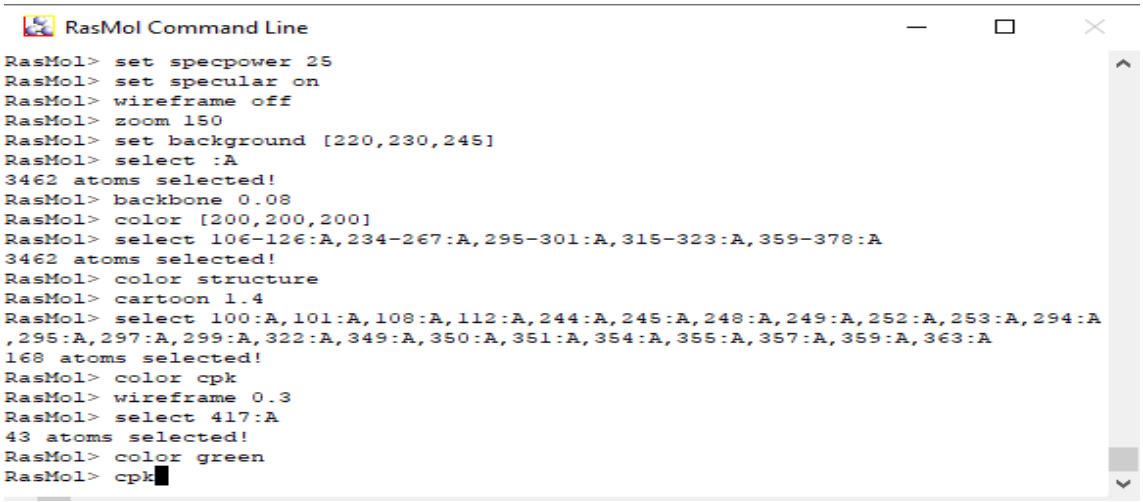


Figure 2.7. Capture of RasMol Command Line.

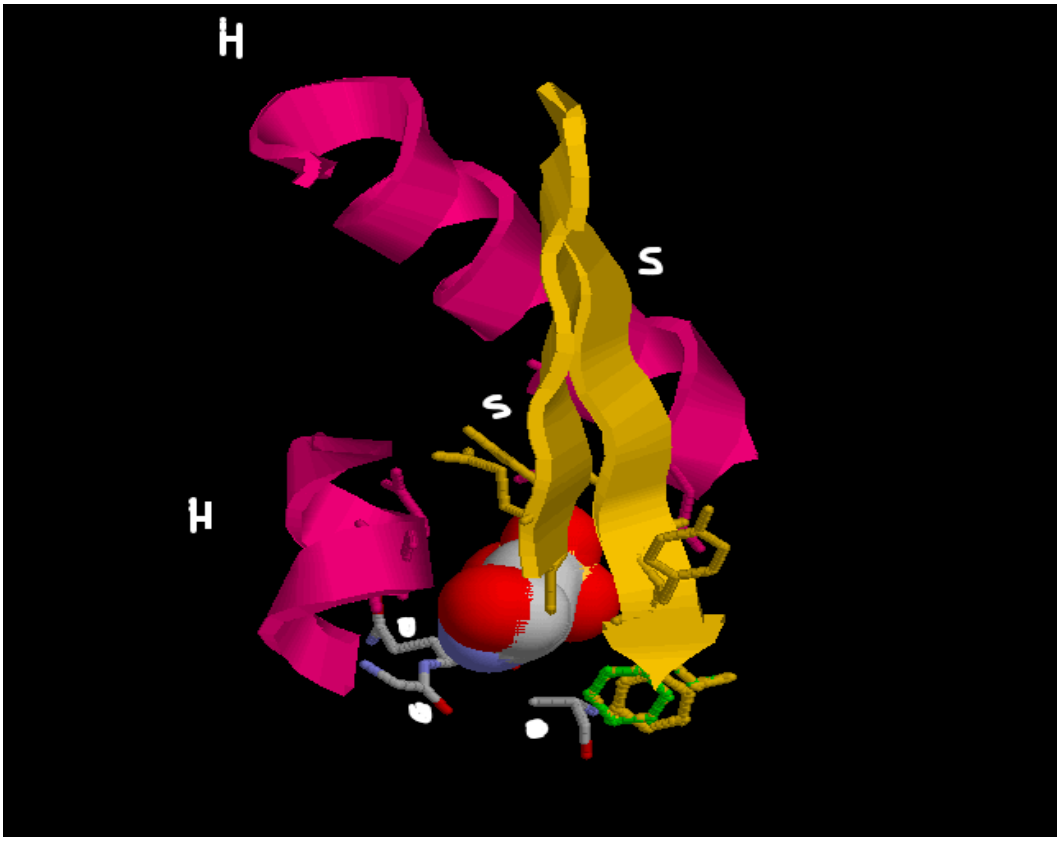


Figure 2.8. 3D-graphical representation of the Binding Motif HH..SS. where the dots (.) are shown as white spots (representing loop region in the motif).

That motif binds the ligand OCS **as described in the binding details data file 1smb_OCSA32.csv** found in the structure PDB id: 1SMB, chain A.

III.Flat-Files Database creation:

For the creation of the Flat File database, Fig 2.9 **.**we using three folders wich are the data_files, PDB files, graphics. The 1st folder data_files contain the all the csv_files of the 21 selected porphyrin proteins, the 2nd folder contain the all the PDB id of the proteins, the 3rd folder contain the the rasmol seans fot the creation of the motifs graphics

For the illustrion we choosing only 6 PDB id, for the rasmol scenes for the rasmol seans we choose chain A for the rasmol seans

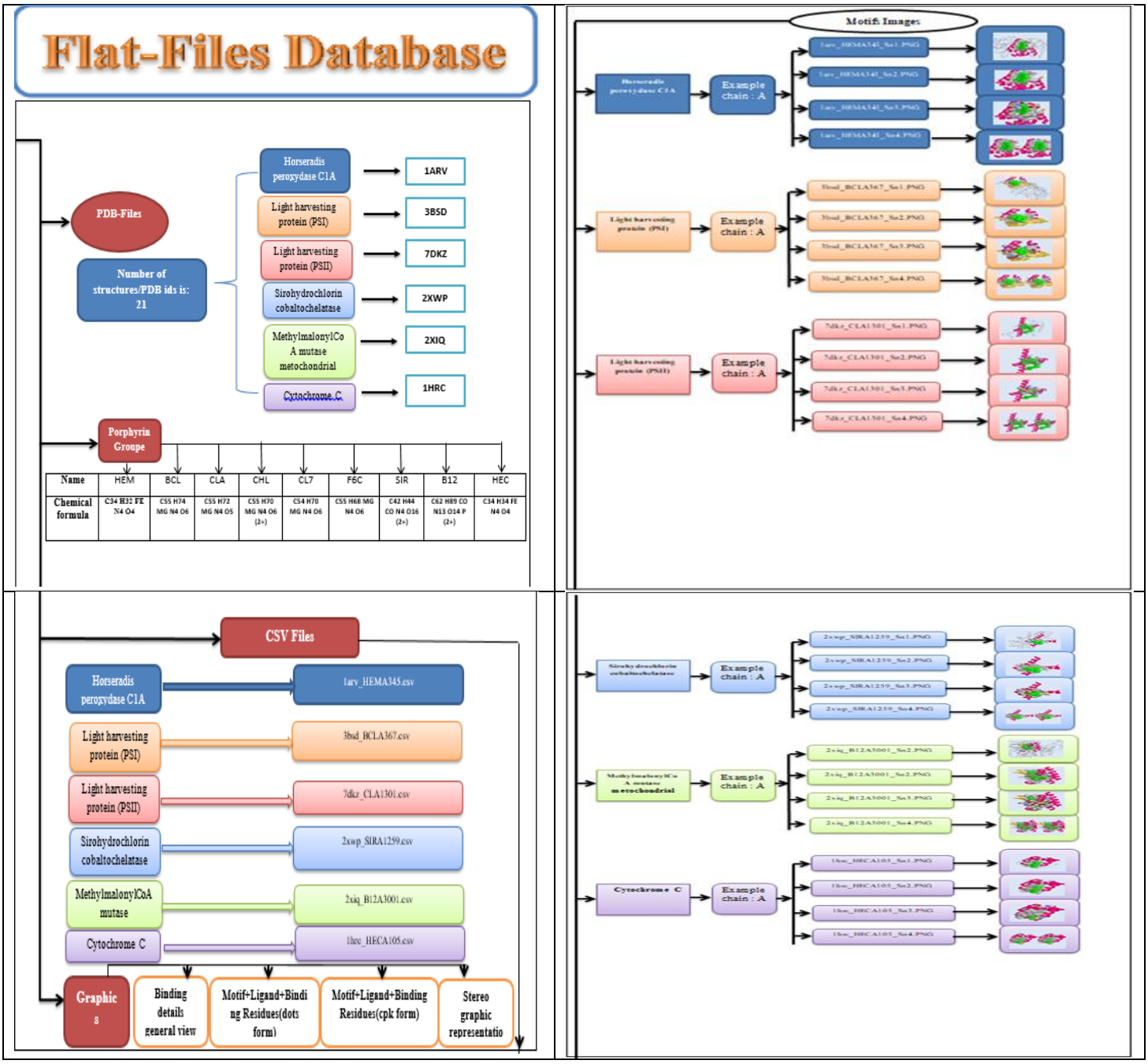


Figure 2.9 . The database schema representing the architecture of the created Flat-File database; the left side show the arrangement and classification of the files containing the csv-files which contain the binding environment details while the right side.

IV. World Wide Web Database :

The supervisor of the project has mounted the database on the "Bioinformatics Tools" server, and developed programming scripts to facilitate searching the database. This was done to enable sharing of the data and results with the scientific community both locally and internationally.

Chapter 3

Results and Discussion

Results and Discussion

I. Introduction

This chapter presents and discusses the importance of the understanding how Pathogenesis-Related proteins (PR-proteins) bind their ligands utilizing these structural motifs. These motifs bound to these ligands would lead to several important outcomes like:

- Inhibition of Pathogen Growth
- Activation of Defence Signalling Pathways
- Enzymatic Activity and Specificity and Recognition.

The binding of PR-proteins is a key aspect of plant immunity, leading to a cascade of defensive actions that collectively enhance resistance to pathogens and stressors. These interactions are fundamental to plant-pathogen coevolution and are a focus of study in agricultural biotechnology for developing disease-resistant crops.

Furthermore, and as shown in Chapter II, this project has resulted in the creation of a Flat-Files database that would be put online in near future. The database would provide details for a constructed set of novel structural motifs, would be referred to as Pathogenesis-Related Proteins Binding Structural Motifs.

II. Pathogenesis Related Proteins Binding Structural Motifs:

The binding motifs are the 3D- structurally arranged secondary structure elements with residue that interact with all the 17 PR- Proteins ligands.

The motifs are represented in string form where α-helixes are represented with **H**, β-strands are represented with **S** and loop-regions are represented with ”**L**”, see Table 3.1.

In addition, residues involved in the binding are presented with single letter code of amino acids, see Fig 3.1.

Tableau 3.1: Structural binding linear presentation, H denotes α-helix and, L denotes loop, S denotes β-strands. Amino acids shown with approximation of their belonging to the secondary structure elements.

PDB	Chain	Ligand ID	Motifs	Binding residue sequences
1FM4	A	**DXC**	**LLLLLLSSLLH**	FPFPTIYHQASM
4A81	A		**HHSLSSSSH**	FIGDFKIFVDYNNSML

III. Binding details:

The binding details sum up the bond lengths and types of each, and every residue atoms involved in the binding of the 17 PR-Proteins ligands.it also show the ranges of the secondary structure and loop regions, see Fig3.1.

Protein Residues					Ligand				Bonds	
Chain	SStr-Elm	Name	Number	Atom	Chain	Name	Number	Atom	Distance/Å	Possible Bond Type
A	32-40 S 1	LEU	38	CD1	A	EDO	2005	O2	3.39	van der Waals
A	No-SSE	TYR	47	C	A	EDO	2005	C2	3.75	van der Waals
A	No-SSE	TYR	47	O	A	EDO	2005	C2	3.57	van der Waals
A	No-SSE	TYR	47	CB	A	EDO	2005	C2	3.97	van der Waals
A	48-49 S 1	HIS	48	N	A	EDO	2005	C2	3.93	H.Bond
A	48-49 S 1	HIS	48	CA	A	EDO	2005	C2	3.95	van der Waals
A	48-49 S 1	LEU	49	N	A	EDO	2005	C2	3.88	H.Bond
A	88-104 H	HIS	98	CE1	A	EDO	2005	C2	3.66	van der Waals
A	88-104 H	HIS	98	CE1	A	EDO	2005	O2	2.91	van der Waals
A	88-104 H	HIS	98	NE2	A	EDO	2005	C2	3.85	H.Bond
A	88-104 H	HIS	98	NE2	A	EDO	2005	O2	3.47	H.Bond
A	88-104 H	TYR	103	CE1	A	EDO	2005	O2	3.73	van der Waals
A	88-104 H	TYR	103	CZ	A	EDO	2005	O2	3.55	van der Waals

Figure 3.1: Screen-capture of SSFS showing binding details.

IV. Structural types of motifs and classification:

The constructed structural binding motifs combine three secondary structure elements, loop, α-helices and β-strands and resulted in 7 classes of motifs summarised below:

1. **Loop based motifs:**

These motifs are composed of only loop regions.

2. **α-Helix based motifs:**

These motifs are composed only of α-helix secondary structure elements.

3. **β-Strand based motifs:**

These motifs are made only of β-strand secondary structure elements.

4. **Loop and α-helix based motifs:**

These motifs are made of loops and α-helices secondary structure elements.

5. **Loop and β-strand based motifs:**

These motifs are made of the secondary structure elements; loops and β-strands.

6. **α-Helix and β-strand based motifs:**

These motifs are made of the secondary structure elements; α-helices and β-stands.

7. **Loop and α-helix and β-strand based motifs:**

These motifs combine three of the secondary structure elements; loops, α-helices and β-stands.

Details of the 7 classes of the motifs' conformation type are highlighted below in Table 2:

Tableau 3.2: The 17 PR- proteins ligands with 03 types of structural motifs

	Ligand Name	Ligand ID	Motifs
Loops only	DI(HYDROXYETHYL)ETHER	PEG	L, LLLL, LL,
	GLYCEROL	GOL	LL, LLLLLLL,
	ZINC ION	ZN	LL, L
	SULFATE ION	SO4	LLLLLL, LLLL, LLL
	CHLORIDE ION	CL	LLL
	2-(N-MORPHOLINO)-ETHANESULFONIC ACID	MES	LL
	1,4-DIETHYLENE DIOXIDE	DIO	LL
	SUCCINIC ACID	SIN	LLLLL
	FORMIC ACID	FMT	LLL, L
	ACETIC ACID	ACY	LLLL
	CHLORIDE ION	CL	LLL
	ACETYL GROUP	ACE	LL, L
	(2S)-2-AMINOPROPANE-1,1-DIOL	ALV	LL
	CHLOROMETHANE	0QE	L
	1,2-ETHANEDIOL	EDO	L, LLLL, LLL,
	ACETATE ION	ACT	LL, L, LLL,
	SODIUM ION	NA	LLL, LL,
	MALONATE ION	MLI	LLLL, LL,
α –Helix Loops	L(+)-TARTARIC ACID	TLA	HLHLHL
	ZINC ION	ZN	LH, HL,
	2-(N-MORPHOLINO)-ETHANESULFONIC ACID	MES	HLLLLHHH, LLLLHHH
	(4S)-2-METHYL-2,4-PENTANEDIOL	MPD	LLLHLL, HHL
	SULFATE ION	SO4	HLLH, LLLH, HLH, LLH, HLL
	1,2-ETHANEDIOL	EDO	LLLLH, HHLH, HLLH, HL, HLL, HLLL, LH, HLH, HLLLL, LLLHH, LLLH, LLHLLH
	CHLORIDE ION	CL	LLLH,
	2-ACETAMIDO-2-DEOXY-ALPHA-D-GLUCOPYRANOSE	NDG	HLLLLLHH
	2-ACETAMIDO-2-DEOXY-BETA-D-GLUCOPYRANOSE	NAG	LLLLHHLH, HLLLLLHH, LLLHHLH,
	GLYCEROL	GOL	HLLLH, LHLL, HLL,
	ACETATE ION	ACT	HLLLLLLL, LHL, HLH, HL, HLL, HHL,
	FORMIC ACID	FMT	LH
	SELENOMETHIONINE	MSE	LH
	CALCIUM ION	CA	HLLLH, HLLLL

	3-(4-HYDROXY-3-METHOXYPHENYL)-2-PROPENOIC ACID	FER	HLLLLLLL
	BENZHYDROXAMIC ACID	BHO	HLLLLL
	CACODYLATE ION	CAC	LHL
β −sheet Loops	CHLORIDE ION	CL	SL, LS
	DI(HYDROXYETHYL)ETHER	PEG	SL
	N-[(3-CARBOXYPHENOXY)ACETYL]-L-GLUTAMIC ACID	TQG	LLSSSLSLS
	4-(2-HYDROXYETHYL)-1-PIPERAZINE ETHANESULFONIC ACID	EPE	LLSSLLL
	SELENOMETHIONINE	MSE	LSSSSLL, LSS,
	BETA-MERCAPTOETHANOL	BME	SL
	1,2-ETHANEDIOL	EDO	LLS, LSL, LSLLS, LLSLL, LLLSLS
	ZINC ION	ZN	LSL,
	ACETATE ION	ACT	SSLSLS
	IMIDAZOLE	IMD	SLSS
	2-AMINO-2-HYDROXYMETHYL-PROPANE-1,3-DIOL	TRS	SLLSLSLS
	HYDROGEN PEROXIDE	PEO	LLSLL
	SODIUM ION	NA	SLL, SL
	(2R,3S)-2-(3,4-DIHYDROXYPHENYL)-3,4-DIHYDRO-2H- CHROMENE-3,5,7-TRIOL	KXN	LLSLL
α −Helix only	CHLORIDE ION	CL	H
	INOSITOL HEXAKISPHOSPHATE	IHP	H
	2-(N-MORPHOLINO)-ETHANESULFONIC ACID	MES	H
	(4S)-2-METHYL-2,4-PENTANEDIOL	MPD	HH
	ZINC ION	ZN	H
	SULFATE ION	SO4	H, HH
	2-ACETAMIDO-2-DEOXY-ALPHA-D-GLUCOPYRANOSE 2-ACETAMIDO-2-DEOXY-BETA-D-GLUCOPYRANOSE	NDG NAG	HH HH, H
	ACETATE ION	ACT	HH, H
	1,2-ETHANEDIOL	EDO	H

	DI(HYDROXYETHYL)ETHER	PEG	H
	GLYCEROL	GOL	H
	GLUTATHIONE	GSH	HH
	3-(4-HYDROXY-3-METHOXYPHENYL)-2-PROPENOIC ACID	FER	HH
	OXYGEN ATOM	O	HH
	OXYGEN MOLECULE	OXY	HH
	SODIUM ION	NA	H
	(2R,3S)-2-(3,4-DIHYDROXYPHENYL)-3,4-DIHYDRO-2H- CHROMENE-3,5,7-TRIOL	KXN	HH
	UNKNOWN LIGAND	UNL	H
β −sheet only	ACETATE ION	ACT	SSS
	CHLORIDE ION	CL	SS
	SELENOMETHIONINE	MSE	SS
	1,2-ETHANEDIOL	EDO	S, SSSSSS
	SODIUM ION	NA	S
	SULFATE ION	SO4	SSS, SS
	UNKNOWN LIGAND	UNL	S
α −Helix + β − sheet	CYSTEINESULFONIC ACID	OCS	SHS
	SULFATE ION	SO4	HS
	DI(HYDROXYETHYL)ETHER	PEG	SHH
	N-(3-METHYLBUT-2-EN-1-YL)-9H-PURIN-6-AMINE	ZIP	SSSSH
	(2E)-2-METHYL-4-(9H-PURIN-6-YLAMINO)BUT-2-EN-1-OL	ZEA	SSSSH, HHSSSSH, SSH
	N-(FURAN-2-YLMETHYL)-7H-PURIN-6-AMINE	H35	SSSH
	N-BENZYL-9H-PURIN-6-AMINE	EMU	SSSSH
	GLYCEROL	GOL	HHSSS, SH, SSH
	1,3-DIPHENYLUREA	BSU	HSSSH, SHSSH
	(3ALPHA,5BETA,12ALPHA)-3,12-DIHYDROXYCHOLAN-24-OIC ACID	DXC	HHSSSSSH
	UNKNOWN LIGAND	UNL	SH, SSH
	N-[2-(5-METHOXY-1H-INDOL-3-YL)ETHYL]ACETAMIDE	ML1	SHSSSSSH
	4-(2-HYDROXYETHYL)-1-PIPERAZINE ETHANESULFONIC ACID	EPE	HSS

	(2R,3S)-2-(3,4-DIHYDROXYPHENYL)-3,4-DIHYDRO-2H- CHROMENE-3,5,7-TRIOL	KXN	HSSHSSH
	8-ANILINO-1-NAPHTHALENE SULFONATE	2AN	SHSSSH
	SULFATE ION	SO4	SH
α –Helix+ β -sheet + Loops	CYSTEINESULFONIC ACID	OCS	HHLLSSL
	O-ACETALDEHYDYL-HEXAETHYLENE GLYCOL	P4C	HLSSSSH, HLSSSH
	N-[(3-CARBOXYPHENOXY)ACETYL]-L-GLUTAMIC ACID	TQG	HLSSSSLS,
	GLYCEROL	GOL	LSLH, LSH, LLHS
	MAGNESIUM ION	MG	LSH
	ACETATE ION	ACT	SLHS, SHLLSH
	SELENOMETHIONINE	MSE	SHLLLSSLH, HLLSL
	1,2-ETHANEDIOL	EDO	SLSHL, SLLLH
	PROTOPORPHYRIN IX CONTAINING FE	HEM	HLLLLLHHHLSHLHH, HLLLLLHHHLSLLH
	3-BENZOYLBENZOIC ACID	A8I	HHLSSSSH
	(3ALPHA,5BETA,12ALPHA)-3,12-DIHYDROXYCHOLAN-24-OIC ACID	DXC	LLLLLLSSLLH, HHSLSSSSH
	CALCIUM ION	CA	HLS
	(2E)-2-METHYL-4-(9H-PURIN-6-YLAMINO)BUT-2-EN-1-OL	ZEA	SLH
	1,3-DIPHENYLUREA	BSU	LSSH, SLSSH,
	SODIUM ION	NA	HLS
	(2R,3S)-2-(3,4-DIHYDROXYPHENYL)-3,4-DIHYDRO-2H- CHROMENE-3,5,7-TRIOL	KXN	HHSHSLH, HHLSLSH, HLLLSH, HHLSLSSH
	(4S)-2-METHYL-2,4-PENTANEDIOL	MPD	LLLSH

V. Binding residues sequences of the motifs:

In the following section, the actual binding residues in the motifs arranged in sequences (non-contiguous) in all of the studied PR-proteins are presented in Table 3.2.

Tableau 3.3: Presents the binding sequences of the motifs

PDB	Chain	Ligand ID	Motifs	Binding Residue Sequences
4DWX	**A**	**ZN**	LL, L	HD, H
4DYG				
6EPB			L, LL, L	D, HD, H
6YLJ			L, LL	D, HD
6YN4			L, L, LL	H, D, HD
7AKQ			L, LL, L, LL	D, HD, H, HW
4J0L	**A**	**SO4**	LLLLLL	QEGARY
3CQL	**B**		LLLL	KNP
4A80	**A**		LLL	SPE
4MST	**B**	**CL**	LLL	FPA
7P20	**A**		LLL	QPH
4MPI	**A**	**MES**	LL	MW
	B	**DIO**	LL	PG
3ZS3	**A**	**SIN**	LLLLL	SLTGA
		FMT	LLL	KNS
			L	V
		ACY	LLLL	PQLP
1LVM	**C**	**ACE**	LL	EN
	D		LL	EN
3I74	**C**		LL	FE
	D		L	F
	C / D	**ALV**	LL	EK
	C / D	**0QE**	L	K
1OM0	**A**	**EDO**	L, L	G, D
1TA3	**A**		L	K
	B		LL	SQ
1TE1	**B**		LLLL, LLL	DCGE, KRV
6EPB	**A**		L, L, LLL, L, L	S, Q, HDW, H, G
	A	**ACT**	LL	SW
6YN4			L, LLL	D, HSW
4GY9	**A**	**NA**	LL, LLL	TG, TEG
4JHG	**A**		LL	TG
		MLI	LLLL	EGIA
4GY9			LLLL	EGIA
4JHH			LLLL	EGIA

7QB6	**B**		LL	F
4AIW	**A**	**IHP**	H	K
2WQL		**CL**	H	SK
		MES	H	PN
2DKV	**A**	**MPD**	HH	TRIT
4DWX	**A**		H, H	HD, H
	B		H	HD
4DYG	**A**		H, H	HD, H
	B		H	HD
6EPB	**A**	**ZN**	H	HG
6YLJ	**A**		H	H
6YN4	**A**		H	HG
7AKQ	**A**		H	HG
4J0L	**A**		HH, H	RR, SDAR
3CQL	**A**	**SO4**	HH	QRI
4A80	**A**		HH	KYH
4A81	**A**		H	EKE
3CQL	**A**	**NDG**	HH	YGRRG
	B		HH	YGRRG
1OM0	**A**	**NAG**	H, HH	QN, NV
1TA3	**A**		HH, HH	QN, NRDVL
1TE1	**A**		HH, HH	KQN, NLRV
3ZS3	**A**		HH	KTE
6YLJ	**A**	**ACT**	H	H
4Y31	**A**		H	K
1OM0	**A**		H	KN
1TA3	**B**	**EDO**	H	NLN
1TE1	**A**		HH, H	RH, A
6EPB	**A**		H	LK
1QGJ	**A / B**	**GSH**	HH	NTDDCAR
1GW2	**A**	**FER**	HH	NNIDA
1GWT	**A**		HH	NIDA
1H55	**A**	**O**	HH	RFHH
1H57	**A**	**OXY**	HH	RFHH
4GY9	**A**	**NA**	H	A
4C9I	**F**	**KXN**	HH	IHKLGYD
5MXB	**A**	**UNL**	H	K
3ZS3	**A**	**ACT**	SSS	SEITYYD
6YLJ	**A**		SSSSS	SHTYNV
7P20	**A**	**CL**	SS	KWT
2WQL			S, S	H, HS
1LVB	**B**	**MSE**	SS	LVTSV
1OM0	**A**	**EDO**	S	H
1TE1	**B**		SS	FPSY
6EPB	**A**		SSSSSS	SHLTYNV

4GY9	**A**	**NA**	S, S	ST, S
4JHG	**A**		S	ST
4C9C	**A**	**SO4**	SSS	LKYK
4A80	**A**		SS	REK
5MXB	**A**	**UNL**	S	FQ
5MXW	**A**		S	I
1SMB	**A**	**OCS**	SHS	KLKNLNKM
4AIW			SHS	KLKNLNKM
5vhg	**A**	**SO4**	SH	CKK
4A80			SH	ETTR
4GY9	**A**	**ZIP**	SSSSH	TQYVFYAGRF
4JHG	**A**	**ZEA**	SSSSH	TQYVIFYYAGRFF
2QIM			HHSSSSH	LVIHYYTVF
			SSSSH	FDYYIFVIGA
			SSH	LFEVRGF
5MXW			SSH	FEVRF
4JHH	**A**	**H35**	SSSH	TQYVIFYGRF
4JHI	**A**	**EMU**	SSSSH	LTQYVIFYAGRFF
1FM4	**A**	**DXC**	HHSSSSSH	FIVIDYYVINNYSL
3E85	**A**	**BSU**	HSSSH	LVLHYYF
			SHSSH	YLFTLGVGF
5MXB	**A**	**UNL**	SH	IG
			SH	IFA
5MXW	**A**		SH	ING
			SSH	FIG
5MXB	**A**	**ML1**	SHSSSSSH	YLVLHYYTVGFF
5MXW			SHSSSSSH	YLVLHYYTVGFF
1XDF	**A**	**EPE**	HSS	NRK
4C94	**B**	**KXN**	HSSHSSH	IVLSHYIR
4A80	**A**	**2AN**	SHSSSH	YFYIINVSKGE
4A81			SHSSSH	TTFYI
1SMB	**A**	OCS	HHLLSSL	LEESGQGERYFA
2WQL	**A, B, D**	P4C	HLSSSSH	IYVYSTTNL
	C		HLSSSH	IYYSTMTNL
6wao	**B**	TQG	HLSSSSLS	RFLRPTVVTPR
5JYS	**A**	MG	LSH	HLH
3ZS3		ACT	SLHS	TRNNI
6EPB			SHLLSH	TWREWE
1LVB	**A, B**	MSE	SHLLLSSLH	TKIRDISPF
	A, B		HLLSL	RLDFS
1OM0		EDO	SLSHL	LYHLHYP
1TE1			LLSH	LGIL
6EPB			SLLLH	NLWSTNF
1QGJ	A, B		HLLLLLHHHLSHLHH	RASIRFSRPSPIFLSAHFGQAKVLLSM

1ATJ	A, B, C D, E, F	HEM	HLLLLLHHHLSLLH	RASRFSRPAPLFLLSGHFGKNQFFIS
1GW2	A		HLLLLLHHHLSLLH	RASLRFSRPALFLLSGHFGKNQFFIS
1GWO	A		HLLLLLHHHLSLLH	RASRFSRPQPLFLLSGHFGKNQFFIS
1GWT	A		HLLLLLHHHLSLLH	RASLRFSRPAPLFLLSGHFGKNQFMIS
1GWU	A		HLLLLLHHHLSLLH	RASLRFSRPGPLFLLSGHFKNQFFIS
1GX2	A, B		HLLLLHHHLSLH	RASLRFSPAPLFLSGHFGKNQFIS
1H55	A		HLLLLLHHHLSLLH	RASLRFSRPAPLFLLSGHFGKNQFFIS
1H57	A		HLLLLLHHHLSLLH	RASLRFSRPAPLFLLSGHFGKNQFFIS
1KZM	A		HLLLLLLHHHLSLLH	RASLSFNSRPAPLFLLSGHFGKNQFFS
7QB6	A	A8I	HHLSSSSH	IYFLQYFFRFFT
1FM4		DXC	LLLLLLSSLLH	FPFPTIYHQASM
4A81			HHSLSSSSH	FIGDFKIFVDYNNSML
2QIM		CA	HLS	PVEI
		ZEA	SLH	FQDPEGK
3E85		BSU	LSSH	TILFIRA
			SLSSH	FGFIGAR
		NA	HLS	PVEI
5MXW			HLS	PVEI
1XDF	A		HLS	PVEI
5MXB			HLS	PVEI
		ML1	HLSSSH	VTILFYRGAF
4C94	A	KXN	HHSHSLH	FAILSHDR
	C		HHLSLSH	FAAILHR
	D		HLLLSH	FAVLHR
	E		HHLSLSSH	FDAILSHIRL
4A81	A	MPD	LLLSH	FFPIQASM
2I0W	A	CL	SL	NP
2wql			LS	QS
6wao	A	TQG	LLSSSLSLS	RFLRRTVVTPR
4BCT		EPE	LLSSLLL	KGEAQDSVYP
1LVB	A, B	MSE	LSSSSLLL	PIVFVIRPKF
	A, B		LSS	LNSS
1**Q31**	A, B	BME	SL	CW
1**OM0**		EDO	LLS	RYW
			LSL	RDD
6EPB	A		LLLSLS	KAANTE
1TE1	A		LSLLS	YTDSW
	A		LLSLL	PARDN
6EPB		ZN	LSL	WDE
		ACT	SSLSLS	YDEMYW
		IMD	SLSS	YEIW
1H57		PEO	LLSLL	ANQRF

4GY9		NA	SLL	FVG
4JHG			SLL	FVG
4JHH			SLL	FVG
4JHI			SL	FG
3E85			SL	TG
4A81			SL	YV
4C94	A	KXN	LLSLL	PQKEG
6wcs	**A**	**TLA**	HLHLHL	SRKSWK
3q2r	**A**	**ZN**	LH	HH
6EPB	**A**		HL	DH
2DKV	**A**	**MES**	HLLLLHHH	EIQLSNFVN
3IWR	**A**		HLLLLHHH	EIQLSNFV
	B		HLLLLHHH	EIQLSNFVD
4DWX	**A**		LLLLHHH	IQLSNFIN
	B		LLLLHHH	IQLSNFIN
4DYG	**A**		HLLLLHHH	EIQLSNFI
	B		HLLLLHHH	EIQLSNFI
2DKV	**A**	**MPD**	LLLHLL	AGWWCF
3IWR	**A**		HHL	TRIP
	B		HHL	TRITP
4DWX	**A**	**SO4**	HLLH	HDRYTY
	A		LLLH	SQWY
	B		HLLH	HDRYTY
4DYG	**A**		HLLH	HDRYTY
	A		LLLH	SQWY
	B		HLH	HDRTY
4J0L	**A**		HLLH	HDRYTY
			LLH	RYD
3CQL	**B**		HLLH	QKDRI
			HLH	RDRTY
4C9C	**B**		HLL	DHPS
3ZS3	**A**	**ACT**	HLLLLLLL	EPQAYDNF
6EPB			LHL	HQVD
6YN4			HLLH	YEHNY
1GWO	A		HL	RFHP
1GWU	A		HLL, HHL	RFHGP, KFPR
1H55	A		HHL	KFPR
1H57	A		HHL	KFPR
3ZS3	A	**FMT**	LH	KA
1LVB	**A**	**MSE**	LH	FNFELLT
	B		LH	FNFELLT
1OM0	**A**		HHLH	KSHD
			HLLH	SMYRDKN
1TA3	**B**		HL	DQ
			HLL	ATKK

		EDO	HLLL	RVPI
			LH	WRAS
1TE1	**A**		LH	LGEG
			HLH	CGK
			HLLLL	QGVPD
6EPB	A		LLLHH	RNFDK
			LLLLH	DGTNY
			LLLH	YADQK
			LH	DANN
			LLHLLH	STPWYQ
			HL	KL
			LLLH	HSWY
			HLLLL	ASVGLDT
			HLL	DNNLN
			LH	DKK
			HLL	YDR
1QGJ	A, B	**CA**	HLLLH	DVGDS
	A, B		HLLLL	TDTTD
1ATJ	A, B, C D, E, F		HLLLL	TDTID
			HLLLH	DVGDSI
1GW2	A		HLLLH, HLLLL	DVGDS, SDTID
1GWO	A		HLLLH, HLLLL	DVGDS, TDTID
1GWT	A		HLLLH, HLLLL	DVGDS, TDTID
1GWU	A		HLLLH, HLLLL	DVGDS, TDTID
1GX2	A, B		HLLLL	TDTID
	A, B		HLLLH	DVGDS
1H55	A		HLLLH, HLLLL	DVGDS, TDTID
1H57			HLLLH, HLLLL	DVGDS, TDTID
1KZM			HLLLH, HLLLL	DVGDS, TDTID
1GW2	A	FER	HLLLLLLL	RHFGSPAFF
1GX2	A, B	BHO	HLLLLL	RFHFGPAP
1H57			LH, HL	GNL, LAG
1KZM	A	CAC	LHL	GRN

VI. Classification of motifs per ligands:

In this part of the studies, in the following, a summary of each ligand with the motifs' structure type, see Table 3.4.

Tableau 3.4: Classification of motifs per ligands

Ligand	Motif 1 Loop only	Motif 2 α_helix only	Motif 3 β_strand only	Motif 4 α & β only	Motif 5 L& α Helix and βstrand	Motif 6 L & α helix	Motif 7 L & β strand
ZN	L, LL	H	/	/	/	LH	LSL
SO4	LLLLLL, LLLL, LLL	HH, H	SSS, SS	SH	/	HLLL, LLH HLH, HLLH	/
CL	LLL	H	SS	/	/	/	LS, SL
NA	LL, LLL	H	S	/	HLS	/	SLL, SL
MG	/	/	/	/	LSH	/	/
MES	LL	H	/	/	/	HLLLLHHH LLLLHHH	/
ACT	L, LL, LLL	H, HH	SSS, SSSSS	/	SLHS, SHLLSH	HLLLLLLL, LHL, ……	SSLSLS
MPD	/	HH	/	/	LLLSH	LLLHL, HHL	/
SIN	LLLLL	/	/	/	/	/	/
FER	/	HH	/	/	/	HLLLLLLL	/
ZEA	/	/	/	HHHHSH, SSH, SSSSH	SLH	/	/
ZIP	/	/	/	SSSSH	/	/	/
ALV	LL	/	/	/	/	/	/
0QE	L	/	/	/	/	/	/
MLI	LLLL	/	/	/	/	/	/
DIO	LL	/	/	/	/	/	/
IHP	/	H	/	/	/	/	/
GSH	/	HH	/	/	/	/	/
O	/	HH	/	/	/	/	/
OXY	/	HH	/	/	/	/	/
NDG	/	HH	/	/	/	/	/
NAG	/	H, HH	/	/	/	/	/
FMT	LLL, L	/	/	/	/	LH	/
ACY	LLLL	/	/	/	/	/	/
ACE	LL, L	/	/	/	/	/	/
KXN	/	/	/	HSSHSSH	HHSHSLH, HLLLSH	/	LLSLL
UNL	/	/	/	SH, SSH	/	/	/

EDO	LLLL, LL	HH, H	SSSSSS	/	SLLLH	LLLH, HL….	LLS, LSL
MSE	/	/	SS	/	HLLSL	LH	LSSSSLLL
OCS	/	/	/	SHS	HHLLSSL	/	/
ZIP	/	/	/	SSSSH	/	/	/
ZEA	/	/	/	SSSSH	SLH	/	/
H35	/	/	/	SSSH	/	/	/
EMU	/	/	/	SSSSH	/	/	/
DXC	/	/	/	HHSSSSH	LLLLLSSLH	/	/
BSU	/	/	/	HSSSH	LSSH	/	/
ML1	/	/	/	SHSSSSSH	HLSSSH	/	/
EPE	/	/	/	HSS	/	/	LLSSSLLL
2AN	/	/	/	SHSSSH	/	/	/
TQG	/	/	/	/	HLSSSSLS	/	LLSSSLSLS
P4C	/	/	/	/	HLSSSSH	/	/
HEM	/	/	/	/	HLLLHLLS	/	/
SIN	LLLLL	/	/	/	/	/	/
BME	/	/	/	/	/	/	SL
IMD	/	/	/	/	/	/	SLSS
TLA	/	/	/	/	/	HLHLHL	/
CAC	/	/	/	/	/	LHL	/
BHO	/	/	/	/	/	HLLLLL	/
A8I	/	/	/	/	HHLSSSSH	/	/

VI.1. Motifs Structure and Evolutionary Relationship:

As show in Table 3.4, show structural conservation of the binding motifs across the different species and biological function. This discovery supports evolutionary relationship between the different functional entities such as the different classes of PR-proteins.

VI.2. Motifs Structural Arrangement and Function relationship:

The groups of motifs reported above and in Table3.4. notably those that bind secondary structure element, α-helices and β-strands constructing these motifs are quite far from each other

in sequence and are separated with large loop regions, however, they group together in close proximity in 3D-space.

Such a structural arrangement of distant regions enables the binding of the ligands that belong to different species thereby insuring their biological function. These cases enforce more the concept of structure-function relations ship.

VII. Graphical Representation of Binding Motifs:

This section provides some examples of the ligands in complex with their motifs. The graphics has been generated using the Rasmol software,

VII. 1. HEM group binding motif mostly α- + β-structure:

The graphical presentation, Fig3.2, and Fig3.3 displays the HEM porphyrin group (Protoporphyrin IX Containing Fe), in van-der-waals representation, binding amostly α-structure + β-structure motif: **HLLLLLLHHHLSLLH** Tableau 3.5. in this example we will study two different PDBs (1KZM, 1ATj). Tableau 3.5 shows the same ligand binding the actual sequence of residues responsible for the function.

Tableau 3.5. Sequence-Structure Correlation of functional residues in PDB entries 1KZM And 1ATJ.

PDB	Ligand ID	Motifs	Residues
1KZM	HEM	HLLLLLLHHHLSLLH	RASLSFNSRPAPLFLLSGHFGKNQFFS
1ATJ		HLLLLLHHHLSLLH	RASRFSPRAPLFLLSGHFGKNQFFIS

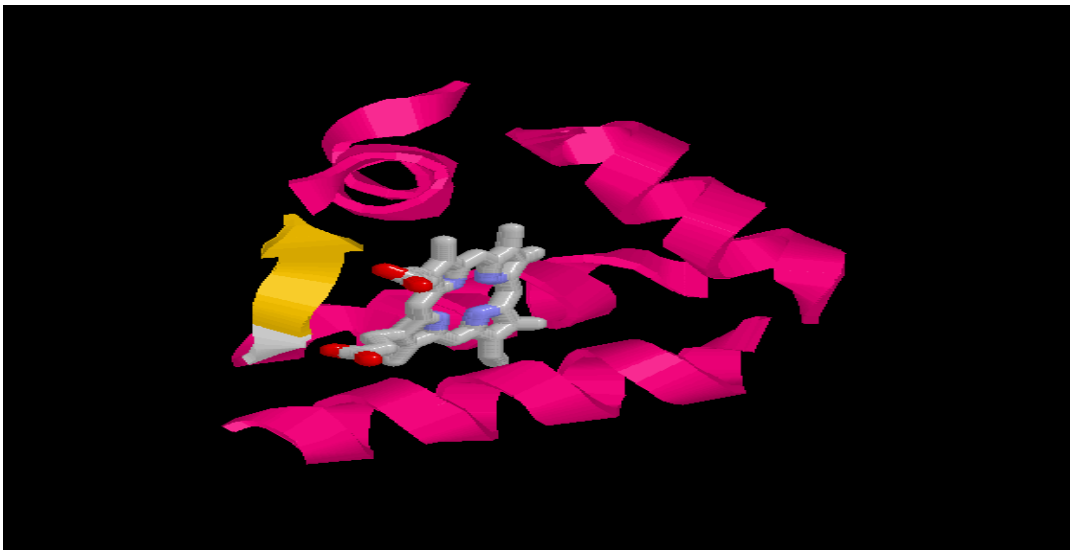


Figure 3.2. The Structural Motif: [**HLLLLLLHHHLSLLH**] – binding the HEM group in van-der-waal display. The HEM is shown binding the residues responsible in the actual binding are: RASLSFNSRPAPLFLLSGHFGKNQFFS (PDB: 1KZM).

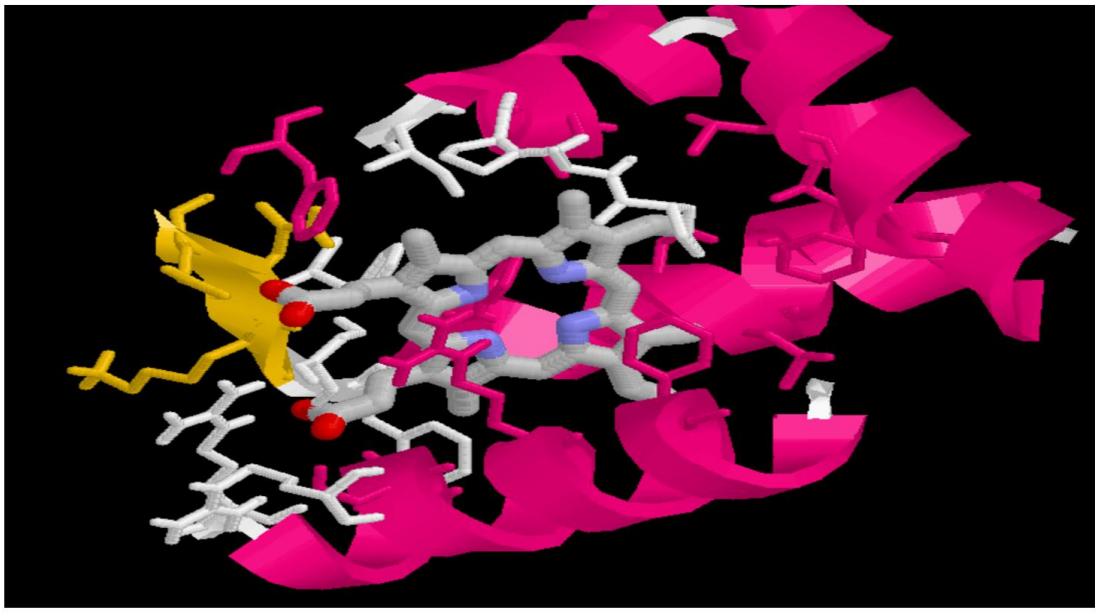


Figure 3.3. the Structural Motifs [HLLLLLHHHLSLLH]-binding the HEM group in van-der-waal display. The HEM is shown binding the residues responsible in actual binding are: RASRFSPRAPLFLLSGHFGKNQFFIS (PDB: 1ATj)

VII.2. OCS binding motif mostly α- + β-structure:

The graphical representation Fig 3.4a, Fig 3.4b and Fig 3.4c in van-der-waals representation, binding amostly α-structure + β-structure motif: HHLLSSL, with sequence of residues responsible for the function. (PDB 1smb) , with different sins.

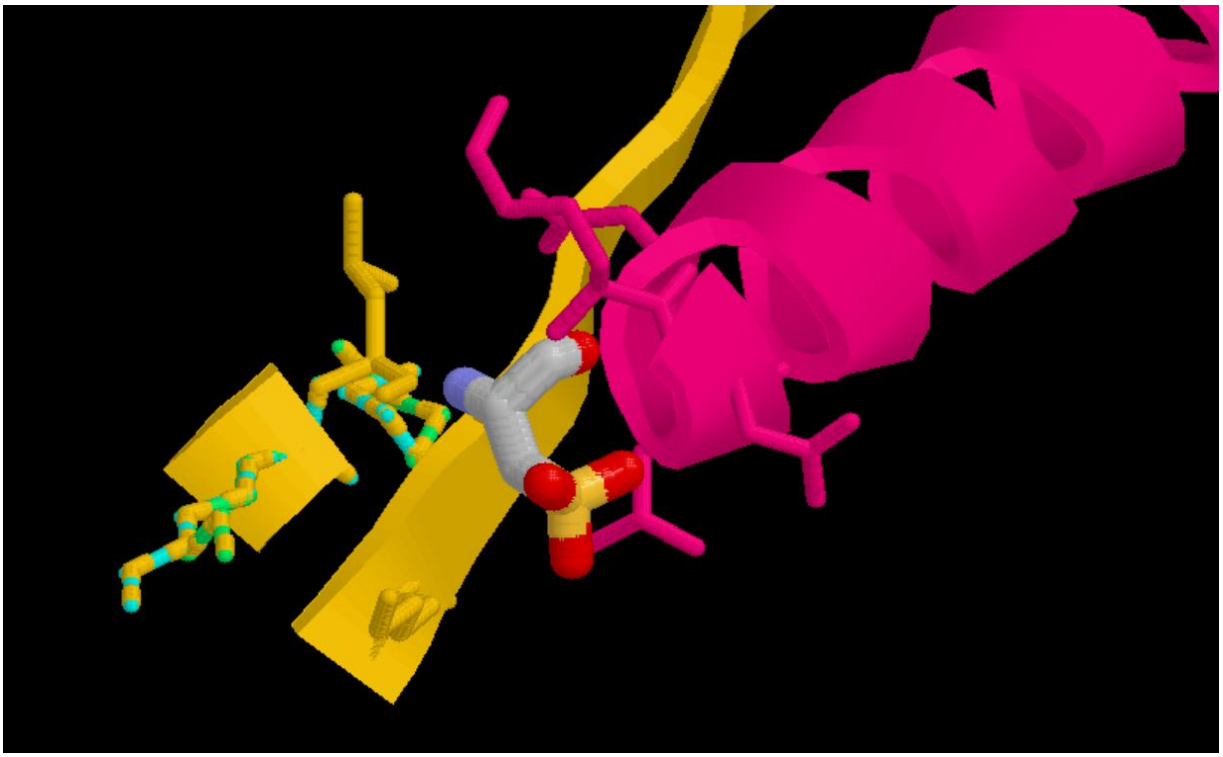


Figure 3.4a The Structural Motifs [HHLLSSL]-binding the OCS in van-der-waal display. The OCS is shown binding the residues responsible in actual binding are: LEESGQGERYFA(PDB: 1 smb) sin1 Batons

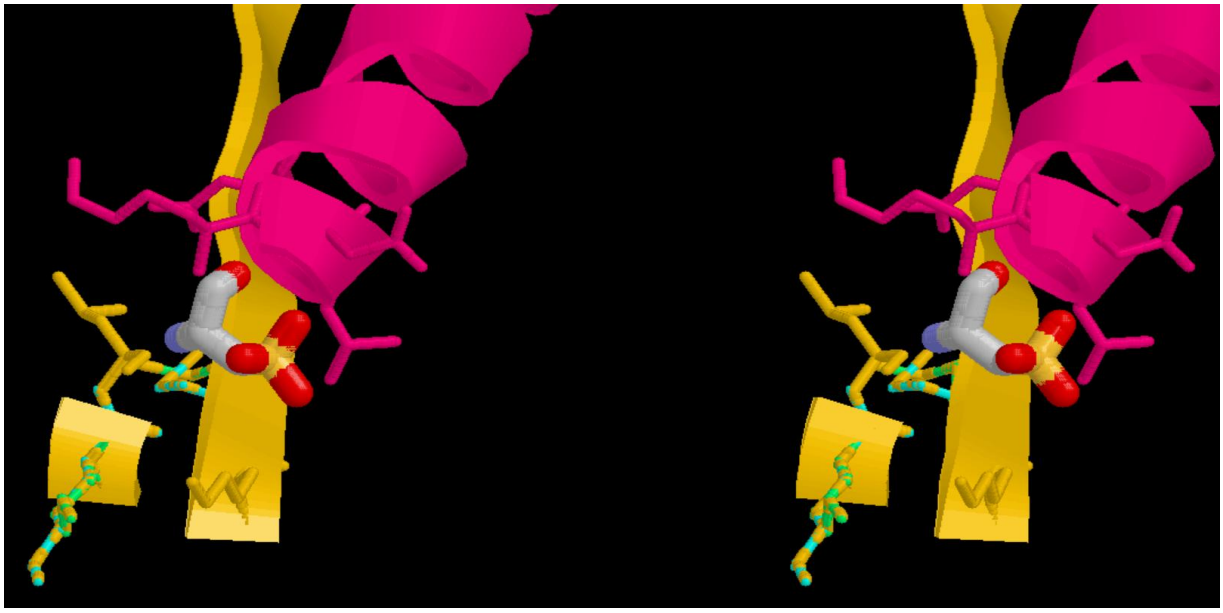


Figure 3.4b the Structural Motifs [HHLLSSL]-binding the OCS in van-der-waal display. The OCS is shown binding the residues responsible in actual binding are: LEESGQGERYFA(PDB: 1 smb) sin2 stereo

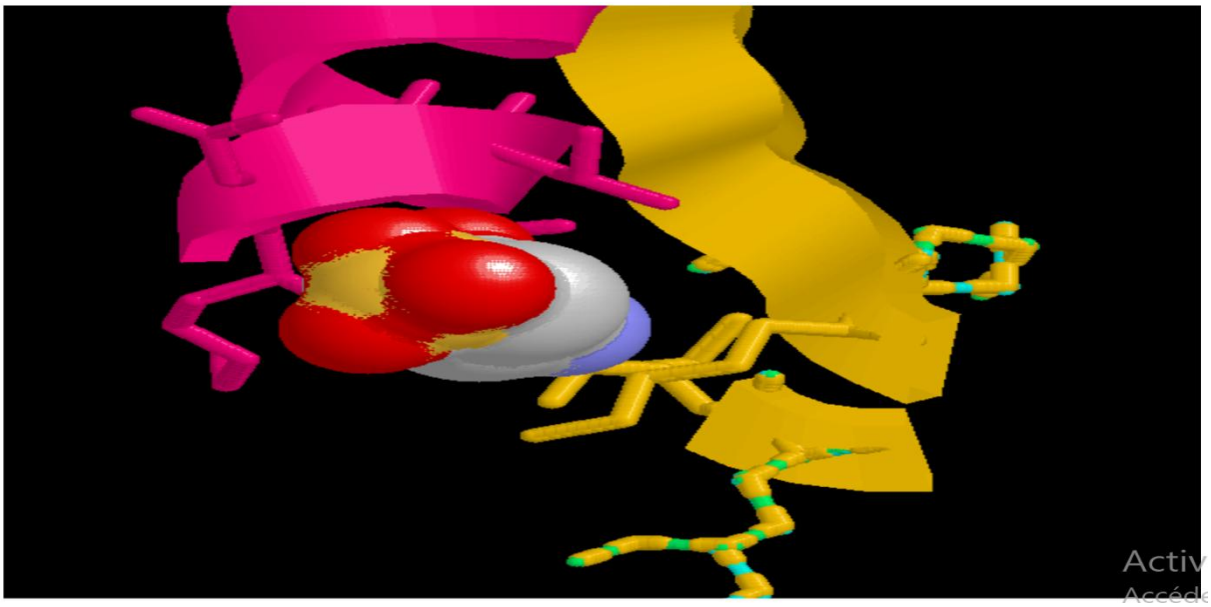


Figure 3.4c The Structural Motifs [HHLLSSL]-binding the OCS in van-der-waal display. The OCS is shown binding the residues responsible in actual binding are: LEESGQGERYFA(PDB: 1 smb) sin3

General Conclusion

General conclusion

This project falls under the theme of Structural Bioinformatics and seeks to explore more the basis behind Structure-Function relationship in biological context of macromolecules; the proteins in the case of this study. Furthermore, the study draws attention to results that would touch upon distant evolutionary relations across species.

As revealed in the various analysis and deductions made in the Results and Discussions (Chapter III), this project has identified, defined and characterised a set of binding structural and functional motifs associated with a set of biologically important ligands/cofactors. These ligands are relevant to vital biological function including immunity of plants against attacks by pathogenic factors.

The project also identified the residues (amino acids) that are directly involved in the binding of a large set of ligands within the set of proteins selected in the study.

The protein structural elements (α-helices and β-strands) and loop regions that compose the structural binding motifs are considered, by this study, as providing important physical support on which the actual functional elements, i.e. the residues, are mounted, with their individual and collective physical and chemical properties, to carry out the specific biological function of the porphyrin proteins.

The discovery of structural similarity between the ligand binding motifs across distant species and functions is indicative of evolutionary relation between these types of proteins that use similar chemical groups such a the porphyrin planar cycles to achieve different biological functions. This would open further venues of research and discovery in this field of study.

The definition of the ligand binding sites, i.e. the binding structural motifs, and construction of a Flat-File database that would be made accessible online, in foreseen future, and that provide such important data and analysis to researchers in the field would be very useful in deeper analysis of the protein function in health and pathological cases, in studies related to phylogenetic analysis, 3D-structure predictions and rational drug design.

However, such conclusions would better confirmed and further explored using larger data sets of PR-proteins and the ligand types used by them. This is to be planned in future studies.

References

References

[1] : KAKOTI, Nilakhi, BORKATAKY, Munmi, et NATH, Ratul. Understanding Plant Immunity Through Plant-Microbe Interaction. In : Plant Microbiome and Biological Control: Emerging trends and applications. Cham : Springer Nature Switzerland, 2025. p. 19-35.

[2] : Wagner, Stephan (2013). Analysing Novel Structures of Protein Complexes Requiring the Central Plant Immune Regulator EDS1. PhD thesis, Universität zu Köln.

[3] : KROKIDIS, Marios G., KOUMADORAKIS, Dimitrios E., LAZAROS, Konstantinos, et al. AlphaFold3: an overview of applications and performance insights. International Journal of Molecular Sciences, 2025, vol. 26, no 8, p. 3671.

[4]: SELS, Jan, MATHYS, Janick, DE CONINCK, Barbara MA, et al. Plant pathogenesis-related (PR) proteins: a focus on PR peptides. Plant physiology and biochemistry, 2008, vol. 46, no 11, p. 941-950.

[5] : HOUTGAST, Ernst Joachim. Scalability of bioinformatics applications for multicore architectures. MSc. Thesis, Delft University of Technology, 2009.

[6] : OVEK, Damla, KESKIN, Ozlem, et GURSOY, Attila. ProInterVal: Validation of Protein–Protein Interfaces through Learned Interface Representations. Journal of Chemical Information and Modeling, 2024, vol. 64, no 8, p. 2979-2987.

[7] : TAGER, Howard S. Principles of biochemistry. JAMA, 1993, vol. 270, no 16, p. 1989-1990.

[8] : https://byjus.com/biology/proteins-structure-and-functions/

[9] : PETSKO, Gregory A. et RINGE, Dagmar. Structure et fonction des protéines. De Boeck Supérieur, 2008.

[10] : MACHO, Alberto P. et ZIPFEL, Cyril. Plant PRRs and the activation of innate immune signaling. Molecular cell, 2014, vol. 54, no 2, p. 263-272.

[11] : SUDISHA, J., SHARATHCHANDRA, R. G., AMRUTHESH, K. N., et al. Pathogenesis related proteins in plant defense response. Plant defence: biological control, 2012, p. 379-403.

[12] : SCHERER, Nicole M. Pathogenesis-related proteins: phylogenetic characterization. 2010. Thèse de doctorat. Düsseldorf, Univ., Diss., 2010.

[13] : SABOKI EBRAHIM, K. Usha et SINGH, Bhupinder. Pathogenesis related (PR) proteins in plant defense mechanism. Sci. Against Microb. Pathog, 2011, vol. 2, p. 1043-1054.

[14] : KAUR, Anroop, KAUR, Sukhpreet, KAUR, Ajinder, et al. Pathogenesis-related proteins and their transgenic expression for developing disease-resistant crops: strategies progress and challenges. IntechOpen, 2022.

[15] : EDREVA, Aglika. Pathogenesis-related proteins: research progress in the last 15 years. Gen Appl Plant Physiol, 2005, vol. 31, no 1-2, p. 105-124.

[16] : BOUTET, Emmanuel, LIEBERHERR, Damien, TOGNOLLI, Michael, et al. UniProtKB/Swiss-Prot: the manually annotated section of the UniProt KnowledgeBase. In : Plant bioinformatics: methods and protocols. Totowa, NJ : Humana Press, 2007. p. 89-112.

[17] : HOKSZA, David. Similarity Search in Protein Databases. 2010.

[18] : https://study.com/academy/lesson/proteins-structure-function-types.

[19] : https://conductscience.com/structure-and-functions-of-proteins/

[20] : https://microbenotes.com/amino-acids-proteins/

[21] : REGUIG, Sabah, TEMAR, Abir, et RACHEDI, Abdelkrim. Bioconversion of Potato Peels Waste into Starch and Bioplastic.

[22] : ALREFAI, R., ALREFAI, A. M., BENYOUNIS, K. Y., et al. An evaluation of the effects of the potato starch on the biogas produced from the anaerobic digestion of potato wastes. Energies, 2020, vol. 13, no 9, p. 2399.

[23] : AHTI, E. Fertiliser-induced leaching of P and potassium from peatlands drained for forestry. In : Proceedings of the Seventh International Peat Congress. 1984. p. 18-23.

[24] : BAYRAM, Berfin et GREIFF, Kathrin. Life cycle assessment on construction and demolition waste recycling: a systematic review analyzing three important quality aspects. The International Journal of Life Cycle Assessment, 2023, vol. 28, no 8, p. 967-989.

[25] : SHRESTHA, Sarita, KHATIWADA, Janak Raj, SHARMA, Hem Kanta, et al. Bioconversion of fruits and vegetables wastes into value-added products. In : Sustainable bioconversion of waste to value added products. Cham : Springer International Publishing, 2021. p. 145-163.

[26] : RACHEDI, Abdelkrim et BAHLOUL, Ouafaa. Auto Protein Homology Modeling (APHM): An Educational and Research Tool for Homology Molecular Modeling.

[27] : Rachedi, A. (2011). Sequence Structure and Function Server, 2011, https://bioinformaticstools.org/ssfs/

[28] : Rehman I, Farooq M, Botelho S. Biochemistry, Secondary Protein Structure. [Updated 2021 Dec 15]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2022 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK470235/

[29] : Rives A. et al. (2021) Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences, Proc. Natl. Acad. Sci., 118 (15) :e2016239118. Available from https://doi.org/10.1073/pnas.2016239118