Decoding The Role of Uncharacterized *Helicobacter Pylori* Proteins on Differentially Expressed Genes in GC Pathogenesis

By

Shajib Dey
22276016

A thesis submitted to the Department of Mathematics and Natural Sciences in partial fulfillment of the requirements for the degree of
Master of Biotechnology

Mathematics and Natural Sciences
Brac University
09/2024

Declaration

It is hereby declared that

1. The thesis submitted is my own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. I have acknowledged all main sources of help.

Student's Full Name & Signature:

Shajib Dey
22276016

Approval

The thesis/project titled “Decoding the role of uncharacterized *Helicobacter Pylori* proteins on differentially expressed genes in GC Pathogenesis” submitted by

1. Shajib Dey (22276016)

of Summer, 2024 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of Master of Biotechnology on [01-09-2024].

Examining Committee:

Academic Supervisor:
(Member)

____
Fahim Kabir Monjurul Haque, PhD
Associate Professor
Department of Mathematics and Natural Sciences
BRAC University

Research Supervisor:
(Member)

____
Mohammad Uzzal Hossain
Senior Scientific Officer & In-charge
Bioinformatics Division
National Institute of Biotechnology

Program Coordinator:
(Member)

____
Munima Haque, PhD
Associate Professor
Department of Mathematics and Natural Sciences
BRAC University

External Expert Examiner:
(Member)

Mohd. Raeed Jamiruddin, PhD
Associate Professor
School of Pharmacy
BRAC University

Departmental Head:
(Chair)

Md. Firoze H. Haque, PhD
Associate Professor and Chairperson
Department of Mathematics and Natural Sciences
BRAC University

Ethics Statement

This research was conducted in accordance with the ethical standards and guidelines of BRAC University and the National Institute of Biotechnology. All procedures and methodologies were approved by the respective ethics committees of both institutions.

The research was carried out under the supervision of my academic supervisor at BRAC University and my research supervisor at the National Institute of Biotechnology. All participants involved in the study provided informed consent, and their privacy and confidentiality were strictly maintained throughout the research process.

No conflicts of interest were identified in relation to this study. The research was conducted with full respect for ethical considerations, ensuring the integrity and validity of the findings presented in this work.

Abstract/ Executive Summary

Gastric cancer (GC) is one of the most frequent and deadly cancers worldwide. *Helicobacter pylori (H. pylori),* a GC-associated bacterium, is critical to gastric pathogenesis. However, the specific molecular mechanisms remain partially understood. Here, we analyze 8 samples of GC transcriptomics of Homo sapiens and 107 samples of the whole genome of *H. pylori* to discover any potential connections between uncharacterized proteins and differentially expressed genes (DEGs) in GC. We identified 11 hypothetical proteins (HPs) that possess potential pathogenic features. Later, microarray analysis revealed 381 DEGs. Subsequently, the genes *CXCL8*, *ICAM1*, and *CXCR2* were identified as hub genes, recognized for their crucial role in inflammatory pathways in GC. Three HPs of *H. pylori,* MLLMICFO_00840, CHOJIKGH_00797, and CHKOBBCO_01290, showed strong interactions with hub genes *ICAM1*, *LFA-1*, and *CXCL8* from the host. The results demonstrate the robust stability and dynamic behavior of these protein complexes, indicating their possible involvement in regulating immunological responses and contributing to the development of GC. This work reveals the linkages between *H. pylori* and host genes in GC and identifies potential proteins that could serve as indicators or targets for therapy.

Keywords: GC, Helicobacter pylori, host, Differentially expressed genes, Pathogen

Dedication

To my family and friends, who have always supported and encouraged me through every step of this journey. A special dedication to ***Mohammad Uzzal Hossain*** sir, whose guidance and wisdom have been invaluable to my progress, and to my beloved friend ***Muntaha Khan***, for always believing in me and giving me strength. This work is a small token of my appreciation for your faith in me and for always being there.

Acknowledgement

First and foremost, I am deeply grateful to God for granting me the strength, patience, and perseverance to overcome the challenges and complete this dissertation. I am very thankful to ***Dr. Fahim Kabir Monjurul Haque*** *and* ***Mohammad Uzzal Hossain*** *sir, my academic and research supervisors, for their endless inspiration, insightful suggestions, scholastic guidance, unflagging assistance, and excellent counsel from the beginning of this research work to the completion of this thesis paper.*

I am also grateful to all the members of the Bioinformatics Division of the National Institute of Biotechnology, Bangladesh, especially ***Zeshan Mahmud Chowdhury****,* ***Arittra Bhattacharjee****,* ***Md. Shahadat Hossain****, and* ***Naimur Rahman****, for being a constant source of inspiration and guidance throughout my research journey. Their expertise and encouragement played a pivotal role in the successful completion of this work.*

I sincerely thank ***Dr. Firoze Haque****, Associate Professor and Chairperson, Department of Mathematics and Natural Sciences, BRAC University, and* ***Munima Haque****, PhD, Associate Professor, Department of Mathematics and Natural Sciences, BRAC University, for their kind cooperation, unreserved help, and financial support in continuing my thesis work.*

Finally, my sincere regards and most heartfelt gratitude goes out to all of my teachers for their direction in each and every step throughout the entirety of my time spent in university, as well as for their assistance, support, and providing me with perceptive guidance whenever it was required.

Shajib Dey

September 2024

Table of Contents

List of Tables

List of Figures

List of Acronyms

GC	GC
H. pylori	*Helicobacter pylori*
HPs	Hypothetical Proteins
DEGs	Differentially Expressed Genes
PPI	Protein-Protein Interaction
MD	Molecular Dynamics
RMSD	Root Mean Square Deviation
RMSF	Root Mean Square Fluctuation
Rg	Radius of Gyration
SASA	Solvent Accessible Surface Area
CXCL8	C-X-C Motif Chemokine Ligand 8
ICAM1	Intercellular Adhesion Molecule 1
LFA-1	Lymphocyte Function-Associated Antigen 1
CXCR2	C-X-C Motif Chemokine Receptor 2
KEGG	Kyoto Encyclopedia of Genes and Genomes
GO	Gene Ontology
NCBI	National Center for Biotechnology Information
WGS	Whole Genome Sequence
VFDB	Virulence Factor Database
SWISS-MODEL	A web-based integrated service dedicated to protein structure homology modelling
PROC	Program for Conformational Search
SAVES	Structure Analysis and Verification Server
RCSB PDB	Research Collaboratory for Structural Bioinformatics Protein Data Bank
CHARMM-GUI	Chemistry at HARvard Macromolecular Mechanics Graphical User Interface
NVT	Isothermal-Isochoric
NPT	Isobaric
CD-HIT	Cluster Database at High Identity with Tolerance
BLAST	Basic Local Alignment Search Tool

Chapter 1

Introduction

1.1 Background

GC is one of the most common types of cancer worldwide, accounting for 4.9% of all new cases and 6.8% of cancers in 2022 [1]. It affects more males than women and is one of the most common cancers in men [2]. The World Health Organization has established *H. pylori* as a class I carcinogen due to its association with the development of GC (GC) and MALT lymphoma, both of which are serious health risks [3] [4]. This gram-negative and microaerophilic bacterium resides in the acidic environment of the stomach [5]. *H. Pylori* infection is one of the major risk factors for GC. The prevalence of *H. pylori* infection is globally noticeable, affecting approximately 50% of the global population. In developing nations, (30 to 50) % of children acquire *H. pylori* infection, which increases to 90% by adulthood [6]. However, the infection rates fluctuate between 30% and 50% in developed countries [7]. In Bangladesh, *H. pylori* is the major risk factor for GC [8].

As previously reported, some virulence factors in *H. pylori*, such as cytotoxin-associated gene A (CagA) or vacuolating cytotoxin A (VacA) are likely to develop GC [9]. Chronic inflammation and the presence of these factors result in DNA damage to the host cell and stimulate certain pathways that aid *H. pylori* in its survival [5]. The processes frequently interact and mutually reinforce each other [5]. Furthermore, *H. pylori* attaches to the lining of the gastric epithelium using several adhesion factors, such as the blood group antigen-binding adhesin (BabA), sialic acid-binding adhesin (Sabal), outer inflammatory protein A (Ipe), and adherence-associated lipoproteins (AlpA/B) [10]. Treatments for *H. pylori* infections are available, including a 14-day triple therapy programme, quadruple therapy based on bismuth, therapy based on levofloxacin, and therapy based on rifabutin; nonetheless, *H. pylori* eradication remains difficult. Possible causes of these multiples include the rise in antibiotic resistance, the prevalence of co-infections, or the interaction between human and *H. pylori* proteins. The proteins that are already identified can be accurately predicted within the framework of their biological processes. However, the uncharacterized or hypothetical protein found in *H. pylori* may have a significant impact on the development of GC.

HPs refer to the large quantities of uncharacterized proteins found in many bacterial genomes [11]. The scientific community frequently disregards HPs, even though they have great promise. Possible critical roles for these proteins in bacterial virulence and survival, especially in *H. pylori,* have been predicted by genome sequencing but their functions have not been experimentally demonstrated [12]. Treatment attempts can be further complicated when these proteins undergo mutations and genetic interactions with hosts, which can enhance virulence and cause antibiotic resistance. More research into the pathogenicity of *H. Pylori* and better ways to fight infections and cancers linked to them could be possible if future studies characterize HPs functionally; this could lead to the discovery of new therapeutic targets and a better understanding of the disease physiology [13]. Also, specific HPs can potentially influence the relationships between bacteria and their hosts [14]. A large number of DEGs have been linked to complicated disorders, according to recent advances in microarray analysis [15]. This high-throughput approach has helped researchers obtain a greater understanding of the physiological differences between health and sickness [16]. Specifically, genes with larger fold changes are more commonly reported in microarray studies than genes with statistically significant differential expression [17]. By integrating this data with other biological sources, more accurate assessments may be made, such as disease target landscapes [17]. HPs are hypothesized proteins that an organism expresses but for which there is no evidence to support their useful functions; hence, they may provide light on how bacteria endure harsh environments [18]. Finding the links between RNA-seq genes and unidentified *H. pylori* proteins may help shed light on GC. Functional modeling of HPs has helped researchers comprehend the relationship between sequence, structure, and function [19]. Hence, GC may be better understood if the connections between RNA-seq genes and uncharacterized *H. pylori* proteins are found.

Hence, we identify the most prevalent HPs from the whole genomes of 107 *H. pylori* isolates selected from the different countries. We perform the microarray analysis to identify DEGs and their link to HPs using the protein-protein interaction (PPI) network. We also perform molecular docking and dynamics to find interaction and their stability between the host and pathogens. Finally, we provide the biological and molecular insights of the uncharacterized *H. Pylori* proteins on DEGs in GC Pathogenesis.

1.2 Objectives

1. To identify and characterize HPs in *Helicobacter pylori* that may play critical roles in GC pathogenesis.

2. To analyze DEGs associated with severe gastritis and their interactions with *H. pylori* hypothetical proteins.

3. To employ molecular docking and molecular dynamics simulations to investigate the stability and interaction quality between key proteins in *H. pylori* and host GC-related proteins.

4. To explore the potential therapeutic implications of the interactions between *H. pylori* hypothetical proteins and host GC genes, aiming to identify novel targets for drug development.

5. To integrate bioinformatics and experimental approaches to provide a comprehensive understanding of the molecular mechanisms by which *H. pylori* contributes to GC development.

Chapter 2

Literature Review

2.1 Historical Perspectives

Globally, GC, which is also referred to as GC, is a complex malignancy that places a substantial burden on public health. GC has a complex etiology, and its incidence and mortality rates exhibit regional variation. Considerable scientific investigation has been devoted to elucidating the intricacies of GC, resulting in seminal revelations that have profoundly influenced our comprehension of this neoplasm [20].

Globally, GC (GC) is among the most prevalent malignancies. Due to the low rate of routine screening and the inconspicuous symptoms of earlier disease, the majority of patients are diagnosed at advanced stages [21]. In the past few years, systemic treatments for GC, such as chemotherapy, targeted therapy, and immunotherapy, have advanced substantially [21]. Geographic variations contribute to the variation in GC incidence; Eastern Asia (specifically Japan and Mongolia) and Eastern Europe exhibit the greatest incidence rates [21]. In recent years, there has been a gradual increase in the prevalence of GC among young adults (aged <50 years) in countries classified as both high-risk and low risk [21]. A bacterium that has infected humans since the early Stone Age, *H. pylori*, is intricately linked to the development of GC [22]. *H. Pylori* was prevalent among the majority of the global population prior to the twentieth century [22]. The identification of *H. pylori* thirty years ago represents a triumph of contemporary medicine [23]. It significantly altered our comprehension of the pathophysiology of gastroduodenal diseases and resulted in advancements in the management of *H. pylori*-associated illnesses [23]. Barry Marshall and Robin Warren established the correlation between *H. pylori* and peptic ulcers in 1984 [24]. In recognition of this discovery, they were bestowed with the Nobel Prize in physiology or medicine in 2005 [24]. Numerous developments in the treatment of GC and other related conditions have resulted from this discovery.

2.2 Global Burden of GC

GC (GC) is a prevalent malignancy that contributes substantially to the overall cancer burden on a global scale [25]. In 2022, the incidence of GC among both sexes varied across different regions. Africa, with a population of 903 million, reported 33,352 cases. Latin America and the Caribbean, with a population of 904 million, had 74,379 cases. Northern America, with a population of 905 million, documented 29,675 cases. Europe, with a population of 908 million, recorded 135,610 cases. Oceania, with a population of 909 million, reported 3,977 cases. Asia, with a population of 935 million, had the highest number of cases at 691,791. In total, the global incidence of GC was 968,784 cases [26]. International Agency for Research on Cancer (IARC) researchers estimate that from 2022 to 2045, the estimated global mortality due to GC for both sexes across all age groups (0-85+) is projected to rise significantly. The world's population is expected to increase from approximately 7.89 billion in 2022 to 9.47 billion in 2045 [26]. Correspondingly, the number of deaths from GC is anticipated to increase from 660,175 in 2022 to 1,170,708 in 2045, reflecting a 77.3% increase [26]. This rise is attributed entirely to population growth, as the change in the number of cases due to risk remains constant.

2.3 Geographical Disparities

The incidence and mortality rates of GC exhibit substantial regional variations, highlighting notable geographical disparities [27]. In the Southeastern United States, for instance, the municipalities with the highest 5% mortality rates for GC were preponderant [28]. The considerable diversity observed in mortality rates may be attributed to an extensive array of determinants, such as socioeconomic standing, healthcare accessibility, and behavioral patterns [27]. The state of affairs in Asia is especially disconcerting. GC has historically exhibited a disproportionate impact on East Asian populations in comparison to Western nations [28]. China is home to fifty percent of all GC cases reported globally, and among men in Japan, this disease is the most prevalent [28]. Eastern Asia bears the greatest burden of GC, which is also the most prevalent form of cancer in China, Bhutan, Cabo Verde, and Tajikistan [2]. Asian populations, residing in the United States, are especially susceptible to developing GC [29]. GC is significantly more prevalent among the Korean, Vietnamese, Japanese, and Chinese populations in California [29]. According to the latest World Health Organization (WHO) data published in 2020, GC accounted for 6,799 deaths in Bangladesh, representing 0.95% of the total deaths in the country [30]. The age-adjusted death rate for GC in Bangladesh is 5.45 per 100,000 population, placing the country at rank 100 globally in terms of stomach cancer mortality [30].

2.4 GC and *H. pylori* Association

The global burden of GC is well-documented, with *H. pylori* identified as a major risk factor for its development. *H. pylori* infection contributes significantly to the incidence of GC, particularly in regions with high infection rates, such as East Asia and developing countries. The bacterium's role in chronic inflammation and its progression to malignancy has been extensively studied, emphasizing the need for ongoing research into the molecular mechanisms underlying this relationship [31] [32]. Historical perspectives on *H. pylori* research have laid a strong foundation for current studies, particularly the identification of *H. pylori* as a Group 1 carcinogen by the World Health Organization (WHO) [33]. The understanding of how *H. pylori* influence gastric carcinogenesis is crucial for developing targeted therapeutic strategies aimed at reducing the global burden of this disease.

2.5 Significance of Hypothetical Proteins in *H. pylori*

Hypothetical proteins (HPs) constitute a significant portion of bacterial genomes, yet their functions remain largely uncharacterized. In *H. pylori*, HPs are increasingly recognized for their potential roles in virulence, antibiotic resistance, and immune evasion, which are critical factors in the bacterium’s ability to persist within the host and contribute to disease [11] [12].

These proteins, despite being labeled as 'hypothetical,' may play key roles in the complex interactions between *H. pylori* and the gastric epithelium, influencing the progression of chronic gastritis to GC. Understanding these proteins is essential for identifying new therapeutic targets and developing strategies to mitigate the impact of *H. pylori* infection on global health.

2.6 Bioinformatics in *H. pylori* Protein Analysis

The integration of bioinformatics tools into the study of *H. pylori* HPs has proven invaluable for predicting protein functions, identifying differentially expressed genes (DEGs), and constructing protein-protein interaction (PPI) networks. These tools facilitate the systematic analysis of large datasets, enabling researchers to generate hypotheses about the roles of uncharacterized proteins in bacterial pathogenesis [34] [35]. In the context of *H. pylori*, bioinformatics has been instrumental in identifying novel interactions between HPs and host cellular mechanisms, providing insights into the molecular underpinnings of gastric carcinogenesis [36]. The use of such computational approaches is essential for advancing our understanding of microbial genomics and for guiding experimental validation efforts.

2.7 Contextualizing *H. pylori* in Microbial Research

Previous research has established a foundation for understanding the role of uncharacterized proteins in bacterial pathogenesis. For example, studies on *Escherichia coli* have demonstrated how hypothetical proteins can be integral to bacterial survival, virulence, and host interaction [37]. These findings highlight the broader significance of studying HPs across different bacterial species, providing a comparative perspective that enriches the current understanding of *H. pylori*. Moreover, research utilizing molecular docking to explore protein-protein interactions has further validated the roles of HPs in pathogenesis, underscoring the need for continued investigation in this area [38]. Such studies are critical for contextualizing new research and for identifying areas where significant knowledge gaps remain.

2.8 *H. pylori* Infection

Infection with *H. pylori* is a significant risk factor in the pathogenesis of GC [39]. In 2018, *Helicobacter pylori* infection was significantly associated with a substantial number of cancer cases [40]. Specifically, non-cardia GC presented 850,000 new incidences, with 760,000 attributable to the infection, demonstrating a higher prevalence in males (490,000) compared to females (270,000) [40]. Cardia GC exhibited 180,000 new cases, with 36,000 linked to the infection, predominantly affecting males (27,000) over females (8,900) [40]. Furthermore, non-Hodgkin lymphoma of gastric location accounted for 22,000 new cases, with 16,000 due to the infection, showing a slight male predominance (8,700) relative to females (7,600) [40].

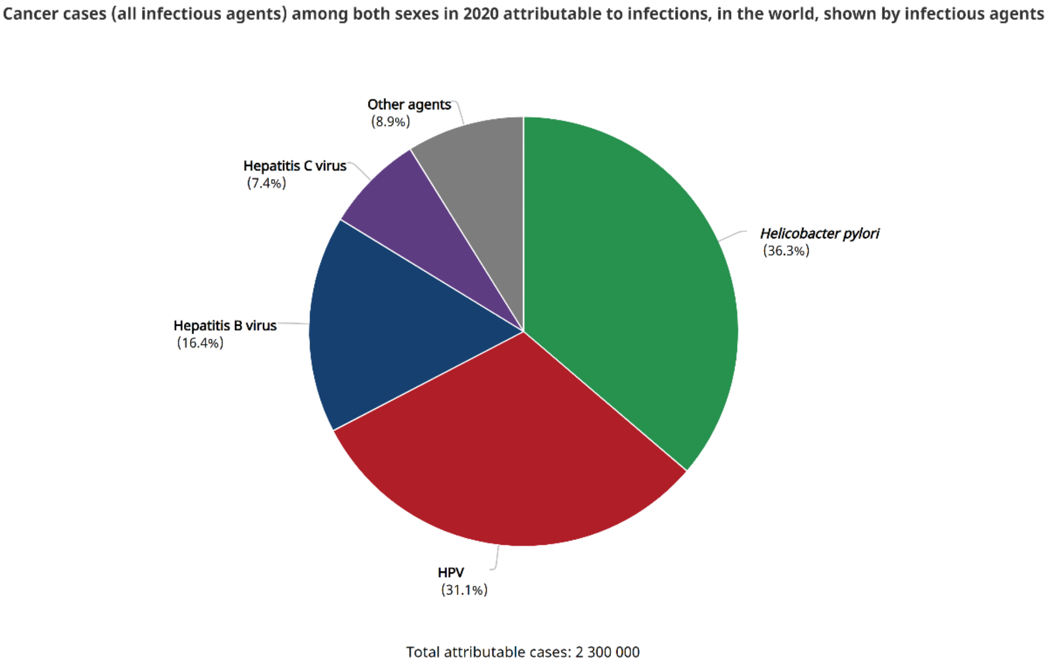

Figure 1: In 2020, the distribution of cancer cases attributable to infections among both sexes globally is illustrated in the pie chart. *The proportions of cancer cases linked to specific infectious agents are as follows: Helicobacter pylori accounts for 36.3% of cases, Human Papillomavirus (HPV) for 31.1%, Hepatitis B Virus for 16.4%, Hepatitis C Virus for 7.4%, and other infectious agents collectively for 8.9% [40] [26].*

H. pylori cause the majority of gastrointestinal ulcers. Additionally, specific strains may increase the tumor risk in the gastric [41]. "The entire process by which *H. pylori* increases the risk of GC is characterized by inflammation," explains Lynch. "*H. pylori* causes an infection that progresses from inflammation to healing to further inflammation." This cycle of constant cell regeneration can eventually lead to errors that develop into cancer [41]. Research indicates that individuals who have contracted *H. pylori* are at a significantly increased risk, up to eight times greater, of developing a specific type of GC [42]. This bacterium, however, is not the only potential cause of GC. Additionally, a history of gastric operations, smoking, and a diet limited in fruits and vegetables can all increase the risk [42].

2.9 Recent Advances and Gaps in Research

Recent advancements in sequencing technologies and bioinformatics have significantly expanded our knowledge of *H. pylori* and its role in GC. These technologies have enabled the identification of novel virulence factors and provided deeper insights into the bacterium's genomic diversity and its interactions with the host [43] [44]. Despite these advancements, there are still considerable gaps in our understanding of the specific functions of many hypothetical proteins within the *H. pylori* genome. Addressing these gaps is essential for developing a comprehensive understanding of *H. pylori* pathogenesis and for identifying new targets for therapeutic intervention. Continued research in this area is crucial for closing these gaps and for advancing the field of microbial pathogenesis.

2.10 Incorporation of Related Pathogens

The study of hypothetical proteins in *H. pylori* can be informed by research on other pathogens where uncharacterized proteins have been implicated in virulence. For example, *Salmonella spp*. and *Staphylococcus aureus* have been shown to utilize HPs in mechanisms critical for infection and survival within the host [45] [46]. These findings suggest that HPs play a universal role in bacterial pathogenesis, making them important targets for study across a range of infectious diseases. By drawing parallels between *H. pylori* and these other pathogens, researchers can better understand the potential roles of HPs and their contributions to disease, thereby broadening the scope of current research.

2.11 Molecular and Genetic Aspects

GC is a group of malignant epithelial tumors that are clinically, biologically, genetically, and microscopically heterogeneous [47]. These tumors arise from a multitude of environmental and genetic factors. Inhereditary GC genes harbor pathogenic variants that are accountable for an estimated 15% to 20% of gastric malignancies [48] [49]. CDH1 and CTNNA1 are two alleles that may increase an individual's susceptibility to hereditary diffuse GC (HDGC) [50] [51].

Hereditary cancer syndromes have the potential to induce additional subtypes of GC. Lynch syndrome, which is induced by pathogenic variants in MLH1, MSH2, MSH6, PMS2, and EPCAM, is among these syndromes [52]. Proximal polyposis of the gastric (GAPPS) and familial adenomatous polyposis (FAP)/gastric adenocarcinoma are both induced by pathogenic variants in APC [53]. Additional syndromes, including Juvenile polyposis syndrome (JPS) and Peutz-Jeghers syndrome (PJS), are attributed to pathogenic variants in SMAD4 and BMPR1A, respectively [54].

The oncogenesis and progression of these malignancies are caused by spurious or inherited mutations in a number of crucial genes that encode the mechanisms accountable for DNA repair and regulate cell growth and differentiation [55]. Chemoresistance induction is a frequent consequence of mutations; this resistance ultimately leads to the failure of therapeutic interventions and the recurrence of tumors [56]. The gastric epithelial cells are eliminated and substituted with cells exhibiting characteristics of the intestines as a result of consecutive mutations. These cells gradually acquire independence, which promotes the progression of malignant alterations (intraepithelial neoplasia) and carcinoma [57].

GC continues to pose a substantial global concern, exhibiting diverse incidence rates and risk factors across distinct geographical areas. Significant advancements have been achieved in the identification of primary causative agents, including *H. pylori* infection and genetic variables, in relation to GC. However, due to the complex and intricate characteristics of this disease, continuous research endeavors are imperative to formulate efficacious approaches for prevention and therapy. The incorporation of molecular insights into therapeutic practice shows potential for tailored methods, presenting a promising opportunity for enhanced results in the battle against GC.

2.12 Molecular Dynamics in Protein Interactions

Molecular dynamics (MD) simulations have become a pivotal tool in the study of protein-protein interactions, offering detailed insights into the stability, flexibility, and dynamics of molecular complexes. In the context of *H. pylori*, MD simulations have been used to validate the interactions between hypothetical proteins and host cellular receptors, providing a molecular-level understanding of how these proteins may contribute to disease progression [58] [59]. The application of MD in this research is particularly valuable for predicting the behavior of uncharacterized proteins under physiological conditions, thus complementing experimental data and guiding the development of therapeutic strategies. This approach enhances the robustness of research findings and underscores the importance of combining computational and experimental methodologies in the study of microbial pathogenesis.

2.13 Therapeutic Potential of *H. pylori* Proteins

The identification of hypothetical proteins as potential virulence factors in *H. pylori* has significant implications for the development of new therapeutic strategies. As antibiotic resistance becomes an increasing concern, alternative approaches that target the underlying mechanisms of bacterial virulence are urgently needed. Studies have shown that targeting bacterial proteins involved in key interactions with the host can be an effective strategy for limiting infection and preventing disease progression [60] [61]. By focusing on the therapeutic potential of HPs, this research contributes to the broader effort to develop innovative treatments for *H. pylori*-associated diseases, including GC. The translational potential of these findings highlights the importance of continued research into the roles of hypothetical proteins in bacterial pathogenesis.

Chapter 3

Methodology

3.1 Microarray data

The gene expression profile of GSE60662 was obtained from the Gene Expression Omnibus database [62]. The GPL13497 platform was used for gene expression profiles. In this study, 8 samples were selected for analysis, comprising 4 control samples (GSM1159807: Control rep1, GSM1159808: Control rep2, GSM1159809: Control rep3, GSM1159810: Control rep4) and 4 severe gastritis samples (GSM1159815: Severe gastritis rep1, GSM1159816: Severe gastritis rep2, GSM1159817: Severe gastritis rep3, GSM1159818: Severe gastritis rep4).

3.2 Identification of differential gene expressions (DEGs)

GEO2R [62] is an interactive web tool for identifying DEGs from the GEO series. Here, the LIMMA package [63] in R software was used to identify up-regulated and down-regulated DEGs between control and severe gastritis groups. Adjustments to P-values were made using the Benjamini & Hochberg method. Auto-detection was applied for the log transformation of the data, while limma precision weights (vooma) were utilized to enhance the precision of the analysis. Normalization was forced. A significant level cut-off of 0.05 was set, with a log2 fold change threshold of 2.

3.3 Protein-Protein Interaction (PPI) network construction and module selection

The PPI network was constructed using the STRING database [64] which identified relationships between proteins encoded by the DEGs. The search was set to the highest confidence score (0.900) to ensure robust and reliable identification of interactions. The PPI network was visualized using the plugin Cytohubba in Cytoscape software [35], which facilitated the identification of genes with high degrees of connectivity, known as hub genes, and bottlenecks genes.

3.4 Functional annotations and pathway enrichment analysis of DEGs

GeneCards is an online database offering researchers a comprehensive set of tools and resources to explore the biological significance of a vast number of human genes [65]. GeneCards integrates data from various sources, such as Gene Ontology (GO), which categorizes genes according to their involvement in biological processes (BP), cellular components (CC), and molecular functions (MF). Additionally, the database provides insights into the Kyoto Encyclopedia of Genes and Genomes (KEGG), facilitating the exploration of gene function annotation in the context of biological pathways and molecular interactions [66]. The hub and bottleneck genes identified and characterized through STRING and Cytoscape analyses were subsequently subjected to functional annotation using GeneCards.

3.5 Retrieval of Whole Genome Sequence (WGS) data

The genomic sequences of 107 *H. pylori* isolates were retrieved from the National Center for Biotechnology Information (NCBI) database [67]. Concerning the Asian continent, a total of five samples were obtained from each of the countries, namely Vietnam (PRJDB3403), Japan (PRJDB4296), China (PRJNA378317), and India (PRJNA419585). In the context of the African continent, a total of nine samples were procured and distributed across Nigeria (PRJEB33903), Egypt (PRJNA689250), and Morocco (PRJNA362473). Data were collected from three European countries, including Denmark (PRJEB37266) (5), Germany (PRJNA490474) (10), and Portugal (PRJNA445654) (5). The North American continent provided a total of 18 samples, consisting of five samples from the United States (PRJNA622860), five from Nicaragua (PRJNA242766), three from Mexico (PRJNA203445), and five from Canada (PRJNA800058). Samples were exclusively obtained from Colombia (PRJNA656306), resulting in a total of 20 samples from the South American continent

Lastly, the Oceania continent was represented by 20 samples obtained from Australia (PRJNA374603).

3.6 Quality control, genome assembly and genome assessment

The Illumina reads were assessed using FastQC to ensure the quality of the samples [68]. Further, the Trimmomatic web tool was used to remove the adapter sequences, and low-quality reads by using a "HEADCROP" value of 15. Simultaneously, throughout the reads, low-quality sections were eliminated by using "SLIDING WINDOW" trimming with a window size of 4 and a threshold of 25 [69]. Unicycler was then used to assemble the bacterial genomes using a mix of short and long reads, producing accurate, comprehensive, and cost-effective assembled reads [70]. The genome of *H. pylori* was annotated using Rapid Prokaryotic Genome Annotation (Prokka) [71].

3.7 Identification of non-paralogous sequences

The paralogous or duplicate HP sequences of *H. pylori* were identified using CD-HIT with a sequence identity cutoff of 0.8 (i.e., 0.8 equivalent to 80%) [72]. CD-HIT is a well-known web tool for clustering and comparing large sets of protein or nucleotide sequences. Among all the proteins, only non-paralogous proteins with a size of >300 amino acids are considered for further analysis.

3.8 Identification of bacterial virulence factor

A customized bash script was used to identify the virulent proteins in *H. pylori*. A new list of virulence proteins was filtered and determined from the Virulence Factor Database (VFDB) [73]. An e-value of 10^{-5} was applied to determine the proteins that exhibited the strongest association with virulence factor. The method assessed possible virulence factors in *H. pylori* and identified proteins that contribute to bacterial pathogenicity. Virulent proteins, as opposed to non-virulent ones, play an important role in initiating serious infection pathways in the body. This results in a wide range of proteome profiles that affect survival, host-pathogen interactions, and virulence levels.

3.9 Validation of *H. pylori* proteins

To confirm that the highly infectious HPs are from *H. pylori*, the proteins were subjected to BLASTp [74] analysis against the reference genome of the *H. pylori* 26695 strain in the NCBI database [67]. In this analysis, a threshold (cutoff value: E-value $\leq$1e-5) was applied to identify the most significant matches in each protein sequence. The *H. pylori* proteins that had the highest number of matches as HPs were selected for further analysis.

3.10 Determination of Physicochemical Properties

The physicochemical properties of HPs were analyzed using ProtParam, an online tool provided by the ExPASy server [34]. This tool calculates various physicochemical parameters of proteins, including molecular weight, theoretical isoelectric point (pI), amino acid composition, and estimated half-life, among others. The analysis aimed to gain insights into the properties of the HPs and their potential roles in biological processes.

3.11 Subcellular Localization

Predicting the location of proteins within cells is vital for genome annotation and the study of bacterial infections since these proteins can be utilized as therapeutic or vaccination targets [75]. The subcellular localization of HPs was predicted using CELLO, an online tool [76] that classifies proteins into different cellular compartments based on their sequence features. This classification helps identify the potential functions of the proteins in specific cellular environments.

3.12 Homology modeling

The three-dimensional (3D) structures of the selected HPs were constructed using the SWISS-MODEL server [77]. To ensure the accuracy and quality of these protein structures, they were evaluated using the PROCHECK [78] and SAVES 6.0 web servers [79], with each structure's integrity being assessed through the generation of a Ramachandran plot. The gene associated with GC that exhibited the highest significance, identified through hub and bottleneck gene analysis via Cytoscape software [35], was chosen for further study. The corresponding ligand, best suited for this gene, was subsequently obtained from the RCSB PDB server. [80].

3.13 Molecular Docking and Molecular Dynamics

Molecular docking analyses were conducted between the top hub and bottleneck genes, *CXCL8* and *ICAM1*, with each HP respectively. This analysis also included the docking of *ICAM1* receptors, such as *LAF-1*, with the HPs. The docking procedures were executed using the GalaxyTongDock server [81], enabling a comprehensive evaluation of interactions between the ligand and each HP separately. Following successful docking, the protein-ligand interactions were visualized utilizing the BIOVIA Discovery Studio Visualizer, which provided detailed insights into the interaction dynamics. For the *CXCL8* and *LFA-1* complexes, 100 ns Molecular dynamics simulation was carried out using the GROningen Machine for Chemical Simulations aka GROMACS (version 2023) [82] for the protein-protein complexes. The large and complex system went under coarse-grained solution simulation using the Martini force field using CHARMM-GUI [83]. Martini3.0.0 model was implemented to conduct the simulation [84]. The protein complex was embedded in a Rectangular box. The water box had edges at a 12 nm distance from the protein surface with 0.15 M NaCl. Following energy minimization, isothermal-isochoric (NVT) equilibration, and Isobaric (NPT) equilibration of the system, a 100 ns molecular dynamic simulation. For the protein complex with *ICAM1*, the protein complexes *ICAM1*_ MLLMICFO_00840, *ICAM1*_CHOJIKGH_00797, and

*ICAM1*_CHKOBBCO_01290 were embedded in 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine (POPC) bilayer using Chemistry at Harvard Macromolecular Mechanics (CHARMM)-Graphical user interface (GUI) [83]. The bilayer system was energetically minimized using CHARMM36m force field [85]. Water box with 11.0 nm length was created on the bilayer surfaces with the TIP3 water model. K+ and Cl- ions were used to neutralize the systems. The results of the simulations were analyzed via the Root Mean Square Deviation (RMSD), Root Mean Square Fluctuation (RMSF), Radius of gyration (Rg), and Solvent Accessible Surface Area (SASA). The plots for each of these studies were produced using the ggplot2 package (https://ggplot2.tidyverse.org/) in R Studio (https://posit.co/). Upon completion of the simulation, the rmsd, rmsf, gyrate, and sasa modules integrated within the GROMACS software were used for the root mean square deviation (RMSD), root mean square fluctuation (RMSF), radius of gyration (Rg), and solvent accessible surface area (SASA) analysis. The ggplot2 package in RStudio was utilized for generating the graphs for each of these analyses. All MD simulations were performed in the high-performance simulation stations running on the Ubuntu 24.04 LTS operating system located at the Bioinformatics Division, National Institute of Biotechnology.

Chapter 4

Result

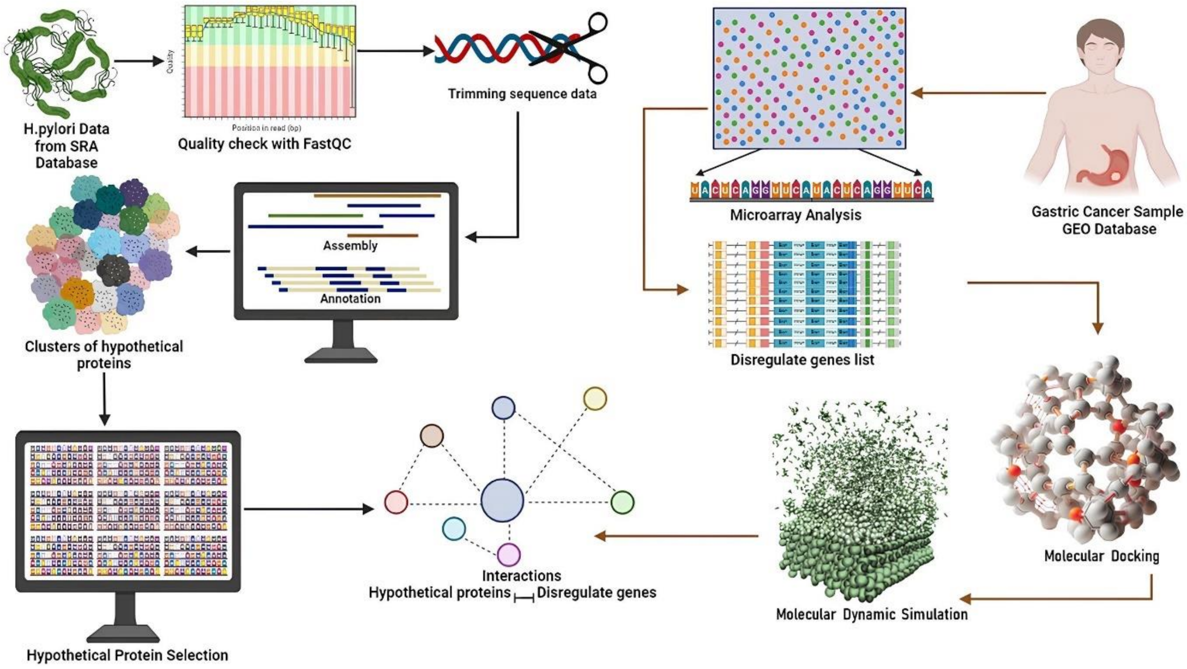
H.pylori Data from SRA Database
Quality check with FastQC
Quality
Position in read (bp)
Trimming sequence data
Assembly
Annotation
Clusters of hypothetical proteins
Hypothetical Protein Selection
Interactions
Hypothetical proteins ⊣ Disregulate genes
Microarray Analysis
Gastric Cancer Sample GEO Database
Disregulate genes list
Molecular Docking
Molecular Dynamic Simulation

Figure 2: Schematic Representation of the work. This figure illustrates the workflow and major findings of the current study. *The process begins with the selection and annotation of the whole genomes of 107 Helicobacter pylori isolated from various regions. The hypothetical proteins (HPs) were identified and subjected to subcellular localization prediction using CELLO. The differentially expressed genes (DEGs) in GC samples were then identified through microarray analysis. The PPI network was constructed using the STRING database and visualized with Cytoscape to identify key hub and bottleneck genes. Molecular docking and molecular dynamics simulations were conducted between these key genes (CXCL8 and ICAM1) and the HPs, along with ICAM1 receptors like LAF-1. The interactions were visualized using BIOVIA Discovery Studio Visualizer, providing insights into the binding affinities and interaction mechanisms.*

4.1 Annotation and assembly lead to high-quality genome sequences

To gain knowledge about the genetic makeup of *H. pylori*, a thorough study was carried out in many steps to create a refined dataset of HP sequences that could have effects on the bacteria's ability to infect others. The data from Prokka [71], showed that 72,268 HPs were found (Supplementary File 1). Strict criteria were used to ensure the quality and usefulness of our dataset. This led to a collection of 61,477 HP sequences (Supplementary File 2). After that, to further optimize the dataset and identify unique sequences, the sequences were clustered into 5,863 effectively (Supplementary File 3).

4.2 Meticulous screening identifies the Putative Virulence Proteins

Virulence factors are the primary cause of bacterial infections. The virulent proteins contribute to the survival of the microbes by facilitating the invasion of the host and the manipulation of the host's immune system [86]. Utilizing the VFDB database [73], we performed a comprehensive analysis that yielded 746 HP sequences with putative virulence traits, suggesting the necessity for further investigation (Supplementary File 4).

To ensure that *Helicobacter pylori* was accurately represented as our target species, a rigorous BLASTp analysis was conducted, resulting in a final dataset containing 11 putative protein sequences [74]. This meticulous screening process resulted in a final dataset containing 11 putative protein sequences that met our stringent criteria as 100% HPs and held substantial promise for further investigation within the context of *H. pylori* virulence (Table 1) and their protein sequences are given in supplementary file 5.

Table 1. Top Hypothetical proteins

1. JGMOFNOI_01064	2. PLLHEGBO_01468
3. FMJNBOFJ_00132	4. ANOHMNDP_00326
5. JALKEJKI_01468	6. PCGEIBGP_00353
7. CHOJIKGH_00797	8. BLHMJNDD_00173
9. CHKOBBCO_01290	10. MLLMICFO_00840
11. LBHCEKMO_01418	

4.3 Physicochemical properties among the HPs suggest diverse potential roles

The ProtParam tool [34] was used to analyze the physicochemical properties of the HPs, providing insights into their characteristics and potential roles in cellular processes (**Table 2**). The proteins varied in terms of amino acid length, molecular weight, isoelectric point (pI), instability index (II), and grand average of hydropathicity (GRAVY). The molecular weights of the HPs ranged from 55,050.91 to 83,015.61 Daltons. Notably, proteins with higher molecular weights included MLLMICFO_00840 (83,015.61 Da), CHKOBBCO_01290 (79,482.71 Da), and BLHMJNDD_00173 (74,425.31 Da), while the lowest molecular weight was observed in JGMOFNOI_01064 (55,050.91 Da). The theoretical isoelectric points (pI) varied significantly, ranging from 4.25 to 9.07. The protein load is determined by the pI. When a direct current passes through the protein at this pH, it has no charge and does not move in the electric field [87]. When conducting in vivo experiments, MW and pI values are crucial for crystallization and purification. The half-life of all the proteins was consistent at 30 hours, indicating similar protein turnover rates across the dataset. The instability index (II) varied among the proteins, with scores ranging from 25.33 to 65.67. A protein is considered stable if its instability index value is less than 40, and unstable if it is greater than 40 [88]. JGMOFNOI_01064, CHOJIKGH_00797, CHKOBBCO_01290, and MLLMICFO_00840 were the most stable proteins. The grand average of hydropathicity (GRAVY) ranged from -0.417 to -1.378, reflecting the proteins' overall hydrophobicity or hydrophilicity. These variations in physicochemical properties among the HPs suggest diverse potential roles and functions within cellular environments, which may have implications for their biological activity and interactions.

Table 2. Physicochemical Properties of Hypothetical Proteins

Hypothetical proteins	No. of Amino acid	Molecular Weight	Half Life	pI	Instability Index (II)	Grand average of hydropathicity (GRAVY)
JGMOFNOI _01064	481	55050.91	30 h	9.07	32.16	-0.790
PLLHEGBO _01468	504	58726.49	30 h	4.25	65.67	-1.315
FMJNBOFJ _00132	551	64247.54	30 h	4.29	63.16	-1.378
ANOHMNDP_00326	481	55589.24	30 h	4.32	58.19	-1.273
JALKEJKI_ 01468	528	61143.02	30 h	4.25	63.73	-1.341
PCGEIBGP_ 00353	484	56195.87	30 h	4.31	55.25	-1.294
CHOJIKGH _00797	662	73650.44	30 h	8.58	25.33	-0.417
BLHMJNDD_00173	672	74425.31	30 h	7.56	26.70	-0.357
CHKOBBCO_01290	711	79482.71	30 h	5.39	28.45	-0.399
MLLMICFO _00840	751	83015.61	30 h	5.47	32.00	-0.380
LBHCEKMO_01418	484	55717.50	30 h	8.91	42.21	-0.909

4.4 Subcellular Localization of identified proteins confer bacterial survival and pathogenicity

The subcellular localization of the HPs was predicted using CELLO [76]. The predicted protein localizations suggested that most of the proteins were located in the outer membrane (45%) and cytoplasm (45%) and had diverse functional roles in different cellular compartments. Proteins JGMOFNOI_01064, CHOJIKGH_00797, BLHMJNDD_00173, and

CHKOBBCO_01290 were predicted to localize to the outer membrane with localization scores ranging from 1.698 to 3.581. The localization to the outer membrane suggests roles in cell-cell communication, environmental sensing, and interactions with external factors, which are vital for bacterial survival and pathogenicity. Another subset of the proteins was localized to the cytoplasm (Table 3), including PLLHEGBO_01468, FMJNBOFJ_00132, ANOHMNDP_00326, JALKEJKI_01468, and PCGEIBGP_00353, with localization scores ranging from 1.413 to 2.068. Cytoplasmic proteins are likely involved in essential intracellular processes such as metabolic pathways, signaling cascades, and structural functions, highlighting their importance in maintaining cellular homeostasis. Additionally, the HP LBHCEKMO_01418 was predicted to localize to the periplasmic space (Table 3), with a localization score of 1.818. Periplasmic proteins often play roles in nutrient transport, enzyme activity, and stress response, which could be critical for bacterial survival and pathogenicity.

Table 3. Subcellular Localization

Hypothetical Proteins	Localization	Localization Score
JGMOFNOI_01064	OuterMembrane	1.698
PLLHEGBO_01468	Cytoplasmic	1.655
FMJNBOFJ_00132	Cytoplasmic	1.413
ANOHMNDP_00326	Cytoplasmic	2.068
JALKEJKI_01468	Cytoplasmic	1.463
PCGEIBGP_00353	Cytoplasmic	1.755
CHOJIKGH_00797	OuterMembrane	3.354
BLHMJNDD_00173	OuterMembrane	3.581
CHKOBBCO_01290	OuterMembrane	3.562
MLLMICFO_00840	OuterMembrane	3.069
LBHCEKMO_01418	Periplasmic	1.818

4.5 DEGs analysis identifies 381 dysregulated genes

A total of 381 DEGs were identified out of 33746 genes between the control and severe gastritis groups. The analysis revealed distinct patterns of gene expression changes between the two groups. Of the 381 DEGs, 20 were upregulated, while 361 were downregulated in the severe gastritis samples compared to the control samples. The complete list of DEGs is presented in Supplementary File 6. Based on the analysis between the control and test groups, the differences between the groups are shown in Figures 3 and 4.

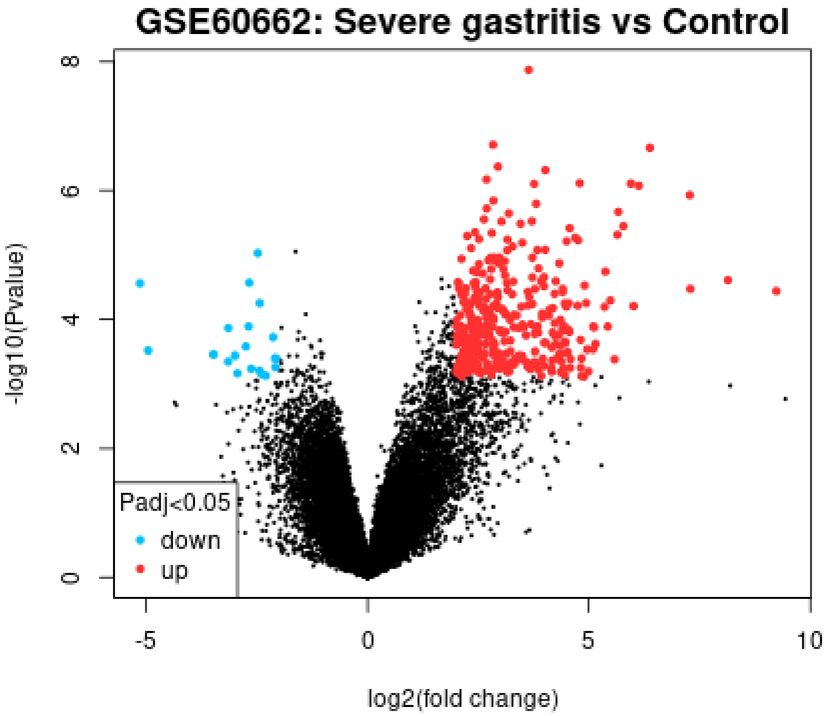

Figure 3: Volcano Plot of DEGs. *The volcano plot presents the relationship between the log2 fold change and the significance (p-value) of each gene. Genes that were significantly upregulated or downregulated in severe gastritis samples compared to control samples are shown as points on the plot. Genes with a log2 fold change threshold of 2 and a p-value cut-off of 0.05 are highlighted, providing a clear visual representation of the most differentially expressed genes.*

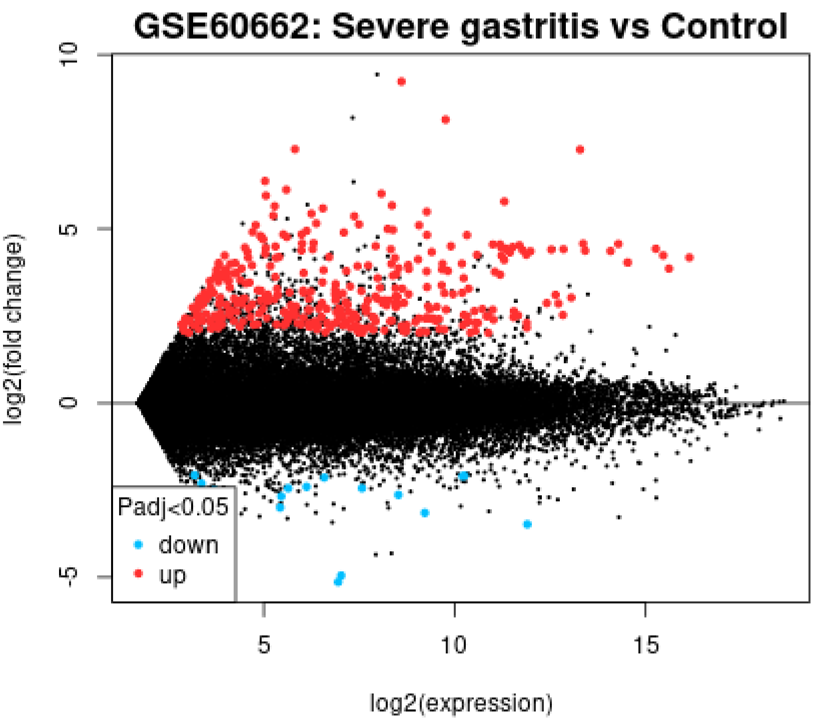

Figure 4: The mean-difference plot. *Also known as an MA plot, it displays the average expression levels (mean) against the fold change (difference) for each gene. This plot helps identify trends in gene expression changes and highlights any genes with particularly high or low expression differences between the groups. The DEGs of interest, determined by the significance and fold change criteria, are emphasized in the plot.*

4.6 PPI Network refers to significant hub genes and bottleneck genes in dysregulated genes

PPI explores complex relationships within the gene set and provides insights into potential regulatory mechanisms in severe gastritis and GC. Hub genes were determined based on their high degrees of connectivity, indicating their central roles in the network (Figure 5 and Table 4). The top 10 hub genes identified were *CXCL8, CXCR2, CCL20, CD74, CCR1, CXCL1, HLA-DMA, CCR7, CCL3 and HLA-DPA1. CXCL8* emerged as the top hub gene with a connectivity score of 12, followed by *CXCR2*, *CCL20*, and *CD74* with a score of 11 each. Bottleneck genes were identified based on their high betweenness centrality, which highlights their critical role in maintaining network connectivity (Figure 6 and Table 5). The top 10 bottleneck genes include *CXCL8*, *CD19*, *ICAM1*, *VAV1*, *CXCR2*, *CD74*, *RAC2*, *PTPRC*, *SELL and ITGB2. CXCL8* was the most critical bottleneck gene with a score of 37. A summary of the 10 hub genes and 10 bottleneck genes are provided in tables 4 and 5 respectively.

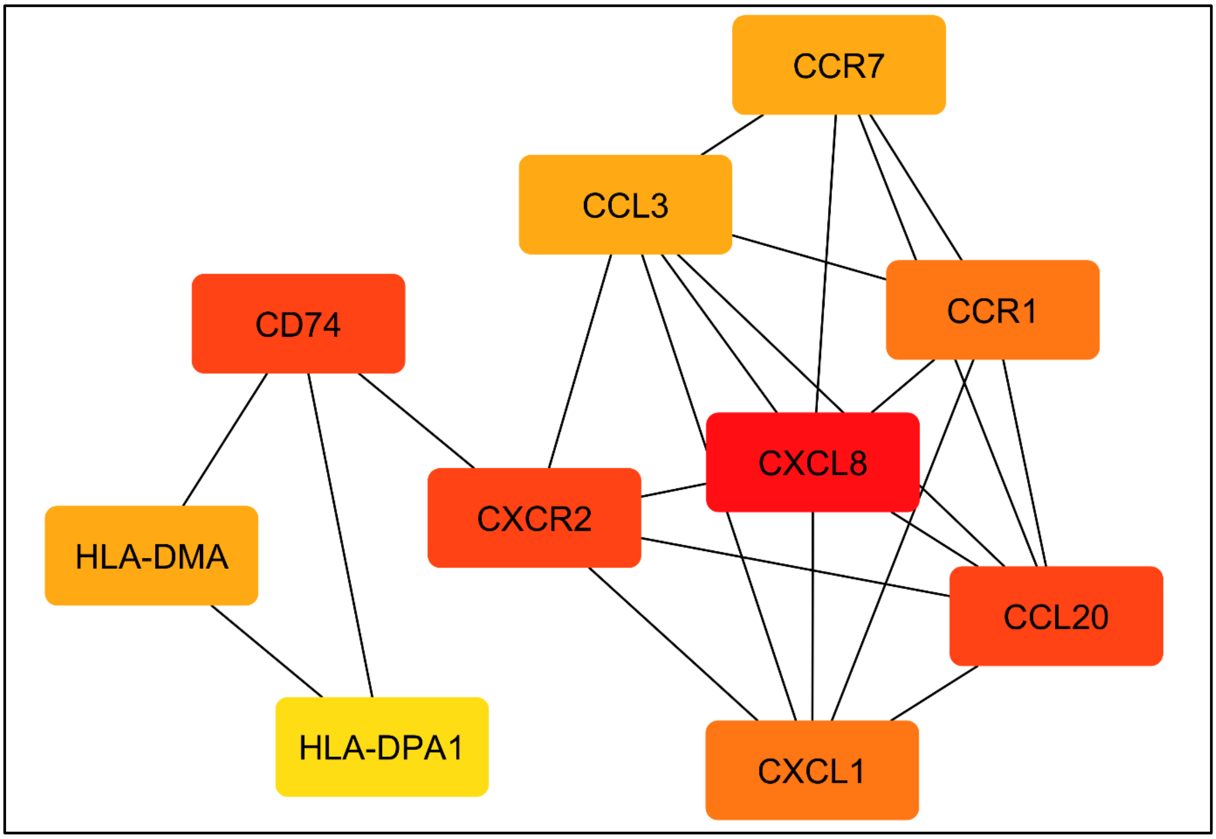

Figure 5: Identification of the hub genes from host. *This figure illustrates the hub genes network identified through the PPI analysis using the STRING database and visualized in Cytoscape. The top 10 hub genes, including CXCL8, CXCR2, and CCL20, are shown with their connectivity scores, highlighting their central roles in the network.*

Table 4. Top 10 Hub Genes and their functions

Rank	**Name**	**Score**	**Full name**	**Function**
1	*CXCL8*	12	C-X-C Motif Chemokine Ligand 8	Plays a significant role in immune regulation and cellular signaling.
2	*CXCR2*	11	C-X-C Motif Chemokine Receptor 2	Activates phosphatidylinositol-calcium signaling pathways.

2	*CCL20*	11	C-C Motif Chemokine Ligand 20	Involves in the recruitment of IL-17 producing Th17 cells and regulatory T-cells to sites of inflammation.
2	*CD74*	11	CD74 Molecule	Plays a crucial role in MHC class II antigen processing.
5	*CCR1*	10	C-C Motif Chemokine Receptor 1	Affects stem cell proliferation.
5	*CXCL1*	10	C-X-C Motif Chemokine Ligand 1	Plays a role in inflammatory processes linked to tumorigenesis.
7	*HLA-DMA*	9	Major Histocompatibility Complex, Class II, DM Alpha	Plays a critical role in catalyzing the release of class II-associated invariant chain peptide (CLIP) from MHC class II molecules.
7	*CCR7*	9	C-C Motif Chemokine Receptor 7	Mediates immune cell trafficking and inflammation within the tumor microenvironment.
7	*CCL3*	9	C-C Motif Chemokine Ligand 3	Binds to CCR1, CCR4, and CCR5 receptors, suggesting a potential role in modulating immune responses.
10	*HLA-DPA1*	8	Major Histocompatibility	Presents antigens derived from endocytosed proteins on the surface of antigen-

			Complex, Class II, DP Alpha 1	presenting cells within the gastrointestinal tract.

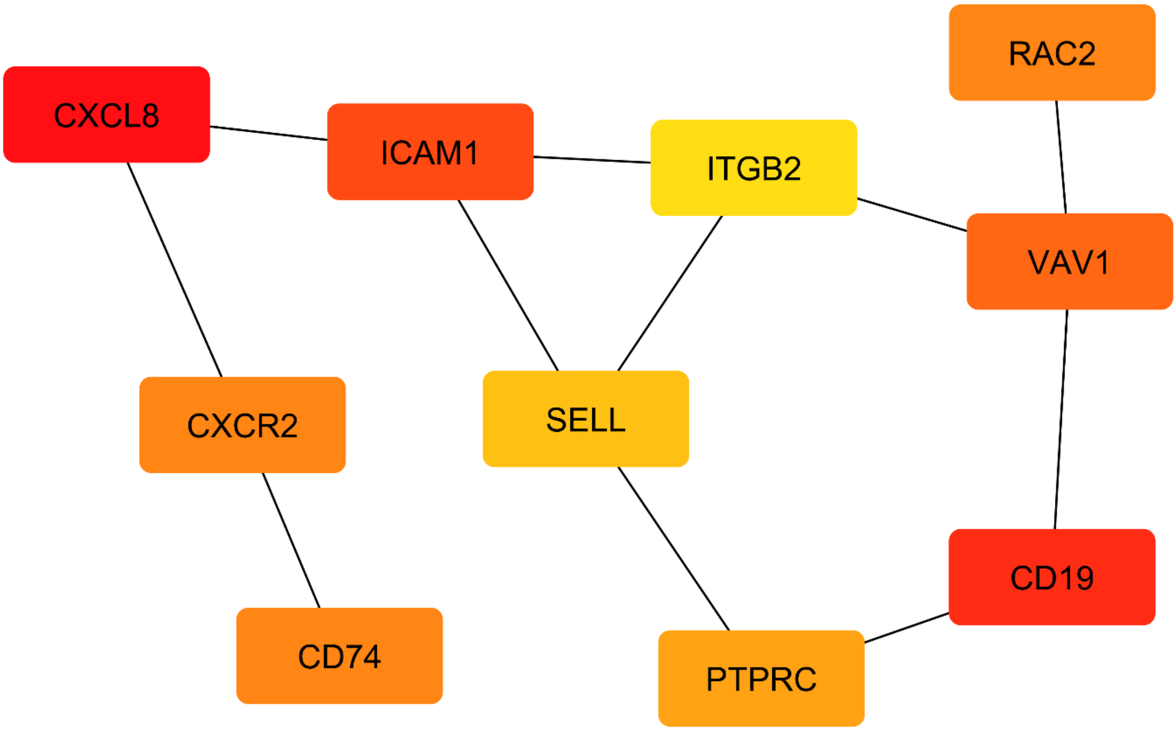

Figure 6: Identification of the Bottleneck Genes from host. *Network figure displays the identified bottleneck genes based on their high centrality, indicating their critical role in maintaining network connectivity. The top 10 bottleneck genes, such as CXCL8, ICAM1, and CD19 are shown. These genes are crucial for understanding the regulatory mechanisms in severe gastritis and gastric cancer.*

Table 5. Top 10 Bottleneck Genes and their functions

Rank	**Name**	**Score**	**Full name**	Function
1	*CXCL8*	37	C-X-C Motif Chemokine Ligand 8	Attracts neutrophils and activates them, playing a significant role in immune regulation and cellular signaling.
2	*CD19*	34	CD19 Molecule	Activates and differentiates B-cells, crucial processes for immune responses.

3	*ICAM1*	30	Intercellular Adhesion Molecule 1	Facilitates leukocyte adhesion and migration across endothelial cells.
4	*VAV1*	20	Vav Guanine Nucleotide Exchange Factor 1	Activates the Rho/Rac GTPases.
5	*CXCR2*	18	C-X-C Motif Chemokine Receptor 2	Activates phosphatidylinositol-calcium signaling pathways.
5	*CD74*	18	CD74 Molecule	Stabilizes peptide-free class II alpha/beta heterodimers and facilitates their transport to the endosomal/lysosomal system for MHC class II antigen processing.
5	*RAC2*	18	Rac Family Small GTPase 2	Regulates cellular processes such as epithelial cell polarization and reactive oxygen species production.
8	*PTPRC*	13	Protein Tyrosine Phosphatase Receptor Type C	Regulates T-cell activation and coactivation.
9	*SELL*	12	Selectin L	Mediates the initial tethering and rolling of leukocytes in endothelial cells.
10	*ITGB2*	11	Integrin subunit beta 2	Facilitates leukocyte adhesion and transmigration.

4.7 Molecular Docking

Among all the eleven virulent HPs, only three MLLMICFO_00840, CHOJIKGH_00797, and CHKOBBCO_01290 interacted with the *CXCL8*, *ICAM1* and the *LFA-1* provided in Table 6. The protein-ligand interactions are represented in Figure 7. The binding affinity data between the hypothetical proteins and GC-related genes, specifically *CXCL8*, *ICAM1*, and *LFA-1*, provide insightful comparisons. For *CXCL8*, the MLLMICFO_00840 protein exhibits the highest binding affinity (1131.364) with a cluster size of 12, indicating a strong interaction, followed by CHKOBBCO_01290 (1071.229) with a cluster size of 7, and CHOJIKGH_00797 (1014.701) with a cluster size of 5. When examining *ICAM1*, CHOJIKGH_00797 stands out with the highest binding affinity (1257.955) and a cluster size of 7, marginally surpassing MLLMICFO_00840 (1250.112) and CHKOBBCO_01290 (1246.485), both of which also show considerable binding affinity with cluster sizes of 9. In the case of *LFA-1*, CHOJIKGH_00797 again demonstrates the strongest interaction with a binding affinity of 1146.776 and a cluster size of 10, followed by CHKOBBCO_01290 (1084.333) with a cluster size of 6, and MLLMICFO_00840 (1031.181) with a cluster size of 7. This analysis highlights the consistent performance of CHOJIKGH_00797 across different interactions, particularly with ICAM1 and *LFA-1*, while MLLMICFO_00840 shows a notable affinity with *CXCL8*.

Table 6. Binding affinity between the hypothetical proteins and GC genes

Ligand	Hypothetical proteins	Binding affinity	Cluster size
CXCL8	CHKOBBCO_01290	1071.229	7
	CHOJIKGH_00797	1014.701	5
	MLLMICFO_00840	1131.364	12

ICAM1	CHKOBBCO_01290	1246.485	9
	CHOJIKGH_00797	1257.955	7
	MLLMICFO_00840	1250.112	9
LFA-1	CHKOBBCO_01290	1084.333	6
	CHOJIKGH_00797	1146.776	10
	MLLMICFO_00840	1031.181	7

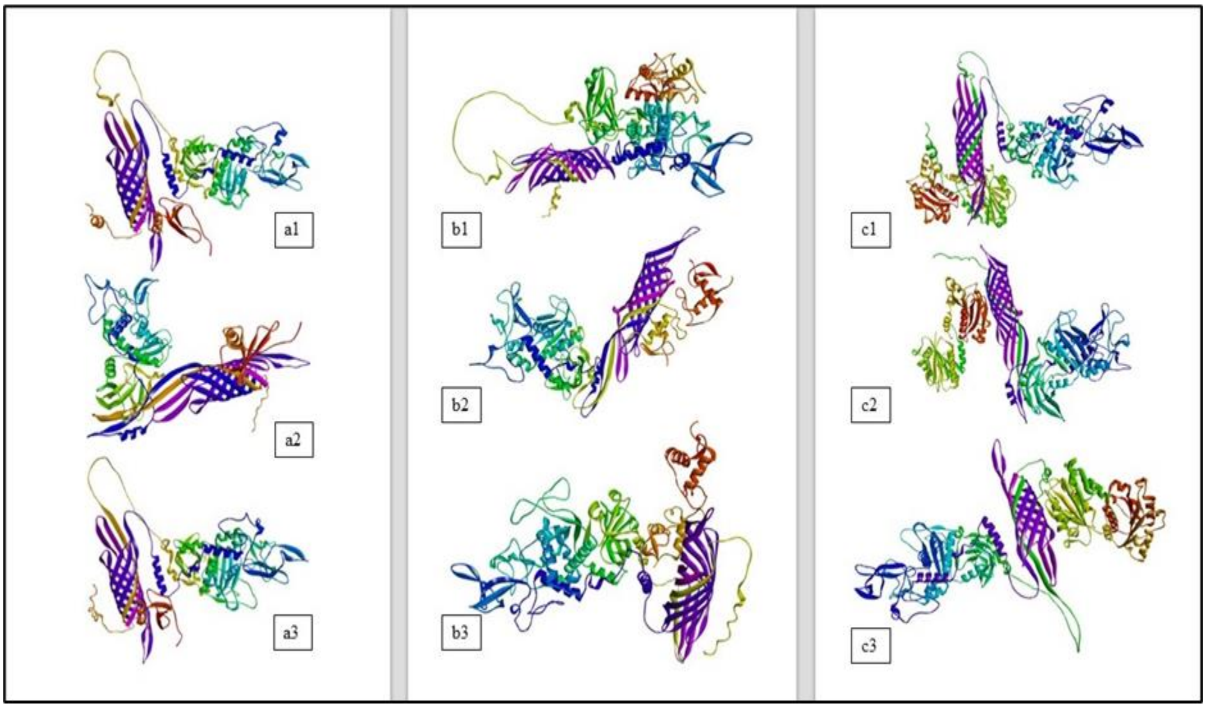

Figure 7: Interacting complexes between host and pathogen. *Interaction between the HPs in H. pylori and CXCL8(a), ICAM1(b) and its receptor, LAF1(c). CXCL8 with CHKOBBCO_01290 (a1), CXCL8 with CHOJIKGH_00797 (a2), CXCL8 with MLLMICFO_00840 (a3). ICAM1 with CHKOBBCO_01290 (b1), ICAM1 with CHOJIKGH_00797 (b2), ICAM1 with MLLMICFO_00840 (b3). LAF1 with CHKOBBCO_01290 (c1), LAF1 with CHOJIKGH_00797 (c2), LAF1 with MLLMICFO_00840 (c3).*

4.8 Molecular Dynamic Simulation

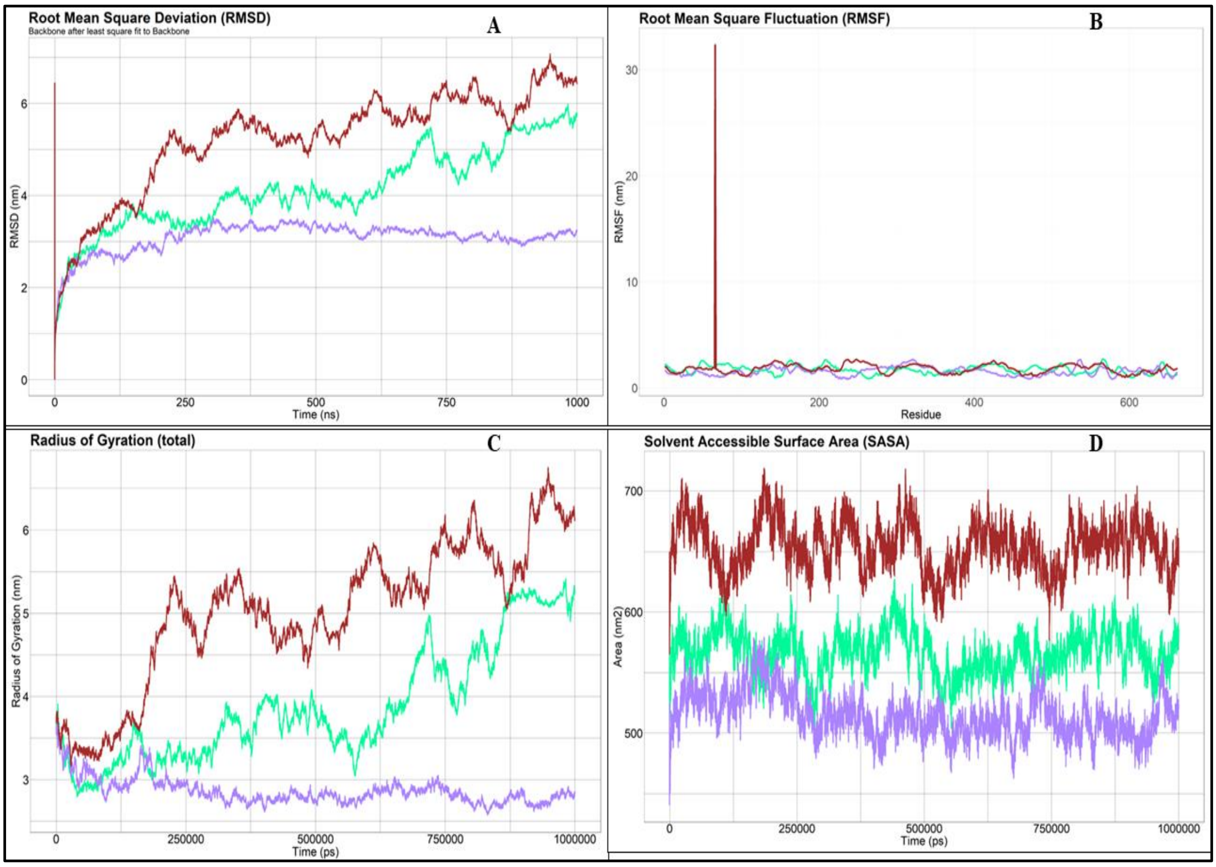

Figure 8: Molecular Dynamics Simulations of CXCL8 Complexed with Hypothetical Proteins. *(A) RMSD profiles of CXCL8 complexes with hypothetical proteins CHKOBBCO_01290, CHOJIKGH_00797, and MLLMICFO_00840, demonstrating the stability of these complexes over the simulation period. (B) RMSF analysis indicating the flexibility of the CXCL8 residues within the protein complexes. (C) Radius of gyration (Rg) showing the compactness of the CXCL8 complexes, reflecting changes in structural integrity. (D) Solvent Accessible Surface Area (SASA) indicating the extent of solvent exposure of the protein surfaces during the simulation.*

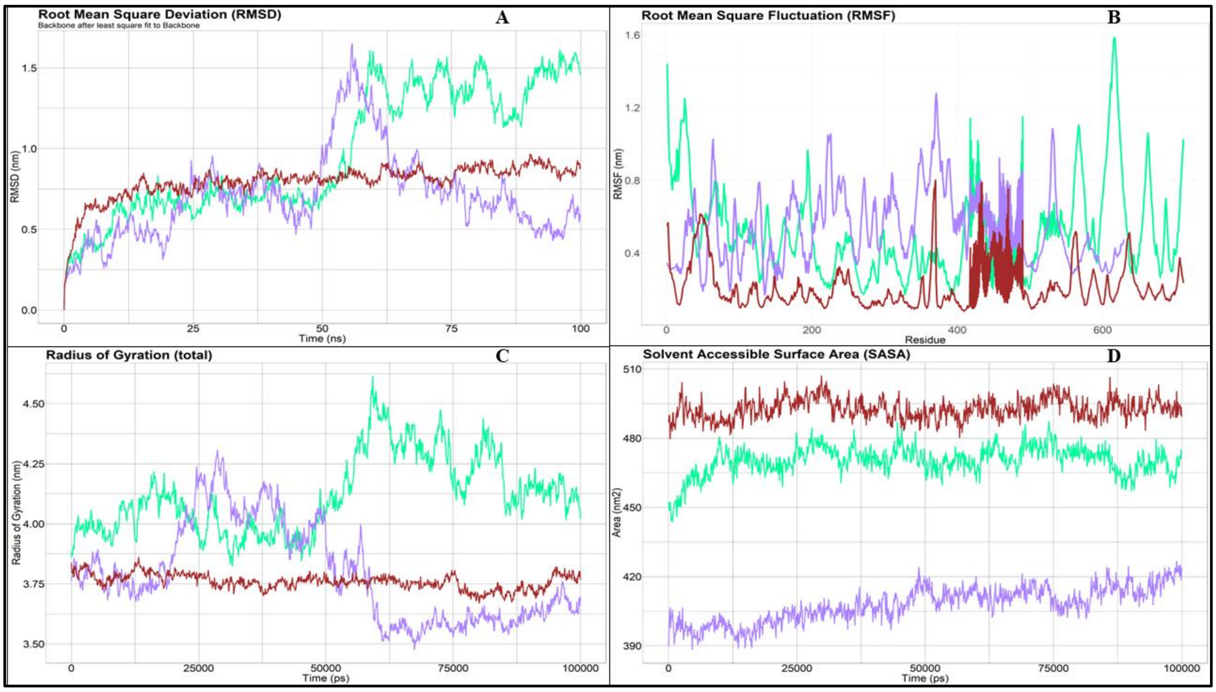

Figure 9: Molecular Dynamics Simulations of ICAM1 Complexed with Hypothetical Proteins. *(A) RMSD, (B) RMSF, (C) Rg, and (D) SASA profiles for Intercellular Adhesion Molecule 1 (ICAM1) in association with hypothetical proteins CHKOBBCO_01290, CHOJIKGH_00797, and MLLMICFO_00840. The data highlight varying degrees of structural stability, flexibility, compactness, and solvent exposure within the ICAM1 complexes throughout the simulation.*

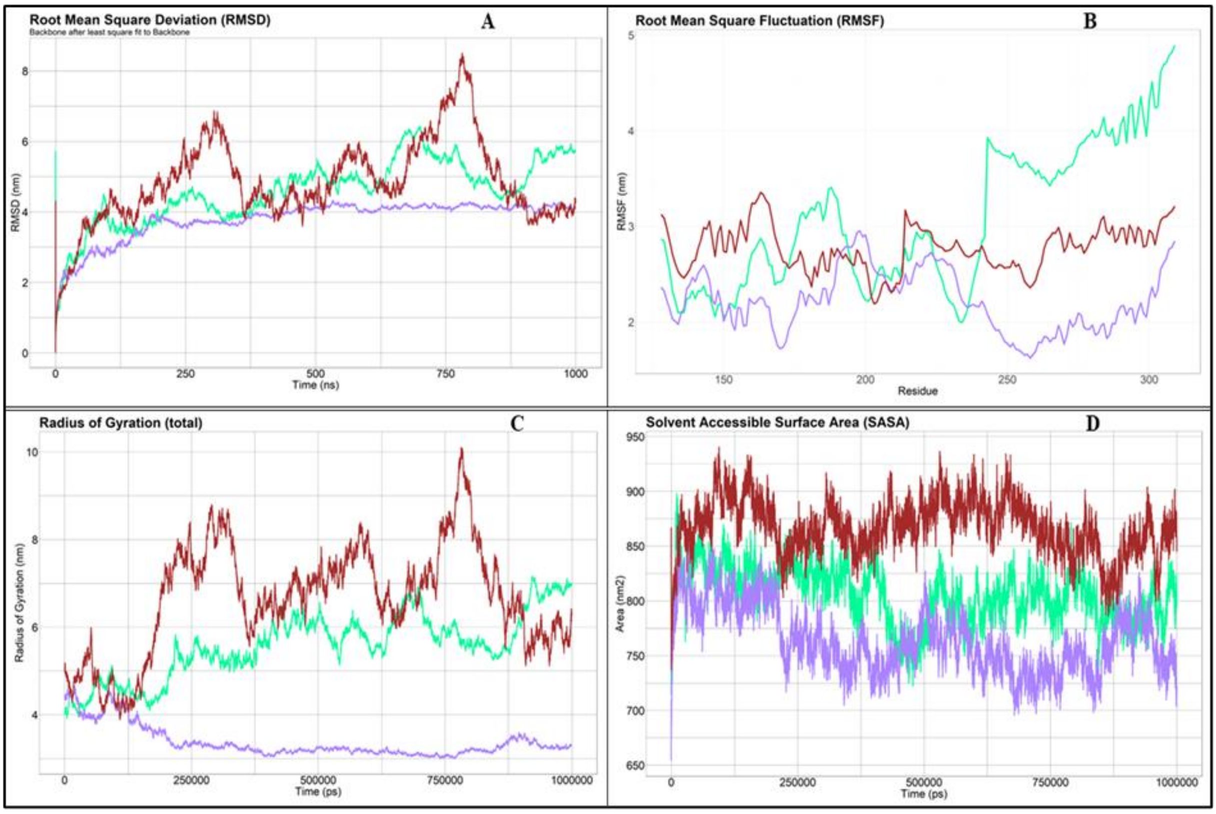

Figure 10: Molecular Dynamics Simulations of LFA-1 Complexed with Hypothetical Proteins. *(A) Root Mean Square Deviation (RMSD), (B) Root Mean Square Fluctuation (RMSF), (C) Radius of Gyration (Rg), and (D) Solvent Accessible Surface Area (SASA) profiles for Lymphocyte Function-Associated Antigen 1 (LFA-1) in complex with hypothetical proteins CHKOBBCO_01290, CHOJIKGH_00797, and MLLMICFO_00840. The figure reveals differences in structural stability, regional flexibility, compactness, and solvent exposure among the complexes.*

The RMSD profile of C-X-C Motif Chemokine Ligand 8 (*CXCL8*) with hypothetical proteins CHKOBBCO_01290 (spring green), CHOJIKGH_00797 (purple), and MLLMICFO_00840 (brown) has been demonstrated in Figure 8 (A). The RMSD values for all three protein complexes are initially increasing, showing that the molecules are deviating from their initial structures. After that, the molecules reached a plateau, indicating that they had reached a stable structure. The RMSD value for *CXCL8*_MLLMICFO_00840 was the highest compared to *CXCL8*_CHOJIKGH_00797 and *CXCL8*_CHKOBBCO_01290, suggesting larger structural fluctuations in *CXCL8*_MLLMICFO_00840. The RMSF profiles of C-X-C Motif Chemokine Ligand 8 (*CXCL8*) with hypothetical proteins CHKOBBCO_01290 (spring green), CHOJIKGH_00797 (purple), and MLLMICFO_00840 (brown) have been demonstrated in Figure 8 (B). The RMSF profile demonstrated several peaks throughout the simulation. It did not show any specific patterns. The RMSF graph displays many peaks, which signify areas of increased flexibility, often located on the protein's surface or inside loop regions. The majority of the protein has low RMSF values, indicating a very stable structure, often associated with alpha-helices and beta-sheets. The Rg profile of C-X-C Motif Chemokine Ligand 8 (*CXCL8*) with hypothetical proteins CHKOBBCO_01290 (spring green), CHOJIKGH_00797 (purple), and MLLMICFO_00840 (brown) has been demonstrated in Figure 8 (C). Rg value increased throughout the simulation. The Rg value of the MLLMICFO_00840 and *CXCL8* complex is gradually increasing, which means the complexes are unfolding over time. The compactness of the *CXCL8*_CHKOBBCO_01290 complex shows a dynamic system with both increasing and decreasing trends. The *CXCL8*_CHOJIKGH_00797 complex exhibits a stable behavior with relatively minor fluctuations. The SASA profile of C-X-C Motif Chemokine Ligand 8 (*CXCL8*) with hypothetical proteins CHKOBBCO_01290 (spring green), CHOJIKGH_00797 (purple), and MLLMICFO_00840 (brown) has been shown in Figure 8 (D). According to the SASA analysis, all six complexes exhibit variations, suggesting that conformational changes in the molecules impact their exposure to the solvent. Throughout the simulation, the *LFA-1*_MLLMICFO_00840 and *CXCL8*_MLLMICFO_00840 show the greatest SASA values, indicating that they have the most solvent-exposed surface area. The *LFA-1*_CHKOBBCO_01290, *LFA-1*_CHOJIKGH_00797, *CXCL8*_CHKOBBCO_01290, and *CXCL8*_CHOJIKGH_00797, on the other hand, show smaller solvent-accessible surfaces due to their lower SASA values.

The RMSD profile of the Intercellular Adhesion Molecule 1 (*ICAM1*) with hypothetical proteins CHKOBBCO_01290 (spring green), CHOJIKGH_00797 (purple), and MLLMICFO_00840 (brown) is depicted in Figure 9 (A). The RMSD values indicate the structural stability of the protein complexes over time. The *ICAM1*_MLLMICFO_00840 complex (brown) shows a significant increase in RMSD value from 200 ns to 850 ns, indicating substantial conformational changes and instability during this period, followed by a plateau, suggesting that the structure stabilized in the later stages of the simulation. The *ICAM1*_CHKOBBCO_01290 complex (spring green) demonstrates a relatively stable RMSD throughout the simulation, indicating that the structure of this complex remained consistent with only minor fluctuations. On the other hand, the *ICAM1*_CHOJIKGH_00797 complex (purple) shows moderate RMSD fluctuations, suggesting that while there were some structural changes, the overall conformation remained relatively stable. The RMSF profile of ICAM1 with hypothetical proteins CHKOBBCO_01290, CHOJIKGH_00797, and MLLMICFO_00840, as shown in Figure 9 (B), provides insight into the flexibility of different regions within the protein complexes. The RMSF values for the *ICAM1*_MLLMICFO_00840 complex (brown) are higher at several points along the residue positions, indicating regions with increased flexibility, especially at the terminal regions. The *ICAM1*_CHKOBBCO_01290 complex (spring green) exhibits relatively lower RMSF values, suggesting a more stable structure with less flexible regions. The *ICAM1*_CHOJIKGH_00797 complex (purple) shows moderate flexibility, with some peaks corresponding to loop regions or areas of structural variation. The overall RMSF trends suggest that the *ICAM1*_MLLMICFO_00840 complex is the most flexible, with *ICAM1*_CHKOBBCO_01290 being the most stable among the three complexes. The Rg profile of ICAM1 in complex with the three hypothetical proteins, as illustrated in Figure 9 (C), shows the degree of compactness of the structures. The *ICAM1*_MLLMICFO_00840 complex (brown) demonstrates notable fluctuations in the Rg values, indicating periodic changes in the compactness of the structure, possibly due to unfolding and refolding events during the simulation. The *ICAM1*_CHKOBBCO_01290 complex (spring green) exhibits a relatively stable Rg value, suggesting a consistently compact structure with minor variations. Meanwhile, the *ICAM1*_CHOJIKGH_00797 complex (purple) shows a gradual increase in the Rg value, indicating a slight unfolding trend over time. These Rg profiles suggest that while *ICAM1*_MLLMICFO_00840 underwent significant structural changes, *ICAM1*_CHKOBBCO_01290 maintained a more compact and stable conformation. The SASA profile, depicted in Figure 9 (D), reveals the extent to which the protein complexes are exposed to the solvent, which is critical for understanding protein stability and interaction with the surrounding environment. The *ICAM1*_MLLMICFO_00840 complex (brown) displays the highest SASA values, suggesting that this complex has the most exposed surface area, which correlates with its higher flexibility and larger conformational changes. The *ICAM1*_CHKOBBCO_01290 complex (spring green) shows the lowest SASA values, indicating that much of the protein is buried within the structure, which aligns with its stable and compact conformation as indicated by the Rg and RMSD results. The *ICAM1*_CHOJIKGH_00797 complex (purple) has moderate SASA values, suggesting a balance between exposure and compactness. The SASA profiles highlight that *ICAM1*_MLLMICFO_00840 is the most solvent-exposed and potentially the least stable complex, while *ICAM1*_CHKOBBCO_01290 is more stable and less exposed to the solvent.

The RMSD profile of Lymphocyte function-associated antigen 1 (*LFA-1*) with hypothetical proteins CHKOBBCO_01290 (spring green), CHOJIKGH_00797 (purple), and MLLMICFO_00840 (brown) has been demonstrated in Figure 10 (A). The RMSD value of the protein complex (*LFA-1*) MLLMICFO_00840 (brown) increased rapidly from 635ns to 875ns. After 875 ns, the RMSD value fluctuates, eventually reaching a relatively stable state. The RMSD value of the protein complex *LFA-1* and CHKOBBCO_01290 (spring green) gradually increased. The conformation of the protein complex LFA-1 with CHOJIKGH_00797 (purple) was relatively stable. The RMSF profiles of Lymphocyte function-associated antigen 1 (*LFA-1*) with hypothetical proteins CHKOBBCO_01290 (spring green), CHOJIKGH_00797 (purple), and MLLMICFO_00840 (brown) have been demonstrated in Figure 10 (B). The C-terminal of CHKOBBCO_01290 is more mobile than the other two complexes. The N-terminal of MLLMICFO_00840 is the most mobile among the two complexes. The Rg profile of Lymphocyte function-associated antigen 1 (*LFA-1*) with hypothetical proteins CHKOBBCO_01290 (spring green), CHOJIKGH_00797 (purple), and MLLMICFO_00840 (brown) has been demonstrated in Figure 10 (C). The Rg profile of the protein complex *LFA-1*_MLLMICFO_00840 (brown) showed fluctuations in compactness. The compactness increased and then decreased throughout the simulation. The compactness of the complex *LFA-1*_CHKOBBCO_01290 was gradually decreasing. The compactness for the complex *LFA-1*_CHOJIKGH_00797 was relatively the same. The SASA profile Lymphocyte function-associated antigen 1 (*LFA-1*) with hypothetical proteins CHKOBBCO_01290 (spring green), CHOJIKGH_00797 (purple), and MLLMICFO_00840 (brown) has been shown in Figure 10 (D). The SASA value insignificantly differed between all three complexes. According to the SASA analysis, all three complexes exhibit variations, suggesting that conformational changes in the molecules impact their exposure to the solvent. Throughout the simulation, the *LFA-1*_MLLMICFO_00840 shows the greatest SASA values, indicating that it has the most solvent-exposed surface area. The *LFA-1*_CHKOBBCO_01290, and *LFA-1*_CHOJIKGH_00797, on the other hand, show smaller solvent-accessible surfaces due to their lower SASA values.

Chapter 5

Discussion

Here, we study the role and functions of uncharacterized proteins and their relationship with DEGs in GC pathogenesis via subtractive genome analysis, transcriptomics, and systems biology to investigate potential drug targets for therapeutics. We identified 381 DEG genes to establish the interactions with 11 HPs, which may shed light on GC pathogenesis. Furthermore, gene ontology supported the biological validity of the genetic interactions. Molecular docking and molecular dynamics provide strong support in terms of their stable interactions. Previous studies identified the DEG genes and also characterized the HPs, however no direct link to the development of GC has been discovered [9]. Our study extended the previous findings and established the genetic interactions between GC and HP proteins.

In the present study, 381 genes were identified as differentially expressed out of the 33746 genes that were analyzed. Among them, 361 genes were found to be upregulated, while 20 genes were observed to be downregulated in severe gastritis samples. The PPI network identified that top 10 hub genes *CXCL8*, *CXCR2*, *CCL20*, *CD74*, *CCR1*, *CXCL1*, *HLA-DMA*, *CCR7*, *CCL3* and *HLA-DPA1* and bottleneck genes *CXCL8*, *CD19*, *ICAM1*, *VAV1*, *CXCR2*, *CD74*, *RAC2*, *PTPRC*, *SELL* and *ITGB2*. *CXCL8* was the most prevalent gene among both hub genes and bottleneck genes. Previously, it was proved that *CXCL8* was a potential biomarker for cancer progression in GC patients [89]. Also, *H. pylori* infection enhances the production of this protein. *CXCL8* proteins generally increase the movement of myeloid-derived suppressor cells which stops the immune system from attacking the cancer in its immediate surroundings [90].

11 HPs were identified which had a size of > 300 amino acids and have potential virulence features. It is important to determine the subcellular localization of a protein since there is a strong relationship between the protein's function and its position within the cell [91], [92]. In addition, it offers valuable information on identifying possible therapeutic or vaccine targets among the virulent proteins. Several proteins, such as JGMOFNOI_01064,

CHOJIKGH_00797, BLHMJNDD_00173, MLLMICFO_00840, and CHKOBBCO_01290, are likely to be found in the outer membrane. These proteins may play a role in communication between cells and interactions with the environment [93], [94]. Conversely, proteins present in the cytoplasm are as follows PLLHEGBO_01468, FMJNBOFJ_00132, ANOHMNDP_00326, JALKEJKI_01468, and PCGEIBGP_00353, indicating their involvement in various cellular processes like metabolism and signaling [95]. Furthermore, LBHCEKMO_01418 is predicted to be located in the periplasmic space which suggests its potential role in transporting nutrients and responding to stress [96], [97]. These findings show that the HPs are quite complex and have a wide range of functions. Therefore, they may play a role in various cellular processes and disease-causing mechanisms.

Furthermore, all eleven HPs were docked against the top interacting *CXCL8*. Only the HP MLLMICFO_00840, CHOJIKGH_00797, and CHKOBBCO_01290 interacted with the protein *CXCL8*. Although these *CXCL8* have been considered an important target for cancer therapy [98], their interactions with these uncharacterized proteins were not taken into account. This may also occur due to the mutation in these HPs. Therefore, these HPs might play a potential role in the progression of GC.

Moreover, the second most interacting protein found was *ICAM1*. *ICAM1* is also recognized as the regulator of GC and a potential biomarker, especially in the early stages of GC [99]. Molecular docking was also conducted between the *ICAM1*, its receptor Lymphocyte function-associated antigen 1 (*LFA-1*) with the HPs respectively. *LFA-1* is an integrin protein present on the surface of the leukocytes and lymphocytes, that helps the leukocytes to move from the circulation to the tissues. Leukocytes are also arrested via this protein [100]. This protein also participates in cytotoxic T-cell and antibody-mediated granulocyte and monocyte death [101]. So, if these adhesion proteins are not present, leukocyte adhesion deficiency (LAD) occurs.

Hence, if the HP of *H. pylori* act as an antagonist, then it might hamper neutrophil, cytotoxic T cell-mediated killing, and antibody-dependent cellular cytotoxicity.

The molecular dynamics simulations conducted on the protein-protein complexes involving Lymphocyte function-associated antigen 1 (*LFA-1*) and C-X-C Motif Chemokine Ligand 8 (*CXCL8*) with the hypothetical proteins CHKOBBCO_01290, CHOJIKGH_00797, and MLLMICFO_00840 have yielded insightful data regarding the stability and structural dynamics of these interactions. Through this analysis, the interaction quality between these proteins has been elucidated, offering a comprehensive understanding of their behaviour under simulated physiological conditions.

5.1 Interaction Quality Between *LFA-1* and Hypothetical Proteins

The interaction of *LFA-1* with the three hypothetical proteins, CHKOBBCO_01290, CHOJIKGH_00797, and MLLMICFO_00840, revealed distinct patterns of stability and conformational changes. The *LFA-1*_MLLMICFO_00840 complex exhibited an intriguing stability pattern, characterized by initial structural fluctuations followed by a stabilization phase. The substantial fluctuations observed initially in the RMSD values suggest that the *LFA-1*_MLLMICFO_00840 complex undergoes significant conformational adjustments before reaching a stable state. This indicates a potentially dynamic binding interface that may be crucial for functional interactions in a biological context. In contrast, the *LFA-1*_CHKOBBCO_01290 complex displayed a more progressive increase in RMSD, indicating a gradual adjustment to a stable conformation. This suggests a more controlled and steady interaction, potentially reflecting a binding interface that is less prone to dramatic conformational shifts. The steady nature of this interaction might imply a stable and specific binding mode, which could be crucial for maintaining the functional integrity of the complex. The *LFA-1*_CHOJIKGH_00797 complex, on the other hand, remained relatively stable throughout the simulation. The minimal fluctuations observed in RMSD indicate a highly stable interaction, suggesting that CHOJIKGH_00797 may have a particularly strong and consistent binding affinity with *LFA-1*. This stability could be indicative of a highly complementary binding interface, which is less susceptible to conformational changes, thus maintaining the structural integrity of the complex.

5.2 Interaction Quality Between *CXCL8* and Hypothetical Proteins

Similarly, the interactions between *CXCL8* and the hypothetical proteins presented diverse stability profiles. The *CXCL8*_MLLMICFO_00840 complex exhibited the highest RMSD values, indicating significant structural fluctuations. This suggests that the interaction between *CXCL8* and MLLMICFO_00840 is more dynamic, potentially involving multiple conformational states before achieving stability. The higher degree of flexibility and structural variation in this complex could point to a more transient or regulatory interaction, where conformational flexibility is key to its functional role. The *CXCL8*_CHKOBBCO_01290 complex showed a different pattern, with RMSD values suggesting a relatively stable interaction, though with a gradual increase over time. This could indicate that the binding interface between *CXCL8* and CHKOBBCO_01290 undergoes slight conformational adjustments as the complex stabilizes, reflecting a binding mode that allows for some flexibility while maintaining overall structural integrity. The *CXCL8*_CHOJIKGH_00797 complex, much like its *LFA-1* counterpart, demonstrated remarkable stability, with minimal RMSD fluctuations throughout the simulation. This stability suggests a strong and consistent interaction between *CXCL8* and CHOJIKGH_00797, potentially indicating a highly specific binding interface that is well-matched to *CXCL8's* structural features. The minor fluctuations in this complex could imply that once the interaction is established, it remains robust, likely contributing to a sustained biological function.

5.3 Structural Flexibility and Compactness

Further insight into the interaction quality was gained through RMSF and Rg analyses. The *LFA-1*_MLLMICFO_00840 complex displayed significant fluctuations in the C-terminal region of MLLMICFO_00840, which might be indicative of a flexible tail that plays a role in the dynamic interaction with LFA-1. In contrast, the N-terminal region of CHKOBBCO_01290 in the *LFA-1* complex showed less mobility, hinting at a more rigid and possibly more structurally integral role in the binding interface. The compactness of these complexes, as evaluated by Rg analysis, further underscored the differences in interaction quality. The *LFA-1*_MLLMICFO_00840 complex showed fluctuations in compactness, suggesting phases of structural rearrangement, which could correlate with the initial RMSD fluctuations. Conversely, the *LFA-1*_CHKOBBCO_01290 and *LFA-1*_CHOJIKGH_00797 complexes exhibited more stable compactness, indicative of a well-formed and consistent interaction interface.

Similarly, the *CXCL8* complexes showed varied compactness patterns, with *CXCL8*_MLLMICFO_00840 and *CXCL8*_CHKOBBCO_01290 complexes exhibiting increased Rg values over time. This suggests that these complexes may undergo a degree of unfolding or loss of compactness, reflecting a dynamic and potentially flexible interaction. In contrast, the *CXCL8*_CHOJIKGH_00797 complex remained relatively compact, further supporting the notion of a stable and consistent interaction.

5.4 Solvent Exposure and Surface Accessibility

The SASA analysis provided additional evidence of the interaction quality. The higher SASA values observed for the *LFA-1*_MLLMICFO_00840 and *CXCL8*_MLLMICFO_00840 complexes suggest that these interactions expose more surface area to the solvent, which could be indicative of less tightly packed complexes. This might correlate with the observed structural fluctuations and dynamic nature of these interactions. On the other hand, the lower SASA values for the *LFA-1*_CHKOBBCO_01290, *LFA-1*_CHOJIKGH_00797, and *CXCL8*_CHOJIKGH_00797 complexes suggest a more deeply embedded hydrophobic core, which is often associated with more stable and energetically favourable interactions.

The molecular dynamics simulations have revealed a spectrum of interaction qualities between *LFA-1* and *CXCL8* with the hypothetical proteins CHKOBBCO_01290, CHOJIKGH_00797, and MLLMICFO_00840. The stability, flexibility, compactness, and solvent exposure analyses have provided a detailed understanding of these interactions, highlighting the diverse structural dynamics and potential functional implications. The differences observed in the interaction patterns are likely reflective of the unique structural features of each hypothetical protein and their specific binding affinities with *LFA-1* and *CXCL8*. These findings pave the way for further experimental validation and functional studies to explore the biological significance of these interactions, potentially informing therapeutic strategies or protein engineering efforts.

5.5 Limitations of the study

While this study has provided significant insights into the roles of hypothetical proteins in *H. pylori* and their interactions with host genes in GC pathogenesis, several areas warrant further investigation. First, expanding the characterization of additional hypothetical proteins not covered in this study could reveal new molecular mechanisms of *H. pylori* pathogenicity. High-throughput functional assays combined with advanced bioinformatics could identify novel therapeutic targets within these uncharacterized proteins.

Moreover, the in vivo validation of the molecular interactions identified through molecular dynamics simulations could substantiate the potential of these hypothetical proteins as therapeutic targets. Animal models of *H. pylori*-induced GC could be employed to assess the biological relevance of these interactions in a physiological context, providing a deeper understanding of the pathogen-host dynamics.

Chapter 6

Conclusion

Helicobacter pylori infection has long been linked to GC. Transcriptomics and genomes were used to discover *H. pylori* HPs that cause GC pathogenesis. Host genes *CXCL8*, *ICAM1*, and *CXCR2* substantially interacted with *H. pylori* putative proteins, supporting the present study statement. The stability of these interacting complexes suggests they may regulate immune responses and cause GC. This study sheds light on the complex relationship between *H. pylori* infection and GC, offering a new avenue to targeting the drug targets. Thus, studying the host proteins and their interactions with hypothetical *H. pylori* proteins may lead to new diagnostic biomarkers and therapies for this cancer.

References

[1] F. Bray *et al.*, 'Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries', *CA Cancer J Clin*, vol. 74, no. 3, pp. 229–263, 2024, doi: 10.3322/caac.21834.

[2] E. Morgan *et al.*, 'The current and future incidence and mortality of GC in 185 countries, 2020-40: A population-based modelling study', *EClinicalMedicine*, vol. 47, p. 101404, May 2022, doi: 10.1016/j.eclinm.2022.101404.

[3] P. Violeta Filip, D. Cuciureanu, L. Sorina Diaconu, A. Maria Vladareanu, and C. Silvia Pop, 'MALT lymphoma: epidemiology, clinical diagnosis and treatment', *J Med Life*, vol. 11, no. 3, pp. 187–193, 2018, doi: 10.25122/jml-2018-0035.

[4] S. Ishaq and L. Nunn, 'Helicobacter pylori and GC: a state of the art review', *Gastroenterol Hepatol Bed Bench*, vol. 8, no. Suppl1, pp. S6–S14, 2015.

[5] P. B. Ernst and B. D. Gold, 'The disease spectrum of Helicobacter pylori: the immunopathogenesis of gastroduodenal ulcer and GC', *Annu Rev Microbiol*, vol. 54, pp. 615–640, 2000, doi: 10.1146/annurev.micro.54.1.615.

[6] B. A. Salih, 'Helicobacter pylori Infection in Developing Countries: The Burden for How Long?', *Saudi J Gastroenterol*, vol. 15, no. 3, pp. 201–207, Jul. 2009, doi: 10.4103/1319-3767.54743.

[7] G. Khoder, J. S. Muhammad, I. Mahmoud, S. S. M. Soliman, and C. Burucoa, 'Prevalence of Helicobacter pylori and Its Associated Factors among Healthy Asymptomatic Residents in the United Arab Emirates', *Pathogens*, vol. 8, no. 2, p. 44, Apr. 2019, doi: 10.3390/pathogens8020044.

[8] K. K. Sarker *et al.*, ‘H. pylori infection and GC in Bangladesh: a case-control study’, *Int J Surg Oncol (N Y)*, vol. 2, no. 10, p. e44, Nov. 2017, doi: 10.1097/IJ9.0000000000000044.

[9] W.-L. Chang, Y.-C. Yeh, and B.-S. Sheu, ‘The impacts of H. pylori virulence factors on the development of gastroduodenal diseases’, *J Biomed Sci*, vol. 25, p. 68, Sep. 2018, doi: 10.1186/s12929-018-0466-9.

[10] S. Fagoonee and R. Pellicano, ‘Helicobacter pylori: molecular basis for colonization and survival in gastric environment and resistance to antibiotics. A short review’, *Infectious Diseases*, vol. 51, no. 6, pp. 399–408, Jun. 2019, doi: 10.1080/23744235.2019.1588472.

[11] S. J. Park, W. S. Son, and B.-J. Lee, ‘Structural Analysis of Hypothetical Proteins from Helicobacter pylori: An Approach to Estimate Functions of Unknown or Hypothetical Proteins’, *Int J Mol Sci*, vol. 13, no. 6, pp. 7109–7137, Jun. 2012, doi: 10.3390/ijms13067109.

[12] A. A. T. Naqvi, F. Anjum, F. I. Khan, A. Islam, F. Ahmad, and Md. I. Hassan, ‘Sequence Analysis of Hypothetical Proteins from Helicobacter pylori 26695 to Identify Potential Virulence Factors’, *Genomics Inform*, vol. 14, no. 3, pp. 125–135, Sep. 2016, doi: 10.5808/GI.2016.14.3.125.

[13] M. Shahbaaz, K. Bisetty, F. Ahmad, and M. I. Hassan, ‘Current Advances in the Identification and Characterization of Putative Drug and Vaccine Targets in the Bacterial Genomes’, *Curr Top Med Chem*, vol. 16, no. 9, pp. 1040–1069, 2016, doi: 10.2174/1568026615666150825143307.

[14] G. Pranavathiyani, J. Prava, A. C. Rajeev, and A. Pan, ‘Novel Target Exploration from Hypothetical Proteins of Klebsiella pneumoniae MGH 78578 Reveals a Protein Involved in Host-Pathogen Interaction’, *Front Cell Infect Microbiol*, vol. 10, p. 109, Apr. 2020, doi: 10.3389/fcimb.2020.00109.

[15] A. L. Tarca, R. Romero, and S. Draghici, ‘Analysis of microarray experiments of gene expression profiling’, *Am J Obstet Gynecol*, vol. 195, no. 2, pp. 373–388, Aug. 2006, doi: 10.1016/j.ajog.2006.07.001.

[16] R. Govindarajan, J. Duraiyan, K. Kaliyappan, and M. Palanisamy, ‘Microarray and its applications’, *J Pharm Bioallied Sci*, vol. 4, no. Suppl 2, pp. S310–S312, Aug. 2012, doi: 10.4103/0975-7406.100283.

[17] R. Rodriguez-Esteban and X. Jiang, ‘Differential gene expression in disease: a comparison between high-throughput studies and the literature’, *BMC Medical Genomics*, vol. 10, no. 1, p. 59, Oct. 2017, doi: 10.1186/s12920-017-0293-y.

[18] F. Shao, Y. Wang, Y. Zhao, and S. Yang, ‘Identifying and exploiting gene-pathway interactions from RNA-seq data for binary phenotype’, *BMC Genet*, vol. 20, p. 36, Mar. 2019, doi: 10.1186/s12863-019-0739-7.

[19] J. Ijaq *et al.*, ‘A model to predict the function of hypothetical proteins through a nine-point classification scoring schema’, *BMC Bioinformatics*, vol. 20, no. 1, p. 14, Jan. 2019, doi: 10.1186/s12859-018-2554-y.

[20] ‘Nivolumab Effective for Advanced Stomach Cancer - NCI’. Accessed: Jul. 30, 2024. [Online]. Available: https://www.cancer.gov/news-events/cancer-currents-blog/2020/stomach-cancer-immunotherapy-nivolumab

[21] ‘GC treatment: recent progress and future perspectives | Journal of Hematology & Oncology | Full Text’. Accessed: Jul. 30, 2024. [Online]. Available: https://jhoonline.biomedcentral.com/articles/10.1186/s13045-023-01451-3

[22] B. Marshall, ‘A Brief History of the Discovery of Helicobacter pylori’, in *Helicobacter pylori*, H. Suzuki, R. Warren, and B. Marshall, Eds., Tokyo: Springer Japan, 2016, pp. 3–15. doi: 10.1007/978-4-431-55705-0_1.

[23] P. Malfertheiner, A. Link, and M. Selgrad, ‘Helicobacter pylori: perspectives and time trends’, *Nat Rev Gastroenterol Hepatol*, vol. 11, no. 10, pp. 628–638, Oct. 2014, doi: 10.1038/nrgastro.2014.99.

[24] G. D. Fock KM, ‘Helicobacter pylori research: historical insights and future directions.’, *Nat Rev Gastroenterol Hepatol*, vol. 10, no. 8, pp. 495–500, 2013, doi: 10.1038/nrgastro.2013.96.

[25] M. Arnold *et al.*, ‘Global Burden of 5 Major Types of Gastrointestinal Cancer’, *Gastroenterology*, vol. 159, no. 1, pp. 335-349.e15, Jul. 2020, doi: 10.1053/j.gastro.2020.02.068.

[26] T. I. A. for R. on Cancer (IARC), ‘Global Cancer Observatory’. Accessed: Aug. 02, 2024. [Online]. Available: https://gco.iarc.fr/

[27] W. S. Shin *et al.*, ‘Updated Epidemiology of GC in Asia: Decreased Incidence but Still a Big Challenge’, *Cancers (Basel)*, vol. 15, no. 9, p. 2639, May 2023, doi: 10.3390/cancers15092639.

[28] ‘East Asians more likely to develop stomach cancer because of lower alcohol tolerance, new study says’, NBC News. Accessed: Jul. 30, 2024. [Online]. Available: https://www.nbcnews.com/news/asian-america/east-asians-likely-develop-stomach-cancer-lower-alcohol-tolerance-new-rcna75329

[29] E. Moskal, ‘Stomach cancer hits Asian populations harder’, Scope. Accessed: Jul. 30, 2024. [Online]. Available: https://scopeblog.stanford.edu/2022/12/16/stomach-cancer-asians/

[30] ‘Stomach Cancer in Bangladesh’, World Life Expectancy. Accessed: Aug. 02, 2024. [Online]. Available: https://www.worldlifeexpectancy.com/bangladesh-stomach-cancer

[31] L. E. Wroblewski, R. M. Peek, and K. T. Wilson, ‘Helicobacter pylori and GC: Factors That Modulate Disease Risk’, *Clin Microbiol Rev*, vol. 23, no. 4, pp. 713–739, Oct. 2010, doi: 10.1128/CMR.00011-10.

[32] D. B. Polk and R. M. Peek, ‘Helicobacter pylori: GC and beyond’, *Nat Rev Cancer*, vol. 10, no. 6, pp. 403–414, Jun. 2010, doi: 10.1038/nrc2857.

[33] I. H. pylori W. Group, *<em>Helicobacter pylori</em> Eradication as a Strategy for Preventing GC*. Accessed: Aug. 12, 2024. [Online]. Available: https://publications.iarc.fr/Book-And-Report-Series/Iarc-Working-Group-Reports/-Em-Helicobacter-Pylori-Em-Eradication-As-A-Strategy-For-Preventing-Gastric-Cancer-2014

[34] M. R. Wilkins *et al.*, ‘Protein identification and analysis tools in the ExPASy server’, *Methods Mol Biol*, vol. 112, pp. 531–552, 1999, doi: 10.1385/1-59259-584-7:531.

[35] P. Shannon *et al.*, ‘Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks’, *Genome Res*, vol. 13, no. 11, pp. 2498–2504, Nov. 2003, doi: 10.1101/gr.1239303.

[36] E. A. Franzosa *et al.*, ‘Relating the metatranscriptome and metagenome of the human gut’, *Proc Natl Acad Sci U S A*, vol. 111, no. 22, pp. E2329–E2338, Jun. 2014, doi: 10.1073/pnas.1319284111.

[37] H. Kaur, V. Singh, M. Kalia, B. Mohan, and N. Taneja, ‘Identification and functional annotation of hypothetical proteins of uropathogenic Escherichia coli strain CFT073 towards designing antimicrobial drug targets’, *J Biomol Struct Dyn*, vol. 40, no. 24, pp. 14084–14095, 2022, doi: 10.1080/07391102.2021.2000499.

[38] A. Shami *et al.*, ‘In Silico Subtractive Proteomics and Molecular Docking Approaches for the Identification of Novel Inhibitors against Streptococcus pneumoniae Strain D39’, *Life (Basel)*, vol. 13, no. 5, p. 1128, May 2023, doi: 10.3390/life13051128.

[39] P. Díaz, M. Valenzuela Valderrama, J. Bravo, and A. F. G. Quest, ‘Helicobacter pylori and GC: Adaptive Cellular Mechanisms Involved in Disease Progression’, *Front Microbiol*, vol. 9, p. 5, 2018, doi: 10.3389/fmicb.2018.00005.

[40] C. de Martel, D. Georges, F. Bray, J. Ferlay, and G. M. Clifford, ‘Global burden of cancer attributable to infections in 2018: a worldwide incidence analysis’, *The Lancet Global Health*, vol. 8, no. 2, pp. e180–e190, Feb. 2020, doi: 10.1016/S2214-109X(19)30488-7.

[41] D. Underferth, ‘H. pylori and your stomach cancer risk’, MD Anderson Cancer Center. Accessed: Jul. 30, 2024. [Online]. Available: https://www.mdanderson.org/cancerwise/h--pylori-and-your-stomach-cancer-risk.h00-159460056.html

[42] S. Bernstein, ‘H. Pylori and Stomach Cancer’, WebMD. Accessed: Jul. 30, 2024. [Online]. Available: https://www.webmd.com/cancer/hpylori-stomach-cancer

[43] C. Figueiredo, J. C. Machado, and Y. Yamaoka, ‘Pathogenesis of Helicobacter pylori Infection’, *Helicobacter*, vol. 10 Suppl 1, pp. 14–20, 2005, doi: 10.1111/j.1523-5378.2005.00339.x.

[44] B. Linz *et al.*, ‘An African origin for the intimate association between humans and Helicobacter pylori’, *Nature*, vol. 445, no. 7130, pp. 915–918, Feb. 2007, doi: 10.1038/nature05562.

[45] S. C. Sabbagh, C. G. Forest, C. Lepage, J.-M. Leclerc, and F. Daigle, ‘So similar, yet so different: uncovering distinctive features in the genomes of Salmonella enterica serovars Typhimurium and Typhi’, *FEMS Microbiol Lett*, vol. 305, no. 1, pp. 1–13, Apr. 2010, doi: 10.1111/j.1574-6968.2010.01904.x.

[46] J. B. Wardenburg and O. Schneewind, ‘Vaccine protection against Staphylococcus aureus pneumonia’, *J Exp Med*, vol. 205, no. 2, pp. 287–294, Feb. 2008, doi: 10.1084/jem.20072208.

[47] ‘GC: Pathology and molecular pathogenesis - UpToDate’. Accessed: Jul. 31, 2024. [Online]. Available: https://www.uptodate.com/contents/gastric-cancer-pathology-and-molecular-pathogenesis

[48] P. L. S. Uson *et al.*, ‘Germline Cancer Testing in Unselected Patients with Gastric and Esophageal Cancers: A Multi-center Prospective Study’, *Dig Dis Sci*, vol. 67, no. 11, pp. 5107–5115, Nov. 2022, doi: 10.1007/s10620-022-07387-x.

[49] G. Y. Ku *et al.*, ‘Prevalence of Germline Alterations on Targeted Tumor-Normal Sequencing of EsophagoGC’, *JAMA Netw Open*, vol. 4, no. 7, p. e2114753, Jul. 2021, doi: 10.1001/jamanetworkopen.2021.14753.

[50] P. D. Pharoah, P. Guilford, C. Caldas, and International GC Linkage Consortium, ‘Incidence of GC and breast cancer in CDH1 (E-cadherin) mutation carriers from hereditary diffuse GC families’, *Gastroenterology*, vol. 121, no. 6, pp. 1348–1353, Dec. 2001, doi: 10.1053/gast.2001.29611.

[51] I. J. Majewski *et al.*, ‘An α-E-catenin (CTNNA1) mutation in hereditary diffuse GC’, *J Pathol*, vol. 229, no. 4, pp. 621–629, Mar. 2013, doi: 10.1002/path.4152.

[52] P. Møller *et al.*, ‘Cancer risk and survival in path_MMR carriers by gene and gender up to 75 years of age: a report from the Prospective Lynch Syndrome Database’, *Gut*, vol. 67, no. 7, pp. 1306–1316, Jul. 2018, doi: 10.1136/gutjnl-2017-314057.

[53] S. C. Abraham, B. Nobukawa, F. M. Giardiello, S. R. Hamilton, and T. T. Wu, ‘Fundic gland polyps in familial adenomatous polyposis: neoplasms with frequent somatic adenomatous polyposis coli gene alterations’, *Am J Pathol*, vol. 157, no. 3, pp. 747–754, Sep. 2000, doi: 10.1016/S0002-9440(10)64588-9.

[54] F. M. Giardiello *et al.*, ‘Very high risk of cancer in familial Peutz-Jeghers syndrome’, *Gastroenterology*, vol. 119, no. 6, pp. 1447–1453, Dec. 2000, doi: 10.1053/gast.2000.20228.

[55] Y. Wang, X. Gao, and J. Wang, ‘Functional Proteomic Profiling Analysis in Four Major Types of Gastrointestinal Cancers’, *Biomolecules*, vol. 13, no. 4, Art. no. 4, Apr. 2023, doi: 10.3390/biom13040701.

[56] A. Biagioni *et al.*, ‘GC Vascularization and the Contribution of Reactive Oxygen Species’, *Biomolecules*, vol. 13, no. 6, Art. no. 6, Jun. 2023, doi: 10.3390/biom13060886.

[57] S. Battista, M. R. Ambrosio, F. Limarzi, G. Gallo, and L. Saragoni, ‘Molecular Alterations in Gastric Preneoplastic Lesions and Early GC’, *Int J Mol Sci*, vol. 22, no. 13, p. 6652, Jun. 2021, doi: 10.3390/ijms22136652.

[58] M. Karplus and J. A. McCammon, ‘Molecular dynamics simulations of biomolecules’, *Nat Struct Biol*, vol. 9, no. 9, pp. 646–652, Sep. 2002, doi: 10.1038/nsb0902-646.

[59] S. A. Hollingsworth and R. O. Dror, ‘Molecular dynamics simulation for all’, *Neuron*, vol. 99, no. 6, pp. 1129–1143, Sep. 2018, doi: 10.1016/j.neuron.2018.08.011.

[60] T. Nishizawa and H. Suzuki, ‘Mechanisms of Helicobacter pylori antibiotic resistance and molecular testing’, *Front Mol Biosci*, vol. 1, p. 19, Oct. 2014, doi: 10.3389/fmolb.2014.00019.

[61] I. Thung *et al.*, ‘Review article: the global emergence of Helicobacter pylori antibiotic resistance’, *Aliment Pharmacol Ther*, vol. 43, no. 4, pp. 514–533, Feb. 2016, doi: 10.1111/apt.13497.

[62] T. Barrett *et al.*, ‘NCBI GEO: archive for functional genomics data sets—update’, *Nucleic Acids Research*, vol. 41, no. D1, pp. D991–D995, Jan. 2013, doi: 10.1093/nar/gks1193.

[63] M. E. Ritchie *et al.*, ‘limma powers differential expression analyses for RNA-sequencing and microarray studies’, *Nucleic Acids Research*, vol. 43, no. 7, p. e47, Apr. 2015, doi: 10.1093/nar/gkv007.

[64] D. Szklarczyk *et al.*, ‘The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest’, *Nucleic Acids Research*, vol. 51, no. D1, pp. D638–D646, Jan. 2023, doi: 10.1093/nar/gkac1000.

[65] G. Stelzer *et al.*, ‘The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses’, *Current Protocols in Bioinformatics*, vol. 54, no. 1, p. 1.30.1-1.30.33, 2016, doi: 10.1002/cpbi.5.

[66] M. Kanehisa and S. Goto, ‘KEGG: Kyoto Encyclopedia of Genes and Genomes’, *Nucleic Acids Res*, vol. 28, no. 1, pp. 27–30, Jan. 2000.

[67] R. Leinonen, H. Sugawara, and M. Shumway, ‘The Sequence Read Archive’, *Nucleic Acids Res*, vol. 39, no. Database issue, pp. D19–D21, Jan. 2011, doi: 10.1093/nar/gkq1019.

[68] S. Andrews, *FastQC: A Quality Control Tool for High Throughput Sequence Data*. (Jun. 2015). [Online]. Available: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[69] A. M. Bolger, M. Lohse, and B. Usadel, ‘Trimmomatic: a flexible trimmer for Illumina sequence data’, *Bioinformatics*, vol. 30, no. 15, pp. 2114–2120, Aug. 2014, doi: 10.1093/bioinformatics/btu170.

[70] R. R. Wick, L. M. Judd, C. L. Gorrie, and K. E. Holt, ‘Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads’, *PLOS Computational Biology*, vol. 13, no. 6, p. e1005595, Jun. 2017, doi: 10.1371/journal.pcbi.1005595.

[71] T. Seemann, ‘Prokka: rapid prokaryotic genome annotation’, *Bioinformatics*, vol. 30, no. 14, pp. 2068–2069, Jul. 2014, doi: 10.1093/bioinformatics/btu153.

[72] W. Li and A. Godzik, ‘Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences’, *Bioinformatics*, vol. 22, no. 13, pp. 1658–1659, Jul. 2006, doi: 10.1093/bioinformatics/btl158.

[73] L. Chen *et al.*, ‘VFDB: a reference database for bacterial virulence factors’, *Nucleic Acids Research*, vol. 33, no. suppl_1, pp. D325–D328, Jan. 2005, doi: 10.1093/nar/gki008.

[74] C. Camacho *et al.*, ‘BLAST+: architecture and applications’, *BMC Bioinformatics*, vol. 10, no. 1, p. 421, Dec. 2009, doi: 10.1186/1471-2105-10-421.

[75] U. Amineni, D. Pradhan, and H. Marisetty, ‘In silico identification of common putative drug targets in Leptospira interrogans’, *J Chem Biol*, vol. 3, no. 4, pp. 165–173, May 2010, doi: 10.1007/s12154-010-0039-1.

[76] C.-S. Yu, Y.-C. Chen, C.-H. Lu, and J.-K. Hwang, ‘Prediction of protein subcellular localization’, *Proteins*, vol. 64, no. 3, pp. 643–651, Aug. 2006, doi: 10.1002/prot.21018.

[77] A. Waterhouse *et al.*, ‘SWISS-MODEL: homology modelling of protein structures and complexes’, *Nucleic Acids Research*, vol. 46, no. W1, pp. W296–W303, Jul. 2018, doi: 10.1093/nar/gky427.

[78] R. A. Laskowski, M. W. MacArthur, D. S. Moss, and J. M. Thornton, ‘PROCHECK: a program to check the stereochemical quality of protein structures’, *J Appl Cryst*, vol. 26, no. 2, pp. 283–291, Apr. 1993, doi: 10.1107/S0021889892009944.

[79] C. Colovos and T. O. Yeates, ‘Verification of protein structures: patterns of nonbonded atomic interactions’, *Protein Sci*, vol. 2, no. 9, pp. 1511–1519, Sep. 1993, doi: 10.1002/pro.5560020916.

[80] H. M. Berman *et al.*, ‘The Protein Data Bank’, *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, Jan. 2000, doi: 10.1093/nar/28.1.235.

[81] T. Park, M. Baek, H. Lee, and C. Seok, ‘GalaxyTongDock: Symmetric and asymmetric ab initio protein–protein docking web server with improved energy parameters’, *Journal of Computational Chemistry*, vol. 40, no. 27, pp. 2413–2417, 2019, doi: 10.1002/jcc.25874.

[82] M. J. Abraham *et al.*, ‘GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers’, *SoftwareX*, vol. 1–2, pp. 19–25, Sep. 2015, doi: 10.1016/j.softx.2015.06.001.

[83] S. Jo, T. Kim, V. G. Iyer, and W. Im, ‘CHARMM-GUI: A web-based graphical user interface for CHARMM’, *Journal of Computational Chemistry*, vol. 29, no. 11, pp. 1859–1865, 2008, doi: 10.1002/jcc.20945.

[84] P. C. T. Souza *et al.*, ‘Martini 3: a general purpose force field for coarse-grained molecular dynamics’, *Nat Methods*, vol. 18, no. 4, pp. 382–388, Apr. 2021, doi: 10.1038/s41592-021-01098-3.

[85] J. Huang *et al.*, ‘CHARMM36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins’, *Nat Methods*, vol. 14, no. 1, pp. 71–73, Jan. 2017, doi: 10.1038/nmeth.4067.

[86] A. K. Sharma *et al.*, ‘Bacterial Virulence Factors: Secreted for Survival’, *Indian J Microbiol*, vol. 57, no. 1, pp. 1–10, Mar. 2017, doi: 10.1007/s12088-016-0625-1.

[87] J. Kirkwood, D. Hargreaves, S. O’Keefe, and J. Wilson, ‘Using isoelectric point to determine the pH for initial protein crystallization trials’, *Bioinformatics*, vol. 31, no. 9, pp. 1444–1451, May 2015, doi: 10.1093/bioinformatics/btv011.

[88] D. G. Gamage, A. Gunaratne, G. R. Periyannan, and T. G. Russell, ‘Applicability of Instability Index for In vitro Protein Stability Prediction’, *Protein Pept Lett*, vol. 26, no. 5, pp. 339–347, 2019, doi: 10.2174/0929866526666190228144219.

[89] W.-Q. Qi, Q. Zhang, and J.-B. Wang, ‘CXCL8 is a potential biomarker for predicting disease progression in gastric carcinoma’, *Transl Cancer Res*, vol. 9, no. 2, pp. 1053–1062, Feb. 2020, doi: 10.21037/tcr.2019.12.52.

[90] Z. Liu *et al.*, ‘H. pylori infection induces CXCL8 expression and promotes GC progress through downregulating KLF4’, *Mol Carcinog*, vol. 60, no. 8, pp. 524–537, Aug. 2021, doi: 10.1002/mc.23309.

[91] M. S. Scott, S. J. Calafell, D. Y. Thomas, and M. T. Hallett, ‘Refining Protein Subcellular Localization’, *PLoS Comput Biol*, vol. 1, no. 6, p. e66, Nov. 2005, doi: 10.1371/journal.pcbi.0010066.

[92] P. Bork, T. Dandekar, Y. Diaz-Lazcoz, F. Eisenhaber, M. Huynen, and Y. Yuan, ‘Predicting function: from genes to genomes and back’, *J Mol Biol*, vol. 283, no. 4, pp. 707–725, Nov. 1998, doi: 10.1006/jmbi.1998.2144.

[93] A. Nasarabadi, J. E. Berleman, and M. Auer, ‘Outer Membrane Vesicles of Bacteria: Structure, Biogenesis, and Function’, in *Biogenesis of Fatty Acids, Lipids and Membranes*, O. Geiger, Ed., Cham: Springer International Publishing, 2019, pp. 593–607. doi: 10.1007/978-3-319-50430-8_44.

[94] G. Magaña, C. Harvey, C. C. Taggart, and A. M. Rodgers, ‘Bacterial Outer Membrane Vesicles: Role in Pathogenesis and Host-Cell Interactions’, *Antibiotics*, vol. 13, no. 1, Art. no. 1, Jan. 2024, doi: 10.3390/antibiotics13010032.

[95] R. C. Baxter, ‘Signaling Pathways of the Insulin-like Growth Factor Binding Proteins’, *Endocrine Reviews*, vol. 44, no. 5, pp. 753–778, Oct. 2023, doi: 10.1210/endrev/bnad008.

[96] A. J. Cumming, D. Khananisho, M. Balka, N. Liljestrand, and D. O. Daley, ‘Biosensor that Detects Stress Caused by Periplasmic Proteins’, *ACS Synth. Biol.*, vol. 13, no. 5, pp. 1477–1491, May 2024, doi: 10.1021/acssynbio.3c00720.

[97] P.-N. Li *et al.*, ‘Nutrient transport suggests an evolutionary basis for charged archaeal surface layer proteins’, *ISME J*, vol. 12, no. 10, pp. 2389–2402, Oct. 2018, doi: 10.1038/s41396-018-0191-0.

[98] D. Gales, C. Clark, U. Manne, and T. Samuel, ‘The Chemokine CXCL8 in Carcinogenesis and Drug Response’, *ISRN Oncol*, vol. 2013, p. 859154, Oct. 2013, doi: 10.1155/2013/859154.

[99] S. Chen *et al.*, ‘ICAM1 Regulates the Development of GC and May Be a Potential Biomarker for the Early Diagnosis and Prognosis of GC’, *Cancer Manag Res*, vol. 12, pp. 1523–1534, Mar. 2020, doi: 10.2147/CMAR.S237443.

[100] K. Ley, Ed., *Adhesion Molecules: Function and Inhibition*. Basel: Birkhäuser Basel, 2007. doi: 10.1007/978-3-7643-7975-9.

[101] ‘Oxford Dictionary of Biochemistry and Molecular Biology’, Oxford University Press, 2006. Accessed: Jun. 20, 2024. [Online]. Available: https://www.oxfordreference.com/display/10.1093/acref/9780198529170.001.0001/acref-9780198529170

Appendix A.

Supplementary Files:

https://drive.google.com/drive/folders/14blG9pF1qz15YBcofRM4vbF3hTKjuGHg?usp=sharing