In silico structure-function studies of *H. pylori* proteome and DNA methyltransferases

A thesis
Submitted for the degree of
DOCTOR OF PHILOSOPHY

By
Swati Singh

School of Chemistry
University of Hyderabad
Hyderabad – 500 046
INDIA
July 2016

I dedicate this thesis to
My Beloved Family

School of Chemistry
University of Hyderabad
Hyderabad – 500 046

STATEMENT

I hereby declare that the matter embodied in this thesis entitled **"*In silico* structure-function studies of *H. pylori* proteome and DNA methyltransferases"** is the result of investigations carried out by me in the School of Chemistry, University of Hyderabad, Hyderabad, under the supervision of **Prof. Lalitha Guruprasad**.

In keeping with the general practice of reporting scientific observations, due acknowledgements have been made whenever the work described is based on the finding of other investigators. Any omission which might have occurred by oversight or error is regretted.

Hyderabad
July 2016 **Swati Singh**

School of Chemistry
University of Hyderabad
Hyderabad–500 046

CERTIFICATE

Certified that the work embodied in this thesis entitled **"*In silico* structure-function studies of *H. pylori* proteome and DNA methyltransferases"** has been carried out by **Mrs. Swati Singh** under my supervision and the same has not been submitted elsewhere for any degree.

Dean **Prof. Lalitha Guruprasad**

School of Chemistry **(Thesis Supervisor)**

School of Chemistry
University of Hyderabad
Hyderabad–500 046

DECLARATION

I, **Swati Singh** hereby declare that the thesis entitled **"*In silico* structure-function studies of *H. pylori* proteome and DNA methyltransferases"** submitted by me under the supervision of **Prof**. **Lalitha Guruprasad** is a bonafide research work which is free from plagiarism. I also declare that it has not been submitted previously in part or in full to this University or any other University or Institution for the award of any degree or diploma. I hereby agree that my thesis can be deposited in Shodganga/INFLIBNET.

A report on plagiarism from the University Library is enclosed.

Name: Swati Singh
Signature:
Reg. No.: 11CHPH01

Prof. Lalitha Guruprasad
(Thesis Supervisor)

ACKNOWLEDGEMENTS

I would like to express my sincere grati[illegible] my research supervisor Prof. Lalitha Guruprasad for her support during my Ph.D study. I thank her for continuous advice and encouragement throughout the Ph.D. I would also like to thank my Research Doctoral Committee members Prof. M. J. Swamy and Prof. A. K. Bhuyan for their insightful comments and encouragement on my thesis.

I express my sincere thanks to the former dean Prof. M.V. Rajasekharan and present dean Prof. M. Durga Prasad for providing infrastructure to carry out the research work in the school. I would also like to thank all other faculty members and non-teaching staff for their help and cooperation.

I thank CSIR for financial support and CMSD for excellent computational facilities.

I deeply thank all my labmates Dr. Karunakar, Dr. Shalini, Rafiya, Rajender, Bala Divya, Satheesh, DJ and Mageed for providing a comfortable and co-operative working environment. I would also like to thank my project students Praveen and Shruthika for helping me to become a good mentor. I am indebted to my batchmates Shruthi, Amla, Rudraditya and Suman for the happy time we spent together during my study.

Words are short to express my deep sense of gratitude towards Ravi Bhaiya, Manisha and Rafiya who selflessly helped me to settle in Hyderabad when I was alone. My stay in Hyderabad could not be possible without their help and support.

I take this opportunity to express the profound gratitude to my beloved parents, I reached this stage only because of their support and their dreams. They have been my pillar of strength through all my ups and downs.

I would like to express my love to my siblings who showered me their love and unconditional support. I also thank my grandparents and relatives for their continued and unfailing love.

The best outcome from these past five years is finding my husband, Prasun. I thank him for his love, support and constant encouragement throughout this journey.

Above all, I thank God for blessing me with such a supportive family. I feel blessed to be a part of their life and give all the credit for what I have achieved and what I will achieve in the future.

Swati Singh

ABBREVATIONS

R-M	Restriction-modification
MTases	Methyltransferases
DNA	Deoxyribo nucleic acid
RNA	Ribo nucleic acid
SAM	S-adenosyl-L-methionine
SAH	S-adenosyl-L-homocysteine
C5mC	C5 specific cytosine
N6mA	N6 specific adenine
N4mC	N4 specific cytosine
MALT	Mucosal-associated lymphoid tissue
VacA	Vacuolating cytotoxin
CagA	Cytotoxin-associated toxin
PIR	Protein Information Resource
HMM	Hidden Markov Model
PDB	Protein databank

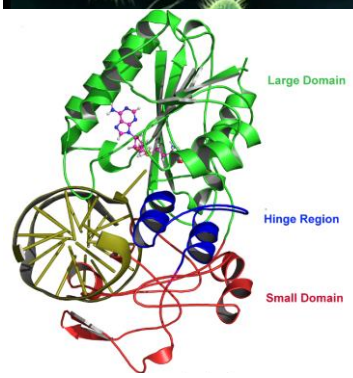
PRF	Protein Research Foundation
UniProt	Universal Protein Resource
3-D	Three-dimensional
BLAST	Basic Local Alignment Search Tool
NCBI	National Center for Biotechnology Information
PSI-BLAST	Position Specific Iterative BLAST
PSSM	Position specific scoring matrix
SSEs	Secondary structure elements
Phyre2	Protein Homology/AnalogY Recognition Engine
TM score	template modeling score
PROCHECK	A program that check the stereochemical quality and geometry of the protein structures
MAPSCI	Multiple Alignment of Protein Structures and Consensus Identification
CD-HIT	Cluster Database at High Identity with Tolerance
MEGA	Molecular Evolutionary Genetic Analysis
SCOP	Structural Classification of Protein
MD	Molecular Dynamics
MSA	Multiple Sequence Alignment
vdW	Van der Waals
RMSD	Root mean square deviation
RMSF	Root mean square fluctuation
GO	Gene ontology
CASTp	Computed Atlas of Surface Topography of proteins

CONTENTS

1

Introduction to *Helicobacter pylori*, DNA MTases and computational methods

- ✓ Reviews the mechanism of colonization, symptoms and therapies used against the *H. pylori* infections.
- ✓ Provides brief introduction about DNA MTases and its various classes.
- ✓ Outlines the computational methods used in the study.



1.1 Pathogenesis of *Helicobacter pylori* infection

1.1.1 History

Helicobacter pylori is a primitive organism and is a predominant pathogen within human populations for over 60,000 years (Moodley et al., 2012). According to World Gastroenterology Organization, approximately half of the world's population is affected by its infection reaching the prevalence up to 70% in developing countries while less in industrialized countries (20–30%). The discovery of *H. pylori* in 1982 by Barry Marshall and Robin Warren (Warren JR, 1983), and its association with peptic ulcer disease was a great achievement in the field of medicine. Both the scientists were jointly awarded Nobel Prize in Physiology or Medicine in 2005 for the unearthing of "The bacterium *H. pylori* and its role in gastritis and peptic ulcer disease" since this was the first time when *H. pylori* was known as a causative agent of peptic ulcer and it brought about major change in the understanding of their etiology.

H. pylori is a spiral shaped (The "H" in the name is short for *Helicobacter* "Helico" means spiral) and Gram-negative bacterium present as leading pathogen of human gastric niche of stomach. It is microaerophilic bacterium having optimal growth conditions at 5–19% O_2, 5–10% CO_2, 37°C and high humidity; similar to conditions present in the gastric mucosa. Being a gastric pathogen, the high acidity of the stomach juice is lethal for the bacteria and it can still survive inside host using different strategies. Without treatment, colonization of *H. pylori* can persist lifelong inside the host.

Before 1982 until *H. pylori* were discovered, ulcers which are common chronic disease are supposed to be caused by dietary habits, stress and lifestyle. In most of the cases, *H. pylori* infections are chronic but asymptomatic *i.e.* patients never experience clinical signs of colonization, but some may experience mild gastric inflammation. Several studies indicated that apart from negative effects, there are also some health benefits associated with *H. pylori* colonization including protection from esophageal adenocarcinoma, allergic airway disease, gastroesophageal reflux disease, diarrheal disease and obesity, suggesting that the relationship between *H. pylori* and human host is complex and dynamic (Arnold et al., 2011; Cover and Blaser, 2009). Conversely, numerous factors have been identified that can contribute

to the development of negative outcomes with respect to *H. pylori* infection (Atherton, 2006). These epsilon proteobacterium are etiologically linked with the increased risk for various harmful disease outcomes in the case of acute infection. These outcomes include chronic active gastritis, peptic ulceration (McColl, 1997), duodenal ulcer, dysplasia, neoplasia, gastric B-cell lymphoma of mucosal associated lymphoid tissue (MALT lymphoma) and in severe cases invasive gastric adenocarcinoma.

Further to this, the pervasiveness and the severity of *H. pylori* infection substantially differs among various populations and individuals (Correa and Piazuelo, 2008). These outcomes are related to the alterations that have occurred in the genome of the bacterial agent along with the ecological changes taking place in the environment of the human host.

1.1.2 Infection, transmission, epidemiology and disease outcome

H. pylori mostly cultivate in the digestive tract and have a tendency to colonize the stomach lining. Mostly *H. pylori* infections are inoffensive, but their persistent infection may cause ulcers in the stomach and small intestine. A brief overview of probable disease outcomes due to *H. pylori* infections are summarized in Figure 1.1. Peptic ulcers can cause an inflammatory condition inside stomach known as gastritis. It is still not known how *H. pylori* is transferred from one person to other or why some patients are symptomatic after its infection while others do not. *H. pylori* is a paradigm for chronic bacterial infection which induces chronic gastritis; a known risk factor for peptic ulcer and distal gastric cancer, and its perseverance in the gastric mucosa is aided by certain mechanisms of immune evasion and immune modulation.

Majority of *H. pylori* infections are attained in childhood and normally persist throughout life if not treated with antibiotics. The *H. pylori* infection is inversely correlated with socioeconomic status and conditions like overcrowding, underprivileged sanitation and low economic standards are main risk factors associated with its infection. No natural reservoir has been found to be responsible for the spread of *H. pylori* and it is probably transmitted from person to person through oral-oral transmission, because bile in the fecal environment is toxic for *H. pylori* fecal-oral infections. The possible environmental reservoir of bacteria is supposed to be contaminated water sources and is likely to spread from person to

person through the above mentioned routes. The most probable way of transmission is during events of gastroenteritis where *H. pylori* can be isolated not only from vomitus but also from the air. During such disease, *H. pylori* can also survive and spread through diarrhea.

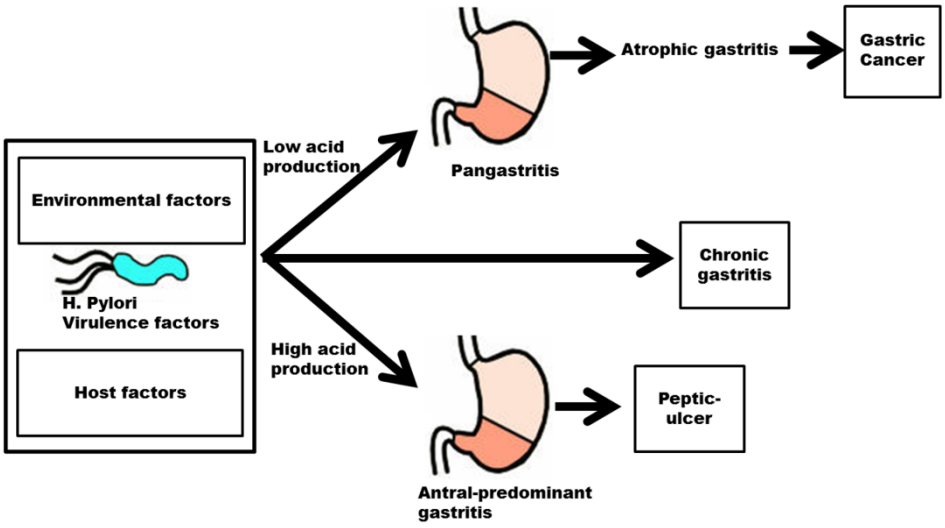


Figure 1.1: Disease outcomes of *H. pylori* infection in the human stomach.

Atrophy can result in a gastric ulcer formation, which is sometimes a precursor to gastric cancer. Even though most of the *H. pylori* gastric colonization are asymptomatic, patient may sometimes complain heartburn, dyspepsia, nausea, vomiting or halitosis while some may suffer with erosive gastritis or ulcers. The infection involves initially the antrum and ultimately may spread proximally to the corpus region of stomach especially in patients receiving antacid medication. Patients who develop duodenal ulcers have low risk of gastric cancer while patients suffering from gastric ulcers usually have multifocal atrophic gastritis with high risk of gastric cancer (Hansson et al., 1996). Some patients with atrophic gastritis may attain intestinal metaplasia and less percentage among them have chance to develop further dysplasia and invasive adenocarcinoma. The intestinal type adenocarcinoma (Lauren, 1965) is more frequent in populations revealing high frequency of gastric cancer and is the final stage in which environmental factors play important etiopathogenetic role in the multistep and multifactorial process of gastric cancer.

1.1.3 Factors affecting successful colonization of *H. pylori* inside host

H. pylori can provoke various strategies to survive and propagate inside the host cells and manipulate their behavior. Disease consequences are greatly affected by host physiology, genotype, dietary habits, bacterial genotype (Hunt, 1996; Labigne and de Reuse, 1996) and host genetic diversity (Peek, 2005). The *H. pylori*

isolates retain significant genotypic diversity, which produces differential host inflammatory responses that impacts the pathological outcome. However, clinical outcomes are not solely dependent upon bacterial virulence factors, but are also influenced by genetic diversity of host, mainly in immune response genes. Some of these factors are summarized in Figure 1.2A.

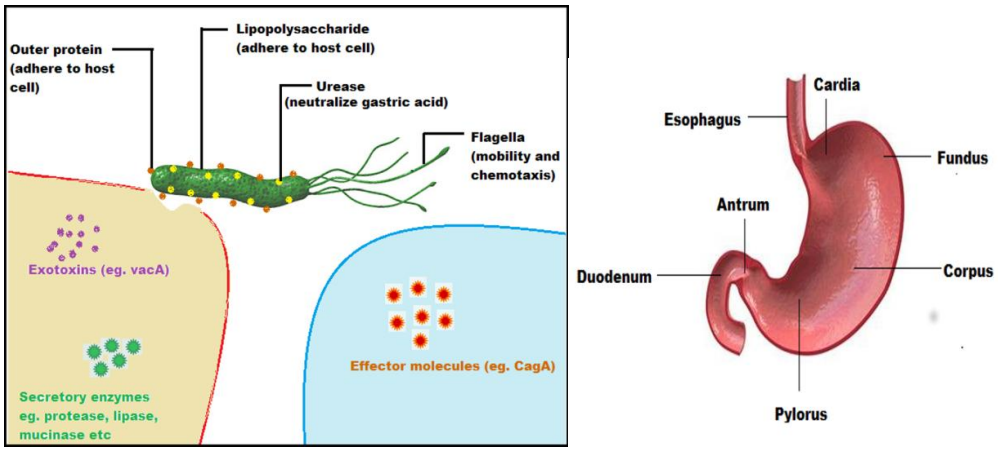


Figure 1.2: (A) Various factors affecting successful colonization of *H. pylori* inside the host. (B) Different regions present in human stomach. (http://www.slideshare.net/sachinpatne570/anatomy-of-stomach-54088362)

1.1.3.1 pH maintenance in the stomach

H. pylori resides in the slimy mucus layer that lines and protects the epithelial cells from the acidic gastric juice. It usually grows in the antrum region of stomach shown in Figure 1.2B. In stomach they mostly occupy gastric mucous layer or may be adherent to the epithelial lining of the stomach. Inside the mucous layer, *H. pylori* is mainly limited to the 100 µm of mucus contiguous to the epithelial cells where pH is relatively neutral. These bacteria also have the ability to change the environment around them and reduce its acidity so that they can survive. To protect themselves from mucus and body's immune cells, they can easily penetrate the stomach lining which is facilitated by their spiral shape. Further they can also interfere with host immune responses to make sure that they are not destroyed. All these factors lead to stomach problems as the disease progresses.

1.1.3.2 Motility and chemotaxis

Motility of *H. pylori* cells along the gastric mucosa is aided by utilization of numerous lophotrichous flagella. Once it enters the gastric mucosa, it interacts with host epithelial cells and elaborates its pili. The *H. pylori* has ability to swim away from the gastric juice and towards the neutral environment close to epithelial cells

since they can sense the pH and bicarbonate gradient in the mucous (Schreiber et al., 2004). *H. pylori* can also penetrate the slimy mucous layer in a cork screw like motion, a function that is essential for survival because of its shape and presence of flagella (Specht et al., 2011; Sycuro et al., 2010). The spiral shape of the bacteria is because of precise peptidoglycan crosslinking and coiled-coil rich proteins along with whirling movements of the flagella that facilitates in this penetration. Basal body, hook and the filament forms flagella structure out of which hook and filament protrude from the basal body. Hook and filament are covered in a membranous sheath that is connecting with the outer membrane. This sheath shields the filament from low pH or against the host immune response (Goodwin et al., 1985; Lertsethtakarn et al., 2011).

During advancement of bacterial growth, the pH initially decreases due to hypersecretion and *H. pylori* might move into the first part of the intestine, the duodenum. This region is less resistant to infection and thus peptic ulcer could develop easily. Long term hypersecretion can cause atrophic gastritis to the mucosa and even loss of the acid producing parietal cells and higher stomach pH. Since the undissociated form of weak acids can pass freely through the cell membrane of any microbe, they possess potent antimicrobial activity. This organism has capability to tolerate the acidic conditions in the gastric environment by generating a positive inner membrane potential in low pH.

Besides pH, *H. pylori* can also sense various environmental clues which include iron, nickel, cobalt and zinc, and responds to these indicators by altering virulence expression (Contreras et al., 2003). These information indicate that the host environment acts as signals and can alter the carcinogenic potential of *H. pylori* thereby increasing the risk of negative disease outcomes. Chronic atrophic gastritis to the mucosa is supposed to be the precursor of gastric adenocarcinoma and MALT lymphoma that is caused by long term hypersecretion and in severe cases, even loss of the acid producing parietal cells and higher stomach pH is observed.

1.1.3.3 Urease

Along with the chemotactic as well as swimming abilities, *H. pylori* uses its cytoplasmic enzyme urease to escape the acidic gastric juice by neutralizing the local environment. Urease is a large 1.1 MDa complex that converts urea to NH_3 and HCO_3^-. NH_3 neutralizes protons in the periplasm and HCO_3^- acts as a buffer to

maintain pH at 6.1 (Scott et al., 2007). Lysis of bacterial cells in the surrounding releases urease and buffers into the microenvironment of the live bacteria, thus the bacterial cells are buffered both internally and externally (Dunn et al., 1997; Marcus and Scott, 2001). This neutralization of the mucin also changes its rheological properties and makes it easier for *H. pylori* to penetrate (Celli et al., 2009).

1.1.3.4 Genomic flexibility and genetic regulation

Stable adaptation of the bacteria is required to survive and grow in hostile ecological niche in the presence of a constant immune response and in substantially changing environment (Suerbaum and Josenhans, 2007). Such adaptive processes include both regulatory mechanisms which act on gene expression and reversible or irreversible genome changes (Fischer et al., 2014). It is hypothesized that the variability within the bacterial genome helps the bacterium to survive and further facilitates chronic colonization, which is again supported by the fact that almost each infected individual is inhabited by a genetically unique strain of *H. pylori*. Recombination events generating allelic diversity are frequent while genome changes involving gain or loss of genes seem to be rare phenomenon in *H. pylori* genome . Allelic diversity is a remarkable property of *H. pylori* strains and is caused by high mutation rates and frequent recombination events (Kraft et al., 2006).

Evolution of clusters of genes in genomic islands from the analysis of various whole genome sequences in different strains of *H. pylori* shows the presence of distinct areas of variability, named as “plasticity zones or plasticity regions”. These regions are genetic locus with high variation and are likely to be involved in horizontal gene transfer (Alm et al., 1999). Horizontal gene transfer is supported by the occurrence of coding regions and short conserved integration motifs. These coding regions are orthologous to integrating conjugative elements (ICEs) and are widely distributed among all sequenced *H. pylori* strains. These ICEs provide benefit to the bacterium by assisting genetic recombination events, which could eventually help *H. pylori* in immune circumvention of host and increased ability to colonize (Fischer et al., 2014). Moreover, mutation rates are higher in *H. pylori* than other bacterial systems and most of the mutations occur in genes encoding putative outer membrane proteins (Linz et al., 2014).

To better understand chronic infectious processes, it is necessary to identify the numerous regulators responsible for mediating complex interactions with the host

which includes features which contribute to changes in gene regulation. These features include modulating expression of ureC (encoding a subunit of the urease complex), cagY (encoding a type IV secretion component), flgE (encoding a flagellar component) (Furuta et al., 2014).

1.1.3.5 Pathogenicity/ toxins encoded in the *H. pylori* genome

The pathogen's chronic infection inside host is successful balance between the host and pathogen that is strongly controlled by expression of virulence factors. These virulence factors are expressed to accomplish vital functions within the host like:

(i) Provoke an immune response which eliminates inhabitant microbiota,

(ii) obtain nutrients,

(iii) permit bacterial penetration inside host tissues and

(iv) allows the bacteria to change the host tissue into a replicative niche.

As a strategy to escape both innate and adaptive immune systems of host body, an immune modulating effect is exerted by *H. pylori*. Several genes of *H. pylori* have been identified as virulence associated and might have significant clinical and epidemiological insinuations. Among them, two toxins, the vacuolating cytotoxin (VacA) and the cytotoxin associated toxin (CagA), have been studied most extensively and involved in perturbing host immunological responses (Cover and Blaser, 1992; Sundrud et al., 2004; Tummuru et al., 1993). *H. pylori* has progressively evolved to adapt external environmental stimuli, such as gastric acid, with a range of regulatory elements such as two component systems which includes sensor kinase (ArsS) and a response regulator (ArsR), which eventually modulates the expression of virulence genes involved in motility and Cag function (Scott et al., 2007).

Type IV secreted effectors play vital roles in the virulence e.g. Cag pathogenicity island (40 kb DNA region) contains the cagA gene and encodes a type IV secretion system which assists in the export of the CagA protein into the epithelial cells (Allen et al., 2000; Censini et al., 1996). The *vacA* gene encodes a vacuolating cytotoxin which damages epithelial cells of host (Palframan et al., 2012). The cagA gene exists in only few strains (>90% of isolates from East Asian countries and in 50-60% of isolates from Western countries) and may possibly have resulted from acquisition of DNA from other bacteria (Censini et al., 1996). Infection with CagA

positive *H. pylori* strains has been linked with greater chances for development of peptic ulcer (Covacci et al., 1993; Tham et al., 2001) and gastric adenocarcinoma (Parsonnet et al., 1997). The *H. pylori* infection in its interaction with the host, is capable to adapt and thus produce new genotypes through mutations and DNA rearrangements. High genetic variability of a strain to another, such as VacA and CagA, not only affects the body's ability to colonize and cause disease but also affects inflammation and gastric secretion.

1.1.3.6 Dietary contribution

The molecules derived from the host diet also help in successful colonization of *H. pylori* in the human gastric niche. Various dietary habits like nitrite, protein, iron deficiency, salt preference and fat intake have been linked with increased risk of *H. pylori* related disease (Zhang et al., 2013). Dietary iron deficiency has been associated with increased risk of *H. pylori* progression (Pra et al., 2009). Low iron conditions induce expression of several virulence factors including urease, VacA and CagA (Noto et al., 2013). Likewise, dietary salt intake has been linked with increased risk of gastric disease and its higher intake has been linked to greater chances of gastric cancer and inflammation. Analysis of bacterial and host transcripts revealed that CagA and IL-1β, respectively, were found to be extremely upregulated in *H. pylori* infected animals in response to dietary salt intake (Gaddy et al., 2013). As a result, variations in dietary iron consumption may lead to alterations of bacterial virulence factor expression that eventually affects the disease progression.

1.1.3.7 Host factors associated with disease

There are many reports which support that the abnormality in the host pathogen interaction, originated through promotion of inflammation, destroys resident microbiota in favor of the pathogen and finally resulting in disease progression (Behnsen et al., 2014; Sassone-Corsi and Raffatellu, 2013). Genomic analyses associated with these, have revealed that both *H. pylori* and human have coevolved in order to promote less severe gastric lesions and disruption of the coevolution (Kodaman et al., 2014).

Other host factors for e.g. MALT lymphoma characterized from microarray analyses are found to contribute in *H. pylori* associated disease outcome. MALT lymphoma has been shown to penetrate gastric tissue with $CD4^{(+)}$ T cells which further expresses CD28 and CD69 with an enhanced expression of calprotectin

(neutrophil associated protein) (Mueller et al., 2005). Supplementing to this study, several subsequent studies have associated Th1, Th17 as host molecules that are induced against *H. pylori* infection (Algood et al., 2007). Interestingly, polymorphisms in IL-1β and the IL-8 promoter region can raise the possibility of *H. pylori* related diseases such as gastric cancer as observed from whole genome expression profiles and sequencing methods (Wang et al., 2009; Xue et al., 2012).

1.1.4 Diagnosis

Endoscopy is commonly used to test the infection caused by *H. pylori* in the stomach by taking a small sample (biopsy) from the lining of stomach. Urea breath test is also used to check the presence of *H. pylori* bacteria in the stomach. A blood test is sometimes used to check whether made antibodies are present in blood against *H. pylori* infection.

1.1.5 Treatment

In most of *H. pylori* infections, the treatment methods are focused on the use of a proton pump inhibitor and antibiotics such as metronidazole and clarithromycin plus either bismuth subsalicylate, ranitidine bismuth citrate or a proton pump inhibitor [http://www.who.int/vaccine_research/documents/Helicobacter_pylori/en/]. H_2 blocker or proton pump inhibitor drugs are used for suppression of acid production. These drugs upon consumption with the antibiotics, help to reduce ulcer related symptoms (abdominal pain, nausea), gastric mucosal inflammation and may increase effectiveness of the antibiotics against *H. pylori* at the gastric mucosal surface.

1.1.6 Evolution of *H. pylori* and geographic distribution of strains

Various tools in genomics such as genome sequencing, restriction fragment length polymorphism (RFLP) genome mapping, as well as analytical methods, like maximum likelihood analysis and multilocus sequence typing (MLST) are currently used to study *H. pylori* pathogenesis. These approaches reveal a notable amount of genetic diversity between various clinical isolates of *H. pylori.* This diversity is supposed to be primarily driven by a high mutation rate, random genetic drift and frequent recombination events along with positive Darwinian selection and fixation of base substitutions.

By the way of migration of human populations across the globe, their endemic *H. pylori* strains diverged together with them leading to phylogeographic

diversity of this pathogen within human populations and can be classified into European, Amerindian, Asian, and African subgroups (Camorlinga-Ponce et al., 2011). Geographic regions like Latin American Andes Mountain region low socioeconomic standards have high *H. pylori* infection rates and higher cases of gastric cancer (Correa and Piazuelo, 2008). Unusually, in Africa, India, Thailand, Bangladesh, Pakistan, Iran, Israel, Malaysia, and Saudi Arabia with comparable socioeconomic situations, infection rates are quiet high, but gastric cancer incidence is relatively low (Hellmig et al., 2003; Misra et al., 2014).

Often the phylogeographic origin of its strain describes specific host adaptive responses through variations in virulence factor expression like European strains of *H. pylori* have lesser virulence compared to the African strains (Sheh et al., 2013). These analyses also suggest that *H. pylori* and human coevolution have been distressed in some geographic areas. *H. pylori* in India shares common evolutionary ancestry with European strains, indicating a possible acquirement of these strains in the period of colonization by European imperial forces (Devi et al., 2007). The changes in sequence of amino acid can be related with increased risk for peptic ulcer and can further be correlated with variations in geographic origin of the *H. pylori* strain (Cao et al., 2005).

For several strains of *H. pylori*, complete genome sequences are available, strains HPAG1 and B128 are isolated from patients with chronic atrophic gastritis (Oh et al., 2006) and gastric ulcer (Israel et al., 2001), strains 26695 as well as J99 from superficial gastritis suffering patients (Israel et al., 2001) and duodenal ulcer (Alm and Trust, 1999) while strain 98-10 from a patient with gastric adenocarcinoma (Ando et al., 2002). From phylogenetic analysis, *H. pylori* strains can be distinguished into three major divisions, the West African (J99), the European (HPAG1, 26695, B128) and East Asian (98-10) (McClain et al., 2009). The comparative genome studies of five *H. pylori* strains (98-10, B128, J99, 26695 and HPAG1) identified 1237 common genes that have been considered to represent the *H. pylori* core genome (McClain et al., 2009). Genes that are conserved in all the *H. pylori* population are involved in various functions related to stomach colonization, adaptation, reproduction, metabolism and etc.

Atrophic gastritis is a premalignant lesion and *H. pylori* infected patients with gastric ulcer disease have increased chances of developing gastric cancer compared to patients suffering from duodenal ulcer (Hansson et al., 1996; Parsonnet, 1996).

Hence strains 98-10, HPAG1 and B128 were isolated from patients having diseases related to gastric cancer. Hence genes found exclusively in these strains are associated with gastric cancer.

1.2 Computational methods for genome annotation

The ultimate aim of genome sequencing projects is to assign structure and function to the protein which is important to understand the biological processes at the molecular level in any organism (Wierenga et al., 1986). Ever since the DNA structure was discovered more than half a century ago, the complete genome sequence of several prokaryotes as well as eukaryotes have been determined till date, thus increasing the number of protein sequences and structures in databases such as UniProt and PDB. With the exponential growth in the determination of protein sequences and structures by genome sequencing and structural genomics efforts, there is an emergent necessity for consistent computational methods to determine the biochemical function of all proteins identified by sequencing methods.

Proteins are the most essential and versatile macromolecules of life, and understanding of their functions is a critical link in the development of new drugs, better crops, and even the development of synthetic biochemicals for example biofuels. The accurate annotation of protein function is important to understand life at the molecular level and has biomedical and pharmaceutical implications. Experimental procedures for protein function prediction are difficult because of low throughput and high expense thus unable to annotate the vast amount of sequence data that are becoming available due to rapid advances in genome sequencing technology.

To address the rising gap between sequence data and protein functional annotations several computational methods have been developed to predict the function of protein over the past few decades (Bork et al., 1998; Friedberg, 2006; Lee et al., 2007; Rost et al., 2003; Sharan et al., 2007). So far the task of computational functional implication frequently depends on traditional methodologies like finding domains or using BLAST hits amongst experimentally determined proteins with known function. Automatic annotation of function based on sequence similarity to a gene from a different species is a risky practice since this can generate false confidence in the annotation. Several other methods have also been suggested to exploit these data, together with function prediction from amino acid sequence (Clark and Radivojac, 2011), protein-protein interaction networks (Nabieva et al., 2005), protein structure data (Laskowski et al., 2005), microarrays (Huttenhower et al.,

2006), inferred evolutionary relationships and genomic context (Gaudet et al., 2011), or a combination of data types (Sokolov and Ben-Hur, 2010). Many large scale efforts have aimed to provide annotations for these sequences. For example, Swiss-Prot (Boeckmann et al., 2005), gene ontology (GO) (Ashburner et al., 2000), the Human Proteomics Initiative (Boeckmann et al., 2005) and the Protein Structure Initiative (Norvell and Berg, 2007), one of the key triumphs of the last decade, thoroughly defines biological function using ontologies which includes molecular function, cellular localization and biological processes.

For about 50-60% of protein sequences from fully sequenced proteome, there is a lack of detailed knowledge about their structure and function. When the function of a protein is unknown, a probable function is assigned by simple bioinformatics analyses which includes sequence and three-dimensional (3-D) structure comparisons using programs like BLAST (Altschul et al., 1990) and DALI (Holm and Rosenstrom, 2010). Comparative modeling analysis is a useful tool to predict the 3-D structures of unknown sequence. In the lack of homology, only the automated prediction of protein function can bridge the gap between available sequences and annotations. From the protein structure, first method is the prediction of protein binding sites and functional hotspots, while the second method includes comprehensive *in silico* mutagenesis experiments. These mutagenesis experiments further improve novel predictions of protein function. Common method used to explore the importance of a residue to a specific interaction involves mutating it, usually with alanine, and then measuring the effect of this substitution on the interaction (Wells, 1991). This is frequently done sequentially on a large scale in a procedure called as an 'alanine scan'. Methods have also been developed to predict the residues important for catalysis and the local spatial arrangements of these residues can be used to identify protein function.

Genome sequence analysis revealed that the circular genome of *H. pylori* strain 26695 consists of 1,667,867 base pairs with 1,590 predicted coding sequences and has well-built systems for motility, scavenging iron, DNA restriction and modification (Tomb et al., 1997). Several putative adhesins, lipoproteins and other outer membrane proteins have been identified in the complete proteome as possible partners for host pathogen interactions in *H. pylori* 26695. In the second chapter of this thesis we have tried to annotate whole proteome of *H. pylori* 26695. This annotation of the proteome would help in identifying better drug targets that can be

exploited to tackle *H. pylori* infections and it is one of the ultimate goals of this sequencing project. Further identification of mechanisms that regulate *H. pylori* and host interactions will not only help in developing targeted diagnostic and therapeutic modalities, but can also provide understanding into other diseases that arise from pathogen initiated inflammatory states.

1.3 DNA methyltransferases and *H. pylori*

Genome analysis of various pathogens has provided understandings into their virulence, host adaptation and evolution. Genetic information provides the blueprint for the assembly of all proteins necessary to perform function, whereas the epigenetic information provides instructions that regulate expression of genes. Both epigenetic alterations as well as genetic mutations are involved in transforming normal cell to cancerous cell, and hence their manipulation holds great potential for cancer prevention, detection and therapy. Since, in different forms of cancer, a range of epigenetic mechanisms are altered like dysregulation of DNA binding proteins, silencing of tumor suppressor genes and activation of oncogenes by transforming methylation patterns of CpG island (short stretch of DNA in which the frequency of the CG sequence is higher than other regions). Cancer epigenetics term is precisely used for the study of epigenetic modifications to the genome of cancer cells without altering the nucleotide sequence.

Epigenetics is the field of genetics and refers to changes in gene expression without changing the original DNA sequence *i.e.* a change in phenotype without a change in genotype. These changes lead to cellular and physiological phenotypic trait alteration in the organism that is affected by external or environmental factors. Epigenetic change is a consistent and natural occurrence, however can also be influenced by numerous factors including age, lifestyle and environmental factors. However three systems which include DNA methylation, histone modification and non-coding RNA associated gene silencing are presently considered to induce epigenetic change and ongoing research is continuously revealing the role of epigenetics in a variety of human diseases. Thus epigenetics has strongly emerged as an important field to better understand the subtle and complex nature of gene regulation.

Based on enzyme subunit composition, DNA specificity characterization, cofactor requirement and reaction product restriction-modification (R-M) systems in bacteria are divided into three types - I, II and III (Wilson and Murray, 1991). Subsequently, all the R-M systems are involved in shielding bacteria from the transformation of DNA from invaders (bacteria or phages) by recognizing specific sequence and cleaving precisely either within or very close to recognition sequences

(Takahashi et al., 2002). Among all R-M systems, type-II is less complex and consists of two polypeptides which perform complimentary function: one is restriction endonuclease and other is DNA methyltransferase (MTase). Restriction endonuclease cleaves DNA while corresponding DNA MTase protects endogenous DNA from endonuclease ingestion by methylating the endonuclease recognition DNA sequence (Roberts, 1990). As discussed above, *H. pylori* is naturally competent and susceptible to take DNA from the surroundings (Suerbaum et al., 1998) and also several bacteriophages are known to infect *H. pylori* (Heintschel von Heinegg et al., 1993). Therefore, multiple R-M systems present in the pathogen might be required to defend its genome integrity. To compensate for the propensity of R-M systems to accumulate inactivating mutations, various R-M systems may be required in *H. pylori*. Biological benefits like escaping the human immune response by changing surface antigens (lipopolysaccharide and cell surface associated proteins) can result from phase variation and mutation. Additionally these R-M systems are believed to be a primitive bacterial 'immune' system of *H. pylori* cells (Kong et al., 2000).

DNA methylation process is catalyzed by enzymes that belong to the category of DNA MTases and has severe effects on both genome architecture as well as gene expression (Kovall and Matthews, 1999). DNA methylation has major biological roles which include: distinction of self and non-self DNA, direction of post replicative mismatch repair, control of DNA replication and cell cycle. MTases are involved in regulating expression of not only DNA, but RNA, proteins and some small molecules, like catechol. DNA MTases are widespread among prokaryotes and are sequence specific DNA binding enzymes that methylate adenine or cytosine residues in recognition sequences of DNA using S-adenosyl-L-methionine (SAM or AdoMet) as the methyl donor forming methylated base of DNA and S-adenosyl-L-homocysteine (SAH or AdoHcy), as shown in Figure 1.3A and B. This covalent addition of a methyl group to the C5 or N4 position of cytosine or N6 position of adenine has several effects like genomic imprinting, transcriptional repression, X chromosome inactivation, chromatin structure modulation and the suppression of the harmful effects of parasitic DNA sequences on genome integrity/stability (Baylin and Herman, 2000; Jones and Laird, 1999; Robertson and Wolffe, 2000).

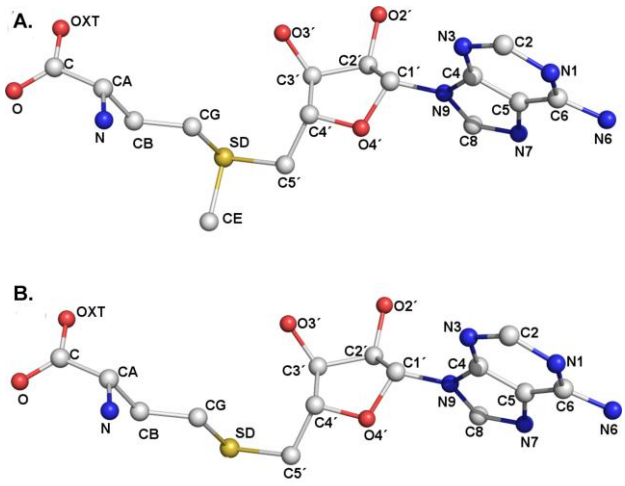


Figure 1.3: Structure of (A) S-adenosyl-L-methionine (SAM or AdoMet) and (B) S-adenosyl-L-homocysteine (SAH or AdoHcy).

Several pathogens, together with *H. pylori*, are well known to introduce virulence factors into eukaryotic host cells by a range of mechanisms. Analysis of the restriction enzymes database (REBASE) depicts that *H. pylori* encodes a large number of known or putative DNA MTases. Only few prokaryotes can possess such a large number of R-M systems or DNA MTases. *H. pylori* possesses an extraordinary large number of R-M system genes having complete restriction and modification activities, or incomplete, *i.e.* orphan MTases (Kong et al., 2000; Vitkute et al., 2001). A majority of the DNA MTases encoded by *H. pylori* irrespective of its class, (both adenine and cytosine specific) belong to type II R-M enzymes (http://tools.neb.com/genomes/index.php?page=H). Using single-molecule real-time sequence analyses of the methylome of closely related *H. pylori* strains, great diversity in the methylation of target sequences have been observed. This result is accredited to dissimilarity in the variation within the methylation target sequence besides specificity of the MTase domain (Furuta et al., 2014; Krebes et al., 2014). Although the link between several viruses and cancer has been widely studied, it is noteworthy that for nearly two decades, *H. pylori* has remained sole bacterial pathogen that has been systematically linked with cancer. Further to this, *H. pylori* is the only bacterium that is classified as a class I carcinogen by the WHO (2000) (Cao et al., 2008; Sue et al., 2015; Vogiatzi et al., 2007). Multiple factors contribute to the etiology of different stages of gastric cancer and a mechanism for carcinogenesis resulting from *H. pylori* triggered inflammation was first proposed (Correa, 1992). The preliminary stages of *H. pylori* infections (gastritis and atrophy) have been found to be associated with excessive salt intake while ingestion of ascorbic acid and nitrate

have been linked to disease progression in intermediate stages. The final stages have been linked with the excessive salt intake as well as with supply of β-carotene (Correa, 1992; Jenab et al., 2006; Kato et al., 2006). Analyses of most studied genome sequences of *H. pylori* strains 26695 (Tomb et al., 1997) and J99 (Alm et al., 1999) revealed that each strain contained more than two dozen genes likely to encode MTases. This MTase coding gene number exceeds far more in *H. pylori* as compared than other bacterial genomes (Kong et al., 2000) and these MTase genes consists of a large fraction of all genes that were specific to one or the other strain. MTases are involved in various cellular processes like regulation of transcription of specific genes (van der Woude et al., 1998), DNA replication (Messer and Noyer-Weidner, 1988), mismatch repair (Lahue and Modrich, 1988), or DNA transposition (Roberts et al., 1985). Some of them can be essential for viability (Stephens et al., 1996) or virulence in pathogens or some may be involved in conferring specificity in the bacterium host interactions (Heithoff et al., 1999). In the sequenced strains 26695, J99 and HPAG1 (Alm et al., 1999; Oh et al., 2006; Tomb et al., 1997), the MTases are situated upstream of the corresponding restriction endonuclease (REase). The MTases of the three strains are closely homologous including their protein lengths (359 amino acids). The J99 and HPAG1 have 97% and 98% identities, respectively, to strain 26695 whereas 26695 and J99 REases shares 93% identity and comprise 290 amino acids.

Effector molecules, such as CagA and VacA, present in *H. pylori* have been studied in great detail and are proposed to be associated with carcinogenesis. *H. pylori* are a unique member of the gastric microbiota that affects its host in various ways like facultative intracellularity, encoding a large amount of functional DNA MTases. The analysis of the effects of DNA MTases as well as various restriction endonucleases and other proteins of the host microbiota can possibly uncover the novel interactions between evolutionarily unrelated species. To adapt the ecological niche by perturbation of host gene expression, possibly these proteins acts as effectors of interspecific epigenetic signals which may assist commensals, symbionts and pathogens to survive.

1.3.1 Structure of DNA MTases

Based on the chemistry of methylation, DNA MTases are divided into two groups: endocyclic DNA MTase or C-MTase: which forms a C-C bond (C5mC MTases) and

exocyclic amino DNA MTase or N-MTase: which forms a C-N bond (N6mA and N4mC MTases) as shown in Figure 1.4.

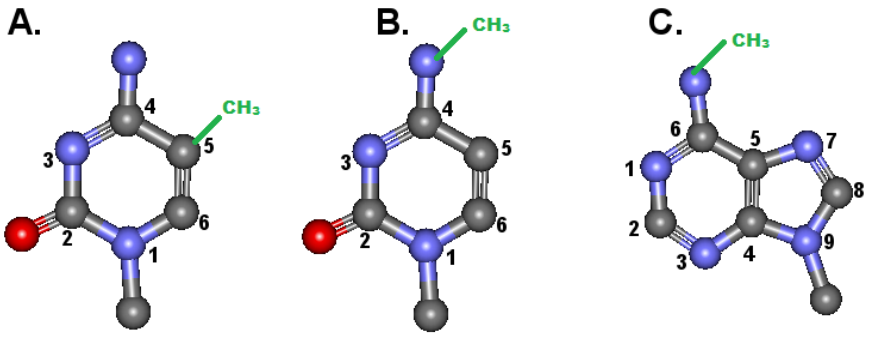


Figure 1.4: Structures of (A) methylated cytosine base at C5 (B) methylated cytosine base at N4 and (C) methylated adenine base at N6.

In DNA MTases, due to the phenomenon of circular permutation of the amino acid sequence of the large domain as well as with respect to the location of insertion of the small domain into the structure of the large domain, N or exocyclic MTases can be further subdivided into six classes: α, β, γ, ζ, δ and ε (Jeltsch, 1999; Malone et al., 1995; Wilson, 1992). Each class has different arrangement of the conserved motifs which are AdoMet-binding domain (FXGXG), the TRD (target recognition domain) and the catalytic domain (DPPY) (Malone et al., 1995) (Figure 1.5.).

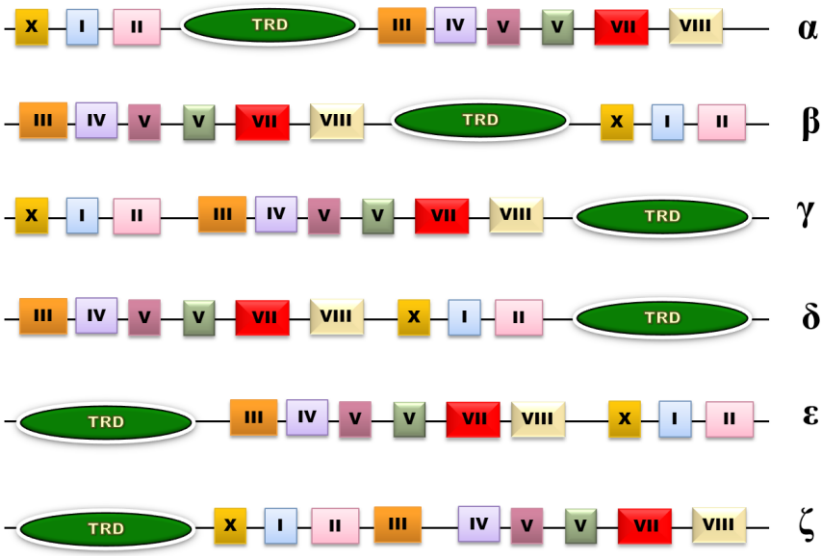


Figure 1.5: Arrangement of the conserved motifs in the large domain and TRD region in different classes of DNA MTases.

Crystal structures of DNA MTases from different organisms pertaining to different classes like *Haemophilus haemolyticus* which is a C5mC DNA MTase (M. HhaI) (Cheng et al., 1993a; Cheng et al., 1993b; Klimasauskas et al., 1994), N6mA DNA MTase from *Thermus aquaticus* (M. TaqI) (Schluckebier et al., 1997), N4mC

DNA MTase from *Proteus vulgaris* (M. PvuII) (Gong et al., 1997a) have been studied in detail till date. In *H. pylori*, all three types (C5mC, N6mA and N4mC) of DNA MTases are observed in different strains. In chapters three and four of the thesis, we have characterized hypothetical proteins from the cancerous strain of *H. pylori* 98-10 as C5mC and N6mA class of DNA MTase using computational methods.

Methyl transfer mechanism is also present in catechol *O*-methyl transferase (COMT) systems which are involved in the inactivation of the catecholamine neurotransmitters (dopamine, epinephrine and norepinephrine). The enzyme introduces a methyl group to the catecholamine, which is donated by SAM.

The structures of DNA MTases from all groups share a common bilobal arrangement having two domains: catalytic and DNA recognition domains. In between these two domains a V-shaped cleft accommodates the DNA substrate and the basic nature as well as size of the cleft is consistent with binding of duplex DNA. The large domain consists of the catalytic and AdoMet binding sites located at the bottom of this cleft, while recognition domain conveys the substrate DNA specificity of each DNA MTase. Catechol MTase on the other hand consists of only a single domain which is comparable to the catalytic domain of the DNA MTases (Vidgren et al., 1994).

1.3.2 Target base flipping

The most remarkable structural and mechanistic feature of DNA MTases is complete flipping out of target base from DNA helix and binding of flipped base into a hydrophobic pocket present in the large domain of the enzyme that creates the active site for methyl group transfer (Goedecke et al., 2001; Klimasauskas et al., 1994; Reinisch et al., 1995). The binding pocket of substrate DNA is formed by residues from motifs IV, VI and VIII which are located in the large domain of the MTase at the ends of β4 and β5 and the beginning of β7, respectively. The reason for this unusual mechanism might be in the catalytic process of endocyclic and exocyclic DNA MTases which requires an intimate contact of aromatic ring (purine or pyrimidine) of the target base with the enzymes; which would only be possible if the base were flipped outside the DNA double helix. Moreover, other DNA interacting enzymes were also recognized that make use of base flipping which include many

DNA repair enzymes like uracil-DNA glycosylase and T4 endonuclease V (Roberts, 1995; Roberts and Cheng, 1998).

In all available structures of different DNA MTases in complex with target DNA, base flipping has been observed. However, the structural adaptations of the DNA after base flipping are dissimilar. In M. HhaI (recognition sequence 5'GC^{m}GC3'), the DNA retains an almost B-DNA like structure with the exception of the flipped target base. The N1, N2 and O6 atom positions of the orphan guanine base which was left unpaired because of base flipping interacts with a glutamine residue from the small domain of the enzyme that is inserted into the DNA and occupies the space of the flipped cytosine (Klimasauskas et al., 1994). In the M. HaeIII DNA MTase structure, the bases of 5'GGC^{m}C3'DNA recognition sequence undergoes extensive rearrangements.

All DNA MTases contain a set of ten or nine conserved amino acid motifs depending upon endocyclic and exocyclic DNA MTases class (Malone et al., 1995; Wilson, 1992). A comparison of topology of large domain and location of conserved motifs in C5mC, N6mA and N4mC is shown in Figure 1.6. The exocyclic amino MTases mostly fall into α, β and γ classes. M. BssHI is the only DNA MTase for which the ζ architecture has been confirmed (Bujnicki, 2002; Xu et al., 1997). As depicted from arrangement pattern of conserved motifs, endocyclic MTases are found to resemble the γ class of exocyclic amino MTases. The N4mC DNA MTases show more flexibility in motif arrangement, and may belong to any of the three classes of exocyclic amino MTases but mostly belong to the β class of exocyclic amino MTases (Ahmad and Rao, 1996a; Schluckebier et al., 1995b). Although the motif arrangements in the sequence are quite different in both exocyclic and endocyclic DNA MTases, but function of these motifs remain almost conserved. For e.g. the AdoMet binding region comprises of motifs I to III and X (Cheng et al., 1993a; Cheng et al., 1993b; Klimasauskas et al., 1994; Labahn et al., 1994) while motifs IV, VI, and VIII are responsible for catalysis (Schluckebier et al., 1995b) in both M. HhaI and M. TaqI.

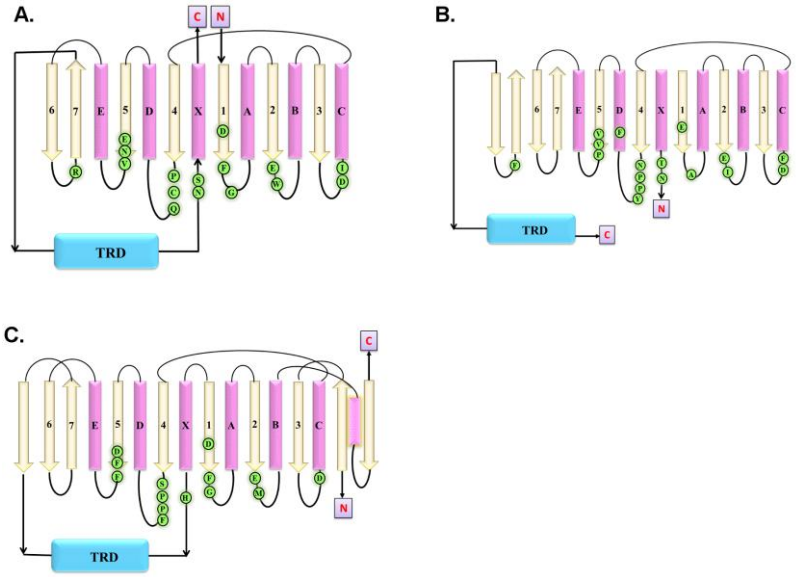


Figure 1.6: Comparison of the topology of large domain and location of conserved motifs in (A) C5mC (B) N6mA and (C) N4mC DNA MTases.

1.3.3 Comparison of conserved motifs among the DNA MTase families

Motif I (FAGxGG in M. HhaI) is shared by other AdoMet dependent MTases and consists of a glycine rich consensus sequence and forms a loop known as 'G-loop' (Ingrosso et al., 1989; Klimasauskas et al., 1989; Lauster et al., 1989). This motif plays an important part in binding of cofactor i.e AdoMet. Motif IV (PCQ in M. HhaI) is a part of the catalytic site in C5mC DNA MTases but not present in both exocyclic amino DNA MTases. Sequence comparison of N6mA and N4mC DNA MTases indicated the presence of (D/N/S)PP(Y/F) motif in the catalytic site (Klimasauskas et al., 1989). The structural comparison of M. HhaI and M. TaqI has also confirmed that both motifs are present at topologically equivalent positions, thus confirming their correspondence with each other (Schluckebier et al., 1995b). In both endocyclic as well as exocyclic amino group of DNA MTase enzymes, the other functionally important residues of the catalytic domain are located at the carboxyl ends of the parallel β-strands that are clustered on the face of the DNA binding cleft in M. HhaI and M. TaqI, and around the catalytic site in COMT. Table 1.1. shows AdoMet recognizing amino acids in the three classes of these enzymes. These amino acids are found in equivalent positions in the secondary structure that are part of the conserved motifs, I, II, III and V, using the nomenclature of C5mC MTases (Posfai et al., 1989). Motifs II and III are termed as less conserved motifs in C5mC MTases (Posfai et al., 1989). Motif II comprises a negatively charged residue at the last position in β2-strand which interacts with the ribose hydroxyls of AdoMet that is followed by a bulky hydrophobic side chain that makes vander Waals (vdW) contacts

with the AdoMet adenine (Glu40-Trp41 in M. HhaI, Glu71-Ile72 in M. TaqI). Motif III contains an Asp/Glu or Asn/Gln (Asp60 in M. HhaI and Asp89 in M. TaqI) in the first position of αC, which interacts with the exocyclic N6 of the adenine moiety of AdoMet. Additionally this motif also provides a hydrogen bond to N1 atom of the AdoMet adenine from a peptide backbone NH group (Ile61 in M. HhaI and Phe90 in M. TaqI).

Motif IV has the consensus sequence Asp-Pro-Pro-Tyr (DPPY) in class α and β for N6mA MTases and Asn-Pro-Pro-Tyr (NPPY) in class γ while in the case of N4mC MTases, Ser-Pro-Pro-Tyr (SPPY) motif is present (Wilson & Murray, 1991; Wilson, 1992). There are notable exceptions to the above mentioned pattern, most remarkably M. BamHI (N4mC in class β, which has a DPPF), M. HhaII (N6mA in class β, DPQY) and M. StsI-a (N6mA in class α, DTPY).

Motif V comprises of the consensus (Asn/Asp)-Leu-Tyr-X-X-Phe-(Leu/Val/Ile). In class γ, this Phe replaces with the G-loop Phe in the MTases of classes α or β. The Leu (Leu100 in M. HhaI, Leu142 in M. TaqI) makes vdW contacts to the AdoMet adenine on the same side as the Phe (Schluckebier et al., 1995b).

Table 1.1: List of AdoMet recognizing amino acids in the three classes of DNA MTases- C5mC, N6mA and N4mC DNA MTases.

Motif	M. HhaI (PDB ID: 5MHT)	M. TaqI (PDB ID: 2ADM)	M. PvuII (PDB ID: 1BOO)	Function
I	AGxGG	PAxAxGP	DXFXG	Forms the loop β1-αA, and the main chain NH of the first residue of αA hydrogen bonds with terminal carboxyl oxygen of AdoMet
II	E40 (β2)	E71 (β2)	E294 (β2)	The last residue of β2-strand; side chain hydrogen bond with the ribose hydroxyls
II	W41 (β2-αb)	I72 (β2-αB)	M295 (β2-αB)	Face-to-face vdW contact with adenine ring
III	D60 (αC)	D89 (αC)	D34 (β3-αC)	The first residue of helix αC; side chain hydrogen bonds to N6 atom
III	I61 (loop)	F90 (αC)	S35(αA)	Main-chain NH hydrogen bonds to N1 atom AdoMet
V	L100 (β4-αD)	L142 (β4-αD)	F273 (β4-αD)	vdW contacts with adenine ring on the same side of F18 in M. HhaI
	F18 (β1-αA)	F146 (αD)	F79 (β4-αD)	Edge-to-face vdW contact with adenine

In all the structurally characterized MTases, motif VI forms β5-strand (Schluckebier et al., 1995b). In β and γ MTases, a conserved Gly is present at the beginning of the strand, and Gly, Pro or Ala are present at the end of the strand, in class α it ends with Ser-Asn. The classes β and γ also vary from class α, in having a conserved hydrophobic residues in β5. Motif VII is not strongly conserved even among C5mC MTases, yet credible candidates can be found within each class. Motif VIII has little resemblance to the motif present in C5mC MTases (Gln161-X-Arg-X-Arg165 in M. HhaI). This probably reveals the fact that the C5mC DNA MTases interact with cytosine *via* hydrogen bonds (through Arg165 in M. HhaI), while the N6mA MTases appear to interact with the adenine of substrate DNA *via* hydrophobic interactions. In the structure of M. TaqI, the corresponding region (the loop connecting strands β6 and β7) contains Phe196, which aligns to a conserved Phe or Tyr in other amino MTases. It has been suggested that Phe196 makes favorable edge-to-face or face-to-face vdW contacts with the substrate DNA adenine (Schluckebier et al., 1995b). The position of motif X in the primary sequence of DNA MTases is one of the key differences present between the endocyclic and exocyclic amino DNA MTases. In the C5mC DNA MTase, this motif originates from the carboxy terminus while in exocyclic amino MTases, this motif is always to the amino side of motif I.

1.3.4 Comparison of large domain structure in DNA MTases

All DNA MTases consists of a conserved large catalytic domain, comprising of two subdomains having AdoMet binding and catalytic function respectively. This conserved domain is supposed to be evolved from a common ancestor and gene fusions have possibly supplemented in the formation substrate recognition domains leading to the evolution of large family of MTases.

Large domain shares common structural core, consisting of α/β type with a central mixed β-sheet around which several α-helices are arranged. The strand is formed by six-stranded parallel β-strand with a seventh strand inserted in an antiparallel manner between the fifth and sixth strands. The β-sheet begins in the middle with strand 1 (6↑7↓5↑4↓1↑2↑3↑) and are reversed once with a switch point between β4 and β1 which divides the domain into two subdomains. The right subdomain (β1-β3) creates the AdoMet binding site, the left subdomain (β4-β7) creates the binding site for the extrahelical target base. Both binding sites are hydrophobic pockets that are located at equivalent positions within the subdomains, a

fact which proposes that the catalytic domain possibly has evolved by gene duplication (Malone et al., 1995). At the carboxyl ends of these two β-strands, the conserved motifs FAGxGG, PCQ in M. HhaI, and PAxAxGP, NPPY in M. TaqI are located, which are part of AdoMet binding and catalytic sites, respectively.

In conclusion, the catalytic domains of the DNA MTases in both endocyclic and exocyclic amino DNA MTases are structurally and functionally analogous (Figure 1.7). This was not expected on the basis of the primary sequences of these three enzymes alone. Only the alignment of secondary and tertiary structures identifies individual amino acids that have comparable functional properties.

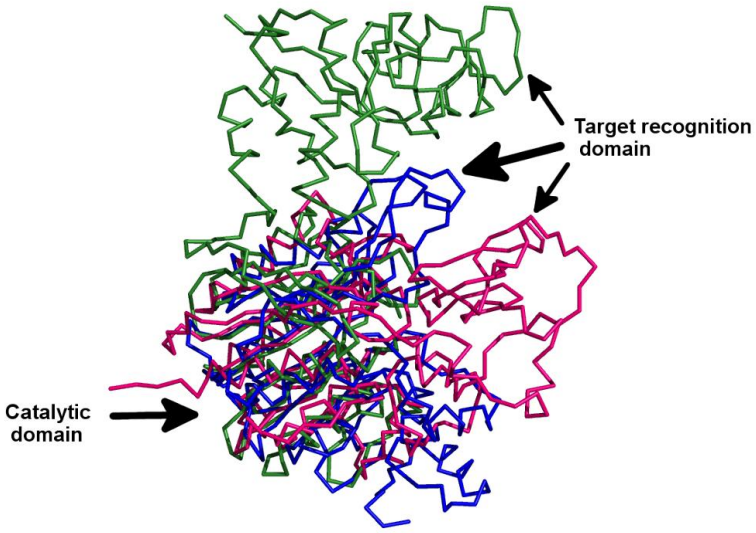


Figure 1.7: Superimposition of catalytic and target recognition domain in M. HhaI (pink), M. TaqI (green) and M. PuvII (blue) DNA MTases.

1.3.5 Small domain structure in DNA MTases

The small domains of different DNA MTases are divergent in terms amino acid sequence, size and structure. Structures of two bacterial C5mC DNA MTases, M. HhaI (Cheng et al., 1993b) and M. HaeIII (Reinisch et al., 1995) are studied in detail. Even though the catalytic domains of both proteins are easily superimposed on each other, their small domains are dissimilar in secondary structure arrangement and cannot be superimposed properly. For e.g. in M. HhaI, the small domain (81 amino acid residues) has 7 β-strands while the M. HaeIII (92 amino acid residues) has no extensive secondary structures. This heterogeneity is even more in the case of exocyclic amino MTases. So far, four crystal structures are known from different organisms: M. TaqI (Labahn et al., 1994), M. PvuII (Gong et al., 1997a), DpnM (Tran et al., 1998) and M. RsrI (Scavetta et al., 2000). The size of the small domain

in M. TaqI MTase is 177 residues. This divergence in structure goes parallel with a difference in function; since many but not all residues of the small domains of DNA MTases are involved in sequence specific contacts with the DNA. So these contacts facilitate the recognition of the DNA sequence in the consensus region that is characteristic for each enzyme. A typical structure of C5mC DNA MTase in complex with cofactor AdoMet and target DNA is shown in Figure 1.8.

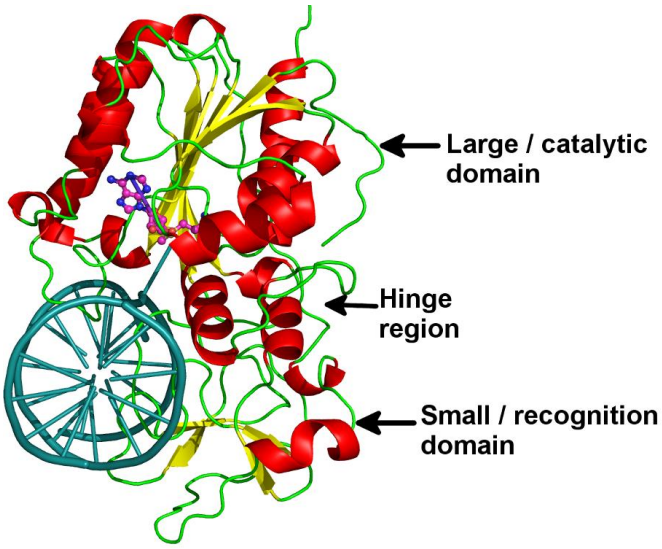


Figure 1.8: Structure of M. HhaI C5mC in complex with cofactor AdoMet (ball and stick representation) and substrate DNA.

1.3.6 Mechanism of DNA MTases

1.3.6.1 Catalytic mechanism of endocyclic DNA MTases

Even though AdoMet is a very effective methyl group donor, still methylation at C5 position of cytosine is a difficult phenomenon. This is due to the presence of an electron poor heterocyclic aromatic ring system in cytosine which makes C5 atom inert and incapable of making a nucleophilic attack on the methyl group of AdoMet. Therefore, the enzymatic reaction catalyzed by DNA MTase follows the reaction pathway of a Michael addition which involves the formation of transient covalent adduct between cysteine of enzyme and C6 of flipped cytosine (Santi et al., 1984; Wu and Santi, 1987). Initially a cysteine SH group present in the invariant ProCys dipeptide (motif IV; Cys 81 in M. HhaI) from active site of the enzyme acts as a catalytic nucleophile, to attack at C6 position of the cytosine ring resulting in the formation of a covalent complex between the enzyme and DNA. This complex serves as an intermediate for the methylation reaction and subsequent to its formation there is a generation of high energy carbanion at C5 position. This carbanion is further

stabilized either by resonance or completely escaping protonation at endocyclic nitrogen atom N3 (Chen et al., 1993b).

It was proposed that methylation occurs favorably on cytosine protonated at N3 position, mediated by Glu119 (motif VI (ENV)) with the help of Arg165 from motif VIII in the case of M.HhaI (O'Gara et al., 1996). This leads to increase charge density at C6 position and activating the C5 of the cytosine for the attack of methyl group (Baker et al., 1988). Increased charge at C6 predisposes it for nucleophilic attack by SH group of cysteine, which results in the formation of a carbanion with the loss of proton from N3 position as shown in Figure 1.9A.

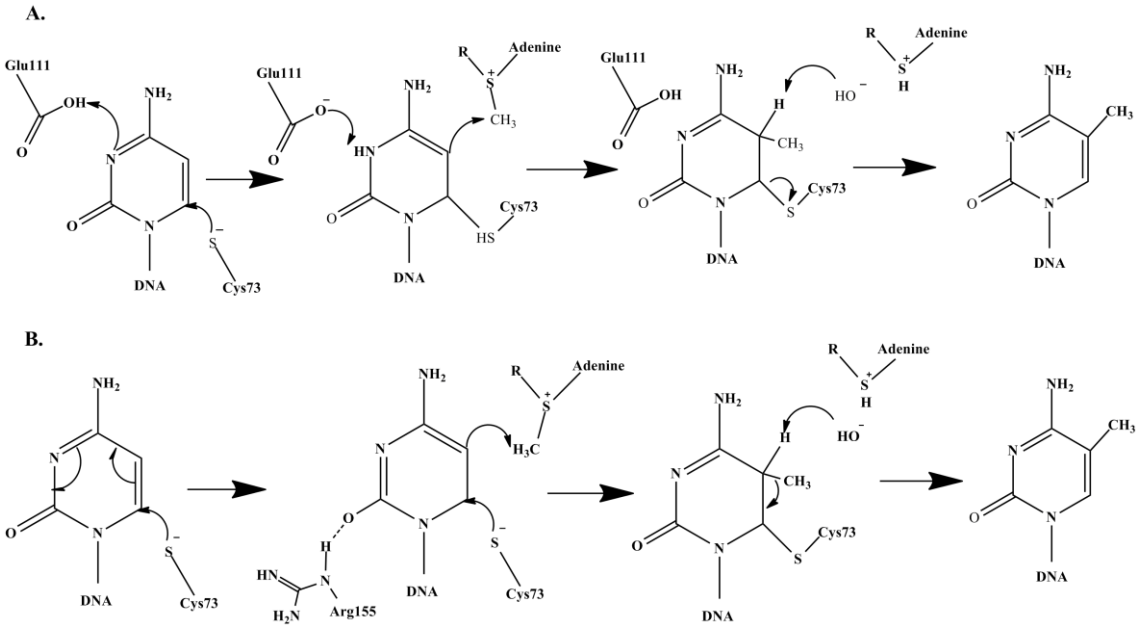


Figure 1.9: Catalytic mechanism of endocyclic C5mC DNA MTases. Mechanism showing stabilization of carbanion by (A) avoiding by the protonation at N3 by glutamic acid or (B) by resonance stabilized

The generated carbanion is resonance stabilized leading to increased charge at C5 and O2 positions of cytosine. Previous studies have shown that Arg165 present in close proximity of O2 atom of cytosine probably accepts increased charge at O2 position (Klimasauskas et al., 1994) as shown in Figure 1.9B. Although it is difficult to propose which reaction is occurring in the system, strong evidence for the protonation at N3 position is available from crystal structure where hydrogen bond is present between Glu119 and N3 (O'Gara et al., 1996).

Activated C5 then attacks the methyl group of AdoMet resulting in the formation of dihydroxycytosine intermediate and AdoHcy molecule. The proton at C5 position is abstracted by basic residue of enzyme with a concurrent elimination of enzyme from C6 position. The covalent enzyme DNA complex is detached by

deprotonation at C5 position which leads to the elimination of the cysteine SH group and restores the aromaticity. The nature of proton abstracting base is still doubtful but involvement of a phosphodiester group has been proposed (O'Gara et al., 1996).

1.3.6.2 Catalytic mechanism of exocyclic amino DNA MTases

Like methylation of cytosine at C5 position, methylation of the exocyclic amino groups of cytosine and adenine is difficult because neither of the two bases (adenine or cytosine) exhibit nucleophilicity at the exocyclic amino group as their free electron pairs is conjugated with the aromatic systems as shown in Figure 1.10. In contrast to endocyclic MTase, the methylation reaction of exocyclic amino groups proceeds with inversion of the configuration of the methyl group in an SN^2 reaction without formation of a covalent intermediate (Pogolotti et al., 1988). In fact, noticeable nucleophilicity can be only detected at N3 position of cytosine and at the N1 and N3 positions of adenine, which in DNA are not the targets for enzyme mediated methylation reactions. Knowing the fact that the specificity of N6mA and N4mC DNA MTases overlap (N6mA can also accept cytosine as a target for methylation and *vice versa* for N4mC) (Jeltsch, 2001; Jeltsch et al., 1999a), it was predicted that the reaction mechanisms of both types of enzymes are very analogous. This conclusion is further aided by the fact that N4mC DNA MTases are found in α, β and γ classes of exocyclic amino DNA MTases (Bujnicki and Radlinska, 1999; Malone et al., 1995). Furthermore, both these results suggest that during the molecular evolution, the target base specificity of exocyclic amino DNA MTases has been altered several times (Bujnicki and Radlinska, 1999; Jeltsch et al., 1999a). N6mA and N4mC DNA MTases are characterized by conserved (D/N/S)PP(Y/F) residues present in motif IV. These conserved residues are located in the active site of the enzyme at topologically equivalent positions to the PCQ motif of the endocyclic DNA MTases.

Further, mutational studies performed in different DNA MTases within this region showed decrease in the catalytic activity (Kong and Smith, 1997; Roth et al., 1998; Sugisaki et al., 1991). According to the structure of the N6mA MTase of M. TaqI (Goedecke et al., 2001), the most important function of this above mentioned conserved tetrapeptide is that the side chain of the D/N/S and the main chain carbonyl group of the proline act as hydrogen bond acceptors for the protons of the exocyclic amino group. As the acceptor groups are presented in a tetrahedral

geometry, a change in hybridization of the nitrogen atom from *sp2 to sp3* is induced which localizes the free electron pair at the N6 position. In the case of M. PvuII, serine is attached to a glutamic acid by a charge relay system, which might allow the serine to act as a base. Therefore it is possible that the D/N/S residue also functions as proton acceptor during the reaction (Gong et al., 1997a; Scavetta et al., 2000).

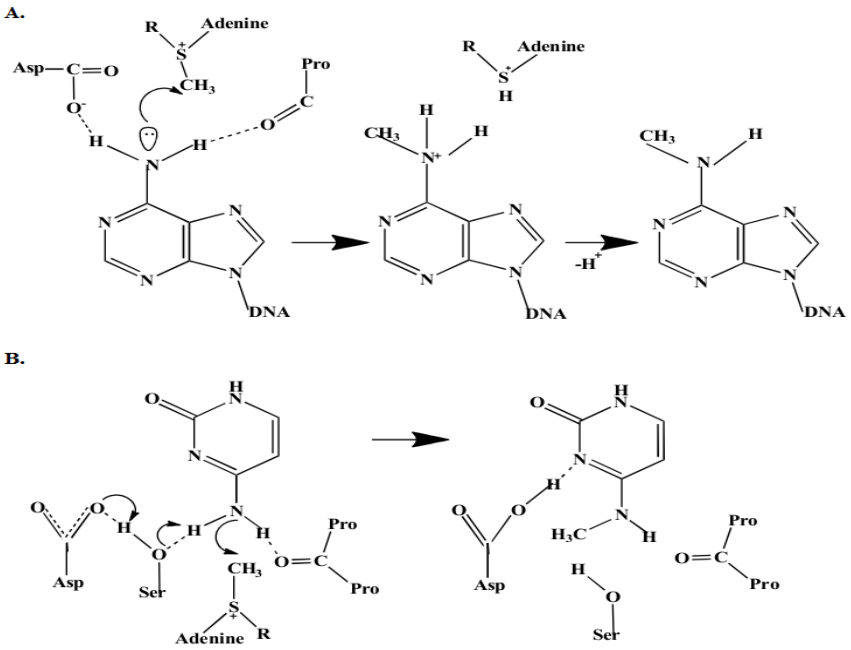


Figure 1.10: Catalytic mechanism of exocyclic amino DNA MTases. The figure is based on the structure of the (A). M. TaqI DNA complex and (B). M. PuvII structure, the aspartate and proline in the figure are from motif IV of the exocyclic amino MTase (D/N/S)PP(Y/F).

Most likely, the D/N/S residue is only momentarily protonated and instantaneously transfers the proton to other residues and lastly to water. Therefore, a cationic transition state has been hypothesized which is further stabilized by cation…π interactions with surrounding aromatic amino acid (Schluckebier et al., 1998). This model is supported by the fact that the active sites of all exocyclic amino DNA MTases contain one or more aromatic amino acids which make contact with the flipped base of substrate DNA (Friedrich et al., 1998; Holz et al., 1999; Jeltsch et al., 1999b) like the Y/F residue in the catalytic tetrapeptide. However C5mC DNA MTases usually lack aromatic amino acid residues near the active site. It has been shown from several mutational analysis that an exchange of these aromatic residues by non-aromatic residue reduces the catalytic efficiency much more than an exchange by another aromatic residue (Friedrich et al., 1998; Kong and Smith, 1997; Pues et al., 1999; Roth et al., 1998; Roth and Jeltsch, 2001).

1.4 Introduction to computational methods

1.4.1 Proteins: databases, sequence and homology search

1.4.1.1 Databases

The aim of most protein databases is to organize and annotate the protein sequences, providing the biological community access to the experimental data in a useful way since protein sequences are the fundamental determinants of biological structure and function.

NCBI (The National Center for Biotechnology Information): NCBI is a part of the United States National Library of Medicine (NLM), a branch of the National Institutes of Health. The Protein database in NCBI is a collection of sequences from several sources, including translations from annotated coding regions in GenBank and RefSeq, as well as records from Swiss-Prot, PIR (Protein Information Resource), PRF (Protein Research Foundation) and PDB (Protein Data Bank).

UNIPROT (The Universal Protein Resource): UniProt is a comprehensive resource for protein sequence and annotation data. The UniProt databases are the UniProt Knowledgebase (UniProtKB), the UniProt Reference Clusters (UniRef) and the UniProt Archive (UniParc). The Universal Protein Resource (UniProt) is the world's most comprehensive catalog of information on proteins. It is a central repository of protein sequence and function created by joining the information contained in Swiss-Prot, TrEMBL and PIR.

1.4.1.2 Protein sequence

The amino acid sequence is the order in which amino acid residues, connected by peptide bonds reside in the protein chain. The sequence is commonly described from the N-terminal end containing free amino group to the C-terminal end containing free carboxyl group. Protein sequence is represented as the primary structure of a protein. The most commonly used sequence format is FASTA, also called as the "Pearson" format. For example;

>gi|216945262|gb|EEC23940.1| hypothetical protein HPB128_151g1 [Helicobacter pylori B128]

MNLLSLFAGAGGLDLGFEKAGFKIVVANEYDKNITPTYRLNHKNTQLLEKDIKNLQTSEINFSVDGIIGGPPCQSWSEA
GNLKGIDDARGQLFYEYLRLLKELKPKFFLAENVRGMLAQRHKTSVKNILNAFKECGYEVNTHLVNAKDYGVAQER
LRVFYIGFRRDLKVNYIFPKGSTHLKKLTLKDVIWDLKDSVVCALAKNKRNPNAINNHEYFIGSYSPIFMSRNRVKNW
DEQAFTIQASGRQCQLHPQAPKMAKFGKNDCRFIENYQHLYRRLSVRECARIQGFDDDFYFVYENLNDAYKMIGNAV
PVSLAKEIAISIYKVLH

A sequence in FASTA format consists of:

First line starting with a ">" and a sequence identification code. It is optionally followed by a textual description of the sequence.

One or more lines containing the sequence.

A file in FASTA format may comprise of more than one sequence.

PIR (Protein Information Resource) format: Example:

```
>P1;gi| 2392799|pdb|5MHT|A
>5MHT:A|PDBID|CHAIN|SEQUENCE
MIEIKDKQLTGLRFIDLFAGLGGFRLALESCGAECVYSNEWDKYAQEVYEMNFGEKPEGDITQVNEKTIPDHDILCAGF
PCQAFSISGKQKGFEDSRGTLFFDIARIVREKKPKVVFMENVKNFASHDNGNTLEVVKNTMNELDYSFHAKVLNALD
YGIPQKRERIYMICFRNDLNIQNFQFPKPFELNTFVKDLLLPDSEVEHLVIDRKDLVMTNQEIEQTTPKTVRLGIVGKGG
QGERIYSTRGIAITLSAYGGGIFAKTGGYLVNGKTRKLHPRECARVMGYPDSYKVHPSTSQAYKQFGNSVVINVLQYI
AYNIGSSLNFKPY*
```

A file in PIR format may comprise more than one sequence. The PIR format is also often mentioned as the NBRF format. A sequence in PIR format consists of:

One line starting with a ">" sign, followed by

A two letter code describing the sequence type (P1, F1, DL, DC, RL, RC, or XX), followed by a semicolon, followed by the sequence identification code (the database ID-code).

One line having a textual explanation of the sequence.

One or more lines comprising the sequence itself. The end of the sequence is marked by a "*" (asterisk) character. Optionally, this can be followed by one or more lines describing the sequence.

1.4.2 Molecular structures and visualizers

In general, the 3-D structures of molecules are obtained by experimental methods in the crystalline state or solution and sometimes bound to target molecules. The structures of small molecules can be predicted by spectroscopic methods and high quantum chemical computation methods to provide highly accurate molecular geometries in the vapor phase. But due to large size and variations in the conformations of the drug molecules, these computational methods cannot provide accurate results. However, with the availability of high speed computing facilities, the computational methods can give sufficiently accurate structures for drugs in the gas or liquid phase.

The 3-D view of the molecule can give proper information and the conformational details. Some of the structural properties can be calculated by using 3-D molecular structures. Several free and commercial software are available for 3-D visualization. Some of them are: Jmol (http://jmol.sourceforge.net/), Rasmol (http://rasmol.org/),

Pymol (http://www.pymol.org/), VMD (http://www.ks.uiuc.edu/Research/vmd/), Chimera (http://www.cgl.ucsf.edu/chimera/), Discovery Studio Visualizer, Swiss PDB viewer, molmol, Tinker (http://dasher.wustl.edu/ffe/), Argus Lab (www.arguslab.com/), Chime, Deep View, PMV and etc.

1.4.3 Basic Local Alignment Search Tool (BLAST)

The BLAST program uses heuristic algorithm to align a query sequence with all sequences in database to find high scoring ungapped segments among related sequences (Altschul et al., 1990). The existence of such segment above a given threshold indicates pairwise similarity beyond random chance, which aids to differentiate related sequence from unrelated sequence in a database. It includes following steps:

Seeding: Create a list of words from the query sequence which is usually three residues for protein and eleven for DNA sequence. This list consists of every possible combination of words from the query sequence.

Database search: A sequence database is searched for the occurrence of these matching words.

Generation of substitution matrix: The extent of matching is scored by a given substitution matrix and a word is considered as a match only if it is above the threshold.

Pairwise alignment is done between query and database sequence by extending from the word in both directions calculating the alignment score by same substitution matrix. The extension continues until alignment score drops below the threshold values because of mismatching (drop threshold is 22 for protein and 20 for DNA). Resulting contiguous aligned segment pair without gap is known as high scoring segment pair (HSP) or maximum scoring pairs.

An improvement of BLAST is ability to provide gapped alignment where dynamic programming is used to introduce gap in HSPs. Process of extension continues in the same way, however overall score is allowed to drop below threshold only if its temporary and rises again to attain above threshold values.

The BLAST output provides a list of pairwise sequence matches which are rated by statistical significance. This significance score helps to distinguish between evolutionary related and unrelated sequences. In BLAST search, this statistical indicator is E-value or expectation value which indicates the probability that the

resulting alignment from a database search are caused by random chance. The E-value is related to P-value used to evaluate significance of single pairwise alignment. E-value is determined by following formula

$$E = m \times n \times P;$$

where m indicates total number of residues in the database, n is the number of the residues in the query sequence and P indicates the probability that an HSP alignment is a consequence of a random chance.

The E-value provides information about the probability that a given sequence matches the database sequence by chance only. Lower the E-value, the less likely is the chance random database match and therefore more significant match.

Another statistical indicator in BLAST search is bit score which measures sequence similarity independent of query sequence length and database size. Bit score is normalized based on the raw pairwise score (S') and is determined by:

$$S' = (\lambda \times S \times \ln K)/ \ln 2;$$

where λ is Gumble distribution constant, S is raw alignment score and K is constant associated with scoring matrix used. Bit score is linearly related to raw alignment score, thus higher the bit score the more significant is the match.

1.4.4 Position Specific Iterative BLAST (PSI-BLAST)

PSI-BLAST program is used to find distant relatives of a protein. The program makes a list of all closely related proteins. PSI-BLAST is the most sensitive BLAST program, making it suitable for searching very distantly related proteins or any novel members of a protein family (Altschul et al., 1997). PSI-BLAST is used when the standard BLAST search either fails to find significant hits, or gives hits with descriptions such as "hypothetical protein" or "similar to" proteins.

For the similarity search of a protein one can use "nr protein database" which searches the available databases like Swiss-Prot, PIR, PDB and the entries with absolutely identical sequences have been merged. PSI-BLAST search against the PDB, a storehouse of the 3-D structural data of proteins and nucleic acids which have been experimentally obtained by X-ray crystallography or nuclear magnetic resonance (NMR), over a number of iterations give the homologous crystal structures.

1.4.5 Multiple Sequence Alignment (MSA)

MSA is an important aspect of sequence analysis to identify and measure similarities between samples of DNA/RNA or protein. An alignment is the vertical arrangement of sequences of 'residues' (nucleotides or amino acids) that maximizes the similarities between them. A MSA arranges three or more sequences, so that residues having common structural positions or ancestral residues are aligned in the same column in a group of sequences and gaps are inserted in the sequences, if required. If two sequences in an alignment share a common ancestor, the mismatches within the sequence can be inferred as point mutations while gaps as insertion or deletion mutations are introduced in one or both lineages, when they diverged from one another. MSA often provides an understanding of evolutionary history of sequences. The function and structure of an unknown protein is predicted by aligning its sequence with others of known function and structure, and also in the prediction of probes for the same family of sequences related to same or different organisms. MSA can build consensus sequences of known families, domains, motifs or sites. Linking these predictions with primary biochemical data can deliver valuable insights into protein structure and function.

Clustal Omega: It is a MSA program which uses seeded guide trees and Hidden Markov Model (HMM) profile-profile method to generate alignments between three or more sequences. Clustal Omega is a fully automated program for global multiple alignment of nucleotide and protein sequences. Clustal Omega generates multiple sequence alignments and a phylogenetic tree, it produces biologically significant multiple sequence alignments of divergent sequences. The alignment in Clustal Omega is achieved *via* three steps: 1) pair wise alignment 2) guide tree generation and 3) progressive alignment. Evolutionary relationships can be observed in a diagrammatic form by viewing Cladograms or Phylograms. It can manipulate existing alignments and carry out profile analysis (Sievers et al., 2011). Clustal Omega provides several options, such as use of slow or fast pair wise alignments, DNA or protein sequences, protein weight matrix, gap open, gap extension, end gaps and gap distances. This program is available for sequence based searches *via* the web server (http://www.ebi.ac.uk/ Tools/clustalw2/).

1.4.6 Phylogenetic trees

A phylogenetic or evolutionary tree, represents the evolutionary relationships among a set of organisms or groups of organisms, called taxa. The tips of the tree

represent groups of descendent taxa (often species) and the nodes on the tree represent the common ancestors of those descendants. Two descendants that split from the same node are called sister groups. A tree is a branching diagram showing the inferred evolutionary relationships among various biological species or other entities and their phylogeny which are based upon similarities and differences in their physical or genetic characteristics.

Molecular Evolutionary Genetics Analysis (MEGA): It is freely available software for conducting statistical analysis of molecular evolution and for constructing phylogenetic trees. It is a user friendly software for mining online databases, building sequence alignments and phylogenetic trees, and using methods of evolutionary bioinformatics in basic biology and evolution (Tamura et al., 2013).

1.4.7 Fold recognition methods

When sequence comparison methods using BLASTP against PDB are no longer sensitive enough to recognize structural homologs for a sequence, fold recognition methods are helpful in assigning the structural fold adopted by the sequence thereby detecting distantly related proteins. Protein folding is the physical process by which a polypeptide folds into its characteristic and functional 3-D structure. Some methods are based exclusively on sequence information and other methods are based on multiple sequence alignment and structural information. Various methods used for fold prediction are FUGUE (Shi et al., 2001), GenTHREADER (Jones, 1999) and ROBETTA (Kim et al., 2004) etc.

GenTHREADER: It is a secondary structure prediction method that incorporates two feed forward neural networks and performs an analysis on the results attained from PSI-BLAST output. The pGenTHREADER method (Lobley et al., 2009) for fold recognition and identification of distant homologs uses profile-profile alignments and predicted secondary structure (using PSIPRED) (McGuffin et al., 2000) as inputs. Output consists of several tabs, each tabs consists of link to tables of the output statistics for each GenTHREADER job. The output table displays the number of structural hits for the query sequence which include full PDB chains for GenTHREADER and pGenTHREADER. For each structure the first portion of the table gives summary statistics which includes:

Confidence : The hit confidence category is based on P-value which is GUESS (<1), LOW (<=0.1), MEDIUM (<=0.01), HIGH (<=0.001), CERT (<=0.0001)

Net Score: is GenTHREADER raw score

P-Value

Pair E: The Pairwise Energy

Solv E: The solvation Energy

Aln Score: Pairwise alignment score

Aln Len: length of the alignment

Str Len: length of the structural hit

Seq Len: The length of the query sequence

The last portion of the table associates to other resources and has the following columns

View Alignment: A button that opens JalView to view an annotated alignment. Known ligand binding residues are annotated on the hit.

SCOP Codes: A link that searches SCOP for the PDB chain (genTHREADER and pGenTHREADER only).

CATH Codes: A link that searches CATH for the PDB chain (genTHREADER and pGenTHREADER only).

Structure: A thumbnail image of the hit, clicking the link will take you to PDBSum.

CATH Entry: A link that searches CATH web services to summarise the hit.

FUGUE: This program is useful for the secondary structure fold recognition and recognizing distant homologs by sequence-structure comparison (Shi et al., 2001). Conventional profile and HMM methods that use both sequence and structure information generally helps to improve the performance of homology recognition by taking into account the extra structural information. It uses environment specific substitution tables and structure dependent gap penalties, where scores for amino acid matching and insertions/deletions are calculated based on the local environment of each amino acid residue in a known structure. The input is given in the form of a query sequence or a sequence alignment, FUGUE scans a database of structural profiles, calculates the sequence-structure compatibility scores and gives output in the form of a list containing potential homologs and alignments. One can get the combined information from both multiple sequences and multiple structures. The prediction is evaluated on the basis of Z score, which has to be ≥ 6.0 for a confident prediction of the fold. The FUGUE program is available at the website (http://tardis.nibio.go.jp/fugue/prfsearch.html). Once the fold of an unknown

sequence is identified, protein 3-D structure modeling methods can be employed to model the same.

1.4.8 Structural Classification of Protein (SCOP) database

This database provides an exhaustive and comprehensive explanation of the relationships of known protein structures (Murzin et al., 1995). This classification is on hierarchical levels: the first two levels, family and superfamily, describe near and distant evolutionary relationships; the third, fold, describes geometrical relationships and the fourth describes the class of protein. Protein sequences of unknown structure are matched to distantly related proteins of known structure using pairwise sequence comparison methods to find homologs. Hierarchical levels in the classification of the proteins in SCOP database is as follows:

Family- Proteins are clustered into families if they satisfy one of two criteria that imply their having a common evolutionary origin: all proteins having residue identities 30% and greater and proteins with lower sequence identities but sharing close similarity in functions and structures. For e.g. globin family.

Superfamily- Families whose proteins share low sequence identities but their similar structures and functional features propose their common evolutionary origin, are placed together in one superfamily; for example, the variable and constant domains of immunoglobulins.

Fold- Superfamilies and families are defined to have common fold if their proteins share common secondary structural arrangement along with identical topological connections. These structural similarities of proteins in the same fold category possibly arise from the physics and chemistry of proteins preferring certain packing arrangements and chain topologies.

Class. The different folds have been grouped into one of the five structural classes which includes:

1. all-α: Secondary structure comprises mainly α-helices;
2. all-β: Secondary structure comprises mainly by β-strands;
3. α/β: Secondary structure comprises of both α-helices and β-strands;
4. α+β: Secondary structure comprises of α-helices and β-strands are largely segregated;
5. multi-domain: Secondary structure comprises of different fold and for which no homologs are known at present.

1.4.9 Protein 3-D structure modeling

1.4.9.1 Protein tertiary structure prediction

Understanding the molecular function of proteins is greatly enhanced by insights gained from their 3-D structures. This structural information provides a basis for understanding protein function and for the design of modified proteins and ligands, including drugs. The structures of proteins are being solved in increasing numbers, particularly as a result of structural genomics projects. Therefore, the number of protein structures that can be modeled are also rising concomitantly. Since experimental structures are only available for a small fraction of proteins compared to the known sequences, computational methods for protein structure modeling plays an increasingly important role. Homology modeling is one such comparative structure prediction method that is widely used to build models of proteins with unknown structures based on the known structures of related proteins. Comparative protein structure modeling is currently the most accurate method, generating 3-D models suitable for various applications which include explanation of biochemical observations, structure guided drug development and virtual screening.

Secondary structures are local stable conformations of a polypeptide chain and are critically important in maintaining a protein 3-D structure. The secondary structure prediction methods are based on two approaches: *ab initio* method or homology based. The *ab initio* method uses single sequence information while homology based methods make use of multiple sequence alignment. The *ab initio* methods predict secondary structure based on statistical calculations of the residues of a single query sequence. It measures the propensity of each amino acid residue belonging to a certain secondary structure element. These propensity scores are derived from known crystal structures. These methods include Chou-Fasman method and Garnier, Osguthorpe, Robson (GOR) methods. Both these methods are developed in 1970s and are first generation methods. Homology based methods are developed in late 1990s and used evolutionary information to predict the secondary structures. It combines the *ab initio* secondary structure prediction of individual sequence as well as alignment information derived from MSA of homologous sequence (>35% similarity).

Three different approaches used to predict 3-D structure of a protein are, sequence-homology modeling, threading and *ab initio* methods. First two are knowledge based methods which rely on the knowledge of already existing protein

structure information in databases. Homology modeling generates model based on experimentally determined structures which are closely related to the query protein sequence. Threading method identifies proteins that are structurally similar but may or may not have sequence similarity. The *ab initio* methods on the other hand uses simulation based approaches and predicts 3-D structure based on physiochemical principles which governs protein folding without using known structural template.

1.4.9.2 Homology modeling

Homology modeling or comparative modeling is a method for building an atomic resolution model of a protein from its amino acid sequence, also called as query sequence. Homology modeling technique is based on the identification of one or more known protein crystal structures (templates or parent structures) possibly resemble the predicted query structure. We can align the secondary structures with each other and map residues of the template sequence on to the query sequence. The template structure and sequence alignment between template structure and query sequence are taken as inputs to generate a structural model of the query sequence. It is generally accepted that proteins with high sequence similarity also possess structural similarity (Marti-Renom et al., 2000). For proteins that share greater than 30% sequence identity, the root mean square deviation (RMSD) of the Cα coordinates is observed to be less than 1Å (Geourjon et al., 2001).

The homology modeling procedure is carried out in four sequential steps:

(i) Template selection: This step involves searching the PDB for homologous structures. The search is performed using a heuristic pairwise alignment search program like BLAST and FASTA. Generally the database protein which is going to be selected as template should have at least 30% sequence identity with the query sequence. Apart from this criterion, E-value and sequence length coverage should also be considered.

(ii) Target- template sequence alignment: The sequences of both query and template proteins need to be realigned using refined alignment algorithm to obtain optimal alignment, so this is the critical step in the homology modeling since it directly affects the quality of the generated model.

(iii) Model construction: This step is further divided into- backbone model building, loop modeling and side chain refinement.

(iv) Model assessment: Final model generated has to be evaluated to ensure that the structural features of the model are consistent with the physiochemical rules which involve checking anomalies in φ-ψ angles, bond lengths and close contacts.

In order to identify the template structures, target sequence is searched against the PDB, using programs like BLAST and FUGUE. The best template structure will be the one with the highest sequence similarity to the target and will serve as the template. Homology modeling is a powerful technique that enhances the value of experimental structure determination by using the structural information of one protein to predict the structures of homologous proteins (Bhattacharya et al., 2008). The homology modeling using MODELLER is the most popular method for comparative protein structure modeling.

MODELLER takes the sequence alignment between the target and template as input and produces a comparative model. MODELLER implements comparative protein structure modeling by satisfying spatial restraints (Sali and Blundell, 1993). The spatial restraints include homology derived restraints on the distances and dihedral angles in the target sequence, derived from its alignment with the template structures; stereochemical restraints such as bond angle and bond length preferences are obtained from the CHARMM22 molecular mechanics force field (Mac Kerell 1998); statistical preferences for dihedral angles and non-bonded interatomic distances are obtained from a representative set of known protein structures (Shen and Sali, 2006). In theoretical protein modeling, misalignment of amino acids with respect to the true position in the fold can seriously mislead the functional interpretation. To surmount these problems various methods have been developed for protein structure validation. These methods evaluate the stereochemical quality and sequence-structure correlation of protein models.

Phyre2: It is an advanced remote homology detection method to build potential 3-D models of unknown query sequence based on alignment to known protein structures and predict ligand binding sites (Kelley and Sternberg, 2009). The method involves: detection of sequence homologues using PSI-BLAST, prediction of secondary structure and disorder using PSI-PRED and DISOPRED, construction of a HMM of query sequence based on the homologs detected from above methods, construction of 3-D models of query protein based on the alignments between the HMM of query sequence and the HMMs of known structure and modeling of insertions and deletions using a loop library. Quality check of generated models are verified using

'Confidence' value which represents the probability (from 0 to 100) that the match between query sequence and the template is an accurate homology. However this confidence score does not indicate the expected accuracy of the model although the two are intimately related. A match with confidence >90%, can generally be very confident that query protein can adopt the fold and the core of the protein is modeled at high accuracy (2-4Å RMSD) from native structure. Next column of result section consists of percentage identity between query sequence and the template. For extremely high accuracy models percentage identity has to be above 30-40%. Although it is also important to understand that even at low sequence identity models can be meaningful provided that the confidence is high.

I-TASSER: The I-TASSER is a protein structure and function prediction server created on the sequence-to-structure-to-function paradigm (Zhang, 2008). Starting from an amino acid sequence, I-TASSER first generates 3-D models using multiple threading alignments and iterative structural assembly simulations. The output consists of full secondary and tertiary structure predictions as well as functional annotations on ligand binding sites, enzyme classification (EC) numbers and GO terms are also assigned to each model. Generated models are assessed by their C-score, template modeling (TM) score and RMSD. C-score is a confidence score for estimating the quality of predicted models by I-TASSER. C-score is calculated based on the significance of template alignments and the convergence parameters of the structure assembly simulations. Its value is in range of [-5,2], where a C-score of higher value signifies a model with a high confidence and *vice versa*. TM score and RMSD are used to measure structural similarity between two structures which are usually used to measure the accuracy of structure modeling when the template structure is known. TM score is a more accurate term to measure the structural similarity between two structures, because RMSD is sensitive to the local errors. RMSD is measured as average distance between all residue pairs in two structures and a local error (e.g. a misorientation of the tail) can give rise to large RMSD value although the global topology is correct. A TM score >0.5 specifies that model is of correct topology while TM score < 0.17 means a random similarity. These cutoffs are independent of the protein length.

1.4.10 3D-BLAST

It is a very fast and accurate method for evolutionary classifications of a newly determined or query protein structure and identifying its homologous proteins (Yang and Tung, 2006). 3D-BLAST has the advantages of BLAST tool for fast protein structure database scanning. It scrutinizes for the longest common substructures, known as SAHSPs (structural alphabet high-scoring segment pairs), present between the query structure and structure in the structural database. The SAHSP is analogous to the HSP of BLAST. 3D-BLAST orders the search homology structures based on both SAHSP as well as E-value computing from the substitution scoring matrix of structural alphabets. The E-value specifies the statistical significance of an alignment to find the reliability of the searching.

1.4.11 Protein structure validation

PROCHECK: PROCHECK is a suite of programs that offers a detailed analysis on the stereochemistry of a protein structure (Laskowski R A, 1993). The program verifies a variety of geometry based criteria such as Ramachandran plot (Ramachandran et al., 1963), main chain, side chain, bond lengths and angles, planarity of rings and end groups, torsion angles, chirality, close non-bonded interactions, main chain hydrogen bonds, disulfide bond geometry and residue by residue analysis. Accordingly, it generates a number of postscript plots analyzing its overall and residue by residue geometry.

Verify_3D: It examines the validity of a preliminary structure or model derived from experimental data or modeling studies. It measures the compatibility between the protein sequences and known protein structures (Bowie et al., 1991). Verify_3D evaluates the 3-D structure by comparing its structural environments with the preferred environments of the amino acids in the known sequences. Environment is defined by the following criteria (i) the area of the residue that is buried; (ii) the fraction of side chain area that is covered by polar atoms (oxygen and nitrogen); (iii) the local secondary structure. If a residue lies in an unusual chemical environment, it will receive a bad score and *vice versa*. Given a 3-D structure, it identifies which amino acid sequences are compatible with that structure (Luthy et al., 1992).

1.4.12 DALI

This server performs structural alignment of query structure and carries out comparative analyses with known crystal structures from publicly available repositories of protein structures (RCSB, PDBe, and PDBj) (Holm and Rosenstrom,

2010). The output generated consists of two blocks with the list of structural neighbours and their corresponding structural alignment, respectively. The summary block in the results has the following columns for each match: rank of the match, PDB and chain identifier of the matched structure, Z-score of the match (ordered by decreasing Z-scores), RMSD of the match, number of aligned positions, number of residues in matched structure, sequence identity of aligned positions and description of the matched structure.

The alignment block has the following information for each match: rank of the match, ID and chain of the query structure, PDB ID and chain of the matched structure, first and last residue of aligned segment in the query structure (residue numbering according to DALI's internal method).

1.4.13 Gene ontology (GO)

It is a major bioinformatics initiative to unify the representation of gene and gene product characteristics across all species. More specifically, the project aims to: 1) Maintain and develop its controlled vocabulary of gene and gene product attributes; 2) Annotate genes and gene products and 3) Provide tools for easy access to all aspects of the data provided by the project, and to enable functional interpretation of experimental data using the GO, for example *via* enrichment analysis. The GO defines concepts/classes used to describe gene function and relationships between these concepts. It classifies functions along three aspects: molecular function, molecular activities of gene products, cellular component where gene products are active in biological process pathways and larger processes made up of the activities of multiple gene products.

1.4.14 3-D Structural databases

PDB: The tertiary structure of a protein is its 3-D structure, defined by the atomic coordinates determined by the protein primary sequence. All the known 3-D structural data of biological macromolecules are deposited at PDB (Dutta et al., 2009) which provides access to the 3-D coordinates and related information of the biological macromolecules that help in understanding the folding pattern, ligand binding etc. of these molecules. This structural information is exploited in protein classification as well as drug design studies. The fast growing RCSB PDB contains the 3-D description of more than 1,20,878 (July 2016) proteins and nucleic acid

structures. The database is made available to researchers worldwide *via* the website (www.rcsb.org/pdb).

1.4.15 Molecular docking

Protein-DNA docking: Although much improvement has been made in the field of protein-protein docking, in the case of protein-DNA complexes, however, progress lags behind. The scarcity of information for proper identification of interaction surfaces on DNA and its inherent flexibility have hampered the development of effective docking methods. To facilitate the development of effective protein-DNA docking methods a set of well-defined test cases that form a common ground for development, validation and comparison of docking methods is necessary. HADDOCK, a two stage docking method, is able to successfully predict protein−DNA complexes from unbound constituents using non-structural experimental data to drive the docking (van Dijk et al., 2006). HADDOCK makes use of available experimental and bioinformatics data to drive the docking process. Global and local DNA flexibility is introduced in the docking by allowing the DNA sugar-phosphate backbone and DNA base pairs to sample conformations during a semi-flexible refinement stage and by starting the docking from a library of pregenerated DNA structures representing various degrees of conformational flexibility. The latter allows for the sampling of a larger conformational space.

1.4.16 Molecular Dynamics (MD) simulations

The primary structures of proteins provide limited information about the structure and function of biological macromolecules. The 3-D structure of a folded amino acid sequence can give more information regarding the functional proteins. Experimental methods, such as X-ray crystallography and NMR decipher the average conformation of a number of proteins, DNA/RNA and as a complex with inhibitor. But they have limited success in describing the conformational heterogeneity and dynamics, which are important for function. Experimental methods have also been largely unsuccessful at characterizing proteins that occupy a large number of highly diverse conformational states under normal conditions, despite the fact that one- third of our proteins are the so-called “intrinsically disordered” proteins. MD is the method of choice to study the dynamical properties of a system in full atomic details and it can provide the properties that are observable within the time scale accessible to simulations. MD simulations describe the time evolution of a molecular system, e.g.

a protein, by numerically solving Newton's equations of motion for all atoms in the system. Such simulations can accurately describe the dynamics of biologically relevant systems by using three approximations; (i) the Born-Oppenheimer approximation, where nuclear and electronic motions are decoupled, (ii) the approximation that nuclei can be treated as classical particles, and (iii) the use of an empirical force field to describe the interactions between particles. The first two methods are computationally intensive due to their high level calculations. So, the third method came to utilization for the large biological molecules. To calculate the dynamics of the system, that is the position of each atom as a function of time, Newton's classical equation of motion are solved iteratively for each atom. MD determines the position r_i and velocity v_i of each and every atom i = 1,.....N, that is contained in a computational cell and subjected to external boundary conditions (force, pressure, temperature or velocities). A differential equation of motion (Newton's) that solves for these 6N (3 positions and 3 momentum components) variables i is

$$F_i = m_i * \partial v_i/\partial t$$
$$v_i = \partial r_i/\partial t$$

where, F_i is the force on an i^{th} atom having mass m_i. In a numerical scheme, ∂t is approximated by Δt. The problem is solved by iteratively computing the states at successive points in time, using a forward time numerical difference approximation and each iteration is referred as a time step. A huge computational effort is required when the state of the system is large, or when the number of time steps is large. The core of any MD algorithm is the calculation of the potential on each atom in the system and the subsequent calculation of the displacement over a very small time step, typically in the order of femtoseconds (10^{-15}s). MD has no defined point of termination other than the amount of time that can be practically covered.

The accuracy of the MD simulations is directly related to the potential energy function used to describe the interactions between particles. The potential energy function is based on a set of interaction functions and parameters, called the force field. In MD, a classical potential energy function is used and is defined as a function of the coordinates of each atom. The potential energy of the system as a function of atomic coordinates is expressed as the sum of simple analytical functions.

The potential energy function is separated into terms representing as bonded and non-bonded interactions. Usually any two atoms which are connected by one,

two or three bonds are treated through bonded interaction terms. The bonded interactions generally include bonds (1-2 interactions), angles (1-3 interactions) and torsion angles (1-4 interactions). The interactions between more distant atoms and between atoms which are not connected, are described using non-bonded terms. In most cases there are two types of non-bonded interactions, namely Coulombic interactions between charged particles and dispersion interactions generally described by a Lennard-Jones (LJ) potential.

$$V(r) = \sum_{bonds} k_b(b - b_0)^2 + \sum_{angles} k_\theta(\theta - \theta_0)^2 + \sum_{torsions} k_\phi[\cos(n\phi + \delta) + 1] + \sum_{\substack{nonbond \\ pairs}} \left[\frac{q_i q_j}{r_{ij}} + \frac{A_{ij}}{r_{ij}^{12}} - \frac{C_{ij}}{r_{ij}^{6}}\right]$$

Above equation is about the simplest potential energy function that can reproduce the basic features of protein energy landscapes at an atomic level in detail, and it has proved to give insights into a remarkably broad range of properties. The combinations of a potential energy function and all the parameters that go into it (k_b, b_o, k_θ, θ_o, *etc*.) constitutes a "force field". Various existing force fields differ in their functional form and the way their parameters were derived (Hünenberger 1997, Ponder 2003). Parameters for the bond lengths and angles are usually derived from quantum chemical calculations or crystal structures. The torsional parameters can be adjusted to fit torsional profiles obtained from quantum-chemical calculations or from experiment. This is done in conjunction with fitting of the non-bonded interaction parameters as the latter have a strong influence on the torsional barriers. In the derivation of non-bonded parameters, many differences exist between the force fields. Force fields such as AMBER (Case et al., 2005) and CHARMM (Brooks et al., 1983; Foloppe and MacKerell, 2000) fit the charges to reproduce the electrostatic potential obtained from quantum-chemical calculations, OPLS (Jorgensen et al., 1996; Jorgensen and Tirado-Rives, 2005) and GROMOS (Oostenbrink et al., 2004; Scott et al., 1999) fit non-bonded parameters (charges and LJ parameters) such that they reproduce the thermodynamic properties like density and heat of vaporization of simple liquids.

Generally, GROMOS, OPLS-AA/L and AMBER force fields are useful for proteins, and the inhibitor molecule force fields are generated by using Dundee PRODRG server or ACPYPE program that uses the GAFF force fields. In the latest

version of the GROMOS force field (Oostenbrink et al., 2004) parameters have been optimized to reproduce the free enthalpy of hydration and apolar solvation. The general AMBER force field (GAFF) 199 for organic molecules is designed to be compatible with existing AMBER force fields. Mobley et al. (Mobley et al., 2009, 2015) have reported hydration free energies for 504 small molecules parameterized using the AMBER Antechamber program (Wang et al., 2006a) to assign GAFF (Wang et al., 2004) parameters.

GROMACS (Groningen MAchine for Chemical Simulation) (Berendsen et al., 1994; Lindahl et al., 2001) software suite (http://www.gromacs.org/) is a versatile package which is primarily designed to perform MD simulations of biochemical molecules such as proteins and lipids. The software, written in ANSI C, originates from a parallel hardware project, and is well suited for parallelization on processor clusters. Since GROMACS is extremely fast at calculating the non-bonded interactions that typically dominate simulations, many groups are also using it for research on non-biological systems such as polymers. GROMACS was initially a rewrite of the GROMOS (Scott et al., 1999) package (http://www.gromos.net/), which itself, like AMBER, was originally derived from an early version of CHARMM.

The main advantages of GROMACS are its ease of use and its exceptional performance on standard personal computers. The authors report that it is normally 3 to 10 times faster than other MD programs. GROMACS is actually a suite of small command line programs each with a simple set of options. In GROMACS all files are plain text based, so they are, in principle, human readable. These plain text formats result in much larger file sizes than binary formats would, so GROMACS transparently utilizes standard UNIX compression tools. Most of the standard types of data analysis can be performed using the set of accompanying tools which can also produce publication-ready plots in a straight forward manner. From a practical point of view, one particularly attractive reason to choose GROMACS is the fact that it is distributed as free software under the terms of the GNU General Public license (http://www.gnu.org/).

RMSD: This parameter is used to measure the similarity in 3-D structure of the Cα atomic coordinates after superposition. RMSD is the measure of the average distance between two equivalent atoms (usually Cα and sometimes C, N, O and Cβ) of superimposed proteins. The program g_rms is used to study the deviation in the RMSD from the reference structure as a function of time during the MD simulations

in GROMACS. It is more representative for structural integrity and can reveal the time at which conformation changes occur in the protein structure.

RMSF: The root mean square fluctuation (RMSF) is the measure of average atomic mobility of backbone atoms during the MD simulations. RMSF is useful for characterizing highly fluctuating regions in the protein molecule. Typically the tails (N- and C-terminal) fluctuate more than any other part of the protein. Secondary structure elements like α-helices and β-strands are usually more rigid than the unstructured part of the protein, and thus fluctuate less than the loop regions.

Energy: Usually, the first step in analysing a simulation is to attempt to determine if the system has reached equilibrium. This is often verified by looking at quantities such as the total energy, temperature and pressure of the system. g_energy is used to examine the total energy of system as a function of time during the MD simulation in GROMACS.

Protein interactions with the ligand or DNA can also be observed throughout the simulation. These interactions are categorized into four types: hydrogen bonds, hydrophobic, ionic and water Bridges.

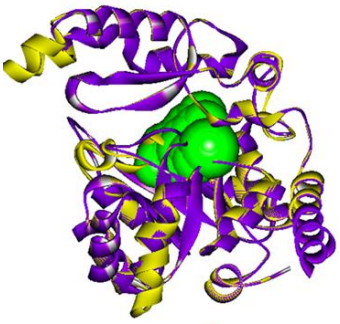
hydrogen bonds plays an important role in ligand binding and is important because of their strong influence on drug specificity, metabolization and adsorption. Geometric criteria for hydrogen bonds considered are: distance of 2.5Å between the donor and acceptor atoms (D—H···A); donor angle of ≥120° between the donor-hydrogen-acceptor atoms (D—H···A); and an acceptor angle of ≥90° between the hydrogen-acceptor-bonded atom atoms (H···A—X).

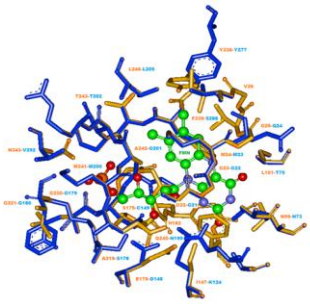
Hydrophobic contacts consist of three subtypes: π…cation, π…π and other non-specific interactions. Mostly these kinds of interactions involve a hydrophobic amino acid and an aromatic or aliphatic group on the ligand. The current geometric criteria for hydrophobic interactions is as follows: aromatic and charged groups within 4.5Å for π…cation, two aromatic groups stacked face-to-face or face-to-edge for π…π interactions while non-specific hydrophobic includes side chain within 3.6Å of a ligand's aromatic or aliphatic carbons. Ionic interactions or polar interactions are between two oppositely charged atoms that are within 3.7Å of each other and do not involve a hydrogen bonds.

2

Structural Annotation of *Helicobacter pylori 26695* proteome

- ✓ Structure-function annotation of *H. pylori* 26695 proteome.
- ✓ Function annotation and binding site analysis of hypothetical proteins.
- ✓ Identification of prominent structural folds and class in whole proteome.





Structure based annotation of *Helicobacter pylori* strain 26695 proteome. Singh S, Guttula PK, Guruprasad L. PLoS One. 2014 Dec 30;9(12):e115020.

2.1 Introduction

H. pylori possess significant genotypic diversity which engenders various strategies to interact with host cells, manipulate their behaviour in order to survive and propagate. In most of the cases, *H. pylori* infections are lingering but asymptomatic, majority of the patients never experience symptoms or some may have only mild gastric inflammation, acute infection may cause clinically relevant chronic active gastritis, peptic ulceration (McColl, 1997), chronic atrophic gastritis that is a sign of gastric adenocarcinoma and MALT lymphomas (Kikuchi et al., 1995).

Disease outcome is greatly influenced by bacterial genotype, host- physiology, genotype, dietary habits (Hunt, 1996; Labigne and de Reuse, 1996), host genetic diversity, particularly within immune response genes (Peek, 2005). Since the undissociated form of weak acids can freely pass through the cell membrane of any microbe, weak acids possess potent antimicrobial activity but *H. pylori* has the ability to tolerate acidic conditions in the gastric environment by creating a positive inner-membrane potential at low pH (Cover and Blaser, 1996). In the absence of treatment, infection can persist lifelong and is frequently transmitted from person to person, probably by oral to oral and/or fecal to oral. Most of the disease treatment methods for *H. pylori* infections are centered on the use of a proton-pump inhibitors and antibiotics such as metronidazole and clarithromycin (Meurer and Bower, 2002).

Genome sequence analysis revealed that the circular genome of *H. pylori* strain 26695 consists of 1,667,867 base pairs with 1,590 predicted coding sequences and has well-built systems for motility, scavenging iron, DNA restriction and modification (Tomb et al., 1997). Several putative adhesins, lipoproteins and other outer membrane proteins have been identified in the complete proteome as possible partners for host pathogen interactions in *H. pylori* 26695. The annotation of this genome/proteome would help in identifying better drug targets that can be exploited to tackle *H. pylori* infections and it is one of the ultimate goals of this sequencing project. Often, functional annotation of a proteome is achieved from comparative sequence analysis. These methods have limitations since they mainly rely on the comparisons based on sequence homology. As a result, all proteomes sequenced so far have ~40-60% unannotated proteins (Boneca et al., 2003).

Focus on 3-D structural organizations of proteins would help in deciphering the protein fold, structure and active site, which have applications in structure based drug design methods that are rational. The availability of complete genome sequences has provided a platform to decipher the structural and functional information of any complete proteome using the computational methods. The results are reliable and provide a solution to the time consuming and expensive experimental methods. The information about function of a protein resides in its structure; the high resolution 3-D structures of proteins are determined using X-ray crystallography and NMR techniques. In the absence of experimental structures, sequence homology methods are employed based on the understanding of proteins which share sequence similarity and also have homologous structure and function, barring a few examples (Pearson, 2013; Pieper et al., 2011). This formalism has a limitation; the numbers of protein sequences available from complete sequencing projects far outweigh the number of available 3-D structures and the functionally characterized proteins experimentally. As a result, alternative procedures such as fold recognition for proteins that share low sequence homology are compared to similar 3-D structures, and *ab initio* modeling methods can also be employed. From the validated 3-D structures, types of folds and the active site can be characterized. The 3-D structures of 145 proteins in *H. pylori* are determined experimentally so far and deposited in protein structure databank (PDB) (Dutta et al., 2008), therefore a wealth of structural information remains to be explored. In this work, we have employed forementioned computational methods to get structural as well as functional insights into the *H. pylori* proteome.

2.2 Method

2.2.1 Directions for structure annotation

In order to understand the biological role of large numbers of linear amino acid sequence data generated through genome sequencing projects, we need to have knowledge of their structure. Even though structures determined by experimental methods provide high resolution data, due to various limitations, structures cannot be determined experimentally for a large proportion of these sequences. Computational structure prediction techniques provide significant and reliable information, and are cost effective as well as less time consuming. Our approach started with finding structural models of the individual Pylorigene database (http://genolist.pasteur.fr/PyloriGene) proteins using different sources in various sequential steps, followed by structure validation. The theoretical models are further subjected to analysis as a way to gain insight into their function. Functional annotation has been assigned through fold to function association as well as by the identification of ligand binding sites and cavities associated with that model. Fold prediction methods attempt to detect structural folds that are compatible with a particular query sequence based on similarities between query sequence and known 3-D structure of protein. Since protein surface dictates the type of interaction it can make with its associated ligand or other interacting partners, we have further analyzed the protein structures through their binding sites. The overall objective is to predict as accurately as possible the probable function of the protein, at sequence and structure level. At amino acid sequence level we have annotated the protein by GO to decipher the function. At the structure level we have assigned structural classification, fold, ligand location (binding site) and ligand type (associated ligand, cofactor, etc.) based on the template structure. The flow chart shown in Figure 2.1 depicts various steps adopted for the annotation of *H. pylori* 26695 proteome.

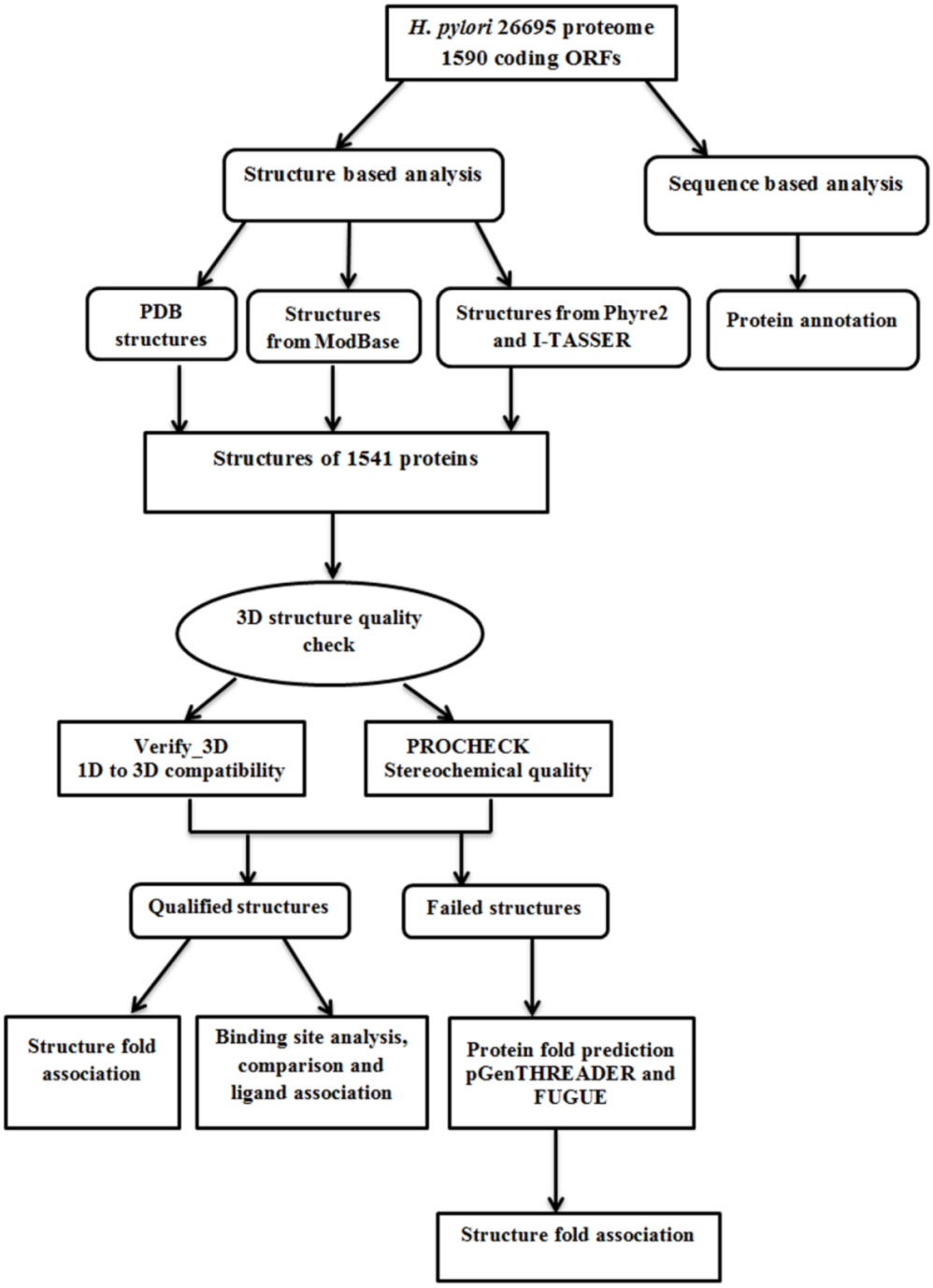


Figure 2.1: Flow chart for annotating individual proteins in *H. pylori* 26695 strain.

2.2.2 Protein models

Functional annotation at the sequence level is performed by using AmiGO gene ontology tool (Carbon et al., 2009). AmiGO is a web based application that allows ontologies and related gene product annotation (association) data. Out of 1590 predicted protein coding genes in *H. pylori*, experimentally determined structures are available for 145 proteins in the PDB, proteins with less than 30 amino acid residues

were excluded from the study and for rest of the proteins structural models were built using various methods described below.

ModBase (Pieper et al., 2011) is a database containing comparative protein structure models of various organisms and relies mainly on MODELLER (Sali and Blundell, 1993) for fold assignment, sequence–structure alignment, model building and model assessment. MODELLER is used for homology or comparative modeling of protein 3-D structures by satisfaction of spatial restraints. The main criteria used to judge the quality of the protein model from ModBase was sequence identity and query length coverage (ranking score (Z) = product of % sequence identity and % length of query sequence in the alignment), DOPE score (Discrete Optimized Protein Energy <0 is reliable) and MPQS (ModPipe Quality Score of >1.1 is considered to be reliable) (Eramian et al., 2006; Shen and Sali, 2006). Proteins for which good models were not available in ModBase were submitted to Phyre2 (Kelley and Sternberg, 2009) which is based on the principles of homology modeling, and followed by I-TASSER (Zhang, 2008) which is a combination of *ab initio* folding and threading methods. Information about all the structures with their respective source are provided in Table S2.1. For some protein sequences, the models built were derived from templates without significant sequence similarities, but these models were accepted because of high compatibility with the structural folds.

2.2.3 Quality estimation of protein structures

Quality of all the protein models obtained through ModBase, Phyre2 and I-TASSER were analyzed using Verify_3D (Bowie et al., 1991) and PROCHECK (Laskowski R A, 1993). Verify_3D checks the compatibility of the model with its own amino acid sequence and PROCHECK validates stereochemical parameters of the protein models by analyzing residue by residue geometry and overall structural geometry. Table S2.1 also provides information about the quality estimation values of all the models which we have obtained from different sources. All the qualified structures can be downloaded from https://sites.google.com/site/lgpscuh/links.

Further the proteins where good quality structures could not be built either by homology modeling or *ab initio* approaches, were further analyzed by PSIPRED which is an accurate secondary structure prediction method that incorporates two feed-forward neural networks and performs an analysis on the output obtained from PSI-BLAST. Here we have used pGenTHREADER method (Lobley et al., 2009) for fold

recognition and identification of distant homologues which makes use of profile-profile alignments and predicted secondary structure (using PSIPRED) (McGuffin et al., 2000) as inputs. The structures whose confidence were certain or high were selected for annotation and are listed in Table S2.2, while rest of the proteins were further subjected to FUGUE (Shi et al., 2001). FUGUE is a method for recognizing distant homologues by sequence-structure comparison. It utilizes environment specific substitution tables and structure dependent gap penalties, where scores for amino acid matching and insertions/deletions are evaluated depending on the local environment of each amino acid residue in known structure. Given a query sequence (or a sequence alignment), FUGUE searches a database of structural profiles and calculates the sequence-structure compatibility scores and provides a list of potential homologues and their sequence alignment. All the sequences with high confidence in prediction were selected for annotation and are listed in Table S2.3. The structures which have high confidence were selected for functional annotation of proteins and were submitted to DALI (Holm and Rosenstrom, 2010) server which performs structural alignment and carries out comparative analyses of newly discovered protein structures with known PDB structures. The output generated gives the list of structural neighbours and their corresponding structural alignment. From the results, the hit with highest Z-score, percentage identity and lowest RMSD were selected for annotating the protein structure. All these validated structures are shown in Table S2.4. While the annotations and a broad functional category were available in the databases for many proteins based on literature and sequence analyses, several new associations have been possible through modeling and structural analyses in the current work. Of the 557 conserved proteins annotated as "hypothetical" in the Pylorigene database, 464 proteins are now associated with fold-based function annotation through the pipeline described above.

2.2.4 Assigning associated fold to each structure

The PDB structures and the validated protein model structures of good quality obtained through ModBase, Phyre2 and I-TASSER along with the associated target structures identified by PSIPRED and FUGUE were submitted to 3D-BLAST (Yang and Tung, 2006) for SCOP classification (Murzin et al., 1995) which searches for the longest common substructure called SAHSPs. This is a fast and accurate method for discovering homologous proteins and evolutionary classifications of newly determined structures. It gives a list of homologous protein structures that are similar to the query,

ordered by E-values. SCOP database has manual classification of protein structural domains based on similarities of their structures and amino acid sequences. The classification of protein structures in the database is constructed on evolutionary relationships and on the principles that govern their 3-D structure. These SCOP-IDs were further submitted to SCOP database in order to retrieve the class, fold, superfamily and family associated with each structure as shown in Table S2.5.

2.2.5 Binding site identification and comparison

Ligand binding is a key aspect of protein function, mediating the ability of proteins to recognize their natural ligands for transport, signal transduction, catalysis etc. This information also aids in the modulation of their function through the discovery of inhibitors. In COFACTOR method (Roy et al., 2012), potential ligand binding sites in various models were identified through a consensus ranking based on C-score, RMSD, identity, TM score and coverage. It analyses conserved surface residues and predicts the functional site located to be around a point without giving any boundary definition to pocket. All the pockets predicted by COFACTOR within 2Å zone of RMSD, and sequence identity greater than 30% were selected and the associated binding sites were compared to known sites for each structure. CASTp (Dundas et al., 2006) (Computed Atlas of Surface Topography of proteins) server was also used to predict the cavities if the structure fails to have a ligand binding site within 2Å. CASTp gives information regarding the cavity or pocket, domain name and also about the residues present in the cavities.

2.3 Results

Pylorigene database, an annotated *H. pylori* proteome consists of large number of weakly annotated proteins termed as 'predicted' since they are largely derived through sequence comparisons. The database also has list of proteins which have not been studied in *H. pylori* but homologs have been characterized in other organisms or gene has not been studied in any other organisms and hence are annotated as 'predicted coding region' and 'predicted coding regions with no homologs in database' respectively. Most of these proteins are uncharacterized in Pylorigene database, but through our annotation pipeline we were able to annotate most of them. Previously uncharacterized proteins (557) were listed in the database, and through this methodology adopted by us, we could annotate 464 proteins as shown in Figure 2.2. We further compared previous annotation from Pylorigene database and new annotation from our work (Table S2.6). Both annotations are agreeable in most of the cases and new annotation has been added to several proteins in *H. pylori* database. Resende et. al., 2013 (Resende et al., 2013) have also annotated the proteins of *H. pylori* 26695 using merlin software tools and online databases. EC numbers and TC numbers to metabolic gene encoding enzymes and transport proteins, respectively have been assigned. In spite of differences in the methodologies, we found that our functional annotation of nearly 95% proteome is agreeable with that of Resende et al., 2013.

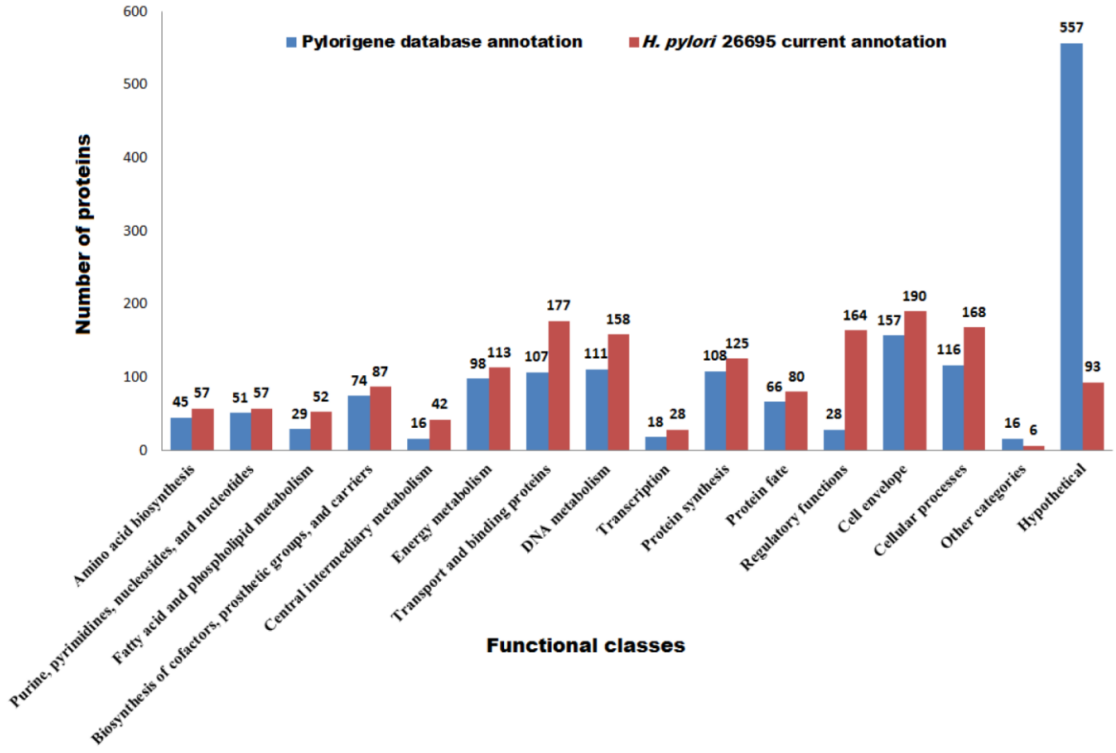


Figure 2.2: Comparison of *H. pylori* 26695 proteome functional classes in Pylorigene database and current annotation studies.

Table 2.1: Each functional class is composed of several subclasses based on the putative role category assigned to each protein of the proteome**.**

Classes	**Subclasses**
1-Amino acid biosynthesis	1.1-Aromatic amino acid family
	1.2-Aspartate family
	1.3-Glutamate family
	1.4-Pyruvate family
	1.5-Serine family
	1.6-Other
2-Purine, pyrimidines, nucleosides, and nucleotides	2.1-2'-deoxyribonucleotide metabolism
	2.2-Purine ribonucleotide biosynthesis
	2.3-Pyrimidime ribonucleotide biosynthesis
	2.4-Salvage of nucleosides and nucleotides
	2.5-Sugar-nucleotide biosynthesis and conversions
3-Fatty acid and phospholipid metabolism	3.1-General
4-Biosynthesis of cofactors, prosthetic groups, and carriers	4.1-Biotin
	4.2-Folic acid

	4.3-Heme, porphyrin and cobalamine
	4.4-Menaquinone and ubiquinone
	4.5-Molybdopterin
	4.6-Pantothenate and coenzyme A
	4.7-Pyridoxine
	4.8-Riboflavin, FMN and FAD
	4.9-Thioredoxin, glutaredoxin, and glutathione
	4.10-Thiamine
	4.11-Pyrimidine nucleotides
	4.12-Other
5-Central intermediary metabolism	5.1-Amino sugars
	5.2-Phosphorus compounds
	5.3-Polyamine biosynthesis
	5.4-Other
6-Energy metabolism	6.1-Aerobic
	6.2-Amino acids and amines
	6.3-Anaerobic
	6.4-ATP-proton motive force interconversion
	6.5-Electron transport
	6.6-Entner-Doudoroff
	6.7-Fermentation
	6.8-Glycolysis/gluconeogenesis
	6.9-Penthose phosphate pathway
	6.10-Sugars
	6.11-TCA cycle
	6.12-Other
7-Transport and binding proteins	7.1-Amino acids, peptides and amines
	7.2-Anions
	7.3-Carbohydrates, organic alcohols, and acids
	7.4-Cations
	7.5-Nucleosides, purines, and pyrimidines
	7.6-Other
	7.7-Unknown substrate
8-DNA metabolism	8.1-DNA replication, recombination, and repair
	8.2-Restriction/modification
	8.3-Degradation of DNA

	8.4-Chromosome-associated proteins
9-Transcription	9.1-Degradation of RNA
	9.2-DNA-dependent RNA polymerase
	9.3-Transcription factors
	9.4-RNA processing
10-Protein synthesis	10.1-tRNA aminoacylation
	10.2-Nucleoproteins
	10.3-Ribosomal proteins: synthesis and modification
	10.4-tRNA and rRNA base modification
	10.5-Translation factors
	10.6-Other
11-Protein fate	11.1-Protein and peptide secretion, and trafficking
	11.2-Protein modification and repair
	11.3-Protein folding and stabilization
	11.4-Degradation of protein, peptides, and glycopeptides
12-Regulatory functions	12.1-General
13-Cell envelope	13.1-Lipoproteins
	13.2-Surface structures
	13.3-Biosynthesis of murein sacculus and peptidoglycan
	13.4-Biosynthesis of surface polysaccharides
	and lipopolysaccharides
	13.5-Other
14-Cellular processes	14.1-Cell division
	14.2-Chemotaxis and motility
	14.3-Detoxification
	14.4-Transformation
	14.5-Toxin production and resistance
	14.6-Pathogenesis
	14.7-Adaptation and atypical conditions
	14.8-Other
15-Other categories	15.1-Plasmid-related fucntions
	15.2-Transposon-related functions
16-Unknown	16.1-General
17-Hypothetical	17.1- *H.pylori* specific with no known function

2.3.1 Proteome analysis based on the functional classes

Sequence as well as structure analysis of 1,590 predicted coding sequences from *H. pylori* 26695 circular genome revealed that the organism has well-built systems for motility, iron scavenging, DNA restriction and modification which are required for successfully inhabiting inside the host. Apart from this, some putative lipoproteins, adhesins and other outer membrane proteins were also described, accentuating the possible intricacy of host–pathogen interaction. *H. pylori*, like other mucosal pathogens, possibly uses recombination and slipped-strand mispairing within repeats as tools for adaptive evolution and antigenic variation. Consistent with its restricted niche, *H. pylori* has a few regulatory networks, restricted metabolic repertoire and biosynthetic capability. As discussed in Chapter 1, survival of *H. pylori* 26695 in acid environment also depends on its ability to establish a positive inner-membrane potential at low pH.

2.3.1.1 Acidity, pH and acid tolerance

The survival of *H. pylori* in acidic environments is probably due to its ability to establish a positive inside membrane potential (Matin et al., 1996) and subsequently to modify its microenvironment through the action of urease and the release of factors that inhibit acid production by parietal cells (Labigne and de Reuse, 1996). A switch in membrane polarity provides an electrical barrier that prevents the entry of protons (H^+). A positive cell interior can be created by the active extrusion of anions or by proton diffusion potential. Proton diffusion potential would need the anion permeability of the cytoplasmic membrane to be low and, so far three anion transporters have been identified. However, it remains to be determined whether anion conductance is associated with other proteins: the MDR-like transporters (HP600, HP1082 and HP1206) or hypothetical proteins. Although it has been suggested that proton translocating P-type ATPases could mediate survival in acid conditions by the extrusion of protons from the cytoplasm (Melchers et al., 1996), this idea is not supported by the identified transporter genes. The P-type ATPase sequences in *H. pylori* (HP791 and HP1503) are more closely related to divalent cation transporters than to ATPases with specificity for protons or monovalent cations. HP0791, is involved in Ni^{2+} supply, an essential component of urease activity (Bayle et al., 1998) in the *H. pylori* 26695 strain.

2.3.1.2 Adhesion and adaptive antigenic variation

Most of the pathogens display tropism to specific tissues or cell types and frequently employ several adherence mechanisms for successful attachment. *H. pylori* also uses at least five different adhesins to adhere the gastric epithelial cells of the host (Labigne and de Reuse, 1996). One of them, HpaA (HP0797), was previously identified as a lipoprotein in the flagellar sheath and outer membrane (Labigne and de Reuse, 1996). HP0256 protein is now characterized experimentally and is confirmed to be involved in motility and cell envelope architecture of *H. pylori* (Douillard et al., 2010). In addition to the HpaA orthologue, we have identified several other lipoproteins in the proteome. Few have an identifiable function, but some are likely to contribute to the adherence capacity of the organism. Two adhesins, one of which mediates attachment to the Lewis^{b} blood antigens, belong to the large family of outer membrane proteins (Boren et al., 1993; Odenbreit et al., 1996). It is possible that other members of these closely related proteins also act as adhesins. Given the large number of sequence related genes encoding putative surface exposed proteins, the potential exists for recombinational events leading to mosaic organization. This could be the basis for antigenic variation in *H. pylori* and an effective mechanism for host defence evasion, as seen in *M. genitalium* (Peterson et al., 1995). Genotypic variation mechanisms may have evolved in bacterial pathogens to increase the frequency of phenotypic variation in genes involved in critical interactions with their hosts (Moxon et al., 1998). These 'contingency' genes encode surface structures like lipoproteins, pilins or enzymes which produce lipopolysaccharides (Moxon et al., 1998). Phenotypic variation at the transcriptional level may also operate in *H. pylori*, for example repetitive DNA mediating transcriptional control have been recognized by the presence of oligonucleotide repeats in the promoter regions (Jonsson et al., 1991). The protein categorized under functional class, cell envelope, consists of adhesion and adaptive antigenic variation related protein. This class is further subdivided into membranes, lipoproteins and porins, murein sacculus and peptidoglycan surface polysaccharides, lipopolysaccharides and antigens and surface structures subclasses based on the exact predicted function of proteins.

2.3.1.3 Virulence

The virulence of various *H. pylori* isolates has been determined by their capability to produce CagA and active VacA (Labigne and de Reuse, 1996). VacA induces the formation of acidic vacuoles in host epithelial cells, and its presence is associated epidemiologically with tissue damage and disease (Atherton et al., 1995). These proteins are supposed to be retained on the outside surface of the cell membrane and contribute to the interaction between *H. pylori* and host cells. The surface exposed lipopolysaccharide (LPS) molecule plays an important role in *H. pylori* pathogenesis (Moran, 1996). The LPS of *H. pylori* is several orders of magnitude less immunogenic than that of enteric bacteria (Baker et al., 1994) and the O antigen of many *H. pylori* isolates is known to mimic the human Lewis^{x} and Lewis^{y} blood group antigen (Moran, 1996). Genes for synthesis of the lipid A molecule, the core region, and the O antigen were identified. Two genes with low similarity to fucosyltransferases (HP379, HP651) were found and may play a role in the LPS-Lewis antigen molecular mimicry. Our analysis also suggests that three genes, two glycosyltransferases (HP208 and HP619) and one fucosyltransferase (HP379), may be subject to phase variation. As with other pathogens, *H. pylori* probably requires an iron scavenging system for survival in the host (Labigne and de Reuse, 1996). Genome analysis suggests that *H. pylori* has several systems for iron uptake. One is analogous to the siderophore mediated iron uptake *fec* system of *E. coli* (Cooksley et al., 2003), except that it lacks the two regulatory proteins (FecR and FecI) and is not organized in a single operon. Other systems for iron uptake present in *H. pylori* consist of the three *frpB* genes which encode proteins similar to either haem or lactoferrin binding proteins. The global ferric uptake regulator (Fur) characterized in other bacteria is also present in *H. pylori*. *H. pylori* motility is essential for its colonization inside the host (Suerbaum, 1995). It enables the bacterium to spread into the viscous mucous layer covering the gastric epithelium. Several proteins in the *H. pylori* genome appear to be involved in the regulation, secretion and assembly of the flagellar architecture.

2.3.1.4 Cell division and protein secretion

The proteome analysis suggests that the basic mechanisms of replication, cell division and secretion along with classical set of bacterial chaperones like DnaJ, DnaK, CbpA, HtpG, GrpE, GroES and GroEL are present in *H. pylori 26695* strain. *H. pylori has* two export systems along with SecA-dependent secretory pathway. One is associated with the Cag pathogenicity island (Censini et al., 1996) while other is the

flagellar export pathway (Macnab, 2004). However type IV signal peptidase and orthologues of the dsbABC system are absent from *H. pylori*, which in other species are required for the maturation of pili and pilin-like structures (Strom et al., 1994).

2.3.1.5 Recombination, repair and restriction systems

In *H. pylori* well organized system for mismatch, excision, homologous recombination and transcription coupled repair are present. Conversely RecBCD pathway, which mediates homologous recombination and double-strand break repair, RecT and RecE orthologues and proteins involved in strand exchange during recombination (Smith, 2012), seems to be absent in *H. pylori*. The ability of *H. pylori* to achieve mismatch repair is confirmed by the presence of methyl transferases, mutS (Mutator S) and uvrD proteins.

Bacteria commonly use R-M systems as their defense mechanism to digest the foreign DNA. In *H. pylori*, this defense system seems to be well developed with eleven R-M systems identified on the basis of gene order and similarity to endonucleases, MTase and specificity subunits. Three type I, one type II, three type IIS, and four type III systems were identified in *H. pylori* (Blanchard and Czinn, 1998; Curnow et al., 1996). In addition to the complete systems, seven adenine-specific, four cytosine-specific MTase, and one MTase of unknown specificity were found. Each of these has an adjacent gene with no database match, suggesting that they may function as a part of R-M systems. DNA metabolism functional classes comprises of these proteins and are divided into degradation of DNA, DNA replication, restriction, modification, recombination and repair, chromosome associated proteins and restriction/modification subclasses.

2.3.1.6 Transcription and translation

H. pylori provides the first example of a bacterial genome apparently lacking an asparaginyl-tRNA synthetase gene (Boneca et al., 2003). Most interesting, is the finding that in *H. pylori* the genes encoding the β and β’ subunits of RNA polymerase are merged whereas in all studied prokaryotes these two genes are adjoining, but separate, and are part of the same transcriptional unit. But whether this gene fusion in *H. pylori* results in a fused protein, or whether the transcriptional or translational product of the fusion is subject to splicing, is still unknown.

2.3.1.7 Metabolism

Metabolic pathway analysis of the *H. pylori* genome suggests that it uses glucose as the only source of carbohydrate and the main source for substrate level phosphorylation. It also derives energy from the degradation of serine, alanine, aspartate and proline. The biosynthesis of peptidoglycan, phospholipids, aromatic amino acids, fatty acids and cofactors is derived from acetyl-CoA or from intermediates in the glycolytic pathway. The metabolism of pyruvate reflects the microaerophilic character of this organism. The conversion of pyruvate to acetyl CoA is performed by the pyruvate ferrodoxin oxidoreductase (POR), a four-subunit enzyme thus far only described in hyperthermophilic organisms (Hughes et al., 1995). The tricarboxylic acid cycle (TCA) is incomplete and the glyoxylate shunt is absent. The analysis of degradative pathways, uptake systems and biosynthetic pathways for pyrimidine and purine suggests that *H. pylori* uses several substrates as nitrogen source, including urea, ammonia, alanine, serine and glutamine.

In *H. pylori*, proton translocation is mediated by the NDH-1 dehydrogenase and the different cytochromes, including the primitive-type cytochrome. Four respiratory electron generating deydrogenases have been identified, glycerol-3-phosphate dehydrogenase, D-lactate dehydrogenase, NADH–ubiquinone oxidoreductase complex, and a hydrogenase complex. Studies also suggest that *H. pylori* is not able to use nitrate, nitrite, dimethylsulphoxide, trimethylamine *N*-oxide or thiosulphate as electron acceptors (Hughes et al., 1995; Mendz and Hazell, 1995).

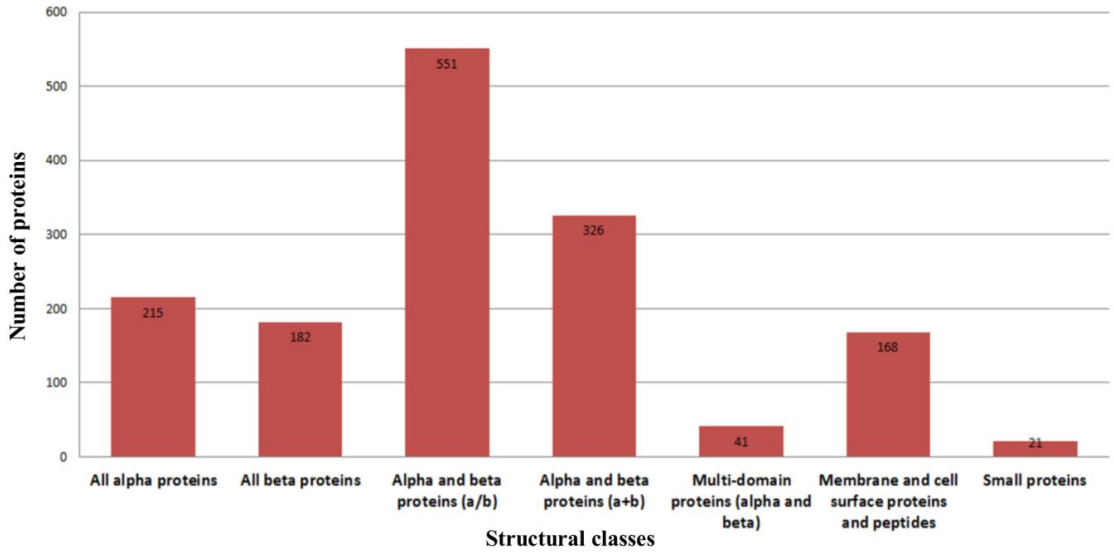
2.3.1.8 Regulation of gene expression

Bacteria control the transcription of their genes in response to several environmental stimuli, like nutrient availability, cell density, pH, connection with the target tissue, DNA damaging agents, temperature and osmolarity. In the case of pathogens, the regulated expression of certain key genes is essential for successful evasion of host responses and colonization, adaptation to different body sites, and survival as the pathogen passes to new hosts. In *H. pylori*, global regulatory proteins are less abundant than in *E. coli*. For example, orthologues of many DNA binding proteins that regulate the expression of certain operons such as OxyR (oxidative stress), Crp (carbon utilization), RpoH (heat shock) and Fnr (fumarate and nitrate regulation) are absent. Two-component regulator systems, consisting of a membrane histidine kinase sensor protein and a cytoplasmic DNA binding response regulator, provide a well-studied mechanism for signal transduction. Four sensor proteins and

seven response regulators were found in *H. pylori*, similar to the number found in *H. influenza* (Fleischmann et al., 1995).

The availability of the structural models covering *H. pylori* complete proteome can be useful in analyzing the fold content and fold preferences of this organism which will be further helpful in identifying important folds that are sufficient for sustaining life and causing infection to the host. Also, the knowledge of structural fold of a protein important in causing disease could be a good target for structure based drug design studies. Assignment of structural folds in the whole proteome is performed through the SCOP database. The fold analysis carried out in this proteome evidently shows the existence of all seven major structural classes (as per SCOP classification) as indicated in Figures 2.3A and 2.3B. Out of 1195 known structural folds present in the SCOP database, 411 (34% of all known folds) are observed in the whole *H. pylori* proteome, with greater inclination for domains belonging to α/β class (36.63%). The top folds (Figure 2.3A) in the modeled *H. pylori* proteome included P-loop containing nucleoside triphosphate hydrolases, TIM barrel, transmembrane helix hairpin, alpha-alpha superhelix, S-adenosyl-L-methionine-dependent MTases and ferredoxin-like, altogether covering almost 75% of all the folds present in this proteome. Predominant folds occurring in various organisms have been studied in their respective proteomes giving rise to a powerlaw distribution of folds (Qian et al., 2001), which is found to be consistent with the pattern seen here in the *H. pylori* proteome.

A



B

Structural Fold	Percentage
P-loop containing nucleoside triphosphate hydrolases	22.60%
TIM barrel	16.70%
Transmembrane helix hairpin	16.05%
Alpha-alpha superhelix	11.10%
S-adenosyl-L-methionine-dependent methyltransferases	10.70%
Ferredoxin-like	3.10%
NAD(P)-binding Rossmann-fold domains	2.95%
Ribonuclease H-like motif	2.84%
BAR/IMD domain-like	2.56%
Adenine nucleotide alpha hydrolase	1.74%

Figure 2.3: (A) Distribution of structural classes in *H. pylori* 26695 proteome according to SCOP classification. (B) Distribution of major structural folds across *H. pylori* 26695 proteome.

2.3.2 Structure based assessment and functional annotation of *H. pylori* 26695 proteome

A significant feature of this annotation pipeline involved obtaining structures from databases of experimental and *in silico* results. In the absence of such data, 3-D structures were either built or at least the fold was predicted. The assessment of each model was made through different estimates of confidence, for example, statistical significance of alignments, extent of sequence similarity, geometry and stereochemistry when compared to high resolution crystal structures, and primary sequence to 3-D structure correlation. We therefore believe that the results from these methods are highly reliable and reproducible.

First example we discuss in this category is HP0773 (UniProt ID: O25465, 363 amino acids) that is labeled as predicted coding region HP0773 in Pylorigene and UniProt databases. Our analyses described the protein as nitroalkane dioxygenase that is also known as nitroalkane oxidase (NAO). NAO is a structural member of the flavoenzyme acyl-CoA dehydrogenase (ACAD) superfamily. These enzymes are mainly involved in catalyzing oxidative denitrification of neutral nitroalkanes to their equivalent carbonyl compounds like aldehydes or ketones, hydrogen peroxide and nitrite using FAD or FMN as a cofactor. These nitroalkanes are further used as intermediates for synthesis in chemical industries (Francis and Gadda, 2006; Ha et al., 2006). Several antibiotics, such as chloramphenicol and azomycin, contain nitro groups, and are also produced by many leguminous plants in the form of nitro toxins such as 3-nitro-1-propionic acid and 3-nitro-1-propanol (Gorlatova et al., 1998). HP0773 was modeled on the template PDB_ID: 2GJL which is the crystal structure of 2-nitropropane dioxygenase by the Phyre2 methodology. Both model and template structures superimposed well with low RMSD (0.08Å) (Figures 2.4A and B). The stereochemical geometry (0.4% residues are present in disallowed regions of Ramachandran plot) and sequence to structure correlation of the model using Verify_3D (88.13% of the residues had an averaged 3D-1D score > 0.2) were confirmed and then subjected to further structural analysis.

NAO comprises of two domains, a triose phosphate isomerase (TIM) barrel domain and C-terminal domain with a novel folding pattern (αααβαβα fold). The TIM barrel domain has eight parallel β-strands and eight α-helices. Between these two domains a cleft is located where flavin mononucleotide (FMN) is bound as a cofactor. In the deep binding pocket situated near the boundary of two domains, one molecule of non-covalently bound FMN is present and majority of the binding site residues are contributed by the main $(\beta/\alpha)_8$ barrel domain. The COFACTOR server (Roy et al., 2012), predicted ligand binding site in the model to be similar to PDB_ID: 2GJL_A that is co-crystallized with cofactor FMN. The CASTp predictions made on the modeled protein also identified pocket that overlaps with the pocket of the template structure. The location of overlapping binding pocket is shown in Figures 2.4C and D. The phosphate moiety of FMN is buried completely inside the pocket and is solvent inaccessible. This phosphate moiety in the modeled structure makes contacts with the backbone amide atoms of Gly180, Gly221, Ala242 and Thr243 (last three residues Gly221, Ala242, and Thr243 constitute the standard phosphate binding motif

characteristic of FMN-dependent oxidoreductase and phosphate binding enzymes family of $(\beta/\alpha)_8$ barrel proteins). The edge of isoalloxazine ring of FMN is somewhat accessible from the protein surface. The FMN binding pocket within 5Å region is lined by amino acid residues, model (template), Gly22 (Gly21), Gly23 (Gly22), Met24 (Met23), Gly25 (Gln24), Val26, Asn99 (Asn73), Leu101 (Thr75), Ile147 (Lys124), Glu175 (Asp145), Ser179 (Cys149), Gly180 (Ala150), Gly181, His182 (His152), Ala219 (Ser178), Gly220 (Gly179), Gly221 (Gly180), Gln240 (Asn199), Met241 (Met200), Ala242 (Gly201), Thr243 (Thr202), Leu246 (Leu205), Tyr338 (Tyr277), Phe339 (Ser288) and Asn343 (Val292) as shown in Figure 2.4E. A loop region comprising Phe339 (Ser288) covers the dimethylbenzene part of the isoalloxazine ring and contributes to the formation of binding site for the substrate, 2-nitropropane. Residues involved in interactions with FMN include Gly23, Val26, Gly181, His182, Gly221 and Thr243. Hydrogen bonds that mediate interactions between FMN and enzyme are Gly23:OH…FMN:O2 (2.17Å), Val26:NH…FMN:N5 (2.33Å), Gly:181:NH…FMN:O4' (2.45Å), Gly:181:NH…FMN:O2P (2.26Å), His182:HD1…FMN:O5' (1.68Å), Gly221:NH…FMN:O3P (1.79Å), Ala242:NH…FMN:O1P (1.91Å), Thr243:NH…FMN:O2P (1.94Å) and Asn343:NH2…FMN:O3P (2.67Å). Further oxidation of the neutral nitroalkanes by NAO needs a catalytic base to initiate oxidation of the neutral substrates by abstracting a proton from the substrate α-carbon (Lehoux and Mitra, 1999). There are many studies in which histidine abstracts a proton from the α-carbon thus exhibiting catalytic base like function and further helps in initiating the oxidation of neutral substrates. In our model structure of HP0773, we propose that His182 (His152) located close to the isoalloxazine ring of FMN in the active site may be involved in enzyme catalysis.

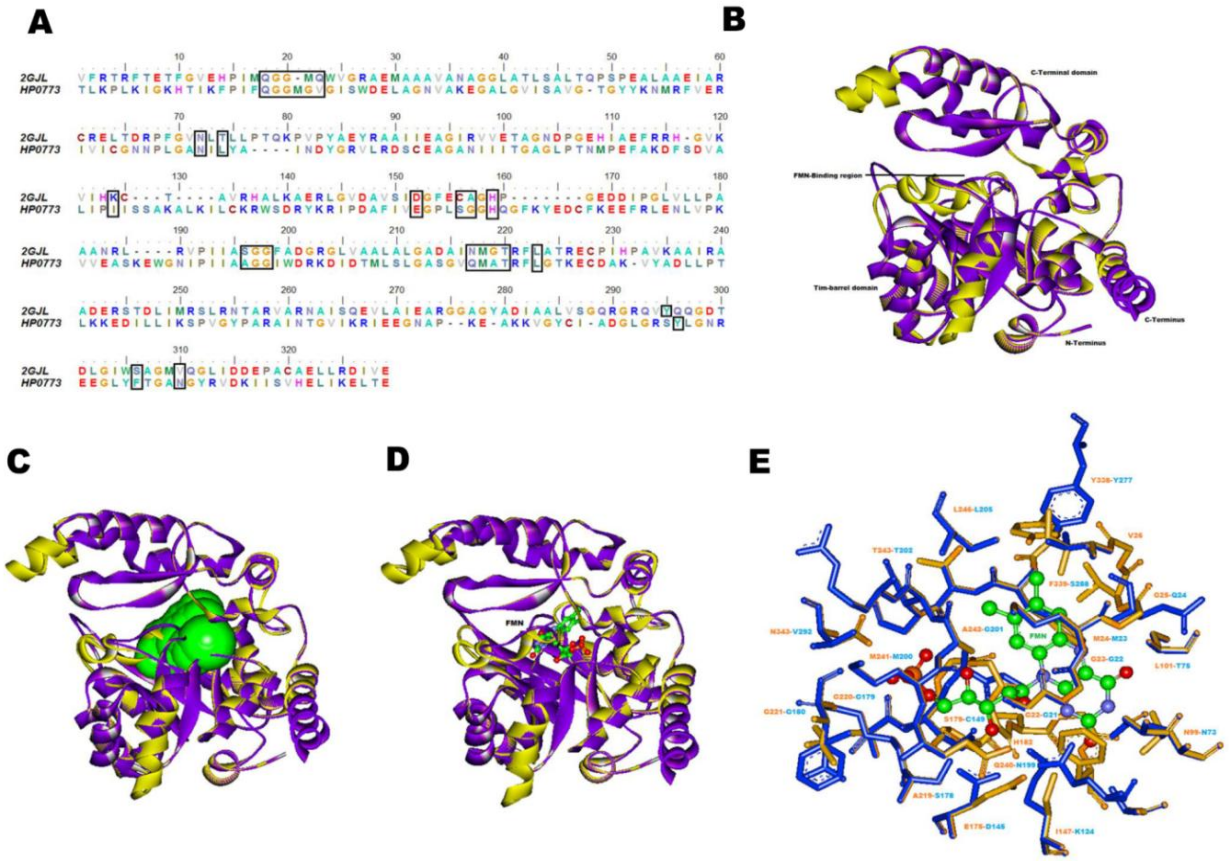


Figure 2.4: Modeling and structural analysis of HP0773. (A) Sequence alignment of HP0773 and its structural template PDB_ID: 2GJL_A used for modeling. Ligand binding residues are shown in black boxes. (B) Superimposition of HP0773 model (yellow) with its template structure PDB_ID: 2GJL_A (purple). (C) In the HP0773 model, CASTp predicted cofactor binding pocket is shown in green surface. (D) Superimposition of the predicted ligand binding site of model with the template structure. (E) Association of the FMN cofactor to the predicted binding site of model (residues in yellow) based on high similarity to a known FMN binding site of template 2GJL (residues in blue). Cofactor FMN is represented in ball and stick model, carbon- green, oxygen- red, nitrogen-blue and phosphorus-orange.

Another example we discuss is HP1214 (UniProt ID: O25813, 240 amino acids) which is an uncharacterized protein in the database and we annotated it as uracil phosphoribosyltransferases (PRTase). This protein was modeled by Phyre2 server using PDB_ID: 1WD5_A as template which corresponds to crystal structure of a predicted phosphoribosyltransferase from *Thermus thermophiles*. Both model and template structures superimposed with low RMSD (0.09Å) (Figures 2.5A and B). According to PROCHECK, the stereochemical quality of the model had no residues in the disallowed regions of Ramachandran plot and Verify_3D showed that 91.47% of the residues had an averaged 3D-1D score > 0.2, indicating good quality of the model

structure. According to SCOP classification, the structural fold of the model structure was found to be PRTase-like.

The DALI homology search of the model showed close similarity with uracil phosphoribosyltransferase from various organisms like *Sulfolobus solfataricus* (PDB_ID: 3G6W), *Aquifex aeolicus* (PDB_ID: 2E55) and *Bacillus caldolyticus* (PDB_ID: 1I5E). The COFACTOR server predicted inhibitor bound crystal structure PDB_ID: 3G6W as template which has similar binding site as that of model. Analysis of ligand binding residues showed that these regions overlap appreciably well with the modeled protein as shown in Figures 2.5C and 2.5D. Structure based sequence alignment of HP1214 and 3G6W is shown in Figure 2.5E Further comparison of ligand binding residues in both structures showed that most of the residues are same and occupy equivalent positions as that of template as shown in Figure 2.5F.

Purine and pyrimidine nucleotides can be synthesized both by *de novo* pathways, from unrelated compounds as well as by salvage pathways by converting the preformed bases and nucleosides to nucleotides. PRTases catalyze the reversible transfer of a phosphoribosyl group from 5-phosphoribosyl-α-1-diphosphate (PRPP) to N1 nitrogen of base resulting in the formation of β-N-riboside monophosphate. The product β-N-riboside monophosphate formed is specified by the base present which can be adenine, guanine, hypoxanthine, xanthine and uracil. Further these PRTases are classified into type I and type II based on the presence and absence of 13 amino acid sequence (PRPP binding motif), respectively (Argos et al., 1983). This well conserved pattern has binding site for PRPP and typically comprises four hydrophobic residues followed by two acidic residues, two hydrophobic residues and four small residues such as glycine (Hove-Jensen et al., 1986).

The salvage enzyme uracil phosphoribosyltransferase (UPRTase) belonging to the type I family of PRTases, catalyzes the conversion of uracil and PRPP to uridine monophosphate (UMP) and diphosphate (PPi) (Schramm and Grubmeyer, 2004). In addition to PRPP binding motif, the sequences of different PRTase reveal little similarity, although a common fold had been predicted for this group of enzymes (Smith, 1995). Type I PRTases are characterized by the presence of common structural core domain, comprising of four or five parallel β-strands enclosed by at least three α-helices with a subdomain called as hood which includes residues critical for pyrimidine binding. Apart from these conserved core regions, presence of two long loops protruding from the core of protein are also a distinctive feature of all Type I

PRTases. Out of the two loops, first one is the β-arm near the N-terminus, which is important for the formation of a stable dimer by embracing a neighboring subunit while the other flexible loop is present close to active site. Like typical type I PRTases, our modeled structure of HP1214 has core region formed by parallel β-strand (β2, β1, β6, β7 and β8) surrounded by α-helices (α1, α2, α5 and α6), containing the conserved PRPP binding motif (residues 147-159) located in the β6–α5 loop. Since our homology modeling is based on a monomer template, the first long loop which is involved in the formation of stable dimer was absent, while the other flexible loop (β3– β4) close to the active site was present with additional 43 residue insertion (β5, α3 and α4), which is not present in the other type I PRTases. The flexible loop present above the active site may be involved in closing the active site during catalysis to protect the intermediate/transition state from hydrolysis. In the inserted region, the β5-strand forms an antiparallel strand (β3, β4 and β5), and the α3-helix interacts with the α6-helix. At the C-terminus, we observed the subdomain region (α7 and α8 helices) that is similar to the hood in type I PRTases but this region has no significant sequence similarity to those of the other type I PRTases, and these hood structures are completely different suggesting that they may be involved in binding of an unknown substrate (Kukimoto-Niino et al., 2005). The active site cleft is situated between the hood and the core harboring the PRPP binding sequence. The COFACTOR server predicted that PRPP binding pocket is enclosed by model (template) Leu51 (Ile78), Ser52 (Leu79), Phe53 (Arg80), Asn54 (Ala81), Asp151 (Asp140), Arg152 (Pro141), Gly153 (Met142), Ile154 (Ile143), Glu155 (Ala144), Thr156 (Thr145), Gly157 (Ala146), Phe158 (Ser147) and Arg159 (Thr148) as shown in Figure 2.5 (E). In the template structure we found that residues Arg80, Asp140, and Thr148 are mostly involved in hydrogen bonding. In HP1214 model structure, side chains of Asp151 and Arg159 are very close to PRPP as compared to template and are involved in hydrogen bonds formation with the PRPP. These include Asp151:OH…PRPP:O2 (2.03Å), Asp151:OH…PRPP:O3 (2.50Å), Arg159:NH…PRPP:O1P (1.76Å), Arg159:NH1…PRPP:O3 (1.2Å), Arg159:NH2…PRPP:O1 (1.8Å), Arg159:NH1…PRPP:O2 (2.2Å), Arg159:NH2…PRPP:O3 (2.4Å), Arg159: NH2…PRPP:O1A (1.01Å), Arg159:NH2…PRPP:O1 (2.03Å), Arg159:NH2…PRPP:O1A (0.86Å) , Arg159:NH2…PRPP:O2A (2.08Å) and Arg159:NH2…PRPP:O3A (1.67Å). Other hydrogen bonding interactions include

Phe53:NH…PRPP:O3B (2.03Å) and Gly157:NH…PRPP:O2P (1.76Å) and Thr156:NH…O3P (2.45Å).

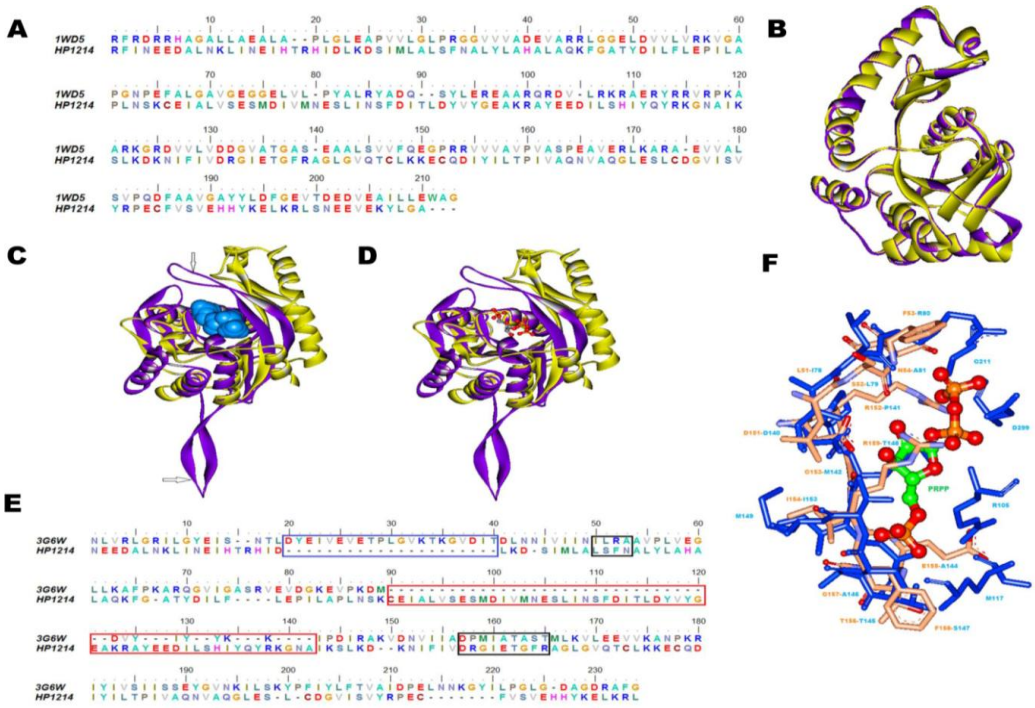


Figure 2.5: Modeling and structural analysis of HP1214. (A) Sequence alignment of HP1214 and its structural template PDB_ID: 1WD5_A. (B) Superimposition of HP1214 model (yellow) with its template structure PDB_ID: 1WD5_A (purple). (C) In the model structure HP1214, CASTp predicted cofactor binding pocket is shown in light blue surface (D) Superimposition of the predicted ligand binding site of model with the template structure. (E) Structure based sequence alignment of HP1214 with PDB_ID: 3G6W_D, residues involved in ligand binding are shown by black boxes, residues of the loop region involved in the dimer formation are shown in blue boxes while red boxes indicate residues involved in the formation of inserted loop. (F) Association of the PRPP ligand to the predicted binding site (residues in yellow) based on high similarity to a known PRPP binding site of template which is PDB_ID: 3G6W_D (residues in blue). PRPP is represented in ball and stick model, carbon-green, oxygen-red, nitrogen-blue and phosphorus-orange.

Another example is HP1504 (UniProt ID: O26034, 238 amino acids) which is unknown in the Pylorigene database but we annotated it as MTase. This protein was modeled by Phyre2 using PDB_ID: 3LPM as template which is the crystal structure of putative MTase small domain protein 2 from *Listeria monocytogenes*. Both model and template structures superimpose with low RMSD (0.13Å) (Figures 2.6A and 2.6B). Modeled protein had 1.6% residues in the disallowed region of the Ramachandran plot and Verify_3D showed that 90.34% of the residues had an averaged 3D-1D score > 0.2, indicating good quality of the model constructed. According to SCOP classification, the predicted fold of the model was AdoMet dependent MTases.

MTases form a large group of enzymes that methylate a variety of substrates but can be segregated into several subclasses based on their structural features. The most

common class of MTases is class I that contains a Rossman fold for binding SAM or AdoMet while class II MTases have a SET domain. Methylation of proteins is found to have regulatory role in protein activation, protein-protein interactions and protein-DNA interactions. They also play an important role in methylation of ribosomal RNA (rRNA) nucleotides, which further helps in the biogenesis and activity regulation of the ribosome, such as fine tuning of local rRNA structure, 30S subunit assembly and antibiotic resistance (Chow et al., 2007; Decatur and Fournier, 2002). The DALI homology search of the model structure showed close resemblance with RNA MTase from various organisms such as *Escherichia coli* (PDB_ID: 2B3T), *Escherichia coli* K-12 (PDB_ID: 4DCM), *Pyrococcus horikoshii* (PDB_ID: 1WY7) and *Thermus thermophiles* (PDB_ID: 3CJT).

Ligand binding site prediction made on the model structure using COFACTOR server identified a pocket that overlaps with the inhibitor bound crystal structure of 16S rRNA MTase belonging to class I of MTases (PDB_ID: 2ZWV_A) which is co-crystallized with ligand AdoMet. These sites are also detected by the pocket predictions by the CASTp server as shown in Figures 2.6C and 2.6D. Structure based sequence alignment of HP1504 and 2ZWV_A is shown in Figure 2.6E. These comparisons further show an extensive overlap with the binding pocket of other MTase. The residues that constituted the SAH binding site were conserved, these residues (within 5Å radius of SAH) were extracted and compared to known binding sites of ligands in PDB_ID: 2ZWV. The topologically equivalent positions in both proteins, enclosing SAH are model (template) Tyr15, Tyr17, Asn18 (Asn213), Ser21 (Ser216), Asp40 (Asp239), Ile41 (Leu240), Gly42 (Gly241), Ser43 (Ala242), Gly44 (Gly243), Leu48 (Leu247), Val64 (Val261), Glu65 (Glu262), Lys66 (Asp263), Met70 (Ser267), Gly89 (Ser287), Asp90 (Asp288), Phe91 (Val289), Asn106 (Asn305), Pro107 (Pro306) and Pro108 (Pro307). The alignment of the binding site residues in both structures can be visualized in Figure 2.6F.

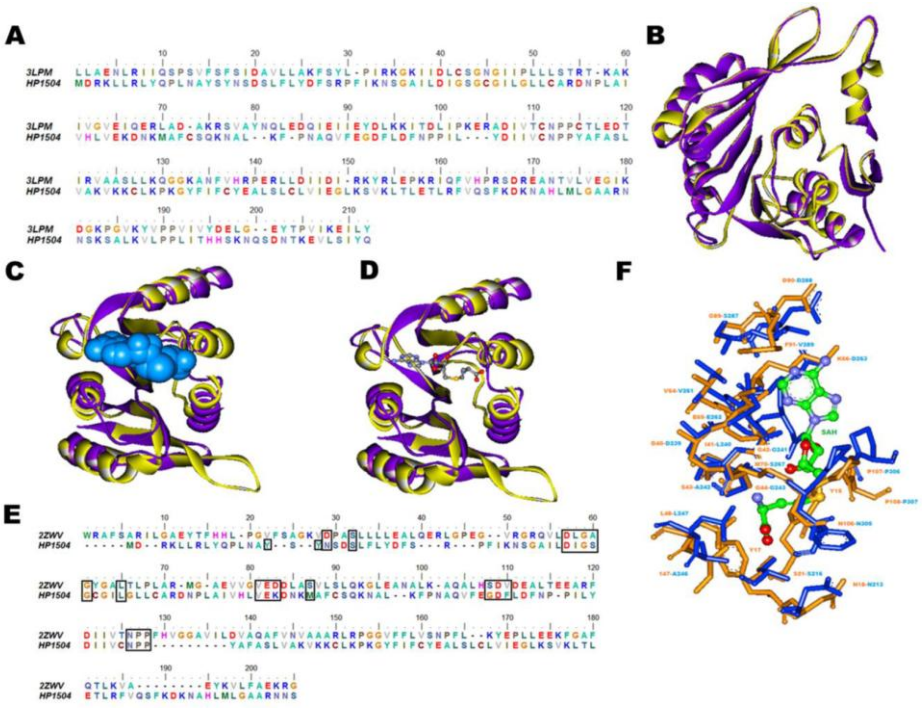


Figure 2.6: Modeling and structural analysis of HP1504. (A) Sequence alignment of HP1504 and its template PDB_ID: 3LPM_A used for modeling. (B) Structural superimposition of HP1504 model (yellow) with its structural template PDB_ID: 3LPM_A (purple). (C) In the model structure HP1504, CASTp predicted ligand binding pocket is shown in blue surface. (D) Superimposition of the predicted ligand binding site of model with the template structure. (E) Structure based sequence alignment of HP1504 sequence with PDB_ID: 2ZWV_A, residues involved in ligand binding are shown in black boxes. (F) Association of the SAH ligand to the predicted binding site (residues in yellow) based on high similarity to a known SAH binding site of template which is PDB_ID: 2ZWV_A (residues in blue). SAH is represented in ball and stick model, carbon- green, oxygen-red, nitrogen-blue and phosphorus-orange.

Proteins, for which reliable and validated model structures could not be built, were further analyzed with high confidence using fold based prediction methods (PSIPRED and FUGUE) and useful annotation could be attributed to some unknown proteins. Some of these examples include, HP0013 (tRNA (5-methylaminomethyl-2-thiouridylate)-MTase), HP0031 (USP like protein, universal stress protein), HP0728 (isoleucyl-tRNA lysidine synthetase), HP1211 (alginate lyase) and HP1413 (NADPH-dependent 7-cyano-7-deazaguanine reductase).

Using the protein sequence based annotation, the function of some proteins could be predicted. Some of these examples include; HP0129 (zinc ion binding protein), HP0130 (DNA binding), HP0158 (N-linked glycosylation), HP0980 (metalloendopeptidase activity), HP0971 (transport protein) and HP0935 (N-acetyltransferase protein).

This systematic structural and functional annotation of *H. pylori* 26695 proteome enhances our knowledge about this pathogenic organism and further provides guidance to find new drug targets.

2.4 Conclusions

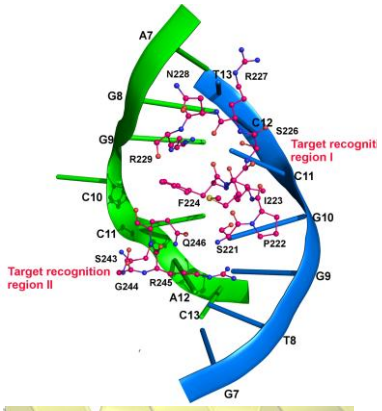
In recent years, the need to design and develop novel antibacterial agents has become crucial due to the global outbreak of infectious diseases. Now and in the immediate future, the understanding of the functions of all proteins in *H. pylori* 26695 is not feasible since there are several hundreds of proteins which have to be biologically or biochemically characterized. This current work provides useful information for better understanding of *H. pylori* 26695 strain proteome by providing information on the 3-D structure of a protein, which is further useful to predict the structure-function association and cellular functions. The structural annotation reported here covers a significant proportion of the *H. pylori* proteome. This annotation has provided insights about the folds for a significant number of proteins and importantly indicates that cellular metabolism in *H. pylori* 26695 can be achieved with only 411 folds and their various combinations. High confidence molecular models can now be obtained for several proteins, which along with the experimental structures available for that species can provide a first glimpse of the structural proteome as well as key residues and motifs present in the functional sites of that protein.

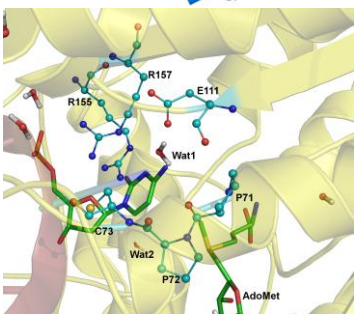
Assigning of structure-function to several unknown proteins that may be probable virulence determinants will allow critical tests of their functions, cellular targets as well as the innate and adaptive immune responses of the host. This will further aid in novel understanding into their mechanisms of early colonization, persistence of this bacterium during long term carriage, and the mechanisms by which it promotes various gastroduodenal diseases, and are therefore novel drug targets. The large scale annotation pipeline used here to derive biological insights about *H. pylori* can be readily applied for other organisms as well and will fill the ‘blank spots’ in respective proteomes. At a later stage, the new drug targets can be exploited to identify novel inhibitors using computational methods, as recently demonstrated for *H. pylori* DapE (Mandal and Das, 2014).

3

Structure and dynamics of *H. pylori* 98-10 C5-cytosine specific DNA MTase in complex with AdoMet and DNA

- ✓ Sequence analyses and 3-D structure modeling of C5mC DNA MTase from *H. pylori* 98-10.
- ✓ MD simulations of protein in complex with AdoMet and DNA.
- ✓ Important residues involved in catalytic activity and specific recognition of DNA.





Structure and dynamics of *H. pylori* 98-10 C5 cytosine specific DNA methyltransferase in complex with S-adenosyl-L-methionine and DNA. Singh S, Tanneeru K, Guruprasad L.
Mol. BioSys. 2016. DOI: 10.1039/C6MB00306K

3.1 Introduction

Cancer is a major public health problem affecting about 14.1 million people worldwide that causes more than 50% deaths according to Cancer Research, UK. Gastric or stomach cancer is the fifth most prevalent cancer with 952,000 new reported cases in 2012 alone according to the world cancer research fund international (http://www.wcrf.org/int/cancer-facts-figures/data-specific-cancers/stomach-cancer-statistics) (Kawabata-Shoda et al., 2015; Lacueva et al., 2010; Parkin et al., 2001). About 90% of the gastric cancer patients have adenocarcinoma and the remaining 10% have lymphoma or gastrointestinal stromal tumour (Nishizawa and Suzuki, 2015). Environmental factors such as aging, diet, chronic inflammation and microbial infection are the causative agents of gastric carcinogenesis. Several factors contribute to the premalignant manifestation of the gastric mucosa that finally leads to gastric neoplasia, these include the accumulation of mutations in oncogenes and tumor suppressor genes, epigenetic alterations such as methylation of DNA, modification of histone, chromatin remodeling and changes in the expression of microRNAs (Miremadi et al., 2007; Sawan et al., 2008). The control of abnormal epigenetic alterations is essential for the treatment of various types of cancers as they can be transmitted from one generation to the next one (Baylin and Jones, 2011).

The Gram-negative microaerophilic bacterium *H. pylori* inhabits the acidic environment of gastrointestinal tract of humans and are a predominant pathogen of human gastric microbiota. In the year 1994, International Agency for Research on Cancer, a subordinate organization of the World Health Organization (WHO), classified *H. pylori* as a class I carcinogen (1994) responsible for gastric cancer. Atrophic gastritis is a premalignant lesion and *H. pylori* infected patients with gastric ulcer have an increased risk of gastric cancer (Parsonnet, 1996; Sue et al., 2015; Vogiatzi et al., 2007). A mechanism for carcinogenesis resulting from *H. pylori* triggered inflammation has been proposed which suggests that multiple factors contribute to the etiology of different stages of gastric cancer (Correa, 1992).

The occurrence of natural competence, high mutation and recombination frequencies in *H. pylori* makes it genetically diverse (Israel et al., 2001). The *H. pylori* strains HPAG1 and B128 were isolated from patients with chronic atrophic gastritis (Oh et al., 2006) and gastric ulcer, respectively, the strains 26695 and J99 were

isolated from patients with superficial gastritis (Israel et al., 2001) and duodenal ulcer, respectively (Alm and Trust, 1999) and the strain 98-10 was isolated from a patient with gastric adenocarcinoma (Ando et al., 2002). Based on the phylogenetic analysis, these *H. pylori* strains have been classified into three major divisions, the West African (J99), the European (HPAG1, 26695, B128) and East Asian (98-10) (McClain et al., 2009).

The structural and functional annotation of *H. pylori* 26695 proteome has been previously studied. The enzymes in this proteome were annotated based on their EC numbers (Resende et al., 2013) and the whole proteome was annotated based on the sequence as well as structure information, thus reducing the number of hypothetical/uncharacterized proteins to a significant extent.

Large number of genes encoding R-M systems are found in the *H. pylori* genome (Skoglund et al., 2007) which protect the bacteria from the transformation of DNA from other bacteria or transduction from phages (Takahashi et al., 2002). The main functions of R-M systems are the DNA modification, which requires enzymes for methyl transfer and restriction modification. During the DNA modification, a methyl group is transferred to a specifically recognized DNA sequence thereby protecting this site from digestion by a corresponding restriction endonuclease.

MTases have been extensively studied because of their widespread occurrence and importance in biological systems and the resulting methylation plays an essential role in numerous cellular processes such as DNA protection from a cognate restriction endonuclease or epigenetic effects on gene expression (Casadesus and Low, 2006). Site specific DNA MTases are divided into two categories: exocyclic amino MTases and endocyclic MTases. All DNA MTases follow a similar pattern of methylation where the methyl group is transferred from AdoMet to the nucleotide base of substrate DNA resulting in the formation of a methylated product and AdoHcy. MTases of the two categories are significantly different from one another, since their targets for methyl transfer are quite different (Jeltsch, 2002). Several C5mC DNA MTases have been structurally characterized with AdoMet as well as in complex with their substrate DNA (Klimasauskas et al., 1994; O'Gara et al., 1996). The catalytic mechanism of these enzymes involves the flipping out of target base from the DNA double helix (Estabrook et al., 2004).

So far, no structural studies have been performed on the DNA MTases from *H. pylori*. We have therefore chosen to study the C5mC DNA MTase from the gastric

cancer causing strain of *H. pylori* 98-10. In this work, we have modeled the 3-D structure of C5mC MTase to provide insights of the protein active sites. Further, we have docked AdoMet as well as DNA in the protein model. This modeled 3-D structure allows us to perform general structural comparison with the other C5mC MTases. Further, we have performed MD simulations with the model of M. Hpy C5mC in complex with AdoMet and DNA to gain insights about the important residues responsible for their specific recognition and to understand the mechanism of methylation.

3.2 Method

The template structure for *H. pylori* 98-10 DNA MTase (M. Hpy) C5mC (NCBI_ID: EEC22533.1) was identified based on the BLAST searches (Altschul et al., 1990) and the fold prediction method FUGUE (Shi et al., 2001). We have used the sequence alignments generated by FUGUE as a guide to build the 3-D model structure of the M. Hpy C5mC using HOMOLOGY module in Discovery Studio 2.5 (Accelerys Inc, USA) that implements the methodology described in MODELLER (Sali and Blundell, 1993; Shen and Sali, 2006). MODELLER is a homology or comparative modeling program for constructing the 3-D model of a protein structure from its amino acid sequence. The program constructs a model for all non-hydrogen atoms by the satisfaction of spatial restraints that includes non-homologous loops and energy optimization of the final model. The stereochemical quality of the best model was validated using PROCHECK and Verify_3D was used to validate the compatibility of the 3-D model structure with its primary amino acid sequence.

The bimolecular and termolecular systems of M. Hpy C5mC were created by extracting AdoMet and AdoMet+DNA, respectively from a homologous structure PDB_ID: 6MHT. The restriction enzyme database, REBASE has a reliable tool for predicting the likely consensus DNA recognition sequences that bind to a given DNA MTase (Roberts et al., 2015). In the termolecular system, the nucleotide bases of extracted DNA from PDB_ID: 6MHT were mutated according to the prediction made by REBASE. To confirm the role of regions responsible for DNA binding, we performed alanine scanning mutagenesis experiments on the proposed sequence motifs. We have used 'Build and edit protein' module present in Discovery Studio 2.5 to mutate the desired residues in M. Hpy C5mC.

3.2.1 Molecular dynamic simulations

The molecular systems, unimolecular system- the modeled structure of M. Hpy C5mC, the bimolecular system- the modeled structure of M. Hpy C5mC bound to AdoMet and the termolecular system- the modeled structure of M. Hpy C5mC bound to AdoMet and DNA.

Four mutated termolecular systems were also generated after mutating all the residues to Ala within the loops; 75-83, 221-229, 243-246 and the system accommodating mutations in all the three loops. As a result of mutagenesis, for instance, the 75-83 loop now comprises the sequence Ala75-Ala76-Ala77-Ala78-

Ala79-Ala80-Ala81-Ala82-Ala83. All these systems were subjected to 50 ns MD simulations using GROMACS 4.5.5 package (Hess, 2009; Van der Spoel et al., 2005). The protein and DNA force fields were generated using the AMBER ff99SB (Hornak et al., 2006) and the force fields for AdoMet were generated in Antechamber (Wang et al., 2006b) using ACPYPE script (Sousa da Silva and Vranken, 2012).

All the three molecular systems were immersed in an octahedron box of extended simple point charge (SPC) water molecules. The termolecular system was neutralized by adding 11 Na^+ and the other two systems were neutralised by adding 11 Cl^- each. To relieve the short range bad contacts, energy minimization was performed using the steepest descent method for 5000 steps followed by the conjugate gradient method for 5000 steps. The MD simulation studies consist of equilibration and production phases. The position restrained simulations were carried out at 298 K for 1 ns. Finally, the three systems were subjected to 50 ns MD simulations production run at 298 K temperature and 1 bar pressure using 0.002 ps time step. The Parrinello–Rahman method was used to control pressure (Parrinello and Rahman, 1981) and the V-rescale thermostat was used to maintain temperature (Bussi et al., 2007). The long range electrostatics were handled using the Particle Mesh Ewald (PME) (Darden et al., 1993) method with a real space cut-off of 10Å, PME order of 6 and a relative tolerance between long and short range energies of 10^{-6}. Short range interactions were evaluated using a neighbour list of 10Å updated every 10 steps while LJ interactions and the real space electrostatic interactions were truncated at 9Å. Hydrogen bonds were constrained using the LINCS algorithm (Hess et al., 1997). The final models in all the three systems were obtained by averaging the snapshots from the trajectories generated by MD simulations after the structure stabilization was achieved. Four mutated termolecular systems (75-83, 221-229, 243-246 and the system accommodating mutations in all the three loops) generated after alanine scanning mutagenesis were also subjected to MD simulations as described above.

To study the conformational variations in the structures of M. Hpy C5mC, the RMSD of the atomic positions with respect to their starting structures were calculated by using g_rms of GROMACS by least-square fitting the structure to the reference structure. The convergence of MD simulations was analysed in terms of the potential energy, RMSD and RMSF. The kinetic, potential and total energies that estimate the variations in the energetics of the systems were also analysed as a course of simulation time.

3.3 Results

3.3.1 Structure

The structural template identified by FUGUE methodology for M. Hpy C5mC, was DNA MTase from *Haemophillus aegyptius* (M. HaeIII) (PDB ID: IDCT) that shares 65% amino acid sequence identity. The sequence alignment used as a basis for homology modeling is shown in Figure 3.1. The homology model was constructed using MODELLER qualified the structure validity test performed by PROCHECK (90.1% residues were in the core and 9.9% residues were in the allowed region of the Ramachandran plot and Verify_3D (84.4% of the residues had an averaged 3D-1D score ≥0.2).

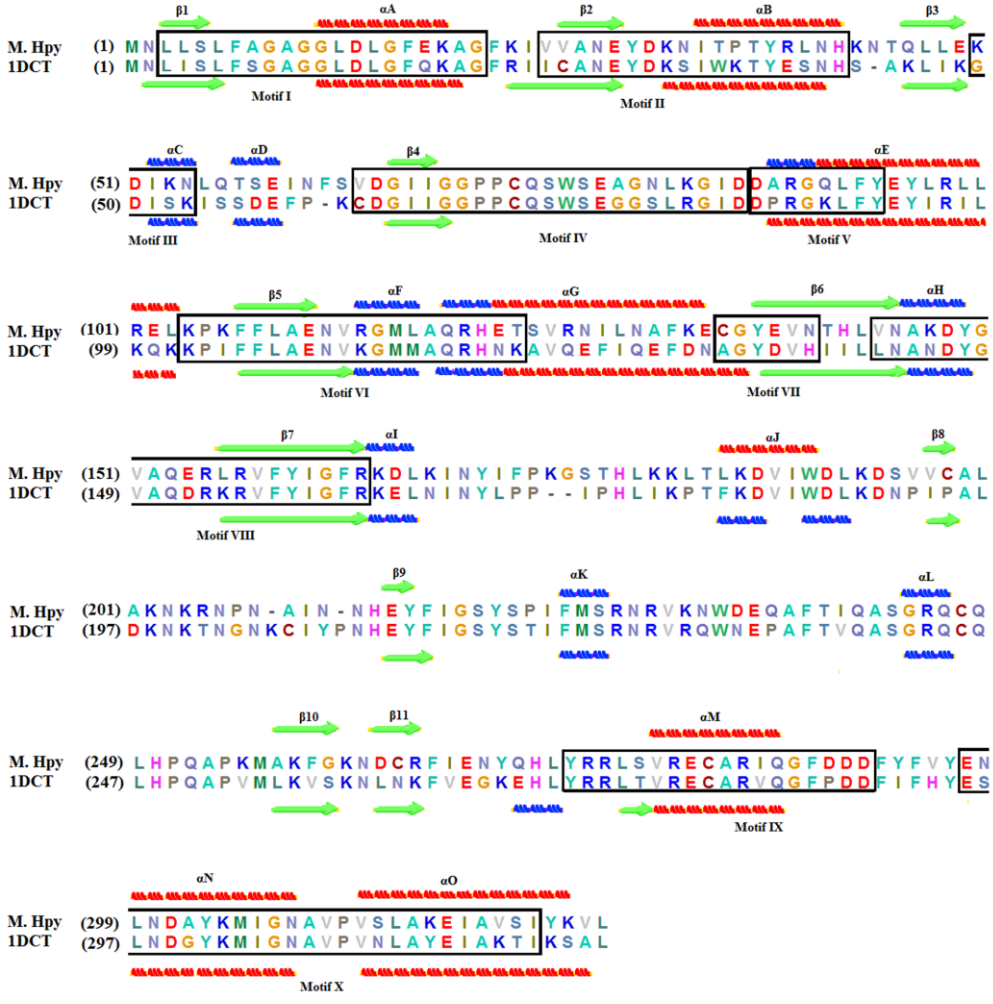


Figure 3.1: Sequence alignment of M. Hpy C5mC with its structural template M. HaeIII (1DCT). Conserved motifs characteristic of C5mC DNA MTases are represented in black boxes. (α-helix, 3_{10} -helix, β-strand).

The template and model structures were highly similar as indicated by low RMSD (0.66Å) from their structure superimposition (Figure 3.2A). The overall fold of the enzyme is a bilobal structure; the large domain has a strong resemblance to the

Rossmann fold and comprises the active site for AdoMet binding and methyl transfer. Rossmann fold is super-secondary structure characterized by alternating motif of β-α-β secondary structures (therefore this fold is also called a βαβ fold). This fold comprises of alternating β-strands and α-helices, where all strands form a central relatively planar β-strand, and helices filling two layers, one on each side of the plane (Gherardini et al., 2010). This domain was connected by a long hinge region to the small domain which recognises the substrate DNA as shown in Figure 3.2B. From the structure visualization, we observed that the DNA is located in an exposed cleft between the large and small domains, and the major groove of DNA faces the small domain while the minor groove faces the large domain that comprises the catalytic site (Figure 3.2C). According to the REBASE predictions, the M. Hpy C5mC recognizes 5'-GGC^{m}C-3'consensus DNA sequence motif.

The conserved motifs along the primary sequence in DNA MTases are the likely sites responsible for the chemistry common to all MTases, while the variable region is responsible for the sequence specificity of the substrate DNA (Ahmad and Rao, 1996a; Madhusoodanan and Rao, 2010). The primary sequences of the C5mC MTases share a constant linear order of ten conserved motifs. The majority of these motifs are responsible for three basic functions, AdoMet binding, catalysis of methyl transfer and sequence-specific DNA binding. The motifs I to V have been implicated in AdoMet binding, motifs IV and VI play a role in the catalysis of methylation, while motifs IV, VI and VIII aid in the correct positioning of the target nucleotide within the active site of the enzyme (Malone et al., 1995). In the large domain of the bilobal protein, an AdoMet dependent MTase fold consisting of six α-helices and seven β-strands (six strands are arranged in parallel and the seventh strand is in antiparallel orientation) are present (Figure 3.2D). The small domain is devoid of extensive secondary structural elements and mostly consists of random coils. The residues 197-199 (β8) form short β-strand and even smaller β-strands are formed by 214-215 (β9), 257-258 (β10) and 264-265 (β11) regions and only one helix 185-190 (αJ) is present in small domain as indicated in Figure 3.1.

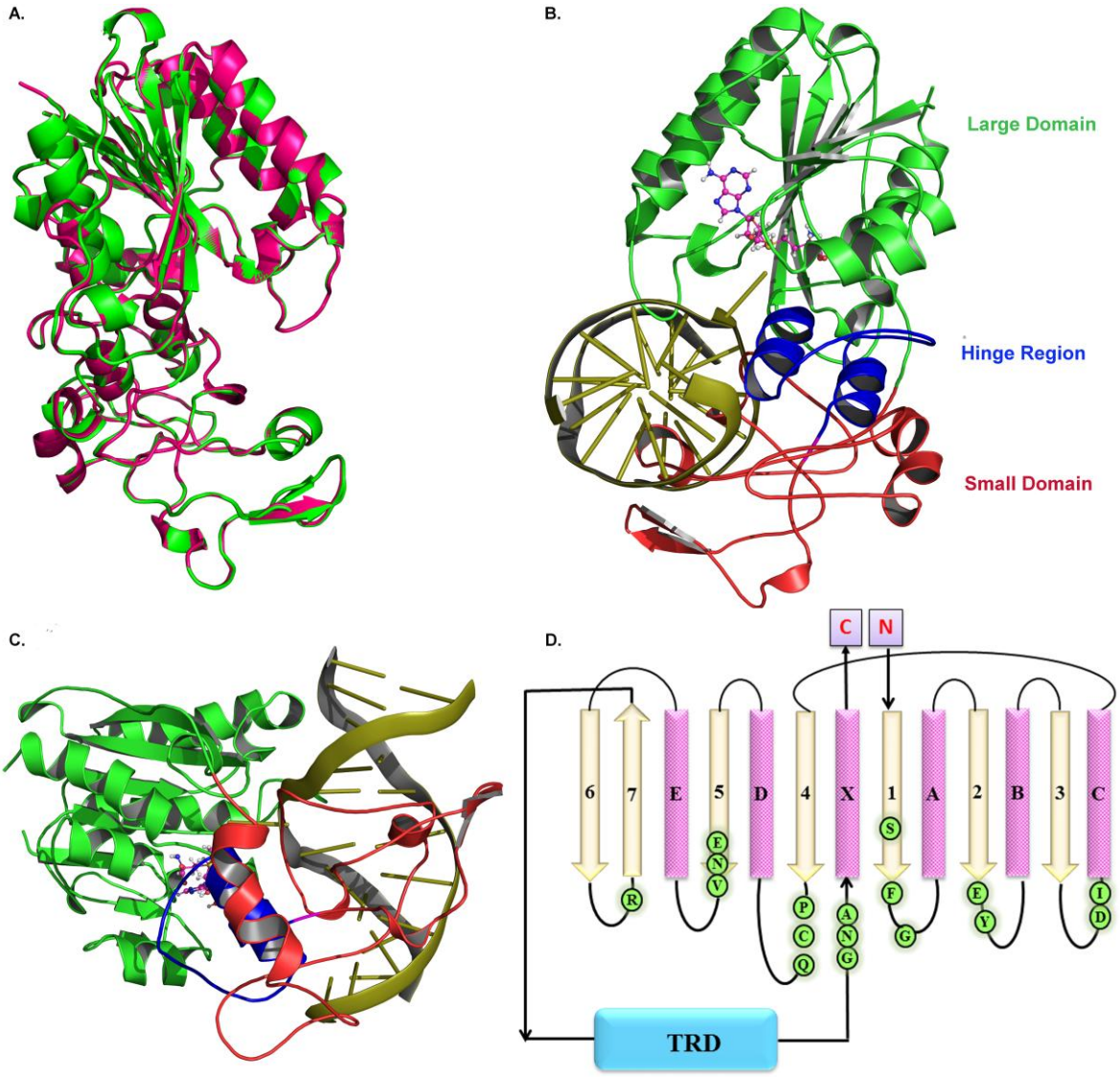


Figure 3.2: (A) Superimposition of template PDB_ID:1DCT (green) with the M. Hpy C5mC model (red). (B) Schematic representation of termolecular complex indicating the hinge region, large and small domains in protein. (C) Location of major and minor groove of DNA in the termolecular complex. (D) Secondary structural arrangement in the large domain indicating the conserved residues in M. Hpy C5mC.

In the modeled M. Hpy C5mC structure (Figure 3.2B), residues from Met1-Lys182 (motif 1-VIII) and Ile306-Tyr327 (motif X) form the large catalytic domain. The intervening region (Leu183-Arg275) constitutes the TRD, and the region from Arg276-Met305 forms the hinge region connecting the two domains. A pronounced cleft between these two domains is the DNA binding site. The large domain comprises motifs essential for catalysis and cofactor binding. The highly conserved sequence motifs, I, IV, VI, and VIII form the core of the large domain. The MTases which recognize the same consensus DNA sequence shows some degree of conservation in their TRD. The complete loss of MTase activity has been reported in the mutational studies of TRD in various DNA MTases (Szegedi and Gumport, 2000).

The highly conserved sequence motifs of C5mC MTases provided important clues regarding the common underlying architecture of these enzymes. Various motifs have been assigned functional roles related to the chemistry that is common to these enzymes. All ten conserved motifs characteristic of C5mC MTases are observed in M. Hpy C5mC modeled structure as shown in Figure 3.1. The probable functional roles of these motifs have been revealed by structure analyses of M. Hpy C5mC and comparison with other C5mC MTases.

The motif I (Leu3-Gly21) forms β1-loop-αA and is the only motif conserved among all AdoMet dependent MTases of class I (R-M systems). The consensus sequence motif FxGxG (FAGAG in the M. Hpy C5mC) in the 'G-loop' is primarily involved in binding to the methionine moiety of the cofactor AdoMet. The Phe7 located at the beginning of G-loop, makes edge to face vdW interactions with adenine moiety of AdoMet in the termolecular complex. The motif II (Val25-His42) forms the structural motif β2-loop-αB; a negatively charged Glu29 present at the end of the β2 interacts with hydroxyl groups of the ribose moiety of AdoMet in the termolecular complex. The bulky and hydrophobic side chains from the loop region make vdW contacts with the adenine moiety of AdoMet. The motif III (Lys50-Asn54) forms αC and a small loop region between β3 and αC; Asp51 and Ile52 also interact with the adenine moiety of AdoMet of the termolecular complex.

The motif IV (Val64-Asp86) forms β4 and loop between β4 and αE that comprises the invariant Pro72 and Cys73 dipeptide. This dipeptide is involved in the catalytic process *i.e.* in the formation of a covalent protein-DNA complex. Previous studies revealed that this nucleophilic Cys undergoes immense conformational change towards the DNA binding cleft and plays an important role in the catalysis (Klimasauskas et al., 1994). Mutational analysis of this residue leads to the abolishment of MTase activity (Chen et al., 1993a). The motif V (Asp87-Tyr94) mainly forms helical structure αE; Leu92 makes vdW contacts with AdoMet adenine. The motif VI (Lys104-Thr123) comprises of several hydrophobic residues and forms β5-αF-αG. The motif VII (Cys135-Asn140) forms a small region of β6 and seems to be less conserved in C5mC MTases. The dipeptide that faces Gly136-Tyr137 away from the DNA binding cleft is involved in the folding of the catalytic region. The motif VIII (Val144-Arg164) forms a small region of β6 and αH-loop-β7. The Arg157 located on β7 forms hydrogen bonds with the flipped cytosine of DNA. The motif IX (Tyr274-Asp291) forms αM and the residues on either sides of the αM regions. The

motif X (Glu297-Ile322) that forms αN-loop-αO is important, because its location in the primary sequence determines the class of DNA MTase. In the C5mC MTase this motif is present in the C-terminus located structurally next to the β1 (motif I) and consists of conserved hydrophobic residues. This motif along with the G-loop of motif I and ProCys loop of motif IV forms the binding pocket of AdoMet methionine.

The AdoMet binding pocket is mainly formed by residues from motifs I-V (Kumar et al., 1994) which are a part of the large domain. This pocket is in the neighbourhood of the catalytic site and faces the cleft present between the two domains. The motif I helps in the correct positioning of AdoMet, while the motifs II-V comprise the highly conserved residues which interact with different parts of AdoMet in both bimolecular and termolecular systems.

The comparison of the average structures from MD simulations revealed that the orientation of AdoMet in the bimolecular and termolecular systems is slightly different. As shown in Figure 3.3A, the adenine moiety of AdoMet in the bimolecular system was surrounded by Pro72, Cys73, Gln74, Arg89 and Arg114. The hydrogen bonding of Arg89 with N1 and C2 atoms of adenine, Arg114 with N7, and Cys73 with both N6 and N7 stabilize the protein-AdoMet complex. The positive charge of S^{+}-CH3 (sulfonium ion) is in a favourable orientation with reference to the π ring of Phe7 and forms cation…π interactions. The pocket surrounding the main chain amino as well as carboxyl group of methionine moiety is formed by the residues Gly9, Ala10, Gly11, Gly12, Leu13, Asn308, Ala309 and Val310. The Asn308 forms hydrogen bonding interactions with O2'and O3' of ribose sugar as well as the NH of methionine moiety from AdoMet, while the rest of the residues interact with the carboxyl group of methionine.

In the termolecular system (Figure 3.3B), the adenine moiety was sandwiched between Tyr30 and Leu92 on one side and Phe7 on the other side. From the trajectory analysis, we observed that the aromatic ring of Phe7 is perpendicular to adenine and also makes π…π stacking interactions with Tyr30 which is seen throughout the MD simulations. The Tyr30 and Asp51 interact with the adenine moiety of AdoMet while the side chain of Glu29 makes hydrogen bonds with O2'and O3' of ribose. The pocket surrounding the methionine moiety of AdoMet is formed by the residues Ala8-Asp14, Asn308, Ala309 and Val310. The NH of AdoMet methionine makes hydrogen bond with main chain carbonyl of Phe7, while its carboxylic group interacts with the Ala10-

Leu13. These interactions are also retained throughout the MD simulations, indicating their important role in stabilising the M. Hpy C5mC termolecular complex.

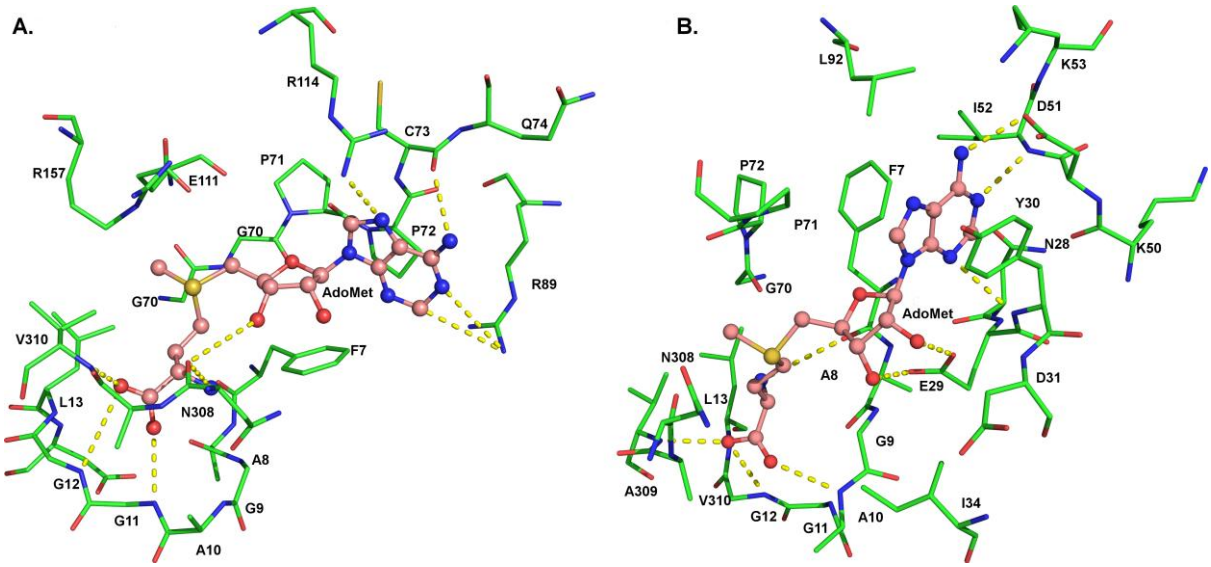


Figure 3.3: The AdoMet binding site in (A) bimolecular and (B) termolecular systems in M. Hpy C5mC.

Previous studies have reported that the invariant ProCys dipeptide present in the loop region of motif IV is important for catalysis and makes favourable contacts with DNA (Kumar et al., 1994). In our model too, the ProCys dipeptide is situated in the flexible loop region of the catalytic pocket and is present in the minor groove of DNA.

In the case of termolecular complex, this loop undergoes movement towards the DNA binding cleft and makes intimate contacts with the minor groove of the DNA, while in the other two systems, this loop is solvent exposed. It is possible that this movement of the catalytic loop in the termolecular complex is necessary to make sure that the catalytic nucleophile is in close proximity to the flipped cytosine, permitting direct attack at C6 position of cytosine.

The flipped cytosine occupies a position close to the catalytic motif, while the AdoMet is located opposite to this flipped base. The chi (χ) torsion angle in DNA that describes the relative orientation of base/sugar, is defined by O4′-C1′-N1-C2 for pyrimidines, and O4′-C1′-N9-C4 for purines. The A and B forms of DNA/RNA duplex have anti-conformation where χ falls within the ranges of +90° to +180°; –90° to –180° (or 180° to 270°) and the χ values for *syn* conformation are in the range of –90° to +90° (Lu and Olson, 2003). In the available crystal structures of C5mC DNA

MTases the χ value ranges from -55° to -85°, for example, χ values are -65° (PDB_ID: 1HMT) and -55° (PDB_ID: 6HMT). Usually, the cytosine base to be methylated has higher values of χ as compared to non-methylated bases (Lau and Bruice, 1999). During the MD simulations of M. Hpy C5mC, the average χ dihedral angle of flipped cytosine (O4'-C1'-N1-C2) was found to be -65.81(±8)° which is close to the previously reported values.

3.3.2 Mechanism of DNA methylation

Cytosine is an electron poor heterocyclic aromatic ring system making its C5 position incompetent for nucleophilic attack on methyl group of AdoMet. AdoMet on the other hand is a very effective donor. In lieu of this, the enzyme follows pathway of Michael addition, which uses cysteine SH group from invariant Cys of ProCys dipeptide (motif IV) from enzyme to make a nucleophilic attack at position C6 of cytosine ring. This results in the formation of a covalent complex between protein and DNA which acts as an intermediate step and further facilitates the nucleolytic attack on C6 by transient protonation of the cytosine ring at N3 position (Chen et al., 1991). This protonation is also mediated by Glu111 from conserved motif VI and Arg157 from motif VIII (O'Gara et al., 1996). Both these residues are required for proper orientation and stabilization of the flipped cytosine by cysteine (Shieh and Reich, 2007; Shieh et al., 2006). Mutational studies have shown that the glutamate residue helps in creating ground state conformation, thus plays a secondary role in catalysis (Zhang and Bruice, 2006). Previous reports on Cm5C showed that water channels are necessary in the active site for deprotonation of Cys-cytosine complex by providing hydroxide anions (Lau and Bruice, 1999). This results in the activation of C5 position of cytosine leading to attack on AdoMet methyl group. An important feature of proposed mechanism is that the enzyme requires contact of the aromatic ring system from both sides of the flipped cytosine, since the first attack of the SH group on the cytosine and the nucleophilic attack of the cytosine ring system on the AdoMet methyl group occur at different sides of the ring system. Therefore, the catalytic mechanism of C-MTases can explain why DNA MTases developed a base-flipping mechanism (Jeltsch, 2002; OGara et al., 1996).

The substitution of conserved valine (motif IV) causes alterations in the methylation activity, although it is not involved directly in the process of catalysis (Estabrook et al., 2004). In the case of *H. pylori*, Val113 was positioned at an average

distance of 3.5Å from the target cytosine (C4), interacting with the base through hydrophobic interactions. Mutation as well as DNA binding studies in *M. HhaI* proposed that this residue could be important for the correct assembly of the catalytic site, and also might take part in the base flipping mechanism.

The M. Hpy C5mC also follows a pathway of Michael addition, which uses cysteine SH group from invariant Cys of ProCys dipeptide to make a nucleophilic attack at C6 position of cytosine ring, and a detailed mechanism of DNA methylation has already been reported (Chen et al., 1993a; Wu and Santi, 1987). The cytosine base flips out of the helix and occupies a position near the catalytic motif. Cys73 and the cofactor AdoMet are positioned on opposite sides of the flipped base, suggesting that the thiolate addition and methyl group transfer occur from opposite sides of the ring. By analogy, the thiol group of cysteine makes nucleophilc attack on C6 of flipped cytosine of the consensus DNA sequence leading to the formation of a carbanion. The distance between SG atom of Cys73 to C6 atom of cytosine (3.2Å) is maintained throughout the MD simulations. Cys73 reacts with the C6 of cytosine resulting in the formation of a covalent complex and activation of C5 position. The active site is also composed of residues Pro71 (motif IV), Arg155 and Arg157 (motif VIII), and Glu111 (motif VI) that interact with the cytosine residue and the MD trajectory analyses also showed that the flipped cytosine established extensive contacts with Cys73, Pro71, Arg157 and Glu111 that constitute the active site of the modeled enzyme (Figure 3.4A). The side chains of Arg155 and Arg157 (motif VIII), and Glu111 (motif VI) form hydrogen bonds with the polar groups of cytosine. The residues Arg157 and Glu111 are well conserved in all C5mC MTases. The Arg155 makes hydrogen bond interactions with O2 (2.09Å) of cytosine as well as with other solvent molecules in the vicinity. The Arg157 makes hydrogen bond interactions with both exocyclic N3 (2.09Å) and O2 (2.07Å) of cytosine. Further, Arg157 also forms hydrogen bond interactions with the cytosine sugar-phosphate backbone (2.11Å). The importance of these interactions have been demonstrated in *M. HhaI*, where a substitution of Arg165 (equivalent to Arg157 of M. Hpy C5mC) by alanine causes adverse impact on the methylation (Shieh et al., 2006), the mutation of Arg230 *M. SssI* with alanine (equivalent to Arg155 of M. Hpy C5mC) had negative effect on MTase activity (Darii et al., 2009). The Arg157 along with AdoMet helps in stabilizing the flipped cytosine by making cation…π interactions. The residues Pro71 and Glu111 make hydrogen bonds with the exocyclic nitrogen N4 (2.09Å and 2.08Å respectively) ensuring that

cytosine is correctly oriented for ideal contacts with Cys73 and AdoMet as can be seen in Figure 3.4B. The conserved Glu111 from the ENV peptide (motif VI) was suggested to protonate cytosine N3 (Shieh and Reich, 2007). However, our trajectory analysis of M. Hpy C5mC depicts that this Glu111 does not interact with the cytosine N3 but interacts with a water molecule close to the N3 atom. In spite of this, cytosine ring could still be activated through water mediated interactions at N3 since water molecules are observed in the vicinity of N3 atom (3.09Å) as shown in Figure 3.4B. The participation of water molecules in the activation of N3 has been proposed in the MD simulation studies on M. HhaI (Lau and Bruice, 1999). In this work, we also observed extensive hydrogen bonds established between cytosine O2 with Arg155 and Arg157. It has been proposed that interactions at O2 can imbalance the electron density in the cytosine ring permitting the attack at C6 by the cysteine residue (Gabbara et al., 1995; Kumar et al., 1997) and may reduce the necessity for direct interactions at N3.

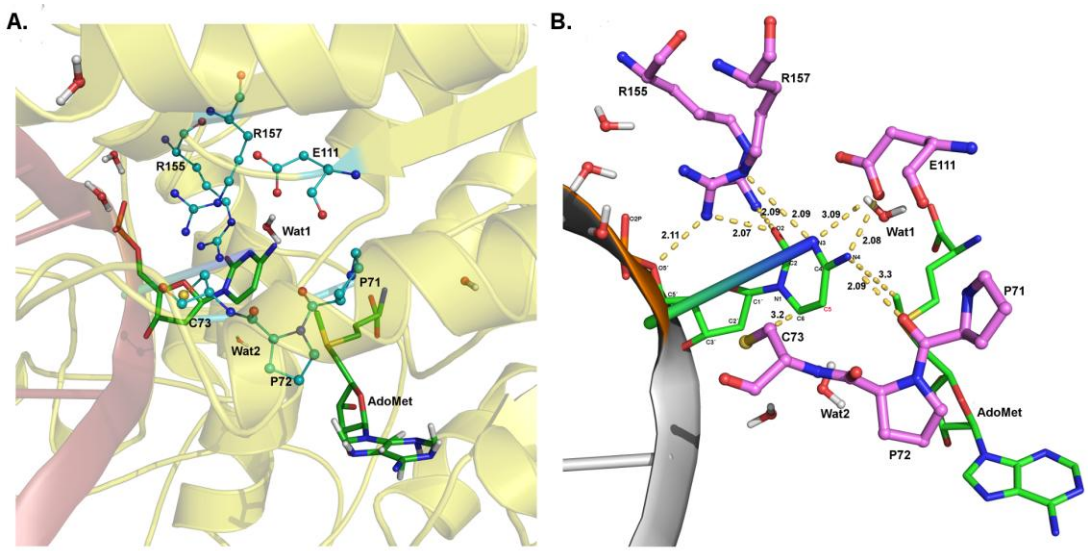


Figure 3.4: (A) The catalytic site in M. Hpy C5mC termolecular system (B) Hydrogen bonding interactions of important catalytic residues along with their distances are indicated in Å.

From these observations, we conclude that the direct interactions of N3 with Glu111 as previously proposed in M. HhaI, may not be an essential criterion for methylation to occur in M. Hpy C5mC due to the absence of direct interaction with N3 atom of cytosine in the proposed structure. An alternative mechanism may operate in M. Hpy C5mC which requires a water mediated or direct interaction of Arg155 and Arg157 with O2 atom of cytosine as proposed (Kumar et al., 1997) (Figure 3.5).

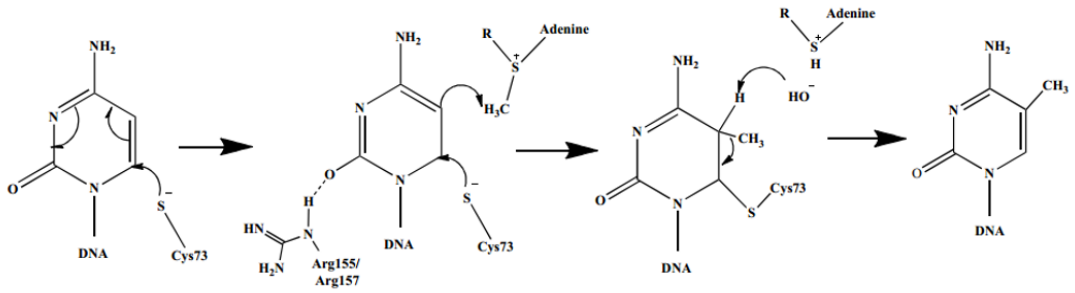


Figure 3.5: Mechanism of methyl group transfer from AdoMet to DNA in M. Hpy C5mC.

The important step in the process of DNA methylation is the transfer of methyl group from AdoMet to C5 position of cytosine ring. The distance between C5 of cytosine to CH_3 of AdoMet was found to be 3.3Å throughout the MD simulations which is sufficiently close enough to permit the transfer of methyl group. The final step in the methylation process is the release of methylated cytosine from the complex that is catalyzed by deprotonation at C5. In the termolecular system, the flipped cytosine is solvent accessible, water molecules are located in the vicinity of C5 and C6 (Figure 3.4B). The MD simulations of M. Hpy C5mC, has shown that a water molecule in the catalytic site may provide hydroxide ion that could deprotonate C5. This is in line with the proposed water mediated interactions for the deprotonation of C5 (Zhang and Bruice, 2006). Based on these results, we believe that the proposed model of M. Hpy C5mC satisfies all the necessary criteria for the methyl group transfer to DNA.

The trajectory analysis of the three molecular systems indicated important residues involved in stabilizing the protein with AdoMet and DNA. As shown in Figure 3.6A, their respective Cα-RMSD values were; 3.0Å (unimolecular system), 2.7Å (bimolecular system) and 1.8Å (termolecular system). From these results we conclude that that protein interacting with both AdoMet and DNA has lowest RMSD as compared to the other two. This was expected, because the surface exposed regions of the protein have large scale motions leading to higher fluctuations in the absence of DNA. In the termolecular complex, these surface exposed areas interact with the nucleotide bases of DNA making it less solvent exposed and stabilising the protein structure on the whole.

The RMSF plots of the Cα atoms of the proteins are shown in Figure 3.6B. Most of the fluctuating regions of protein include those parts which interact with the DNA or the loop regions which reside on the surface of the protein and the hinge

regions between the two domains. For the unimolecular system, large fluctuations are seen in the amino acid regions 71-95 (part of motif IV and region connecting β4 and αE), 113-127 (part of motif VI) and 190-196 (consists of αJ and small loop connecting αJ-β8). The 71-95 region has catalytic residues Pro72 and Cys73 and is close to the flipped cytosine and minor groove of DNA. Similar fluctuation is also observed in the bimolecular system, while the termolecular system exhibits very less fluctuation due to the presence of DNA (around 0.9Å) indicating the structural stabilization. The superposition of bimolecular and termolecular systems illustrate the involvement of active site loop in binding to the substrate DNA (Figure 3.6C). The binding of DNA also effects the relative location of the two domains in the protein, which can be seen as a small shift in the TRD towards the DNA binding cleft (Figure 3.6C). The 113-127 region close to DNA mainly comprises of helices and the 190-196 region is solvent exposed and present on the surface. This region has fluctuations up to 3Å for unimolecular system, but unusually this region is stabilised in the other two systems. The regions corresponding to 165-183 and 257-273 show fluctuations in all the three systems. The 165-183 region is present after motif VIII and is connected to the small domain, while 257-273 forms a loop region along with β10 and β11 and is present on the surface of the protein, close to the hinge region. Unusually, this region showed more fluctuations in the bimolecular system. Apart from these differences, the rest of the trajectories followed the same pattern *i.e.* more fluctuations are observed in the unimolecular system, less in the bimolecular system and least in the termolecular system.

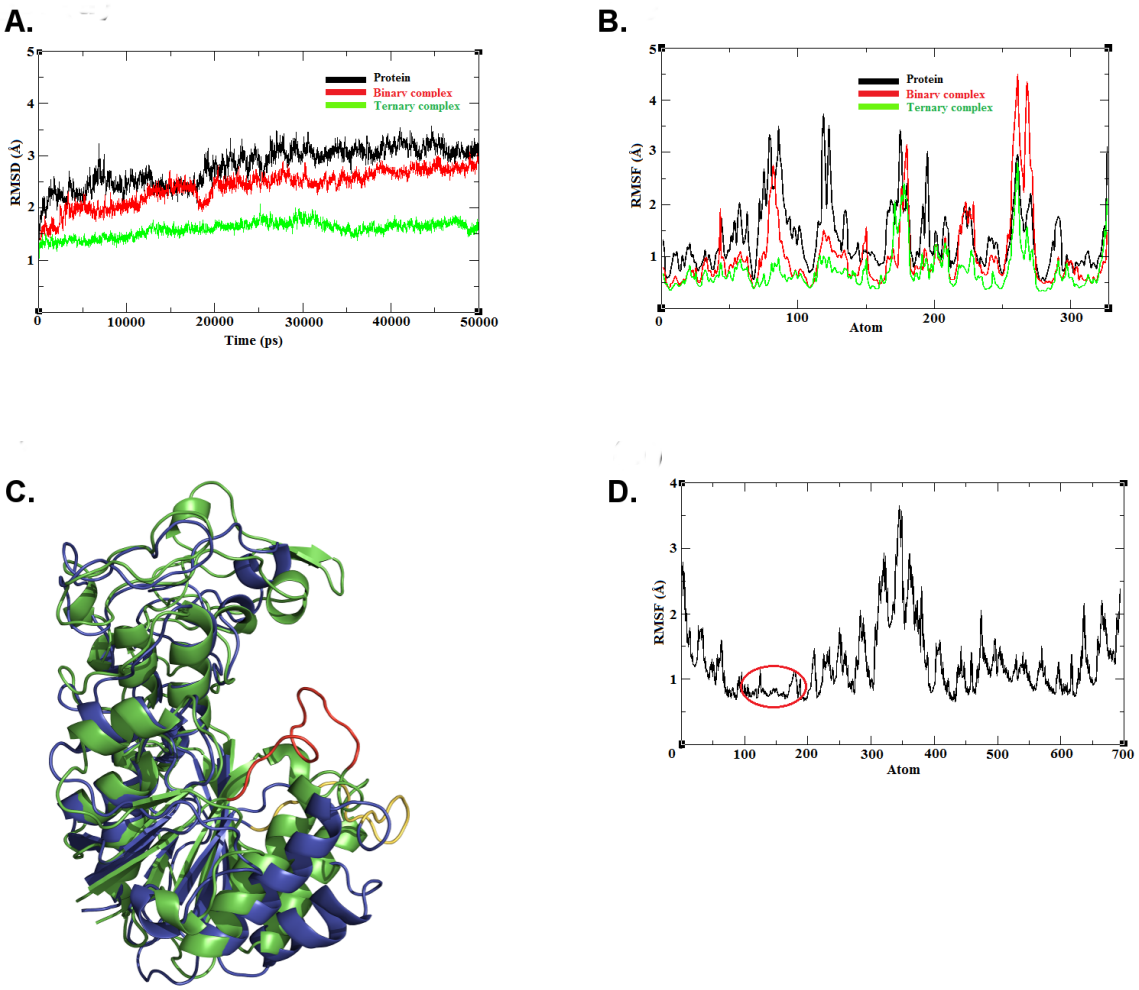


Figure 3.6: (A) RMSD of M. Hpy C5mC Cα atoms in unimolecular, bimolecular and termolecular systems. (B) RMSF of M. Hpy C5mC Cα atoms in unimolecular, bimolecular and termolecular systems. (C) Schematic representation of the superimposition of bimolecular (blue) and termolecular (green) systems. ProCys loop of bimolecular (yellow) and termolecular systems (red) are indicated. (D) RMSF of the DNA (all atoms) in the termolecular system. Target base is highlighted.

The potential energy plots of three molecular systems during the MD simulations were obtained from the trajectory files for 50ns. The potential energies of all the systems stabilized during the MD simulations indicating the stability of the three molecular systems and were found to be -5.232 x 10^{-5} kJ mol^{-1} (unimolecular), -2.089 x 10^{-5} kJ mol^{-1} (bimolecular) and -1.832 x 10^{-5} kJ mol^{-1} (termolecular) systems. The measurements of DNA fluctuations by means of RMSF showed that the fluctuations of flipped cytosine are restricted by Cys73 present in the active site resulting in low fluctuations for this nucleotide. Throughout the MD simulations, hydrogen bonds between the base pairs are well maintained in the DNA double helix. Also, in comparison, the nucleotides of DNA that face the enzyme have much lower fluctuations as compared to the solvent exposed nucleotides (Figure 3.6D). From the RMSF plot we can infer that bases of DNA which are in close vicinity of protein for

example 62-119, 416-417, 569-630 (all atom numbering in the DNA) have less fluctuation. The 62-119 region consists of atoms from consensus bases of DNA and exhibits least fluctuations. These minor positional fluctuations as compared to the rest of the DNA, explained the influence of the enzyme on the stability of DNA.

The specific recognition of DNA by the class specific MTases remains to be explored till today. This is due to large sequence divergence in the small TRD which makes comparative studies difficult. To understand the amino acid sequence determinants in the interactions of M. Hpy C5mC with DNA, we performed the sequence alignments of M. Hpy C5mC with other closely related MTases that recognise the same consensus DNA sequence. The sequences related to putative C5mC enzymes were obtained from the REBASE.

The interactions that contribute to the specificity of DNA MTases include substrate recognition/consensus sequence of DNA. The sequence of M. Hpy C5mC small domain is divided into two regions as shown in Figure 3.7; a scaffold region and a specific DNA recognition region. Scaffolds are structurally conserved regions (Val230-Ala242 (I) and Cys247-His250 (II)) that support the residues which are involved in the binding of DNA backbone, while the later region constitutes residues which actually interact with the specific consensus bases Ser221-Arg229 (I) and S243-Q246 (II)), these are therefore considered as specific base recognition regions as shown in Figure 3.7.

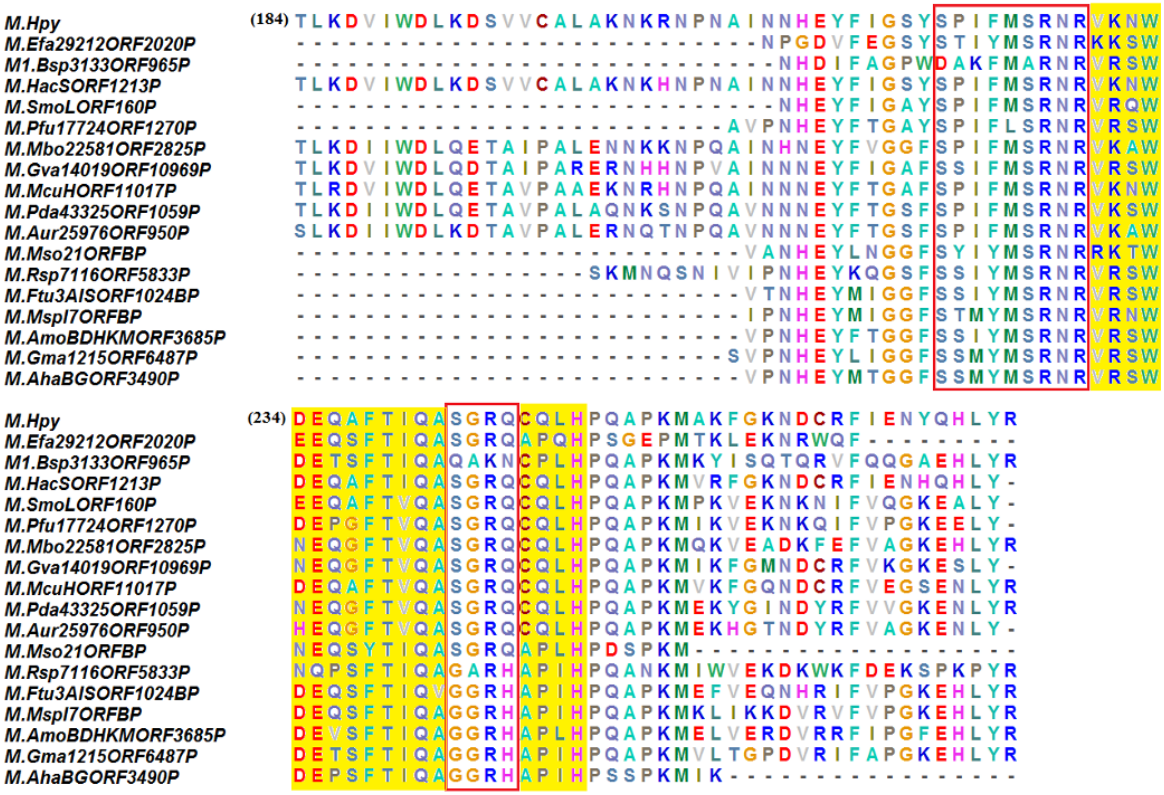


Figure 3.7: Sequence alignment of small domain region in M. Hpy C5mC with other putative C5mC DNA MTase sequences.

A conserved dipeptide "ThrLeu" in the scaffold region I, has been reported to constrain and position the flipped cytosine in the catalytic pocket. In the modeled M. Hpy C5mC, residues Thr239 and Ile240 are at topologically equivalent positions. Further to this, the conservation of residues surrounding the ThrLeu dipeptide that identified same DNA sequences, indicated the probability of using this region as a predictor of DNA sequence specificity (Neely and Roberts, 2008; Vilkaitis et al., 2000). The residues located towards the carboxyl terminus of Thr239 interact with the sugar phosphate backbone of the three bases of DNA consensus sequence (5'GC^{m}C3') while residues located towards amino terminus interact with the rest of the DNA. All the contacts between the protein and DNA are calculated using MolBridge (Kumar et al., 2014) and listed in Table 3.1.

Table 3.1: Mean (standard deviation) values for different interactions between atoms of amino acids and nucleotides/ water molecules present in the active site. An interaction with at least 50% occurrence during the simulation has been considered. The donor atoms are shown as (Residue Number: Residue Name: Chain ID: Donor Atom: Hydrogen Atom), while the acceptor atoms are presented as (Residue Number: Residue Name: Chain ID: Acceptor Atom)
*H: Hydrogen; D: Donor atom; A: Acceptor atom

SL. No.	Donor Atoms	Acceptor Atoms	HA* (Å)	DA* (Å)	DHA* (°)
1	89:ARG:C:NH1:2HH1	12:DA:A:O1P	1.99 (0.27)	2.91 (0.19)	153.30 (15.34)
2	157:ARG:C:NE:HE	10:DC:A:O2	2.1 (0.24)	2.97 (0.18)	147.92 (12.96)
3	157:ARG:C:NH1:2HH1	10:DC:A:O2P	2.05 (0.24)	2.96 (0.18)	152.91 (14.22)
4	155:ARG:C:NE:HE	333:SOL:X:OW	2.04 (0.24)	2.96 (0.18)	152.95 (14.15)
5	10:DC:A:N4:H41	71:PRO:C:O	2.09 (0.27)	3.01 (0.23)	154.61 (14.01)
6	9:DG:A:N2:H22	78:GLU:C:O	2.07 (0.26)	3.01 (0.23)	155.96 (13.66)
7	155:ARG:C:NH2:1HH2	10:DC:A:O2	2.09 (0.27)	2.99 (0.23)	152.85 (16.44)
8	242:ALA:C:N:H	10:DC:A:O3'	2.12 (0.28)	3.03 (0.25)	153.05 (15.96)
9	157:ARG:C:NH2:2HH2	10:DC:A:O5'	2.11 (0.28)	3.02 (0.24)	153.11 (15.55)
10	242:ALA:C:N:H	11:DC:A:O2P	2.08 (0.26)	3.00 (0.23)	154.39 (15.04)
11	78:GLU:C:N:H	10:DC:A:O1P	2.09 (0.27)	3.01 (0.23)	153.61 (15.55)
12	239:THR:C:N:H	9:DG:A:O1P	2.07 (0.27)	3.00 (0.23)	155.05 (15.10)
13	157:ARG:C:NH2:1HH2	10:DC:A:O2	2.07 (0.26)	2.99 (0.23)	154.82 (14.85)
14	155:ARG:C:NH2:1HH2	333:SOL:X:OW	2.08 (0.27)	3.00 (0.23)	154.21 (15.15)
15	157:ARG:C:NE:HE	10:DC:A:N3	2.09 (0.28)	2.99 (0.23)	152.41 (17.08)
16	229:ARG:C:NH2:2HH2	9:DG:A:O6	2.09 (0.28)	2.99 (0.23)	152.49 (16.99)
17	120:ARG:C:NH2:2HH2	13:DT:B:O1P	2.10 (0.29)	3.00 (0.23)	152.15 (17.39)
18	74:GLN:C:NE2:1HE2	11:DC:A:O1P	2.09 (0.29)	2.99 (0.23)	152.91 (17.19)
19	111:GLU:C:N:H	331:SOL:X:OW	2.09 (0.29)	2.99 (0.22)	153.30 (17.02)
20	10:DC:A:N4:H42	111:GLU:C:OE1	2.08 (0.28)	2.99 (0.22)	153.85 (16.66)
21	157:ARG:C:NH2:1HH2	10:DC:A:O4'	2.07 (0.28)	2.97 (0.23)	152.41 (17.08)
22	77:SER:C:OG:HG	DC10:A:O1P	2.11 (0.28)	3.00 (0.23)	154.21 (15.15)
23	245:ARG:C:NH1:1HH1	DG9:B:O6	2.07 (0.27)	3.00 (0.23)	155.05 (15.10)
24	246:GLN:C:NE2:1HE2	DG10:B:O6	2.07 (0.28)	2.99 (0.23)	152.41 (17.08)
25	221:SER:C:OG:HG	DG10:B:O6	2.11 (0.28)	2.96 (0.18)	154.21 (15.15)
26	204:LYS:C:NZ:HZ3	DG6:B:O1P	2.12 (0.28)	2.99 (0.23)	152.91 (17.19)
27	231:LYS:C:NZ:HZ2	DG8:A:O2P	2.11 (0.26)	2.96 (0.16)	153.39 (15.04)
28	298:ASN:C:ND2:1HD2	DG6:B:O2P	2.08 (0.24)	2.99 (0.20)	152.41 (17.05)
29	227:ARG:C:NE:HE	DA7:A:O2P	2.08 (0.24)	2.71 (0.21)	153.44 (16.05)

From the table we observed that base specific interactions of TRD region are more prominent within four consensus recognition nucleotide bases of DNA. The water mediated interactions between the protein and DNA are found in several DNA binding proteins as a common mechanism in the specific recognition of DNA (Rhodes et al., 1996). These water molecules also play a key role in stabilizing protein-DNA interactions by reducing the electrostatic repulsion. In this work also, we found that many water molecules act as a bridge in the extension of the protein-DNA contacts (data not shown).

As indicated in the Table 3.1, the carbonyl group of Glu78 and its preceding residue Ser77 interact with the backbone of the consensus bases in the minor groove of DNA. The regions, Pro71-Lys83 from large domain, Ser221-Arg229 and Ser243-Gln246 from the small domain make base specific interactions with the DNA. Apart from these consensus base specific interactions, contacts which cover the whole protein play an important role in the proper positioning of DNA in the protein cleft. The DNA backbone interactions are supposed to play an important role in the base flipping mechanism (O'Gara et al., 1998). In M. Hpy C5mC, the residues from both domains that include Gln74, Ser77, Arg89, Arg120, Arg157 (large domain), Lys204, Lys231, Thr239 and Ala242 (small domain) and Asn298 (hinge region) interact with the backbone phosphate groups of DNA.

The rearrangement in the base pairing of DNA recognition sequence has been observed when the substrate cytosine is adjacent to another cytosine (Reinisch et al., 1995). In M. Hpy C5mC also, since the recognition sequence comprises of a second cytosine adjacent to the flipped one (5'-GGC^{m}C-3'), we observed similar phenomenon of base rearrangement based on the results from the MD simulations. The regular guanine partner of the flipped cytosine in the protein-AdoMet-DNA complex (termolecular system) does not remain unpaired, but translates along the DNA axis so that it can form hydrogen bond with cytosine in the 3' site. The new base pair digresses significantly from coplanarity but retains Watson-Crick geometry. As a consequence of the altered pairing, the outer guanine in the 5'of the complementary strand is left unpaired. This unpaired guanine is stabilized by hydrogen bonding interactions with Arg245. The shift in base pairing opens a cleft in the DNA into which the side chains of Pro222 and Ile223 from the small domain protrude as shown in Figure 3.8A.

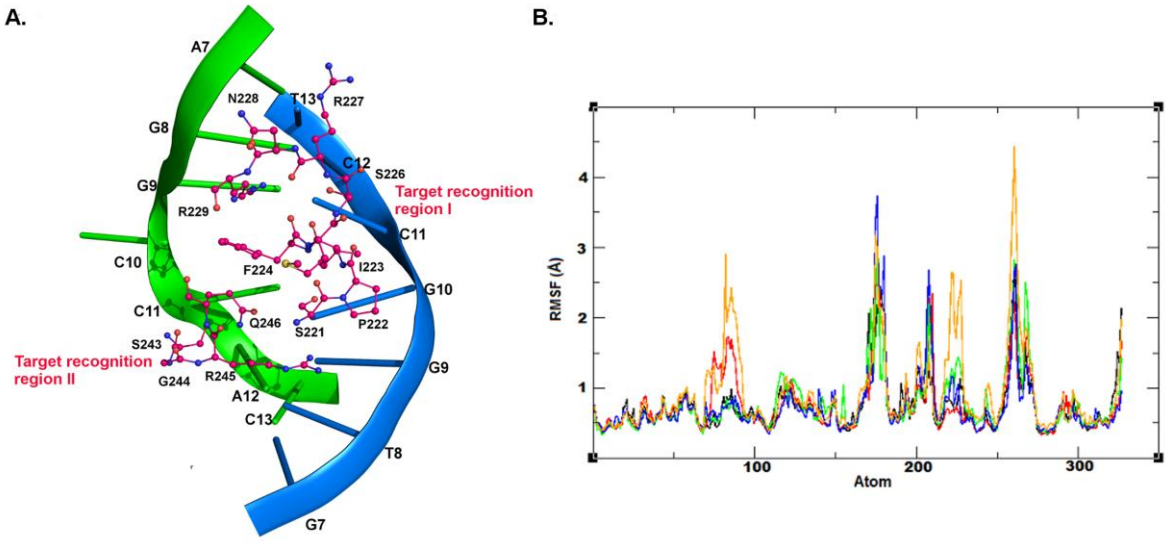


Figure 3.8: (A) Residues in red box are involved in consensus base DNA recognition; residues in yellow background represent the scaffold region. (B) Important residues that interact with consensus bases of DNA.

3.3.3 Mutational analysis

To see the effect of proposed structural motifs 75-83 (large domain), Ser221-Arg229 and S243-Q246 (small domain) on the specificity in target DNA recognition, we performed mutational analysis by alanine scanning of the above mentioned regions in the termolecular system of M. Hpy C5mC. As expected, highest fluctuations were observed in the M. Hpy C5mC termolecular complex that accommodated mutations in all the three regions (Figure 3.8B). From the RMSF plots, the fluctuations in the regions 165-183 and 257-273 for all the mutations is considerably high which is consistent with the native RMSF plots. To determine whether the mutations had any effect on DNA binding ability, we analysed their hydrogen bonding pattern. Mutation of 75-83 region leads to loss in helix-turn structure and also loss of important hydrogen bonds between Ser77 and Glu78 with DNA. Mutation in the proposed TRD region (221-229) which recognizes the target DNA leads to increased fluctuations in that region along with the loss of important hydrogen bonds which were required to hold the DNA backbone for example Ser221, Arg227, Arg229, Arg245 and Gln246. Alongside the protein, free movement of DNA within the binding cleft was observed as shown in Figure 3.9.

On comparison of all the RMSF plots of mutated regions of Ser75-Lys83 and Ser221-Arg229 we observed more fluctuations in their respective regions as compared to native whereas Ser243-Gln246 motif mutation does not cause any drastic change in the RMSF of Cα atoms. However, fluctuations in DNA were much higher for both 221-229 and 243-246 regions compared to the 71-83 mutation. These results clearly

indicated that the 221-229 and 243-246 regions as proposed by us form the target recognition regions and are responsible for binding the target DNA with high specificity in our model.

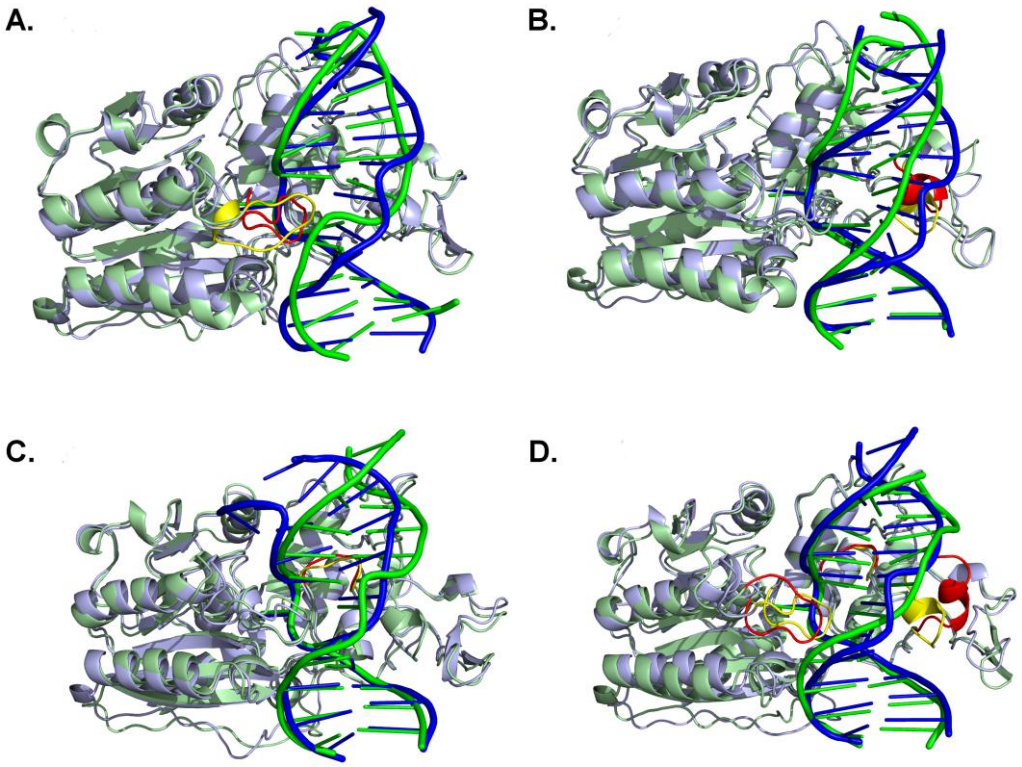


Figure 3.9: Superimposition of native (protein shown in light blue and DNA is represented in dark blue) and mutant protein(protein shown in light green and DNA is represented in dark green) structure having mutation in the region (A) 75-83 (B) 221-229 (C) 243-246 and (D) structure accommodating all the mutations. Specific mutated region are represented in yellow (native) and red (mutant) colours in all the superimposed structure.

Based on the above detailed studies, we identified conserved residues specifically involved in making base specific interactions that contribute to the specificity of substrate DNA and propose the mechanism of C5 methylation in the M. Hpy C5mC model in complex with AdoMet and DNA.

3.4 Conclusions

The *H. pylori* is a primitive human pathogen that has evolved to persistently colonize the gastric microbiota and cause diverse consequences ranging from asymptomatic gastritis to invasive adenocarcinoma. We report here the sequence as well as 3-D structure modeling and analysis of *H. pylori* 98-10 C5mC DNA MTase (M. Hpy C5mC) from cancerous strain. This enzyme catalyzes the transfer of methyl group from cofactor AdoMet to C5 position of flipped cytosine from consensus DNA sequence 5'GGC^{m}C3' as predicted from REBASE. To understand the importance of catalytic site residues and mechanism of methylation, we docked the cofactor (AdoMet) as well as the flipped DNA in to C5mC DNA MTase. In total, three molecular systems of the enzyme were generated *i.e.* protein alone, protein + AdoMet, and protein + AdoMet + DNA. The conservation of ten sequence motifs as well as the 3-D structure is in accordance to the other C5mC DNA MTases. The MD simulation has been carried out for all three molecular systems in order to have insights into the structural changes in the protein on AdoMet and DNA binding, and to understand the mechanism of methyl group transfer to DNA. To the best of our knowledge, this is the first MD simulation performed in all three systems of M. Hpy C5mC. Stable interactions of Glu111, Arg155, Arg157 and water molecules with cytosine revealed their role in the DNA methylation. In M. Hpy C5mC, the interactions of cytosine O2 with Arg155, Arg157 and water mediated interactions with N3 are responsible for stabilizing the carbanion intermediate. Further, involvement of water molecules in the deprotonation at C5 position was also observed in our model.

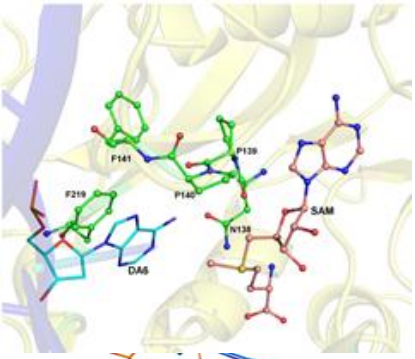
We observed that the specificity between M. Hpy C5mC and DNA is guided by sequence motifs Ser75-Lys83 (large domain), Ser221-Arg229 and Ser243-Gln246 (small domain) and the involvement of ThrIle dipeptide in the correct positioning of the flipped cytosine. These regions in the small domain are conserved in C5mC DNA MTases from other organisms that recognise 5'GGC^{m}C3'. Alanine scanning mutagenesis confirmed the role of proposed sequence motifs in specific recognition of the target DNA. Since the C5mC DNA MTases are present in both eukaryotes and prokaryotes, the identification of regions of dissimilarity in TRD could possibly contribute in designing of novel direct inhibitors of this enzyme. This homology suggests that HP9810_873g29 utilizes the same reaction mechanism for methyl-group

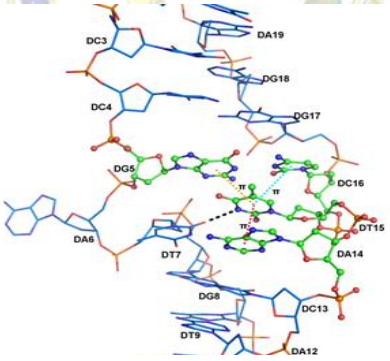
transfer as described for M.HhaI. The observed AdoHcy conformation as well as the DNA-binding model presented in this study, further supports this hypothesis.

4

N6-adenosine DNA MTase from *H. pylori* 98-10 in complex with DNA and AdoMet: Structural insights from MD simulation

- ✓ Sequence analyses and 3-D structure modeling of γ class N6mA DNA MTase from *H. pylori* 98-10.
- ✓ MD simulations of protein in complex with AdoMet and DNA.
- ✓ Important residues involved in catalytic activity and specific recognition of DNA.





Structure and dynamics of N6 adenine specific DNA methyltransferases from H. pylori 98-10: in complex with S-adenosyl-L-methionine and DNA. Singh S, Guruprasad L. (*Manuscript communicated*).

4.1 Introduction

The *H. pylori* is a primitive, microaerophilic bacterium which inhabits the acidic environment of human gastrointestinal tract. These Gram-negative bacteria belong to ε-proteobacteria phylum, has implications in human health affecting around one-half of the global human population and are reported to cause diseases like gastric ulcer, gastritis, duodenal ulcer, gastric cancer and MALT lymphoma (Graham, 1991; Peterson, 1991).

H. pylori is genetically diverse due to the existence of natural competence, high mutation and recombination frequencies (Israel et al., 2001) making it suitable for survival alongside human microbiota. A majority of *H. pylori* infected individuals remain asymptomatic and less than 20% of infected patients develop peptic ulcer, gastric cancer or malignant lymphoma, proving potential host defense mechanisms against *H. pylori* infection. In the year 1994, International Agency for Research on Cancer, a subordinate organization of the World Health Organization, classified *H. pylori* as a class I carcinogen, since several strains of *H. pylori* are linked with gastric malignancies (Hansson et al., 1996; Parsonnet et al., 1991) and gastric lymphoma (Nakamura et al., 1997). For example, *H. pylori* 98-10 was isolated from a patient with gastric adenocarcinoma and is responsible for causing gastric cancer (Ando et al., 2002). Multiple factors have been suggested to contribute to the etiology of gastric cancer in different stages and a mechanism for carcinogenesis caused by *H. pylori* has been proposed (Correa, 1992).

Abundant genes coding R-M systems are present in the *H. pylori* genome (Skoglund et al., 2007) for defense from the transformation of foreign bacterial DNA or transduction from phages (Takahashi et al., 2002). The R-M systems pose a key obstacle to DNA transformation and genetic engineering of bacterial species and are classified on the basis of their subunit composition as well as cofactor requirements (Bujnicki et al., 2001). The most studied type II R-M systems comprise of two enzymes, a monomeric MTase (MTase) and a dimeric restriction endonuclease, both these components usually recognize a specific palindromic DNA sequence. The MTases are involved in transferring a methyl group from cofactor SAM or AdoMet to the specific base (cytosine or adenine) of the host target recognition DNA sequence

forming a methylated base and SAH or AdoHcy, while the endonuclease cuts the foreign DNA at the same specific DNA sites which lack the methylation.

DNA methylation is pervasive and plays crucial part in several cellular mechanisms which includes gene regulation, genetic imprinting, immunity and cancer (Jeltsch, 2002; Paulsen and Ferguson-Smith, 2001). DNA methylation in bacteria and archaea kingdoms acts as a defense 'immune response' to destroy foreign DNA (Bickle and Kruger, 1993; Cheng, 1995; Wilson, 1991) and protect the host from foreign DNA through the action of R-M systems (Paulsen and Ferguson-Smith, 2001). Based on the position of methyl group transfer on bases in DNA, MTases are classified into two classes: exocyclic amino (N4mC and N6mA) and endocyclic MTases (C5mC).

Endocyclic and exocyclic amino DNA MTases have ten and nine conserved motifs, respectively, since the homologue of motif IX in C5mC MTases could not be identified in amino MTases. Further, on the basis of linear arrangements of the AdoMet-binding domain, the catalytic domain and the TRD, the exocyclic amino MTases are further subdivided into- α, β, γ, ζ, δ and ε subclasses (Malone et al., 1995). The majority of exocyclic amino MTases fall into, the α (MTases such as DpnII from *Streptococcus pneumoniae*), β (MTases such as *Proteus vulgaris*) and γ (MTases such as *Thermus aquaticus*) subclasses. Subclass α contains N6mA MTases, subclass β contains both N4mC MTases and N6mA MTases, and subclass γ consists of N6mA MTases.

The spatial positioning of functional regions required for AdoMet binding and catalysis of methyl transfer are maintained in the large domain of all MTases (Gong et al., 1997a; Schluckebier et al., 1997). The catalytic domains of both endocyclic and exocyclic classes revealed analogous 3-D folding, and out of all conserved motifs, the catalytic motif IV, which comprises a conserved ProCys dipeptide for the endocyclic (Wu and Santi, 1987) and a conserved (Asp/Asn/Ser)ProPro(Tyr/Phe) tetrapeptide for the exocyclic amino DNA MTases (Malone et al., 1995), is related in catalysis of methyl transfer (Schubert et al., 2003). Structural studies of different MTases elucidate that despite the differences in sequential motif orders, large domain shares a conserved catalytic core structure, while the TRD, which is sequence-specific DNA binding domain, differs among the MTase classes in sequence as well as in its spatial position (Malone et al., 1995). As compared to C5mC, the knowledge about the methylation

mechanism of N6mA DNA MTase is still limited. So far, no structural studies have been performed on the exocyclic amino DNA MTases in *H. pylori*.

In this work we focus on the structure and function relationship of an enzyme that catalyzes the transfer of a methyl group from AdoMet to the N6 position of adenine within the recognition sequence 5'TCCGA^{m}3', from a cancer causing strain *H. pylori 98-10*.

To achieve this, we have modeled the 3-D structure of N6-adenine specific DNA MTase from *H. pylori* 98-10 (M. Hpy N6mA) to provide insights of the protein active sites and confirmed that it belongs to γ subclass exocyclic N6mA. Further, we have docked the cofactor AdoMet as well as substrate DNA in the protein model. We have also performed MD simulations of M. Hpy N6mA complexed with AdoMet and DNA to gain insights about the important residues responsible for specific recognition and to understand the mechanism of N6-adenine methylation in *H. pylori*.

4.2 Method

The template structure for M. Hpy N6mA (NCBI_ID: EEC22779.1) was identified based on the homology searches, BLAST (Altschul et al., 1990) and the fold prediction searches, FUGUE (Shi et al., 2001). We have used the sequence alignments generated by FUGUE as a guide to build the 3-D model structure of the M. Hpy N6mA using HOMOLOGY module in Discovery Studio 2.5 (Accelerys Inc, USA) that implements the MODELLER method (Sali and Blundell, 1993; Shen and Sali, 2006). MODELLER is a homology or comparative modeling program for constructing the 3-D model of a protein structure from its amino acid sequence. The program constructs a model for all non-hydrogen atoms by the satisfaction of spatial restraints that includes non-homologous loops and energy optimization of the final model. The stereochemical quality of the best model was validated using PROCHECK and Verify_3D was used to validate the compatibility of the 3-D model structure with its primary amino acid sequence.

The bimolecular and termolecular systems were created by extracting AdoMet and AdoMet+DNA, respectively from a homologous template structure PDB_ID: 1G38. The restriction enzyme database, REBASE has a reliable tool for predicting the likely consensus DNA recognition sequences that bind to a given DNA MTase (Roberts et al., 2015). In the termolecular system, the nucleotide bases DNA were mutated according to the prediction made by REBASE. To confirm the role of regions responsible for DNA binding predicted from this work, we performed alanine scanning mutagenesis experiments on the proposed sequence motifs. We have used 'Build and edit protein' module present in Discovery Studio 2.5 to mutate the desired residues in M. Hpy N6mA.

4.2.1 Molecular dynamic simulations

Three molecular systems (1-3), unimolecular system, the modeled structure of M. Hpy N6mA (1), bimolecular system, the modeled structure of M. Hpy N6mA bound to AdoMet (2) and termolecular system, the modeled structure of M. Hpy N6mA bound to AdoMet and DNA (3) were subjected to 50ns MD simulations using GROMACS 4.5.5 package (Hess, 2009; Van der Spoel et al., 2005).

For all the molecular systems, protein and DNA force fields were generated using the AMBER ff99SB (Hornak et al., 2006). The force fields for AdoMet were

generated in Antechamber (Wang et al., 2006b) using ACPYPE script (Sousa da Silva and Vranken, 2012).

The three molecular systems were immersed in an octahedron box of extended simple point charge (SPC) water molecules. The termolecular complex was neutralized by adding 14 Na^+, the unimolecular and bimolecular complexes were neutralised by adding 14 Cl^- each. To relieve the short range bad contacts, energy minimization was performed using the steepest descent method for 5000 steps followed by the conjugate gradient method for 5000 steps. The MD simulation studies consist of equilibration and production phases. The position restrained simulations were carried out at 298 K for 1ns. Finally, the three systems were subjected to 50 ns MD simulations production run at 298 K temperature and 1 bar pressure, using 0.002 ps time step. The Parrinello–Rahman method was used to control pressure (Parrinello and Rahman, 1981) and the V-rescale thermostat was used to maintain temperature (Bussi et al., 2007). The long range electrostatics were handled using the PME (Darden et al., 1993) method with a real space cut-off of 10Å, PME order of 6 and a relative tolerance between long and short range energies of 10^{-6}. Short range interactions were evaluated using a neighbour list of 10Å updated every 10 steps while LJ interactions and the real space electrostatic interactions were truncated at 9Å. Hydrogen bonds were constrained using LINCS algorithm (Hess et al., 1997). The final models in all three systems were obtained by averaging the snapshots from the trajectory generated by MD simulations after the structure stabilization was achieved. To establish the role of various loops in N6mA for DNA binding, six mutated termolecular systems (143-155, 183-189, 212-220, 280-293, 308-325 and one system accommodating all the mutations) were generated after alanine scanning mutagenesis. This implies that in the specified loop all amino acids were mutated to alanine. These mutated termolecular systems were also subjected to MD simulations as described above.

To study the conformational variations in the structures of M. Hpy N6mA, the RMSD of the atomic positions with respect to their starting structures were calculated by using g_rms of GROMACS by least-square fitting the structure to the reference structure. The convergence of MD simulations was analysed in terms of the potential energy, RMSD and RMSF. The deviations relative to the Cα backbone atom were calculated, with the aim to evaluate the stabilization of the systems throughout the MD simulations. The kinetic, potential and total energies estimate the variations in the energetics of the system and were also analysed as a course of MD simulation time.

4.3 Results

4.3.1 Sequence analysis

The conserved motifs along the primary sequence in DNA MTases are the likely sites responsible for the chemistry common to all MTases, while the variable regions are responsible for the sequence specificity of substrate DNA (Ahmad and Rao, 1996a; Madhusoodanan and Rao, 2010). The majority of these motifs are responsible for three basic functions; the motifs I to IV have been implicated in AdoMet binding, motifs IV and VI play a role in the catalysis of methylation and motifs IV, VI and VIII aid in the correct positioning of the target nucleotide within the active site of the enzyme (Malone et al., 1995).

As discussed above, various groups of exocyclic amino MTases differ with respect to the location of insertion of the small domain into the large domain as well as by a circular permutation of the amino acid sequence of the large domain (Jeltsch, 1999; Malone et al., 1995; Wilson, 1992). The exocyclic amino MTases belonging to the subclass γ have a motif order which include N- terminus - AdoMet binding region - catalytic region-TRD-C-terminus. Sequence alignment of M. Hpy N6mA with the template revealed the presence of nine conserved motifs indicating that it belonged to the γ subclass of MTases.

Sequence alignment between M. Hpy N6mA and the template shown in the Figure 4.1 indicated a 19 amino acid insertion in the template sequence between motifs IV and V with respect to the target sequence. Further analysis revealed that this region is missing in all Helicobacter genus. Motif arrangement in the sequence alignment is shown in Figure 4.1. A brief illustration of motifs present in M. Hpy N6mA is discussed as follows.

Motif I (Ile72-Leu86) forms a part of the AdoMet-binding site and is conserved among all AdoMet dependent enzymes including all DNA, RNA and several protein MTases (Ingrosso et al., 1989; Kagan and Clarke, 1994; Smith et al., 1990; Wilson and Murray, 1991). Structurally this motif forms β-loop-α in all DNA MTases and side chains of hydrophobic residues present in β-strand are required for packing against the αA helix. In N6mA this region is formed by secondary structure β1-loop-αB and is involved in accommodating the methionine moiety of AdoMet in M. Hpy N6mA. This loop is called as ‘G-loop’ since it mainly consists of Gly and less frequently Ala or Pro that binds to the methionine moiety. In M. Hpy N6mA, the G-loop; Gly78-Asn79-

Gly80 is present, and the last residues of β1-strand is Cys75, while majority of amino MTases consists of a Pro residue at this position.

Motif II (Asn94-Asn101) forms secondary structure β2, which consists of a negatively charged amino acid (Glu99) at the end of the strand. This negatively charged residue is involved in interacting with hydroxyl group of ribose moiety in AdoMet and Ile100 makes vdW contacts with adenine moiety of AdoMet. Motif III (Leu118-Phe126) consists of β3 and a turn that connects to αD. Asp122 present in the turn is involved in direct interaction with exocyclic NH_2 group of AdoMet adenine moiety (N6) and the main chain NH of Phe123 forms hydrogen bond with N1 of adenine moiety of AdoMet in both bimolecular and termolecular systems.

In M. Hpy N6mA, motif IV (Tyr132-Arg149) forms β4 and loop that connects to αE. This consensus sequence Asn138-Pro139-Pro140-Phe141, has a diprolyl moiety and is hence called as 'P-loop'. The P-loop along with motifs VI and VIII forms the protein active site in amino MTases (Schluckebier et al., 1995b). In the motif V (Asn155-Lys169), region from Leu156-Leu167 forms αE that is involved in the stabilization of AdoMet. Leu156 makes vdW contacts with AdoMet adenine on the same side as Phe160, further it makes edge to face vdW contacts with adenine moiety of AdoMet in both bimolecular and termolecular systems as discussed below.

The motif VI (Asp170-Met183) forms secondary structural motifs β5 and αF, the conserved Gly171 present at the beginning of the strand and Pro179 at the end of the strand are characteristic of γ subclass of MTases. Motif VII (Leu191-Ala208) is not strongly conserved in MTases but certain degree of conservation in residues can be seen in each subclass. This motif includes some region of αG and β6 and the connecting loop formed by Gln199 and Phe200. These two residues face away from DNA binding cleft and is assumed to be involved in proper folding of the catalytic region (Cheng, 1995). Motif VIII (Lys209-Gly216) consists of loop connecting β6 and β7, Phe212 makes face to face vdW contacts with the flipped adenine moiety of substrate DNA in the initial modeled structure.

In N6mA, motif X (Ile46-Ile60) forms αA with small extension on either side of the helix. This motif has conserved hydrophobic residues which are required for packing against β1 and a loop preceding the helix (Malone et al., 1995). Location of this motif in the primary sequence is responsible for major difference between endocyclic and exocyclic amino DNA MTases. Both G loop of motif I and P loop of

motif IV with some region of motif X forms the binding pocket to accommodate the methionine moiety of AdoMet in M. Hpy N6mA.

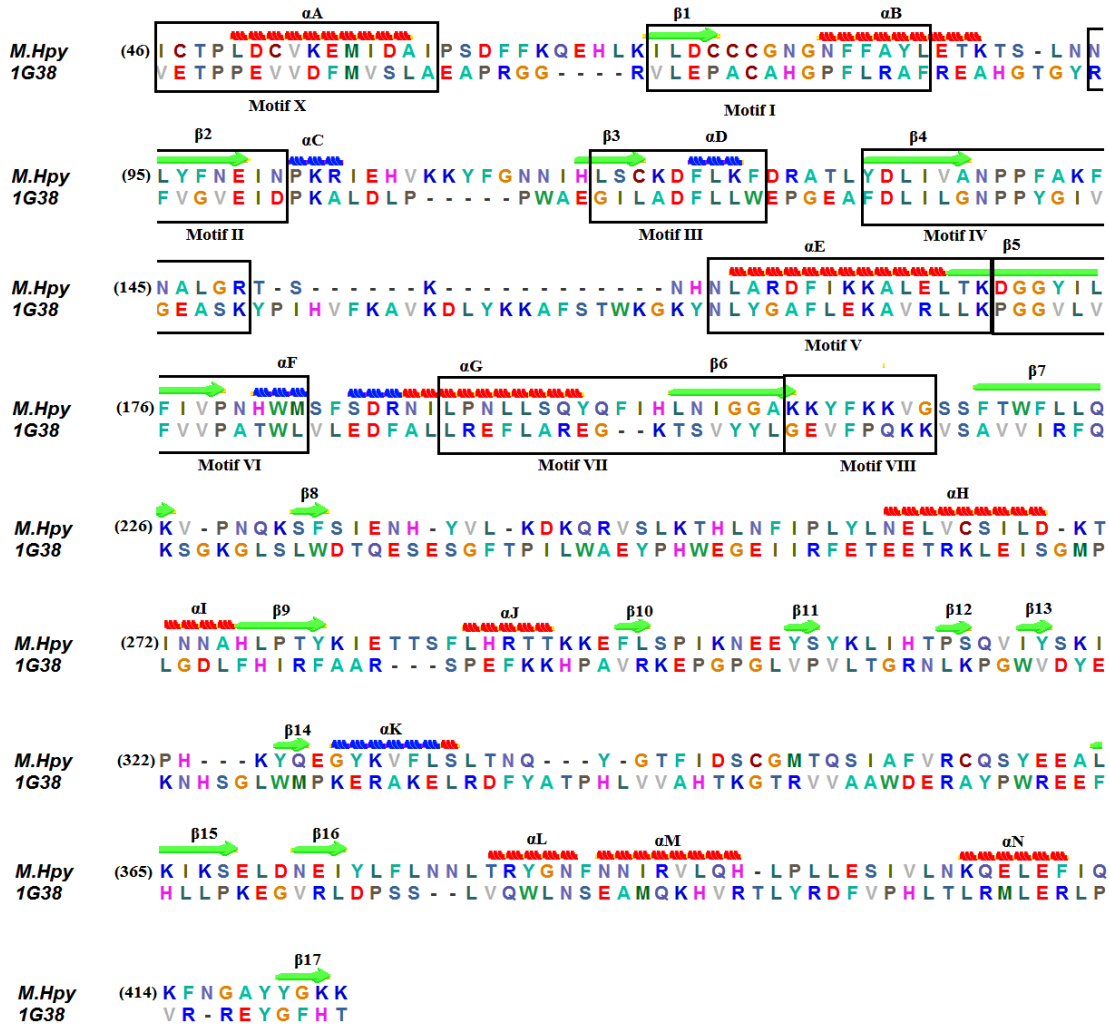


Figure 4.1: Sequence alignment of M. Hpy N6mA with its structural template M. TaqI (PDB_ID: 1G38). Conserved motifs characteristic of N6mA DNA MTases are represented in black boxes. (- α-helix, - 3_{10} helix, - β-strand).

4.3.2 Structure

In *H. pylori* 98-10 strain, HP9810_2g36 (EEC22779.1) an uncharacterized protein is N6mA DNA MTase confirmed by sequence and structural modeling studies. HP9810_2g36 belongs to γ subclass of exocyclic amino DNA MTases. Detailed analyses using homology and fold prediction (43% sequence homology) searches revealed that this protein shares similarity with the crystal structure of DNA MTase from M. TaqI, PDB_ID: 1G38. The homology model of this protein was constructed based on the template structure, a DNA MTase from *Thermus aquaticus* (M. TaqI) using MODELLER. The sequence alignment used as a basis for homology modeling is shown in Figure 4.1. The homology model qualified the structure validity test performed by PROCHECK (86.2% residues were in the core and 9.8% residues were

in the additionally allowed region of the Ramachandran plot) and Verify_3D (82.4% of the residues had an averaged 3D-1D score ≥0.2). The homology model of M. Hpy N6mA and the template structures are highly similar as indicated by low RMSD (0.45Å) from their structure superimposition (Figure 4.2A). The 19 amino acid insertion region in the template forms an α-helix between the secondary structures β4 and αE and is located close to DNA. According to REBASE, M. Hpy N6mA catalyzes the transfer of methyl group from AdoMet to the exocyclic amino (N6) nitrogen of the adenine in its recognition sequence 5′-TCNGA^{m}-3′, where N could be any base (A, T, G and C), and in the present case, N is considered as cytosine, making recognition sequence 5′-DT2-DC3-DC4-DG5-DA^{m}6-3′. M. Hpy N6mA model structure consists of large catalytic domain (Ile46-Ile235) and small domain (Leu260-Lys423). Central core of large or N-terminal domain is formed by eight-stranded β-sheet surrounded by α-helices on both sides. The large domain consists of a seven-stranded β-strand (6↑7↓5↑4↑1↑2↑3↑) formed by six-stranded parallel β-sheet with the seventh strand inserted in an antiparallel fashion between the fifth and sixth strand (Figure 4.2B).

The overall fold of the enzyme is a bilobal structure; the large domain comprises the active site for AdoMet binding and methyl transfer, while the small domain recognises the substrate DNA. The protein is dominated by an open α/β-sheet structure with a prominent V-shaped cleft where AdoMet and catalytic amino acids are located at the bottom of this cleft as shown in Figure 4.2C. The size and basic nature of the cleft between the two domains are consistent with duplex DNA binding. From the structure visualization of the termolecular systems, we observed that DNA is bound in the cleft between the two domains flanked by several loops, and the major groove of DNA faces the small domain while the minor groove faces the large domain that comprises the catalytic site (Figure 4.2C). The common feature of all DNA MTases is the presence of a structural core known as AdoMet dependent MTase fold which constitutes large domain (Cheng and Roberts, 2001). On the other hand, the small domain of different MTases are dissimilar in amino acid sequence, size, structure and mostly possess loops involved in sequence specific DNA recognition and the intrusion of the DNA to flip the target base.

The small or C-terminal domain consists of 162 amino acid residues and mainly consists of 7 α-helices and 9 β-strands. The two domains are connected covalently by the hinge region (Glu236-Tyr259) between β8 in the N-terminal domain

and αH in the C-terminal domain. The small domain is mainly responsible for sequence specific contacts between the DNA and MTases which further help in the recognition of DNA sequences characteristic of each class of enzyme.

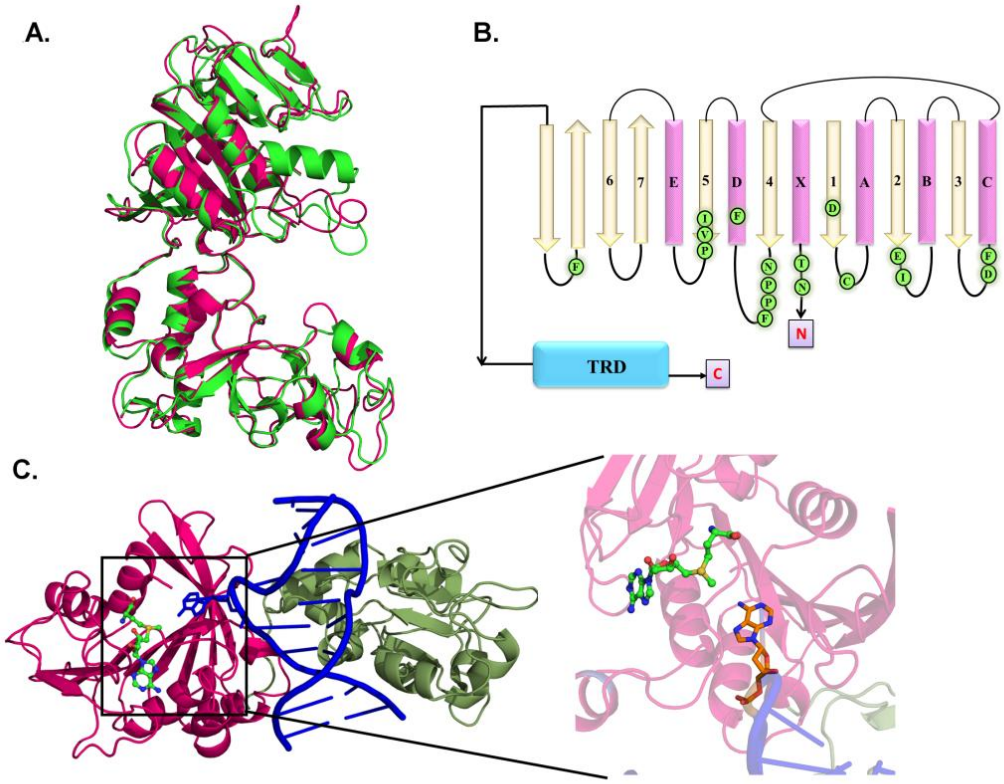


Figure 4.2: (A) Superimposition of template PDB_ID:1G38 (green) with the M. Hpy N6mA model (pink). (B) Secondary structural arrangement indicating the conserved residues in the large domain. (C) Schematic representation of termolecular systems indicating the large (pink) and small domains (green) in protein. Location of major and minor groove of DNA in the termolecular systems is also represented along with the location of AdoMet.

4.3.3 AdoMet binding

The orientation of AdoMet in bimolecular and termolecular systems is marginally different (Figure 4.3A and 4.3B). This is consistent with the previously reported results which confirms that AdoMet is catalytically competent for methylation process in bimolecular complex and does not require reorientation upon DNA binding in N6 MTases (Schluckebier et al., 1997). The AdoMet binding cleft is surrounded by six regions of the polypeptide chain belonging to the conserved motifs I- IV and X in both bimolecular and termolecular complexes. In M. Hpy N6mA, both AdoMet binding and catalytic site consists of hydrophobic pockets. The cofactor AdoMet is inserted into a cavity such that AdoMet is accommodated in an extended V-

shape conformation (Figure 4.2C) in the large domain formed by the conserved regions which are a part of classical Rossmann fold (β1-αB-β2) (Wierenga et al., 1986). This Rossmann fold requires the conserved Cys76 for the sharp bend connecting the first β-strand to the following α-helix.

In bimolecular and termolecular systems, the N6 of adenine atom in AdoMet is hydrogen bonded to OD1 and OD2 of Asp122 (motif III) with an average distance of 2.09Å and 2.04Å, respectively. While Phe123 (motif III) interacts with N1 atom of adenine with average distance of 2.50Å and 2.03Å, respectively. The binding of adenine ring of AdoMet is further supported by vdW contacts with Ile100 (motif II) and Phe160 (motif V). Ribose moiety in AdoMet is hydrogen bonded to residues of motif II. O2' of ribose forms hydrogen bond with side chain carboxylate of Glu99 OE2, with average distance of 2.02Å and 2.05Å, respectively. The interactions of ribose ring are further improved by vdW contacts of Arg104 and Cys76 which are present in close contact with C4'-C5' of ribose. The binding pocket for methionine moiety of AdoMet is formed by Cys47, Thr48, Cys77, Cys78, Asn79, Asn138 and Pro139. The carboxylate group (OXT) of AdoMet is hydrogen bonded to N atom of Thr48 with an average distance of bimolecular (1.99Å) and termolecular (2.07Å) systems. The ammonium group of AdoMet forms ionic interactions with main chain carbonyl O of both Cys77 and Gly78 (motif I). The active site residues Asn138-Phe141 (motif IV) are also involved in exocyclic amino DNA methylation process (Labahn et al., 1994). The CH_3 group from S^{δ}-CH_3 moiety, to be transferred from AdoMet to adenine in DNA substrate is surrounded by residues Asn138-Phe141 and is pointing away from NPPY segment which is in agreement with previous results (Schluckebier et al., 1995a).

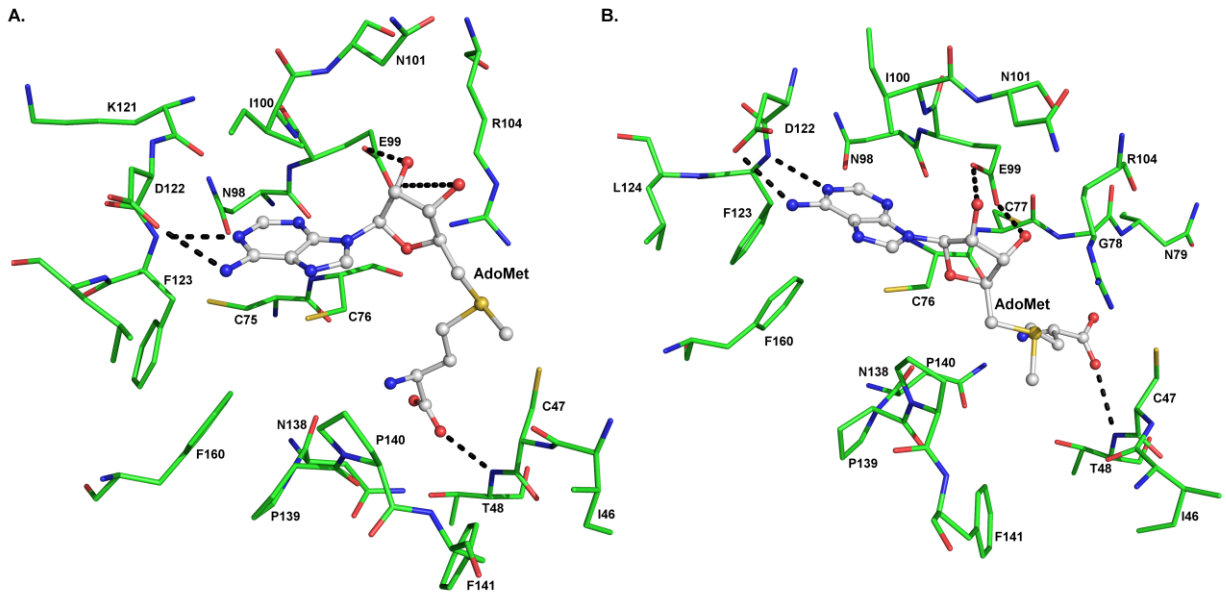


Figure 4.3: The AdoMet binding site in (A) bimolecular and (B) termolecular systems in M. Hpy N6mA.

4.3.4 Active site interactions

The substrate B-DNA is docked into the concave side of M. Hpy N6mA bound to AdoMet. Translocation of target base from double stranded DNA to the catalytic pocket of protein is a key step for methyl group transfer in all DNA MTases. The flipped DA6 base from substrate DNA is inserted into the active site located in the hydrophobic pocket in the large domain as shown in Figure 4.2C. This flipped DA6 is stabilized by means of hydrogen bond interactions with residues Asn138, Pro139 and Phe141. Residues of motifs IV, VI and VIII which are located in large domain form the binding pocket for DA6. The distance between AdoMet methyl group and N6 atom DA6 is 3.71Å in the termolecular systems and is maintained throughout the MD simulations. The corresponding distance in M. TaqI is 4.14Å (Goedecke et al., 2001).

The DNA methylation mechanism at the exocyclic amino groups involves direct transfer of methyl group from AdoMet to N6 position *via* SN^2 reaction by inversion of the methyl group configuration without formation of a covalent intermediate. A detailed mechanism of DNA methylation in the exocyclic DNA MTases has been proposed (Pogolotti et al., 1988). M. Hpy N6mA are characterized by a conserved Asn138-Phe141 (motif IV), which is located in the active site of the enzyme. Several groups have performed mutational studies on amino acids in motif IV which in turn have revealed the importance of this motif in catalysis (Ahmad and Rao, 1996b; Friedrich et al., 1998; Roth et al., 1998). In M. Hpy N6mA, the reaction mechanism in

the presence of Asn138 may operate as shown in the Figure 4.4. This NPPF motif is most significant among the conserved motifs with the consensus sequence (S/N/D)PP(Y/W/F) for exocyclic DNA MTases and the side chains of this tetrapeptide residues and the main chain carbonyl group of Pro139 serve as hydrogen bond acceptors for the protons of the exocyclic amino group (Gong et al., 1997b). Apart from this, Asn138 also helps in correct positioning of the target adenine, while the methylation would result from a direct attack of the AdoMet methyl group on the adenine N6 with a general base assisting the proton transfer that occurs at N6 (Alison McCurdy, 1992).

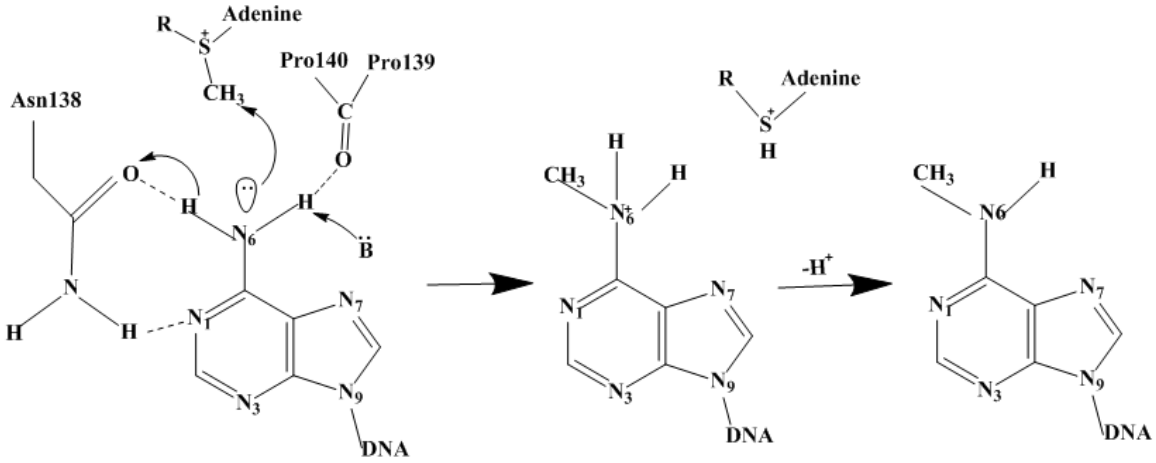


Figure 4.4: Mechanism of methyl group transfer from AdoMet to DNA in M. Hpy N6mA.

The hydrogen bond distance between Asn138 (OD1 atom) and Pro139 (main chain O) with N6 atom of DA6 is 2.08Å and 2.09Å, respectively. The N1 atom of flipped adenine also interacts with the ND2 atom of Asn138 with an average distance of 2.06Å. All these distances are maintained throughout the MD simulations. These hydrogen bonds appear to increase the partial negative charge of the N6 atom of adenine and activate it for direct nucleophilic attack on the methyl group of the cofactor (Goedecke et al., 2001). Further to this, these two residues are present below the plane of adenine ring and so could pull the hydrogen atoms at N6 position out of coplanar arrangement. The lone pair of N6 is no longer conjugated with aromatic ring and this process is further assisted in changing of hybridization of N6 from sp^2 to sp^3. This hybridization change also helps in placing the lone pair at N6 in ideal geometry to favour the attack of activated methyl group of AdoMet (Schluckebier et al., 1998).

A cationic transition state was proposed to be stabilized by the presence of aromatic residues, that make cation…π interactions with the flipped base (Holz et al.,

1999; Roth et al., 1998; Schluckebier et al., 1998; Wong and Reich, 2000) and predicted to be directly involved in the catalytic process (Malone et al., 1995). Furthermore, a previous mutational study (Pues et al., 1999) on Tyr108 and Phe196 of M. TaqI (equivalent to Phe141 and Phe212 in M. Hpy N6mA) revealed the importance of aromatic residues in proper orientation of the flipped adenine. Replacement of Tyr108 and Phe196 with equivalent aromatic residues retained or enhanced the enzymatic activity, while the corresponding mutations with Ala or Gly strongly reduced the enzymatic activity (Pues et al., 1999). From these experiments it is evident that aromatic amino acid in motif IV and VIII are required for enzymatic activity.

The M. Hpy N6mA initial model is consistent with the structural template, in which the flipped adenine interacts with the aromatic amino acids Phe141 (*via* face to face π-stacking interaction) and Phe212 (*via* edge to face π-stacking interaction). However, during MD simulation we observed that Phe212 is pushed back by the nearby bulky aromatic residues (Phe219, Try221 and Phe222). This position is now occupied by Phe219 that makes face to face π stacking with the flipped adenine and Phe141 makes edge to face π interactions with the flipped adenine. In the average structure of M. Hpy N6mA obtained after MD simulations, the flipped adenine in the active site is surrounded by residues of the active site forming several hydrogen bonds (Asn138, Pro139 and Pro140), π-stacking interactions (Phe141 and Phe219) as shown in Figure 4.5A and B. Further to this, the flipped base is sandwiched between Phe219 on one side (π…π interaction) and CH_3-S^+group of AdoMet on another side (cation…π interaction) and thus further stabilizing the flipped adenine in the hydrophobic active site.

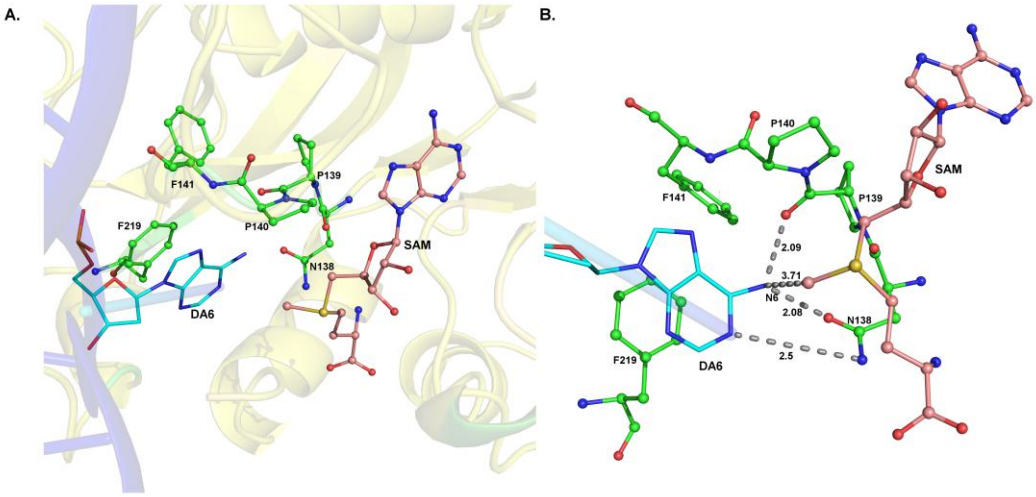


Figure 4.5: (A) The catalytic site in M. Hpy N6mA termolecular system (B) Hydrogen bonding interactions of important catalytic residues along with their distances are indicated in Å.

Trajectory analysis of the three molecular systems- protein, protein-AdoMet and protein-AdoMet-DNA showed important residues involved in stabilizing protein with AdoMet and DNA. Their respective Cα RMSD values are; 3.5Å, 3.0Å and 2.0Å (Figure 4.6A). From these results we conclude that protein interacting with both AdoMet and DNA has lowest RMSD as compared to the other two molecular systems. This is expected because, the surface exposed regions of the protein have large scale motions that leads to higher fluctuations in the protein and when bound to AdoMet alone. When it is complexed with DNA, these surface exposed regions interact with the DNA making it less solvent exposed and overall stabilising the structure. The potential energy plots of three molecular systems were obtained from the trajectory files of 50 ns MD simulations. The potential energies during the MD simulations were found to be -2.926 x 10^{-5} kJ mol^{-1} (unimolecular), -2.921 x 10^{-5} kJ mol^{-1} (bimolecular) and -2.932 x 10^{-5} kJ mol^{-1} (termolecular) systems, indicating their stability.

The RMSF plots of the Cα atoms of the proteins are shown in Figure 4.6B. Most of the fluctuating regions of protein include those parts which interact with the DNA or the loops which reside on the surface of the protein and the hinge region connecting the two domains. Apart from these differences, the rest of the trajectories followed the same pattern *i.e.* more fluctuations are observed in the unimolecular system (protein alone), less in the bimolecular system and least in the termolecular system. The regions 61-70, 100-125 and 315-339 are exposed to solvent, while 234-246 is a part of the hinge region between two the domains. The regions in close proximity to DNA are; 138-155 (loop region), 170-192 (helices along with turns), 205-

220 (strands along with small turn), 280-308 (helices), 308-337 (strands with loop) and 392-412 (part of helices connected with turns). Above mentioned regions show more fluctuations in protein, less in bimolecular while least in termolecular systems. The solvent exposed region 365-380, has almost same fluctuations in unimolecular and bimolecular systems but less in termolecular system.

4.3.5 Effect of DNA binding on the protein structure

DNA binding persuades the movement of the large and small domains closer to each other when compared to unimolecular and bimolecular systems. From trajectory analysis, we found that region (143-155) is well ordered and is predicted to play a vital role in sequence specific DNA recognition as well as in preventing the target adenine from flipping back into the DNA helix. The region (212-220) changes its conformation in the presence of DNA and participates in the formation of binding pocket for the flipped adenine. The region (308-325) is close to DNA and also solvent exposed. Apart from these, regions (280-293) and (183-189) also show greater movement in the presence of DNA. All these regions show higher fluctuation in their Cα atoms in protein only, less in bimolecular and least in termolecular systems of M. Hpy N6mA as depicted from the RMSF plots (Figure 4.6D). We propose that these structural motifs are involved in the recognition of substrate DNA.

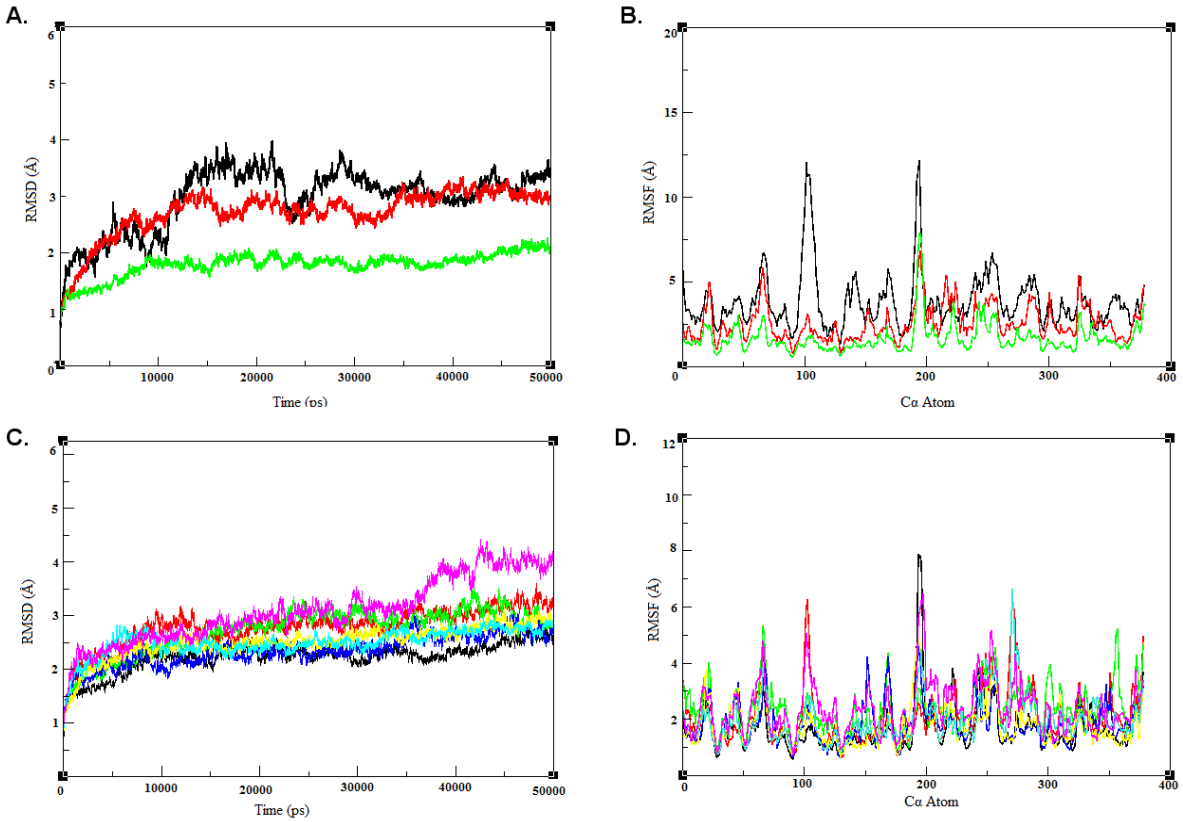


Figure 4.6: (A) RMSD of M. Hpy N6mA Cα atoms in unimolecular (black), bimolecular (red) and termolecular (green) systems. (B) RMSF of M. Hpy N6mA Cα atoms in unimolecular (black), bimolecular (red) and termolecular (green) systems. (C) RMSD of M. Hpy N6mA Cα atoms termolecular system (black) with the 143-155 (red), 183-189 (green), 212-220 (blue), 280-293 (yellow) 308-325 (cyan) and structure accommodating all mutations (magenta) (D) RMSF comparison of M. Hpy N6mA Cα atoms termolecular system (black) with the 143-155 (red), 183-189 (green), 212-220 (blue), 280-293 (yellow) 308-325 (cyan) and structure accommodating all mutations (magenta).

4.3.6 Alanine scanning mutation experiment

To further support the role of proposed structural motifs in DNA binding, we have analysed the results of mutant M. Hpy N6mA termolecular systems obtained from alanine scanning mutagenesis of the above mentioned regions (143-155, 183-189, 212-220, 280-293, 308-325 and one system accommodating all the mutations).

As anticipated, highest fluctuations and highest RMSD were observed in the mutant M. Hpy N6mA termolecular systems that accommodated mutations in all the five regions (Figure 4.6C and 4.6D). This system also showed large displacement of DNA in the cleft affecting the position of flipped adenine in the catalytic pocket (data not shown). From RMSD plots (Figure 4.6C), we observed that mutant termolecular system with 98-110 mutation has highest deviations that is followed by 280-293

mutation; while 212-220 and 308-325 mutations have deviations of their Cα atoms close to the native termolecular system.

From the RMSF plots of all mutant termolecular systems, it can be seen that most of the fluctuations in the regions is consistent with the native termolecular system and is considerably high compared to the native. To determine whether the mutations had any effect on DNA binding ability, we analysed their hydrogen bonding pattern.

The M. Hpy N6mA 143-155 mutation affects the DNA binding, as large fluctuation in the DNA backbone within the cleft is observed (Figure 4.7A) along with loss of β8-strand in the hinge region. This mutation also affects the hydrogen bonding between the flipped adenine and catalytic residues in the active site. Apart from the regular fluctuations seen in the native protein, more fluctuations are also seen in the region 143-155 which is expected since this region is mutated, along with the region 213-217 (turn region), 312-325 (solvent exposed region) and 394-402 (turn region). This implies that mutation in this loop region severely affects the interaction of protein with DNA as large fluctuations are also observed in other DNA binding regions.

The M. Hpy N6mA 183-189 mutant shows more fluctuations as compared to native termolecular systems with the exception in the regions 235-245 and 265-270 that show lower fluctuations. Trajectory analysis depicts that the mutation in this region affects the DNA binding as large movement of DNA backbone is seen in the cleft (Figure 4.7B). Also the changes in torsion angle of flipped adenine is observed due to which the flipped base shows a reversal tendency and reorients itself inside the DNA helix. Apart from this, loss of helicity in the mutated region and small domain were also observed. This mutation also induces the disruption of important interactions responsible for catalytic activity.

The M. Hpy N6mA 212-220 mutation leads to loss of important interactions required to hold the DNA at proper position in the C-shaped cleft and loss of β-strands in hinge region Figure 4.7C. RMSF plots also indicate higher fluctuations compared to the native termolecular system with the exception of the region 235-245.

The M. Hpy N6mA mutations in the regions 280-293 and 308-325 induced initial movement of DNA from its position but after 10ns it stabilized without affecting the DNA interaction with the catalytic residues (Figure 4.7D and 4.7E). However, the regions 3’ to recognition sequence (DT7 to DC10) and the complementary strand show higher displacement in 280-293 compared to 308-325. The 280-293 mutation induces gain of helicity in the mutated region while 308-325

mutation leads to loss of β-strands in the hinge region. Both the mutations also increased the disordered region in the small domain. RMSF plots indicated that the mutation in both regions leads to overall increase in the fluctuations compared to the native termolecular systems, but the mutated region does not show much changes in RMSF plots in 280-293 as compared to 308-325 mutated region.

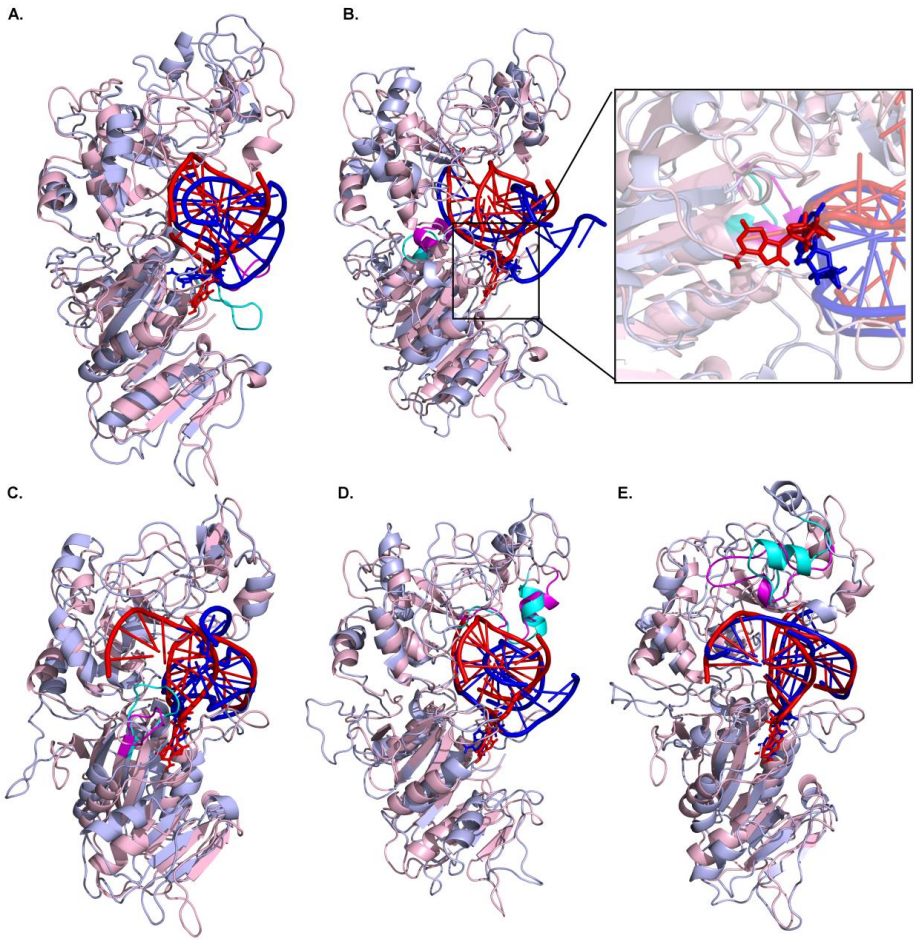


Figure 4.7: Superimposition of native (protein shown in light pink and DNA is represented in red) and mutant protein (protein shown in light blue and DNA is represented in dark blue) structure having mutation in the region (A) 143-155 (B) 183-189 (C) 212-220 (D) 280-293 and (E) 308-325. Specific mutated region are represented in magenta (native) and cyan (mutant) colours in all the superimposed structure. Flipped adenine base is shown in sticks with their respective colours in native and mutant.

In the termolecular systems of M. Hpy N6mA, we analyzed the hydrogen bond interactions held between base pairs in the DNA duplex. All the base pairs uphold an average number of canonical hydrogen bonds which are close to 1.7 or 2.5 for A-T or C-G base pairs, respectively and the values are consistent with those observed for simulations of B-DNA in aqueous solution (Perez et al., 2007), with the following exceptions.

In the DA6-DT15 base pair, the flipped base DA6 is in the protein environment, as a result the adjacent neighbour DT7 forms hydrogen bond with unpaired DT15 (2.03Å). DT15 is stabilized by means of interactions within the DNA appearing as the deformed DNA structure. Analysis of the evolution of the distance between the centroid of the unpaired base DT15 and the surrounding nucleic bases revealed that DT15 base is shifted towards the complementary strand forming a π-stacking interaction with nearby bases like DG5, DT14 and DA16, and these interactions are maintained throughout MD simulations (Figure 4.8A). DT15 also makes stable hydrogen bond with Lys324 (2.11Å) which is also maintained throughout the MD simulations. We also observed a conformational change in the backbone of DNA, reflected in the RMSD plot presented in Figure 4.8B.

The measurements of DNA fluctuation by means of RMSF showed that the fluctuations of DNA recognition sequence especially flipped adenine are restricted by residues from motif (IV, VI and VIII) which are present in the protein active site resulting in low fluctuations for this nucleotide (Figure 4.8C). These less positional fluctuations as compared to the rest of DNA, explain the stabilizing influence the enzyme has on the DNA. Also in comparison, the nucleotides facing the enzyme have much lower fluctuations as compared to the solvent exposed side.

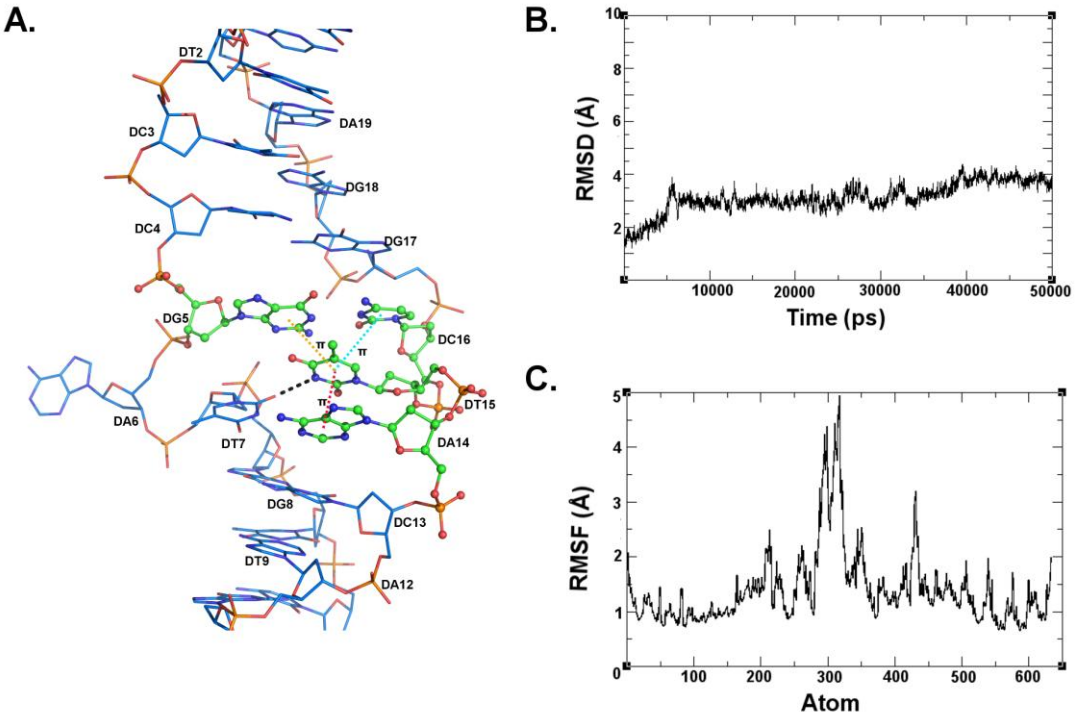


Figure 4.8: (A) Interactions between unpaired thymine base (DT15) with surrounding nucleicbases showing various base stacking conformation. (B). RMSD of the DNA (all atoms) in the termolecular system. (C). RMSF of the DNA (all atoms) in the termolecular system

4.3.7 Target recognition domain

In different classes of MTases, sequence specific DNA binding domain *i.e.* TRD differs with respect to sequence and spatial position due to large sequence divergence in the TRD making the comparative studies among different MTases a difficult task.

All the M. Hpy N6mA amino acid residues which interact with the substrate DNA are calculated using MolBridge (Kumar et al., 2014) and listed in Table 1. We observed that small domain of the protein interacts with the major groove of DNA. These interactions are important for recognizing and stabilizing the DNA in the termolecular systems. As expected, the most stable hydrogen bond interactions are those established with the recognition sequence (5'TCCGA^{m}3') and these interactions are responsible for the specificity of DNA binding.

Important interactions of the flipped adenine with large domain are already discussed in detail since it forms the catalytic site of the protein. Other bases of the recognition sequence mostly interact with amino acids in the large domain (Lys143, Asn145, Ala146, Ala147, Gly148, Ser186, Asp187 and Arg188) and small domain (Tyr280, Thr285, Lys324, Glu362 and Lys406).

Apart from these recognition base specific contacts, DNA backbone interactions play an important role in the proper positioning of DNA in the protein cleft (O'Gara et al., 1998). In M. Hpy N6mA, the residues from both domains (Lys143, Asn153, His154, Lys214, Gly216 and His311) interact with the bases and backbone phosphate groups of DNA.

Table 4.1: Mean (standard deviation) values for different interactions between atoms of amino acids and nucleotides/ water molecules present in the active site. An interaction with at least 50% occurrence during the simulation has been considered. The donor atoms are shown as (Residue Number: Residue Name: Chain ID: Donor Atom: Hydrogen Atom), while the acceptor atoms are presented as (Residue Number: Residue Name: Chain ID: Acceptor Atom). *H: Hydrogen; D: Donor atom; A: Acceptor atom

Donor Atoms	**Acceptor Atoms**	**HA* (Å)**	**DA* (Å)**	**DHA* (°)**
311:HIS:C:NE2:HE2	713:DC:B:O5'	2.01	2.94	157.20
214:LYS:C:NZ:HZ3	609:DT:A:O4	2.02	2.94	155.37
146:ALA:C:N:H	719:DA:B:O1P	2.02	2.92	152.43
147:LEU:C:N:H	719:DA:B:O1P	2.01	2.92	154.51
187:ASP:C:N:H	603:DC:A:O3'	2.07	2.97	152.74
280:TYR:C:N:H	602:DT:A:O2P	2.06	2.97	153.70
143:LYS:C:NZ:HZ2	607:DT:A:O3'	2.1	2.98	150.51
143:LYS:C:NZ:HZ3	605:DG:A:O3'	2.12	2.99	149.19

143:LYS:C:NZ:HZ1	607:DT:A:O3'	2.15	2.99	146.83
719:DA:B:N6:H61	285:THR:C:O	2.15	3.0	146.61
145:ASN:C:ND2:1HD2	718:DG:B:O3'	2.12	2.98	148.57
154:HIS:C:NE2:HE2	720:DC:B:O1P	2.12	2.98	148.47
145:ASN:C:ND2:1HD2	719:DA:B:O1P	2.12	2.99	148.77
143:LYS:C:NZ:HZ2	608:DG:A:O1P	2.12	2.98	148.99
153:ASN:C:ND2:2HD2	720:DC:B:O1P	2.11	2.98	149.70
606:DA:A:N6:H61	139:PRO:C:O	2.09	2.97	151.07
143:LYS:C:NZ:HZ1	608:DG:A:O1P	2.08	2.97	151.08
606:DA:A:N6:H62	138:ASN:C:OD1	2.08	2.97	151.38
143:LYS:C:NZ:HZ1	606:DA:A:O1P	2.08	2.97	151.33
709:DA:B:N6:H61	285:THR:C:O	2.08	2.97	151.37
188:ARG:C:NE:HE	604:DC:A:O1P	2.08	2.97	151.65
406:LYS:C:NZ:HZ2	603:DT:A:O2P	2.07	2.96	152.12
138:LYS:C:NZ:HZ3	607:DT:A:O3'	2.08	2.97	151.84
138:LYS:C:NZ:HZ1	605:DG:A:O3'	2.08	2.97	151.61
216:GLY:C:N:H	607:DT:A:O2P	2.08	2.97	151.60
186:SER:C:N:H	604:DC:A:O2P	2.09	2.97	151.39
143:LYS:C:N:H	606:DA:A:O1P	2.09	2.98	151.74
146:ALA:C:N:H	719:DA:B:O5'	2.1	2.99	151.80
148:GLY:C:N:H	719:DA:B:O1P	2.09	2.99	152.24
188:ARG:C:NH1:2HH1	605:DG:A:O2P	2.09	2.99	151.96
214:LYS:C:NZ:HZ2	609:DT:A:O4	2.09	2.98	151.94
406:LYS:C:NZ:HZ3	603:DT:A:O2P	2.09	2.98	151.83
311:HIS:C:NE2:HE2	713:DC:B:O2P	2.09	2.98	152.04
188:ARG:C:NH2:2HH2	605:DG:A:O2P	2.09	2.98	152.03
187:ASP:C:N:H	604:DC:A:O1P	2.09	2.99	151.99
143:LYS:C:NZ:HZ3	606:DA:A:O1P	2.09	2.98	151.98
216:GLY:C:N:H	606:DA:A:O3'	2.1	2.99	151.85
716:DC:B:N4:H41	362:GLU:C:OE2	2.1	2.99	151.94
716:DC:B:N4:H41	362:GLU:C:OE1	2.09	2.99	152.04
143:LYS:C:NZ:HZ2	605:DG:A:O3'	2.1	2.99	151.75
188:ARG:C:N:H	604:DC:A:O1P	2.1	2.99	151.88
188:ARG:C:NH2:1HH2	604:DC:A:O5'	2.1	2.99	151.83
153:ASN:C:ND2:2HD2	719:DA:B:O3'	2.11	3.0	151.91
324:LYS:C:NZ:HZ1	715:DT:B:O1P	2.11	3.0	151.81
606:DA:A:N6:H62	138:ASN:C:ND2:2HD2	2.06	2.98	151.83

To better characterize the interactions of M. Hpy N6mA with DNA, we further analysed the conserved amino acids in the TRD of MTases in some of its closely related sequences. We have extracted all the putative N6mA MTase sequences which recognize 5'TCCGA^{m}3'from the REBASE database and performed multiple sequence alignment in order to find the sequence similarity (Figure 4.9). This allows us to find important residues involved in DNA binding in 5'TCCGA^{m}3'recognizing DNA MTases. The protein sequence alignment indicated two conserved sequence motifs even in very distantly related sequences as shown in the Figure 4.9. The conserved sequence motifs form structurally conserved regions which are mainly responsible to support the residues that are involved in the binding of DNA backbone. The residues surrounding the conserved sequence motifs are actually interacting with the DNA as shown by *in silico* mutation experiments in the small domain region (280-293 and 308-325). This indicates that these DNA MTases which recognize the same sequence could interact with DNA in a similar fashion as in M. Hpy N6mA.

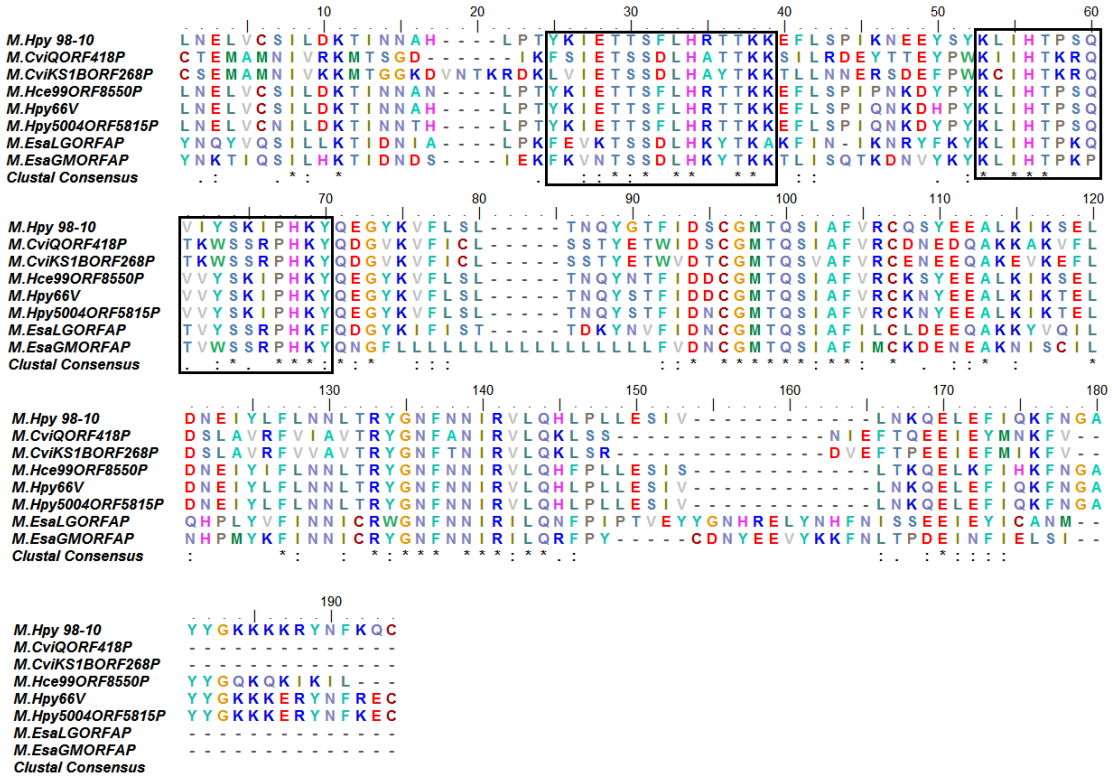


Figure 4.9: Sequence alignment of small domain region in M. Hpy N6mA with other putative N6mA DNA MTase sequences. Residues in black box are predicted TRD region in M. Hpy N6mA

4.4 Conclusions

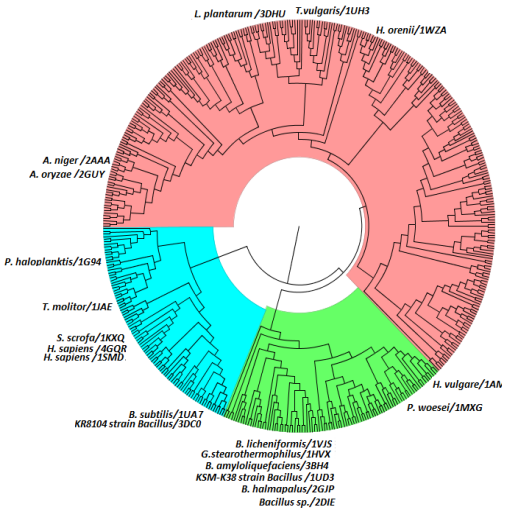
M. Hpy N6mA is a DNA MTase from a cancer causing strain of *H. pylori* 98-10 which catalyzes the transfer of a methyl group from AdoMet to the N6 position of an adenine belonging to recognition sequence 5'TCCGA^{m}3'. Sequence as well as structural analysis of this DNA MTase confirmed it in γ subclass of exocyclic DNA MTase.

A 50ns long MD simulations have been carried out for the termolecular system formed by the cofactor AdoMet, the protein, and a DNA decamer which comprises the recognition sequence 5'TCCGA^{m}3'. This 3-D structure of M. Hpy N6mA MTase in complex with AdoMet (bimolecular systems) as well as complex of protein-AdoMet with specific DNA (termolecular systems) discloses a previously unrecognized stabilization of the extrahelical target nucleotide. In addition, structural comparisons with the other N6mA specific DNA MTase structures like M. TaqI in complex with the natural cofactor as well as DNA provide information about the catalytic mechanism of M. Hpy N6mA. To the best of our knowledge, this is the first MD simulation performed in three systems of M. Hpy N6mA. Five structural motifs were observed to be involved in stabilizing DNA in the protein cleft. Further, to confirm the role of these motifs we have studied the stability of the mutant complex using MD simulations.

The specific recognition of DNA by the class specific MTases remains to be explored till today. In M. Hpy N6mA MTase, we observed that the unpaired T15 base is stabilized not only by means of protein− DNA interactions but also by means of π-stacking interactions with neighbouring bases DG5, DC16 and DA14, and Lys324. Apart from this, structure motifs involved in DNA recognition are also identified. N6mA have been found only in prokaryotes, and therefore this enzyme is a potential target for the development of antibacterial drugs.

Sequence and structure based analysis of α-amylase evolution

- ✓ Comprehensive structure and sequence based analysis of α-amylase evolution.
- ✓ Differences in the profile specific conserved and insertion/deletion regions.
- ✓ Overview of various factors responsible for the Ca^{2+} and Cl^{-} binding and the disulfide connectivity pattern.



Structure and sequence based analysis of alpha-amylase evolution. Singh S, Guruprasad L. Protein Pept Lett. 2014;21(9):948-56

A.1 Introduction

α-Amylase (EC 3.2.1.1) forms one of the largest families within the glycosyl hydrolases and are distributed between families 13 (archaea, bacteria, and eukaryota) and 57 (archaea and bacteria) of glycoside hydrolases (Henrissat and Bairoch, 1996). This enzyme catalyzes the hydrolysis of α-1,4-glycosidic bonds of glycogen, starch, related polysaccharides (mainly linear amylose and the branched amylopectin) and some oligosaccharides, releasing α-maltose, α-glucose and α-limit dextrin in a stepwise manner by means of retaining mechanism (the resulting hydroxyl group retains α configuration). They play an essential role in the metabolism of plants by hydrolyzing starch in the germinating seed and in other tissues by catalyzing the hydrolysis of α-amylase (1,4) glycosidic linkages of starch components (amylose and amylopectin), glycogen and other various oligosaccharides.

Apart from playing a crucial role in digesting carbohydrates or polysaccharides into smaller disaccharide units and eventually converting them into monosaccharides, α-amylases are also involved in the digestion of dead white blood cells (pus). Low amylase content in the blood may cause abscesses (areas with pus and no bacteria). Amylases are also involved in anti-inflammatory reactions triggered by the release of histamine and similar substances. The inflammatory response usually occurs in organs which are in contact with the outside world such as the lungs and skin. Therefore, an amylase deficiency can include skin problems such as psoriasis, eczema, hives, insect bites, allergic bee and bug stings, atopic dermatitis, and all types of herpes. Lung problems, including asthma and emphysema, require amylase plus other enzymes depending on the particular condition. Hyperamylasemia and hyperamylasuria of varying degrees are frequently observed in patients with lung and ovarian cancers (Shimamura et al., 1976; Sudo and Kanno, 1976).

Amylases are found in all organisms across the evolution. Amylases mainly of bacterial and fungal origin are widely used in starch processing, paper manufacture, detergent industry and pharmacology, while cereal amylases are used in the production of beer and alcoholic beverages. Thermostable enzymes secreted from bacillus species are very important industrial enzymes in the production of corn syrups and dextrose. In humans α-amylase is one of the major secretory product of the pancreas and salivary glands which produce different α-amylases with molecular weight of about

54.5 kDa for pancreatic enzyme and about 56 kDa for salivary enzyme (Matsuura et al., 1978; Stiefel and Keller, 1973). Pancreatic α-amylase is one of the key targets for drug therapies against diabetes mellitus which is a metabolic disorder characterized by chronic hyperglycemia. Pancreatic α-amylase is a target in drug therapies because it is known to spike post prandial blood glucose levels (Ferey-Roux et al., 1998; Mizuno et al., 2008). Diabetes and obesity lead to a significantly reduced quality of life, with an increased risk of serious complications including cardiovascular disease, hypertension, stroke, kidney failure and nerve damage. Therefore, human pancreatic α-amylase provides a unique opportunity for the development of potential therapeutic agents for the treatment of these conditions.

The sequences of several thousands of α-amylases from genome sequencing projects and otherwise are known so far. The 3-D structures of α-amylases from a variety of sources have been solved with and without inhibitors. The evolution of this conserved protein across various organisms would indicate the interesting events of alterations the protein has undergone. Therefore we studied the 3-D structure analysis of this important class of enzymes. Here we present a structure and sequence based alignment of α-amylases from 19 different organisms representing broad phylogenetic diversity that include higher eukaryotes, plants, insects, microbial and fungal proteins. Using this structure-based alignment, an analysis was performed to evaluate the degree of site specific variability or conservation of both sequences as well as structures. Further the phylogenetic tree of all representative α-amylase sequences that share less than 50% sequence identity was generated to verify the observations made from structure based sequence alignment. Additionally, these trees were compared with the phylogenetic trees resulting from sequence alignment. The goal of this work is to gain insights into the structure and sequence based evolution of α-amylases that belong to the conserved class of enzymes.

A.2 Method

A.2.1 Selection of data

The PDB was searched to find all deposited α-amylase structures. A total 155 X-ray crystal structures and associated protein sequences were downloaded in the FASTA format. This information was imported into a spreadsheet where they were sorted by species and their classification as either a mutant or a wild type protein sequence. Only wild type structures were selected from each species. In the organisms having more than one α-amylase structure, the best PDB structure was selected based on their completeness, resolution, B-factors and uniqueness. In the case of *H. sapiens* both pancreatic α-amylase and salivary α-amylase were considered since they were two distinct structures with different primary sequences.

The NCBI non redundant (NCBI nr) protein sequence database was searched using PSI-BLAST (Altschul et al., 1997) to find all α-amylase sequences. This was carried out for several rounds of iterations and by performing reciprocal searches. Since this dataset was large, it was sorted using CD-Hit (Li and Godzik, 2006) to select sequences that share less than 50% sequence identity. Further this dataset was manually verified to remove all sequences that contained only fragments. This allowed us to collect all representative α-amylase sequences that share less than 50% sequence identity.

A.2.2 Structural alignment

From each of the 19 unique prokaryotic and eukaryotic species one crystal structure was selected. Using Discovery Studio 2.5, each structure was cleaned by removing all waters and any heteroatoms. In addition, each PDB file was truncated to include only one chain of the α-amylase. The structures were aligned using MAPSCI (Ilinkin et al., 2010). The structure based aligned sequences were then exported as a .pir file for importing into BioEdit for further analysis.

A.2.3 Sequence alignment

The 20 FASTA sequences of known structures from the 19 species were imported into MEGA 5.0 (Tamura et al., 2011). A full sequence alignment was performed using the Muscle module in MEGA 5.0 with the assignment of parameters, gap opening penalty (-2.9), gap extension penalty (0), hydrophobicity multiplicity (1.2) and the clustering method used was UPGMA (Unweighted Pair Group Method

with Arithmetic Mean). Similar procedure was adopted for the sequence dataset obtained from NCBI nr database.

A.2.4 α-Amylase phylogenetic tree

The phylogenetic tree of all the species was built from publically available tools at www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/wwwcmt.cgi by entering the names of 19 organisms. The resulting phylogeny was saved in Phylip tree format (.phy) and imported into MEGA 5.0 for further formatting. For structure and sequence based alignments, phylogenetic trees were generated with MEGA 5.0 using the 'Maximum Likelihood' method with the following parameters: uniform rates of site specific mutation; partial deletion of gaps and missing data; and nearest-neighbor interchange heuristic tree inference method. WAG amino acid substitution model was used for PDB sequences and LG amino acid substitution model for NCBI nr sequences. The best fit amino acid substitution model was selected using BIC (Bayesian information criterion) and AIC (Akaike information criterion) scores, the model with lowest BIC and AIC values are considered to describe best substitution pattern. Each phylogeny was evaluated using the bootstrap method with 500 replications.

A.2.5 Determination of structure conservation and variability

The structural alignment of the representative α-amylases was produced using the JOY method (Mizuguchi et al., 1998). JOY is an analysis and formatting program for multiple protein sequence alignments to annotate the alignments with 3-D structural features. It was developed to display 3-D structural information in a sequence alignment and help understand the conservation of amino acids in their specific local environments.

A.3 Results

The dataset in our analysis has 20 α-amylase structures from 19 different species. In humans, α-amylase gene is encoded on chromosome 1 as part of a multi gene family but is regulated by different isozymes which are synthesized in either salivary glands or pancreas. These two human α-amylases are highly homologous in terms of primary sequence (Nishide et al., 1984; Nishide et al., 1986) but do exhibit somewhat different cleavage patterns. The functional differences observed arise from the 15 amino acid substitutions between these sequences, some of which occur in the active site region (Brayer et al., 1995). Hence, we have chosen two distinct protein structures from *H. sapiens* for our study.

NCBI nr database searches identified 492 α-amylases that share less than 50% sequence identity. The numbering of amino acid residues in this chapter is as per the structure of human pancreatic α-amylase (PDB_ID: 4GQR) unless otherwise mentioned.

A.3.1 Structural alignment

The α-amylases contain three structural domains - A, B and C as shown in Figure A.1A. Domain A is the largest and forms an eight-stranded parallel β-barrel. It has the location of three active site residues. Usually, the loops that link β-strands to the adjacent α-helices carry amino acid residues of the active site. Domain C is only loosely associated with the other two domains and its function is unknown (Li et al., 2005). Domain B forms the Ca^{2+} binding site which is positioned adjacent to the wall of the β-barrel of domain A. Members of the α-amylase family are known to bind one or more Ca^{2+} (Buisson et al., 1987). The Ca^{2+} preserves the structural integrity of the active site by linking the two fragments; the catalytic $(\beta/\alpha)_8$ barrel and domain B, thus maintaining the correct conformation, thermostability and activity of α-amylases (Janecek, 1997). This is an invariant structural Ca^{2+} in all α-amylase structures. The structural superposition of α-amylases obtained from MAPSCI is shown in Figure A.1B and their overall RMSD of the Cα atoms was found to be 0.86 Å.

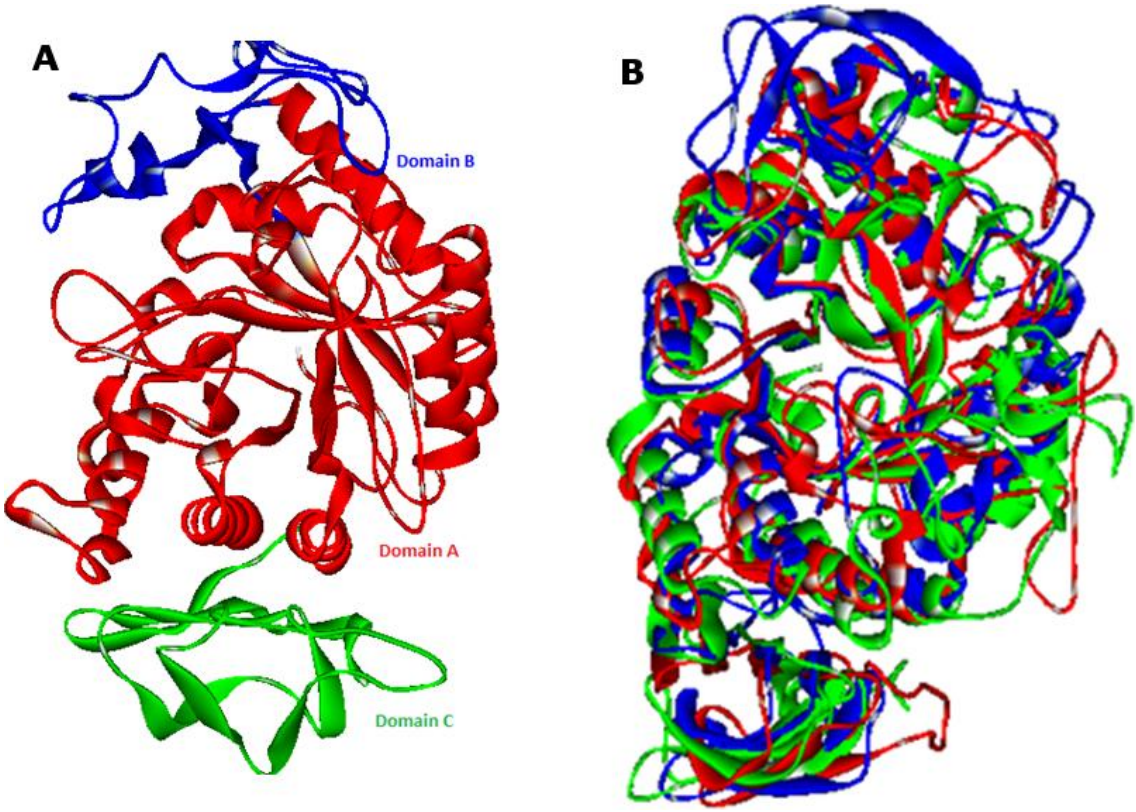
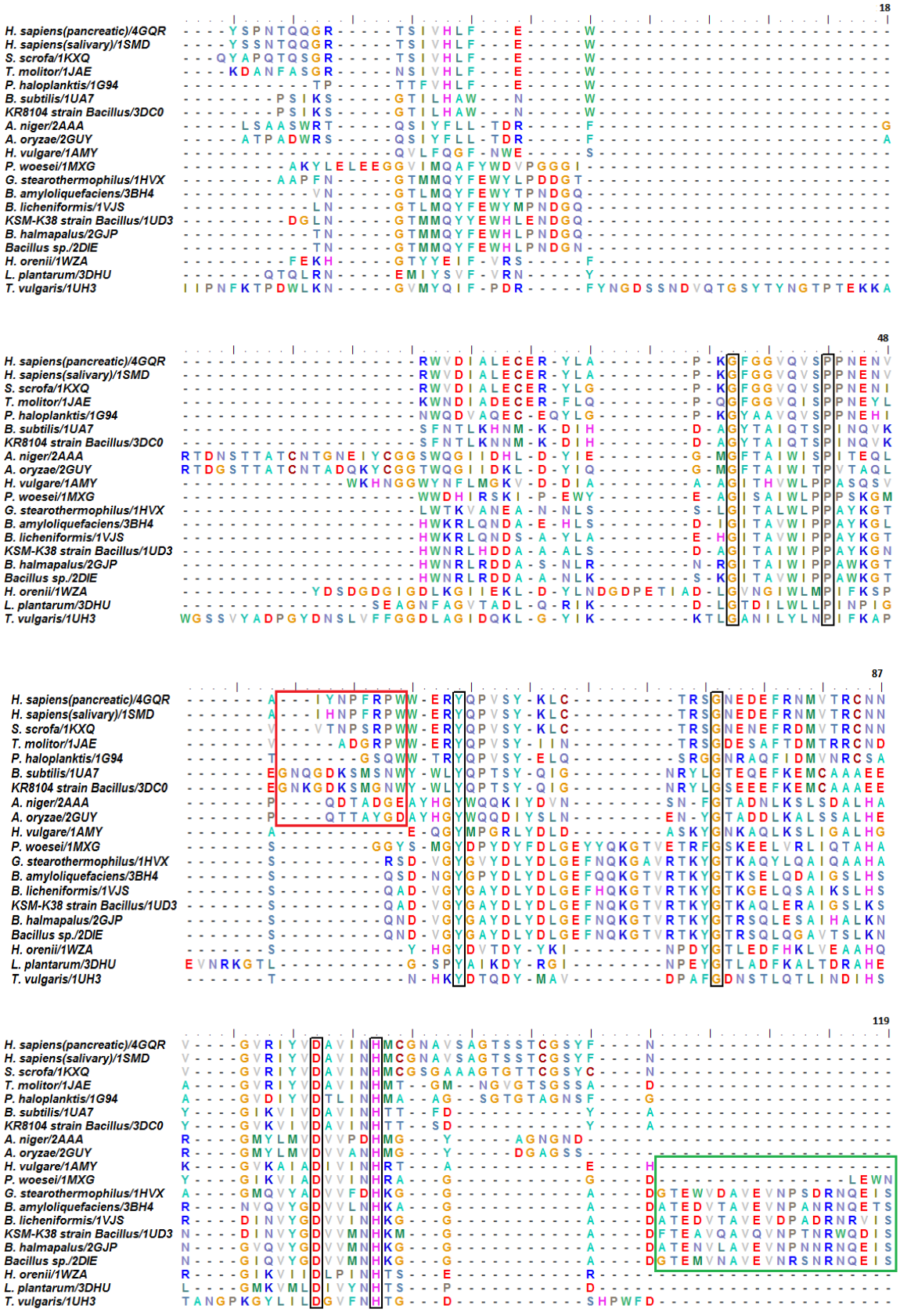
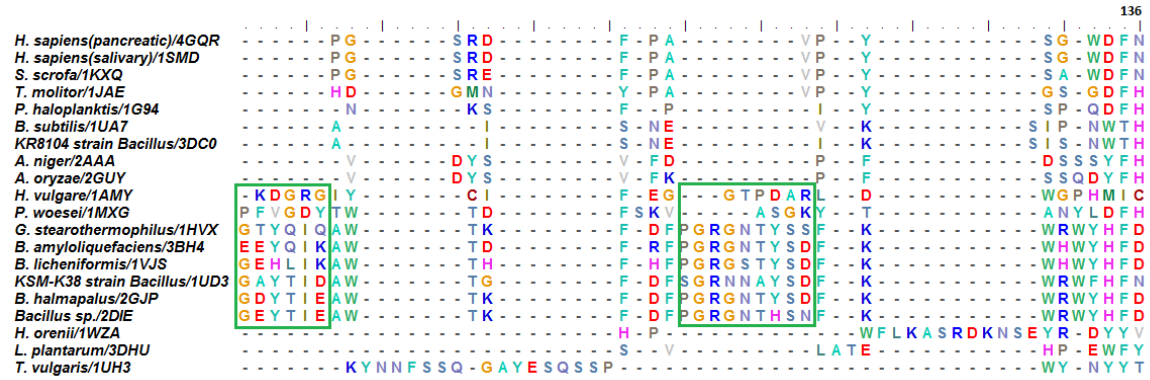


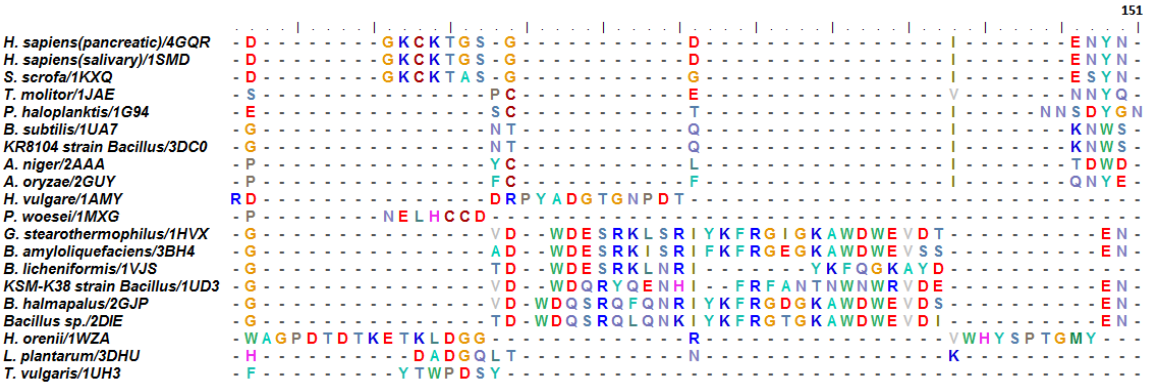
Figure A.1: (A) Solid ribbon representation of α- amylase showing three structural domains - A, B and C. Domain A with the $(\beta/\alpha)_8$ barrel structure is shown in red, domain B in dark blue and domain C in green.(B) Representative 3-D structural superimposition of α-amylases.

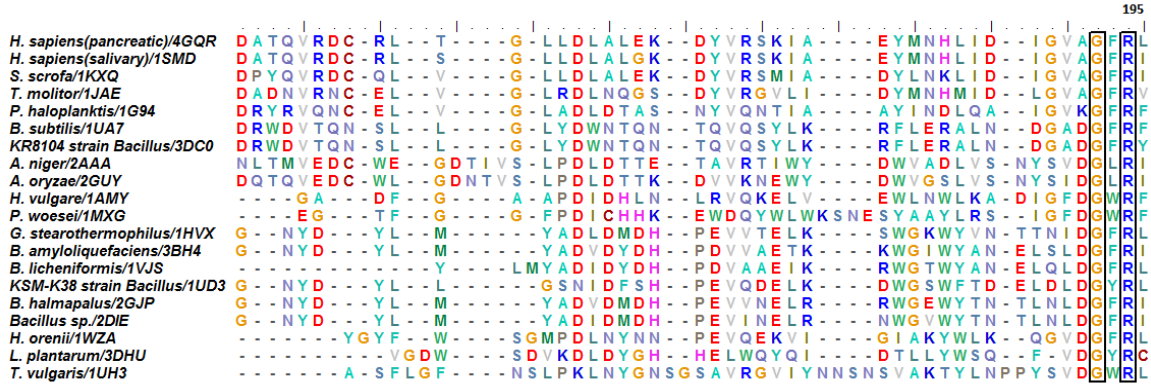
Structure based sequence alignment (Figure A.2) indicated the presence of conserved residues. Three acidic amino acid residues (Asp197, Glu233 and Asp300) were identified as the catalytic sites of α-amylase family enzymes based on the results obtained by X-ray crystallographic analysis (Katsuya et al., 1998; Matsuura et al., 1984), chemical modification (Plant et al., 1987) and site directed mutagenesis (Kuriki et al., 1991; Nagashima et al., 1992). The structure based sequence alignment shown in Figure A.2 indicated some degree of similarity even between distant organisms such as bacteria and vertebrates and revealed four highly conserved regions (95-100, 193-200, 232-235 and 294-299) that correspond to sites necessary for enzyme structure and/or function as reported earlier. In addition to these, we observed that regions corresponding to 240-247 and 323-340 are also highly conserved in all the sequences. This would suggest common ancestry for the α-amylases and conservation of critical sequences over evolutionary time. These stretches of conserved regions being more or less easily identifiable in all species in the sequence alignment, comprise the catalytic residues that play a role in the enzyme catalysis. In all the species analysed, Gly36, Pro44, Gly74, His101, Gly193, Arg195, Asp197, Glu233 and

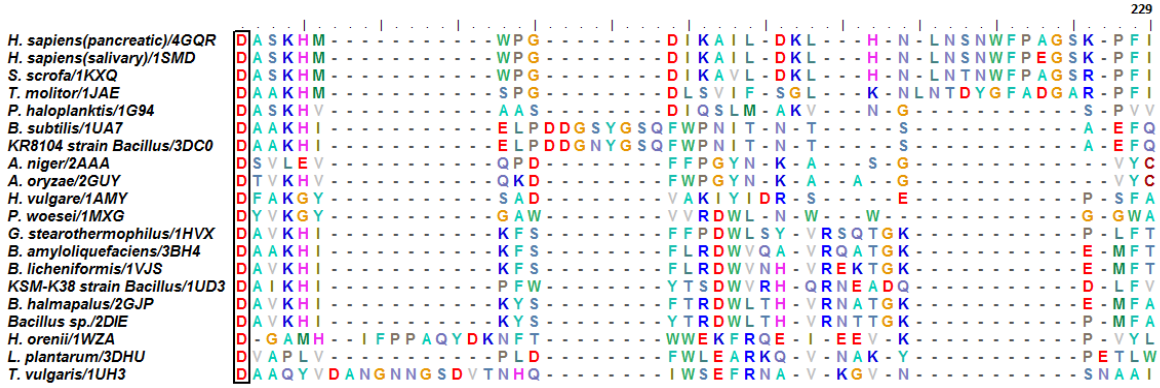
Asp300 are found to be identical along the alignment. The location of conserved residues in the protein 3-D structure is shown in Figure A.3.

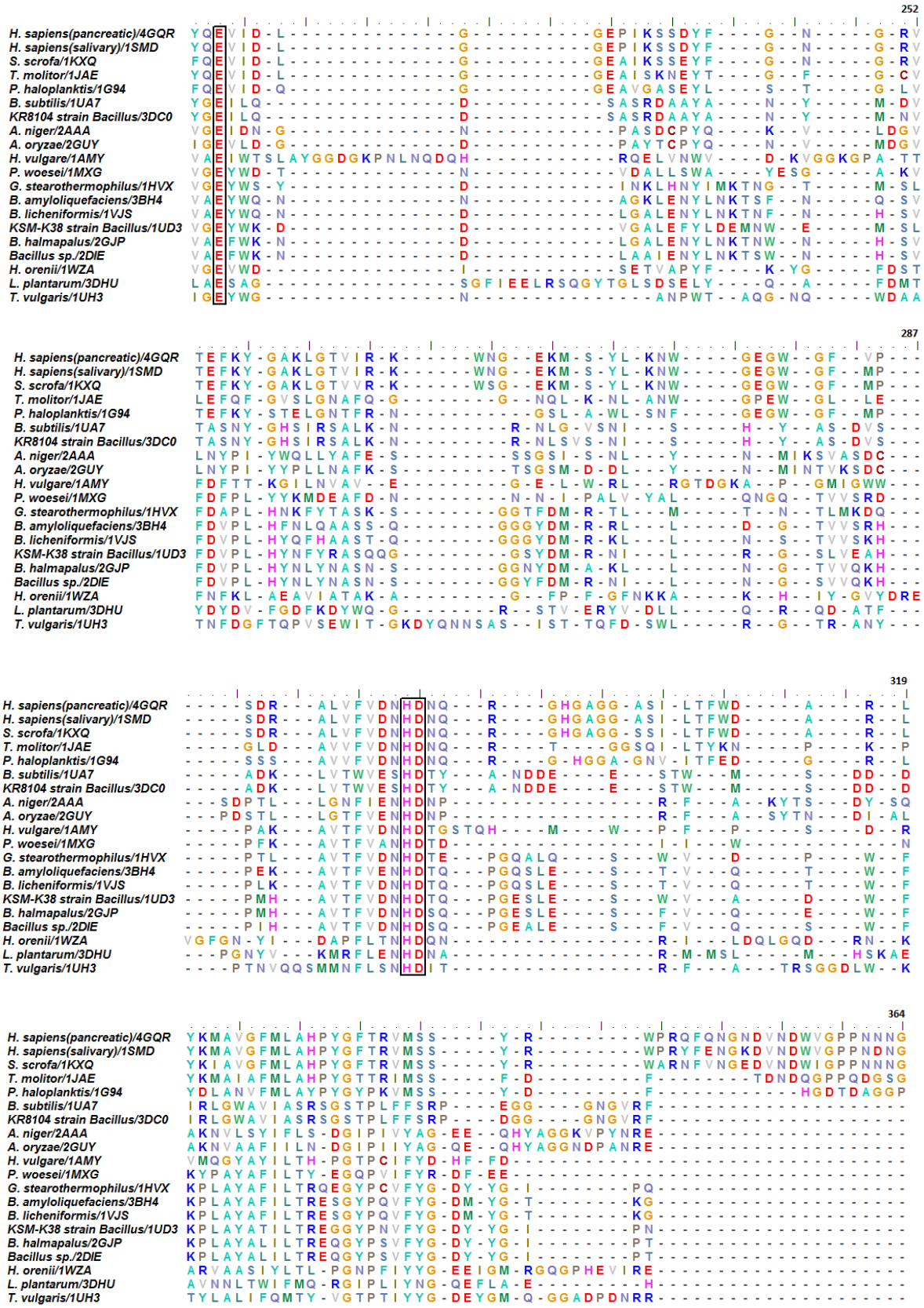


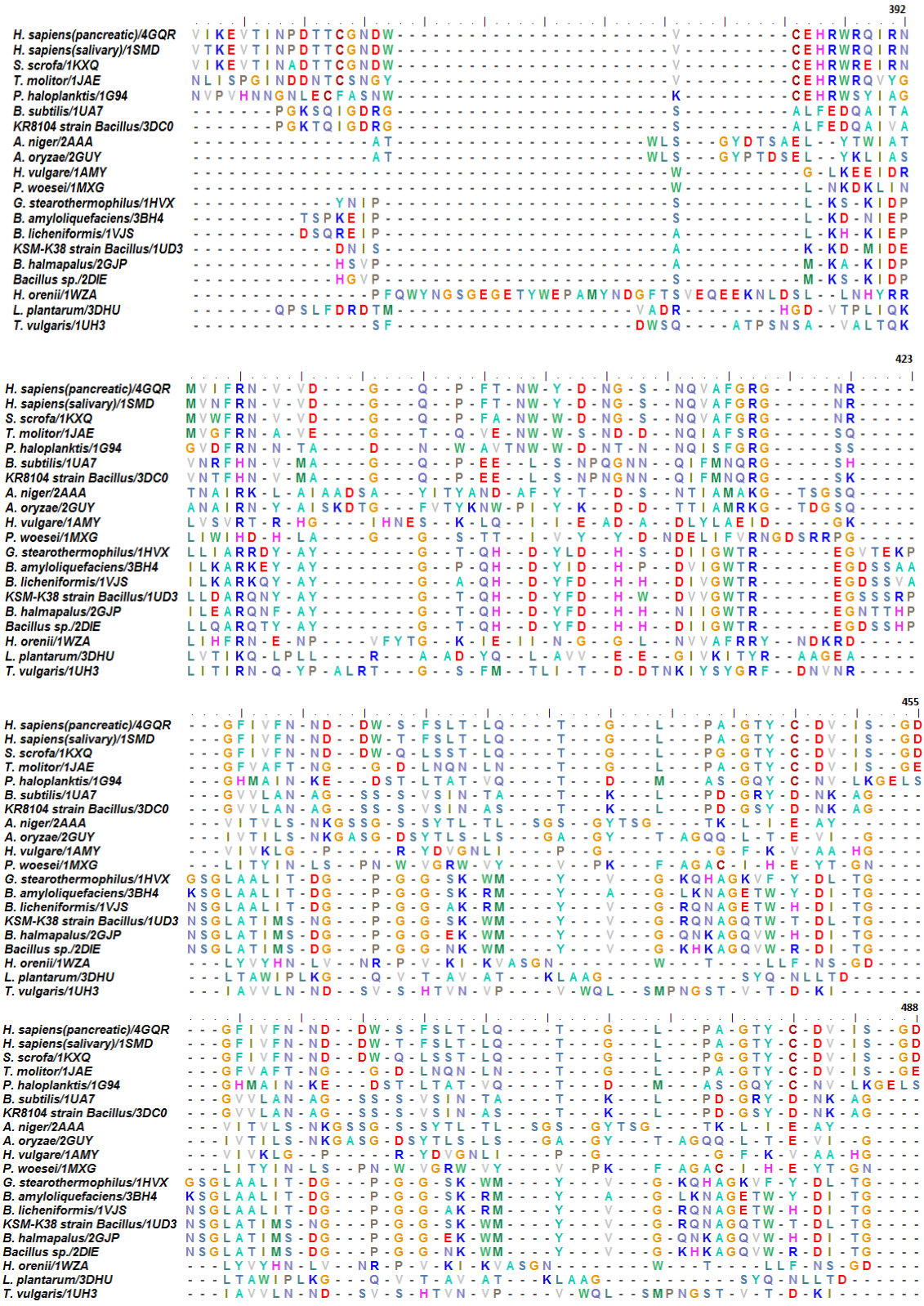

136
H. sapiens(pancreatic)/4GQR
H. sapiens(salivary)/1SMD
S. scrofa/1KXQ
T. molitor/1JAE
P. haloplanktis/1G94
B. subtilis/1UA7
KR8104 strain Bacillus/3DC0
A. niger/2AAA
A. oryzae/2GUY
H. vulgare/1AMY
P. woesei/1MXG
G. stearothermophilus/1HVX
B. amyloliquefaciens/3BH4
B. licheniformis/1VJS
KSM-K38 strain Bacillus/1UD3
B. halmapalus/2GJP
Bacillus sp./2DIE
H. orenii/1WZA
L. plantarum/3DHU
T. vulgaris/1UH3


151
H. sapiens(pancreatic)/4GQR
H. sapiens(salivary)/1SMD
S. scrofa/1KXQ
T. molitor/1JAE
P. haloplanktis/1G94
B. subtilis/1UA7
KR8104 strain Bacillus/3DC0
A. niger/2AAA
A. oryzae/2GUY
H. vulgare/1AMY
P. woesei/1MXG
G. stearothermophilus/1HVX
B. amyloliquefaciens/3BH4
B. licheniformis/1VJS
KSM-K38 strain Bacillus/1UD3
B. halmapalus/2GJP
Bacillus sp./2DIE
H. orenii/1WZA
L. plantarum/3DHU
T. vulgaris/1UH3


195
H. sapiens(pancreatic)/4GQR
H. sapiens(salivary)/1SMD
S. scrofa/1KXQ
T. molitor/1JAE
P. haloplanktis/1G94
B. subtilis/1UA7
KR8104 strain Bacillus/3DC0
A. niger/2AAA
A. oryzae/2GUY
H. vulgare/1AMY
P. woesei/1MXG
G. stearothermophilus/1HVX
B. amyloliquefaciens/3BH4
B. licheniformis/1VJS
KSM-K38 strain Bacillus/1UD3
B. halmapalus/2GJP
Bacillus sp./2DIE
H. orenii/1WZA
L. plantarum/3DHU
T. vulgaris/1UH3


229
H. sapiens(pancreatic)/4GQR
H. sapiens(salivary)/1SMD
S. scrofa/1KXQ
T. molitor/1JAE
P. haloplanktis/1G94
B. subtilis/1UA7
KR8104 strain Bacillus/3DC0
A. niger/2AAA
A. oryzae/2GUY
H. vulgare/1AMY
P. woesei/1MXG
G. stearothermophilus/1HVX
B. amyloliquefaciens/3BH4
B. licheniformis/1VJS
KSM-K38 strain Bacillus/1UD3
B. halmapalus/2GJP
Bacillus sp./2DIE
H. orenii/1WZA
L. plantarum/3DHU
T. vulgaris/1UH3


252
H. sapiens(pancreatic)/4GQR
H. sapiens(salivary)/1SMD
S. scrofa/1KXQ
T. molitor/1JAE
P. haloplanktis/1G94
B. subtilis/1UA7
KR8104 strain Bacillus/3DC0
A. niger/2AAA
A. oryzae/2GUY
H. vulgare/1AMY
P. woesei/1MXG
G. stearothermophilus/1HVX
B. amyloliquefaciens/3BH4
B. licheniformis/1VJS
KSM-K38 strain Bacillus/1UD3
B. halmapalus/2GJP
Bacillus sp./2DIE
H. orenii/1WZA
L. plantarum/3DHU
T. vulgaris/1UH3
287
319
364


392
H. sapiens(pancreatic)/4GQR
H. sapiens(salivary)/1SMD
S. scrofa/1KXQ
T. molitor/1JAE
P. haloplanktis/1G94
B. subtilis/1UA7
KR8104 strain Bacillus/3DC0
A. niger/2AAA
A. oryzae/2GUY
H. vulgare/1AMY
P. woesei/1MXG
G. stearothermophilus/1HVX
B. amyloliquefaciens/3BH4
B. licheniformis/1VJS
KSM-K38 strain Bacillus/1UD3
B. halmapalus/2GJP
Bacillus sp./2DIE
H. orenii/1WZA
L. plantarum/3DHU
T. vulgaris/1UH3
423
455
488

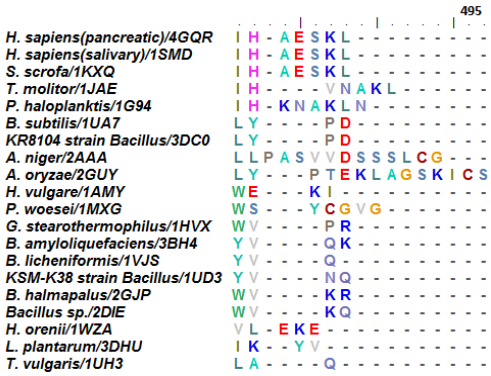


Figure A.2: Structure based sequence alignment of 20 α-amylases showing conserved regions (black) and specific insertion regions in different profiles- red: profile one, green: profile two. The amino acid numbering shown in the first line is according to PDB_ID: 4GQR.

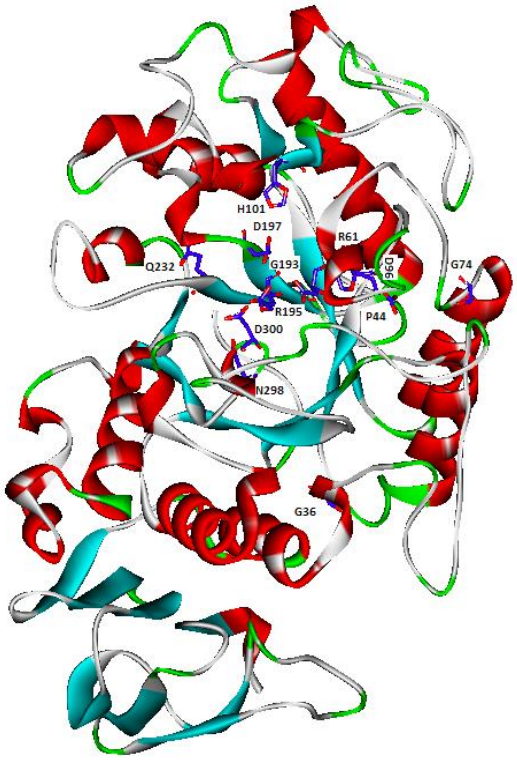


Figure A.3: The 3-D Structure of human pancreatic α-amylase (PDB_ID: 4GQR) indicating the location of the conserved residues.

A.3.2 Sequence alignment

The FASTA sequences of 20 α-amylase structures from all 19 species were imported into MEGA 5.0 and a multiple sequence alignment was performed using the module Muscle. Similarly alignment of 492 NCBI nr sequences was generated.

A.3.3 Consensus structure in the catalytic center of the α-amylase family

The essential criteria of an enzyme to be in α-amylase family is a single polypeptide chain folded into three domains, with the first domain belonging to $(\beta/\alpha)_8$ barrel protein family. The three catalytic residues; two aspartic acids and one glutamic

acid, reside on the C-terminal end loops of the 4th, 5th and 7th β-strands of the barrel, respectively (MacGregor, 1988) and are strictly conserved in both primary sequence and 3-D structure. These three catalytic residues form a triad with distances between carboxylate groups ranging from 5 to 7Å with no direct hydrogen bond to each other. The sequence alignment of 492 NCBI nr sequences also showed the conservation of these three catalytic residues.

A.3.4 α-Amylase phylogenetic tree

The phylogenetic tree of the 19 species generated from NCBI (Figure A.4) was used as a reference tree to evaluate the phylogenetic trees resulting from the sequence alignments of structure based (Figure A.5) and 492 NCBI nr α-amylase dataset (Figure A.6).

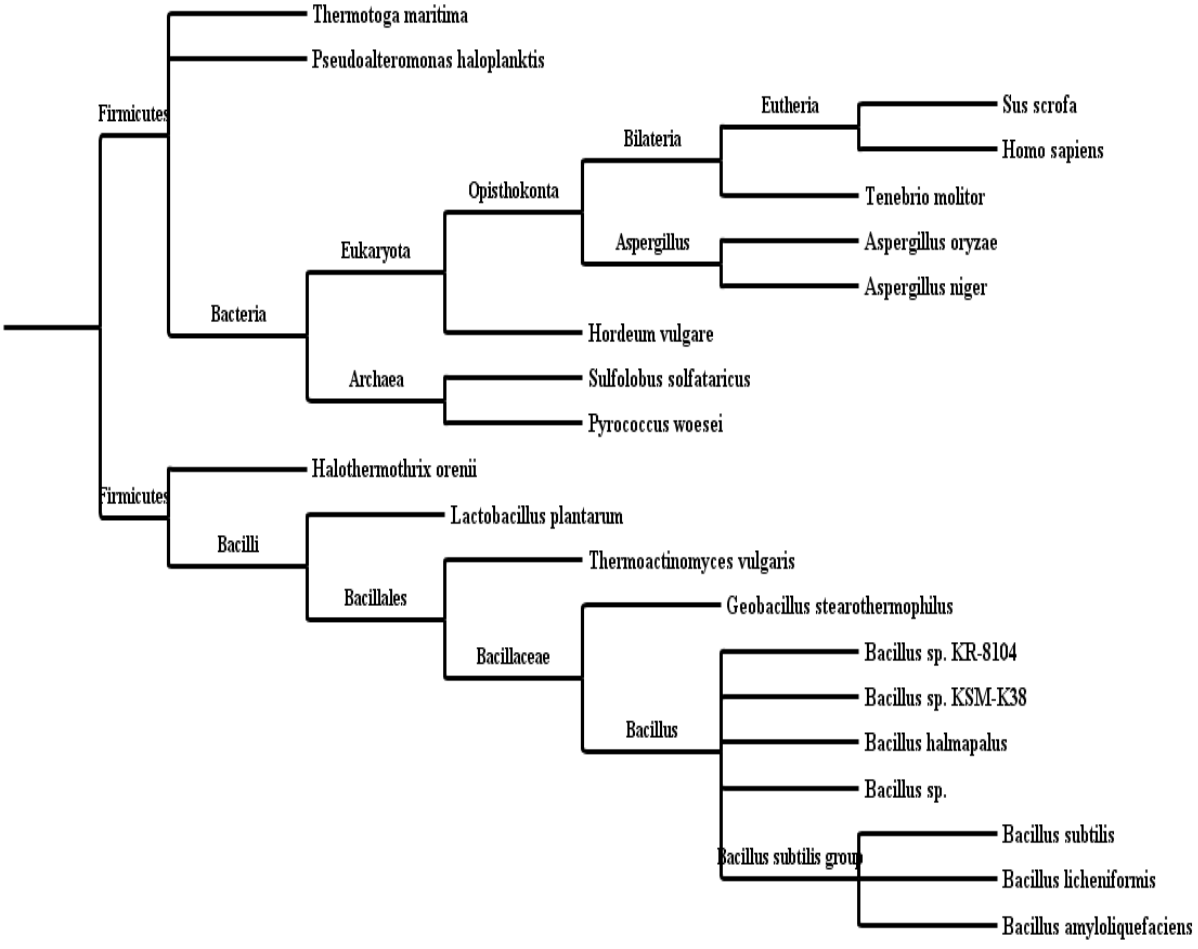


Figure A.4: Taxonomic tree of 19 representative species used as reference tree.

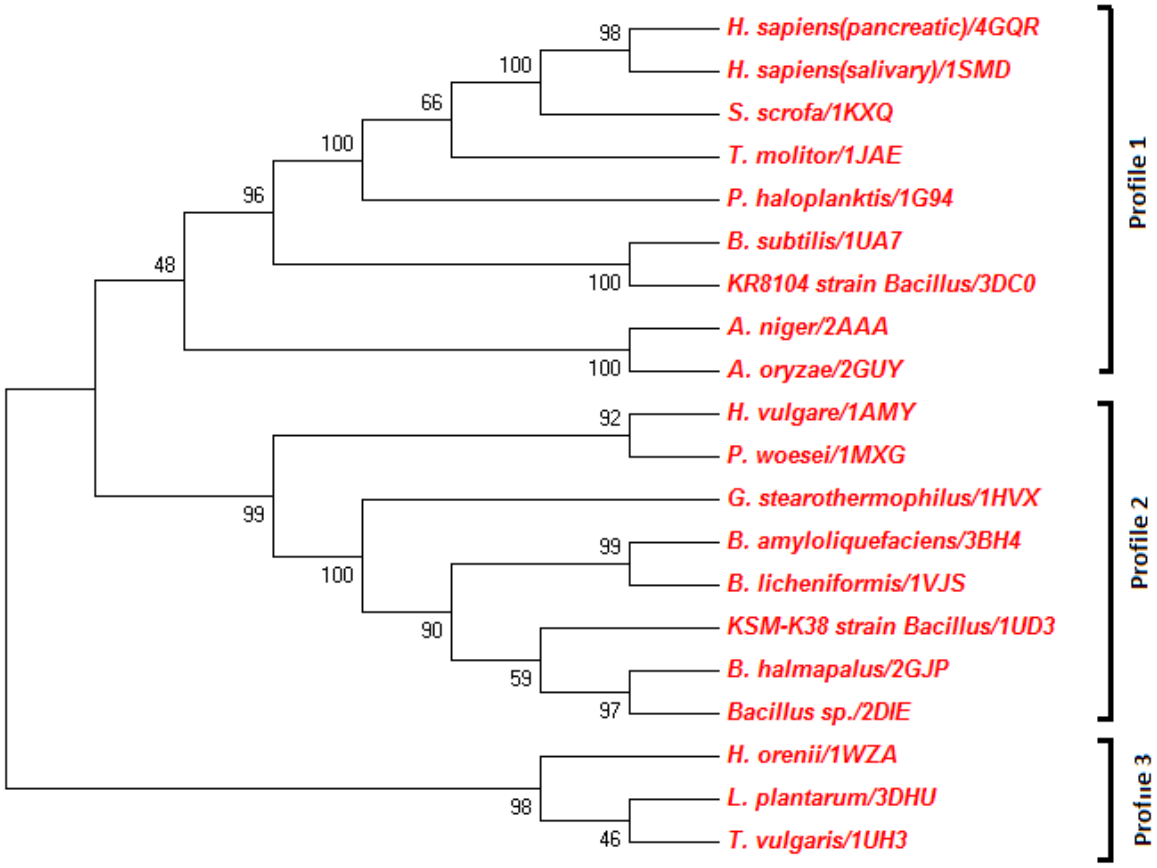


Figure A.5: Phylogenetic tree generated from structural alignment of 20 α-amylases.

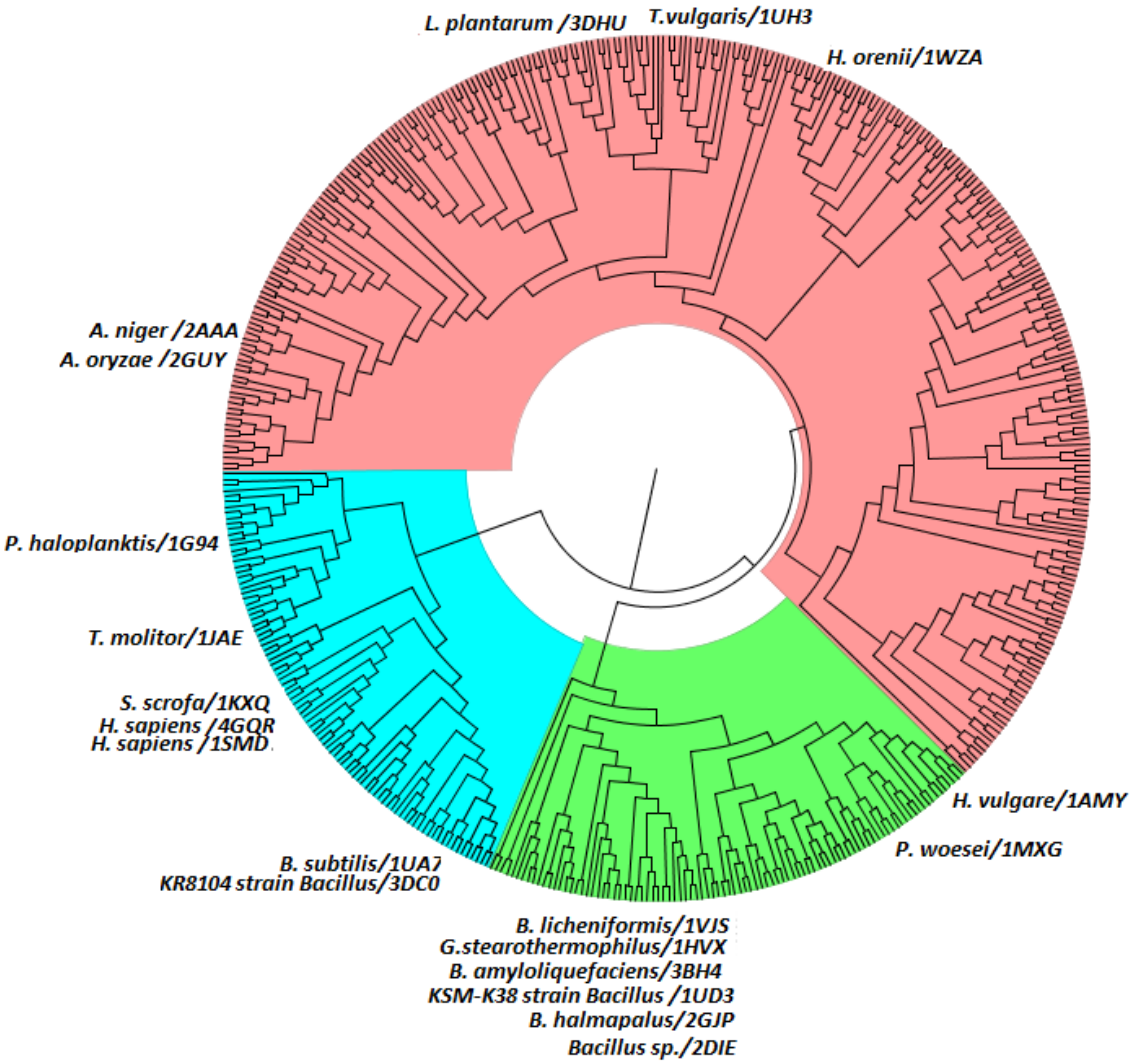


Figure A.6: Phylogenetic tree showing position of 20 representative PDB structures in the phylogentic tree generated by 492 α-amylases from NCBI nr dataset.

Comparison of the taxonomic tree (Figure A.4) with structure (Figure A.5) and sequence based phylogenetic trees (Figure A.6) revealed that the location of α-amylase from *B. subtilis, KR8104 strain Bacillus, H. vulgare* and *P. haloplaktis* were different in both the trees. From the analysis of α-amylase phylogenetic trees we observed that both trees were slightly different to each other with respect to the positioning of *Aspergillus sp.* On comparing both structure and sequence based phylogenetic tree with the taxonomic tree, we observed that in sequence based phylogenetic tree the *Aspergillus sp.* is coming close to *L. plantarum* and *T. vulgaris* while in the taxonomic and structure based phylogenetic trees *Aspergillus. sp*. is close to Bilateria gp. Except the positioning of *Aspergillus sp*. all the patterns in both trees are same, thus we can

conclude that phylogenetic tree generated from the sequence alignment differs more with respect to the taxonomic phylogeny tree than the structure based alignment phylogeny.

From the above, we believe that the structure based sequence alignment truly represents the functional aspects of proteins than from the sequence information alone, so we have considered the structure based phylogenetic for further studies.

In the α-amylase phylogenetic tree, the eukaryotic clade was fairly close to reference phylogenetic tree with few exceptions. The α-amylase of *P. halopanktis* of bacterial origin was promoted close to eukaryotic class as compared to NCBI reference tree. The α-amylase from *H. vulgare* which is expected to be close to eukaryotic class as per the reference tree got demoted to the bacterial clade in the α-amylase phylogenetic tree. These results show that there is significant structural homology between distantly related α-amylases which means that distant α-amylases with correspondingly dissimilar primary sequences are also likely to fold into structures comparable to that of *H. sapiens*, *S. scrofa* and *A. oryzae* α-amylases. Thus, for these enzymes, polypeptide chain folding is the overriding constraint over the course of molecular evolution as opposed to the preservation of primary sequence identities.

For the analysis of α-amylase phylogenetic trees, three major profiles were created based on branching pattern as shown in Figure A.5. Profile one is of eutherian, bacterial, insect and fungal origin, second is a joint profile of plant and bacteria while third profile is of bacterial origin only.

Profile one is again divided into three clusters I, II and III. Cluster I has four species (*H. sapiens, S. scrofa, T. molitor, P. haloplanktis*). Cluster II has two bacterial species (*B. subtilis* and KR8104 strain *Bacillus sp*.) while cluster III has Aspergillus sp. only. This profile has insertion between positions 50-55 of variable lengths in all organisms. The presence of this insertion is distinctive to members of profile one. Eutherians or mammalian clades have unique insertion region from 138-144 while insertion from 216-225 is present only in all the eukaryotes of cluster I. Cluster I has unique insertion region 344-371 and 454-462, and conserved regions between 237-239, 383-387 and 331-340 (PDB_ID: 4GQR). Cluster II has insertion at 182-189 (PDB_ID: 1UA7) and various conserved regions like 210-221, 245-257, 268-278 and 384-396. Cluster III has insertion 168-171(PDB_ID: 2AAA) and conserved regions from 107-117 and 332-352.

Profile two is of bacterial and plant origin that showed two major clusters. All the species of this profile have insertions between 109-136 and 142-151 of variable lengths. Cluster I consists of two distantly related species (*H. vulgare* and *P. woesei*) that are otherwise not a part of the same cluster in the NCBI reference tree. Long insertions between 109-136 and 142-151 (PDB_ID: 3BH4) are present in whole profile but in these two organisms this insertion is of few amino acid residues only. Unique insertion region is present in *H. vulgare* from 210-223 (PDB ID: 1AMY) which may be plant specific only.

Cluster II consists of six species (*G. stearothermophilus, B. amyloliquefaciens, B. licheniformis, KSM-K38 strain Bacillus sp. B. halmopalusand, Bacillus sp*.). Members of this cluster have insertions between 164-172, 174-189, 276-279, 332-337 and 417-424 and conserved residues from 190-197 and 343-349 (PDB_ID: 3BH4).

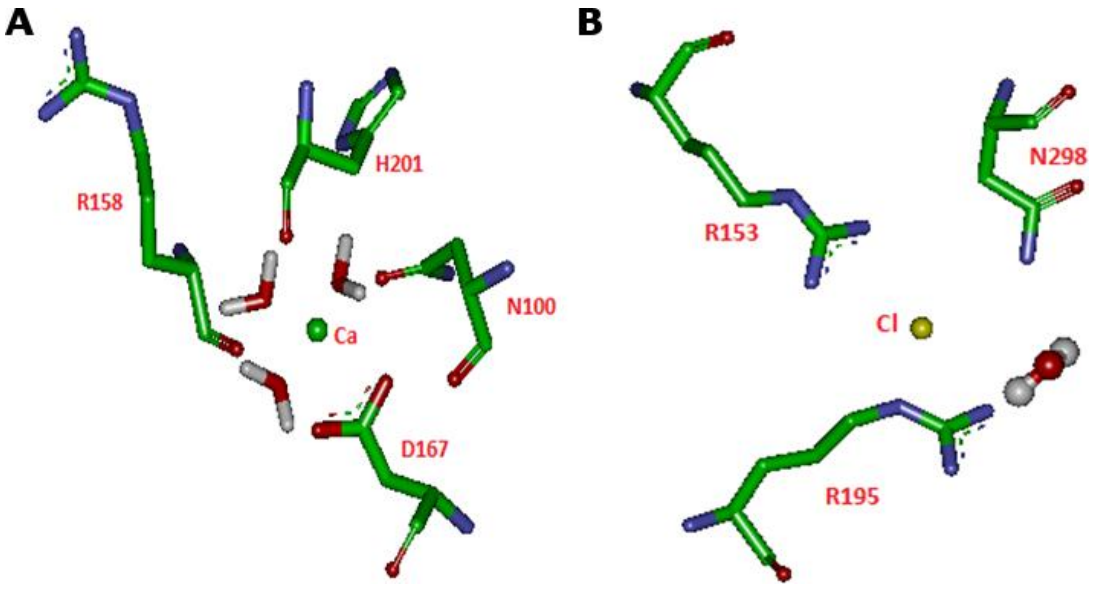
Third profile has only bacterial α-amylases from *L. plantarum, H. orenii and T. vulgaris* and is quiet distinct from the reference phylogenetic tree. This profile lacks insertion region between 149-158 which is present in all other species in the structure based sequence alignment. Common insertion regions specific to this profile were not seen. Intriguingly, in the sequence based phylogenetic tree these organisms are in the same clade as *A. oryzae* and *A. niger*. The locations of these specific regions mentioned above are shown in Figure A.2.

A.3.5 Homology of ion binding sites

All α-amylases have at least one Ca^{2+} binding site which is involved in preserving the structural integrity of the active site by linking the two fragments *i.e.* the catalytic $(\beta/\alpha)_8$ barrel and domain B (Brayer et al., 1995) of the enzyme, thus organizing and stabilizing the structure of domain B. From our analysis we observed that the α-amylases are highly homologous in terms of the binding of Ca^{2+} near the active site and the conformations of active site residues irrespective of their origin. In spite of the presence of differences in the identities of some of the amino acids involved, the positioning and binding of ligands within this Ca^{2+} binding site are almost identical in all the proteins suggesting that Ca^{2+} binding has similar roles in the animal, bacterial and fungal enzymes. The Ca^{2+} binding site is located between the central barrel and domain B. This cation is coordinated by eight ligands: the side chain carboxylate oxygen atoms of Asp167 in a bidentate mode, the main chain carbonyl oxygen of Arg158 and His201, the side chain carbonyl oxygen of Asn100 and three

water molecules as shown in Figure A.7A. We observed that this Ca^{2+} coordination is conserved in all α-amylases in spite of some amino acid residue mutations at the specific positions. In the PDB_ID: 1UD3, KSM-K38 strain *Bacillus* species, the Ca^{2+} is replaced by Na^{+} and the protein retained its activity (Nonaka et al., 2003). This implies that Ca^{2+} binding is only a structural feature and not a necessary condition for the amylase activity. It might be possible that the cluster of negative charge from the side chain and main chain of amino acids at this region is compensated by positively charged cation such as Ca^{2+} and Na^{+}. It has been reported that additional Ca^{2+} binding sites are found in the plant and fungal α-amylases (Boel et al., 1990; Kadziola et al., 1994), but they are absent in the animal enzymes, implicating that they do not perform any primary role in the function of α- amylases.

Figure A.7: (A) Coordination of Ca^{2+} binding in α-amylase structures. (B) Coordination of Cl^{-} binding in α-amylase structures.



Although the binding of a Ca^{2+} is considered to be a more general feature common for several different enzyme specificities from the α-amylase family, the binding of a Cl^{-} seems to be a feature characteristic of few α-amylases only. Previous studies identified the Cl^{-} dependent α-amylases in all animals (vertebrates and invertebrates) as well as in some Gram-negative bacteria (D'Amico et al., 2000; Feller et al., 1996). The Cl^{-} binding site is present at center of the $(\beta/\alpha)_8$ barrel close to the catalytic and Ca^{2+} binding sites. Here, we have observed the presence of Cl^{-} in α-amylases from *H. sapiens*, *T. molitor*, *S. scrofa* and *P. haloplanktis*. These organisms belong to the cluster I of profile one. The psychrophilic

P. haloplanktis α-amylase has been found to be structurally closely related to the eukaryotes based on phylogenetic trees. The Cl^- is coordinated by six ligands: in a unidentate mode to Asn298 (ND2); in a bidentate mode to Arg195 (NH2 and NE) and Arg337 (NH1 and NH2) and a water molecule as shown in Figure A.7B. Further, we analyzed the residues associated with the binding of Cl^- and found that side chains and main chains of Arg195, Asn298 and Arg337 (PDB_ID: 4GQR) are responsible for compensating the negative charge of Cl^-. These Cl^- binding residues are highly conserved in all higher eukaryotes but in *P. haloplanktis* α-amylase (PDB_ID: 1G94), Arg337 is substituted by a Lys300 thus providing only a unidentate coordination *via* its side chain Nζ. Several specialized α-amylases having Cl^- binding site are allosterically activated by this anion (Aghajari et al., 1998a). Absence of Cl^- binding in bacteria infers that the development of enzymatic control afforded by Cl^- is an evolutionarily recent event that does not seem to play a role in enzymes from lower organisms.

The presence of disulfide linkages prevents the site from developing ionic interactions. Importantly, they enhance the conformational stability of a protein by decreasing the flexibility and entropy of the unfolded state (Pace et al., 1988). Five disulfide bonds are observed in vertebrates, four in α-amylases from *A. niger, A. oryzae*, *P. halopalnktis* and *T. molitor*, and only one disulfide bond is present in archaebacteria *P. woesei* (thermophilic and anaerobic). The location of these disulfide bonds in α-amylase structures is shown in Figure A.8. In pancreatic human α-amylase, five disulfide bonds are present connecting cysteines at positions 28–86, 141–160, 378–384, 70-115 and 450–462 (PDB_ID: 4GQR). The α-amylase from *T. molitor* (PDB_ID: 1JAE) and *P. haloplanktis* (PDB_ID: 1G94) have four disulfides each that superimpose on the human α-amylases and the 70-115 connecting disulfide bond is missing. The function of this bond (70-115) has been related to temperature adaptation as it restricts thermal motion around the active site (Aghajari et al., 1998b). The fungal α-amylases also have four disulfide bonds, of which two of them overlay with the human α-amylases (450–462 and 141–160), two disulfide bonds at different positions that connect cysteines at 30-38 and 240-283 (PDB_ID: 2AAA). The *P. woesei* α-amylase has a single disulfide bond at positions 388-432 at an altogether different location. Based on the above, we believe that all Cl^- dependent α-amylases retain eight conserved cysteines that form four disulfide bonds in all eukaryotes as well

as psycrophilic bacterium *P. haloplanktis*. Further, the formation of disulfide bonds is not an essential criteria for the structural stability of α-amylases.

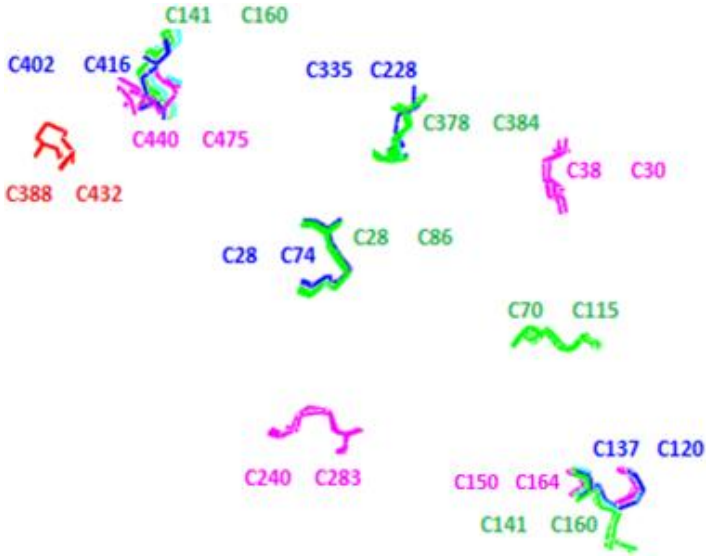
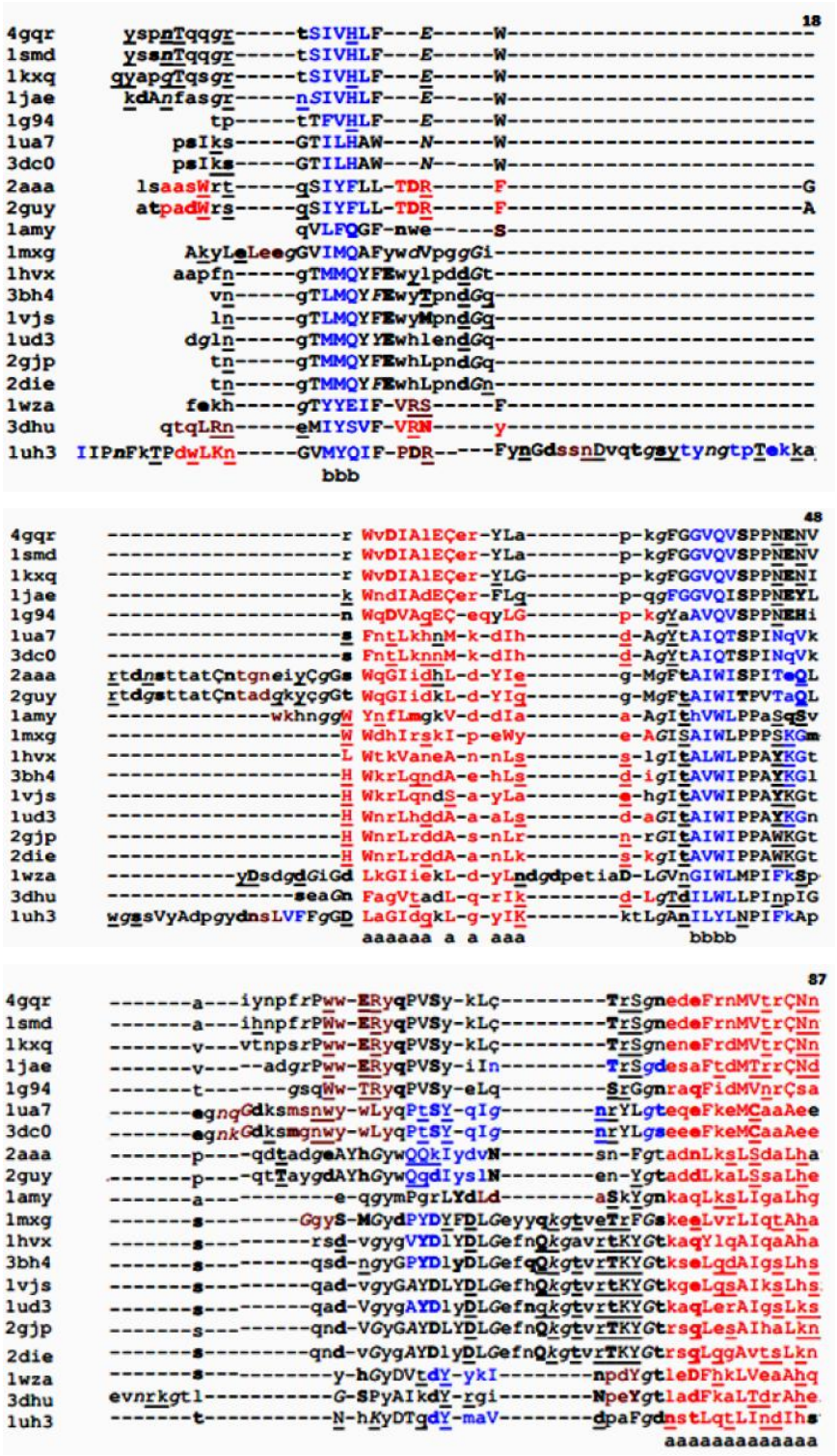


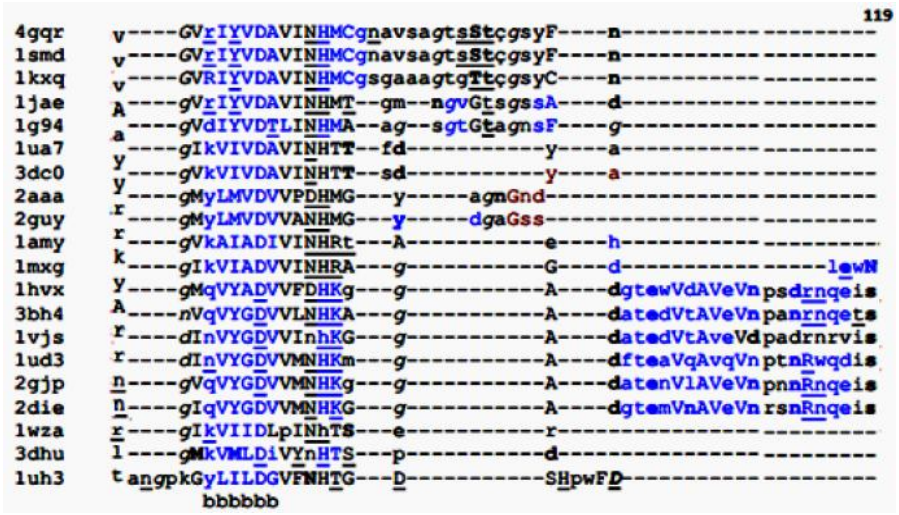
Figure A.8: Superimposition of the location of disulfide bonds in α-amylase structures. Colour coding is green: *H. sapiens* and *S. scrofa.* pink: *Aspergillus sp*. red: *P. woesei*, blue: *T. molitor* and *P. haloplanktis*.

A.3.6 Joy analysis

The Joy output represents the 3-D structure of a protein according to the secondary structures and additional features such as solvent accessibility and inaccessibility, hydrogen bond to main chain amide or main chain carbonyl and disulfide bond. The JOY structural alignment of all α-amylases is shown in Figure A.9. Out of the three active site residues, Asp196 is in solvent inaccessible region while the other two residues Glu232 and Asp299 reside in the solvent accessible region.

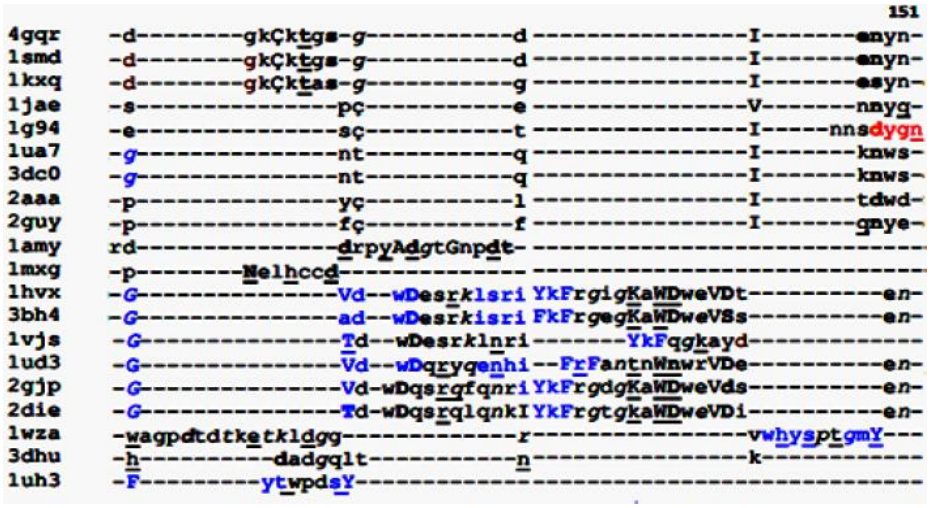
In all the α-amylases, it can be seen that all β-strands from $(\beta/\alpha)_8$ barrel are solvent inaccessible. As shown in Figure A.9, six proteins, belonging to the cluster II of profile two have insertion regions (Ala110 to Gln133 and Trp164 to Ser189) (PDB_ID: 3BH4) forming a unique three stranded antiparallel β-strand. A long insertion (Pro345 to Thr376) in proteins of the profile one, cluster I shown in Figure A.9, is present on the $(\beta/\alpha)_8$ barrel. Another insertion is present as a beta hairpin turn from Gly455 to Gly464. This cluster also has an insertion region Leu217 to Ser226 that is present only in species such as *H. sapiens*, *S. scrofa* and *T. molitor*. This can be considered as a unique sequence motif present only in animal kingdom. The proteins in profile one, cluster II has a unique insertion region from Asp185 to Gln192 (PDB_ID: 1UA7) that forms a loop region connected to α-helix.

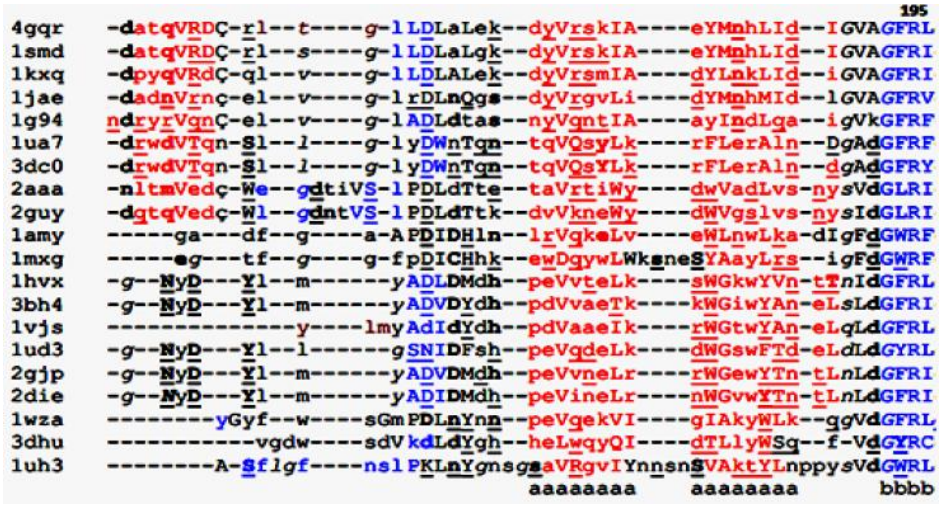


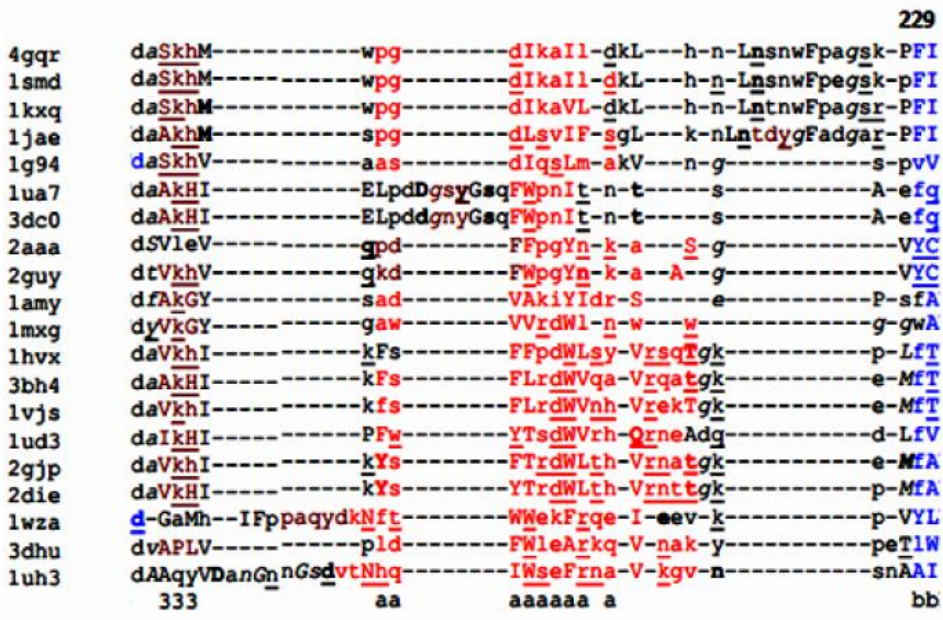

119
4gqr
1smd
1kxq
1jae
1g94
1ua7
3dc0
2aaa
2guy
1amy
1mxg
1hvx
3bh4
1vjs
1ud3
2gjp
2die
1wza
3dhu
1uh3
bbbbbb

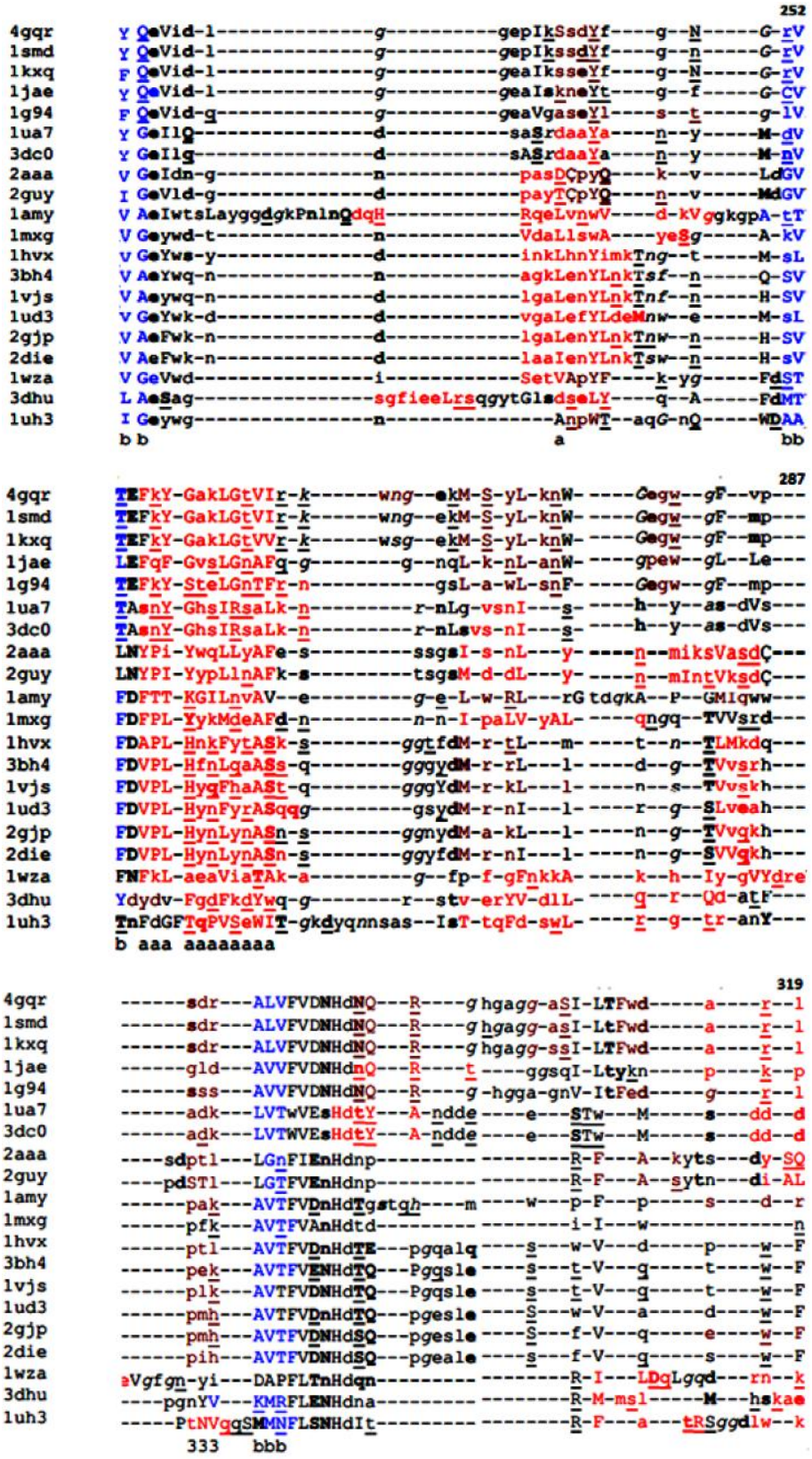

136
4gqr
1smd
1kxq
1jae
1g94
1ua7
3dc0
2aaa
2guy
1amy
1mxg
1hvx
3bh4
1vjs
1ud3
2gjp
2die
1wza
3dhu
1uh3
33

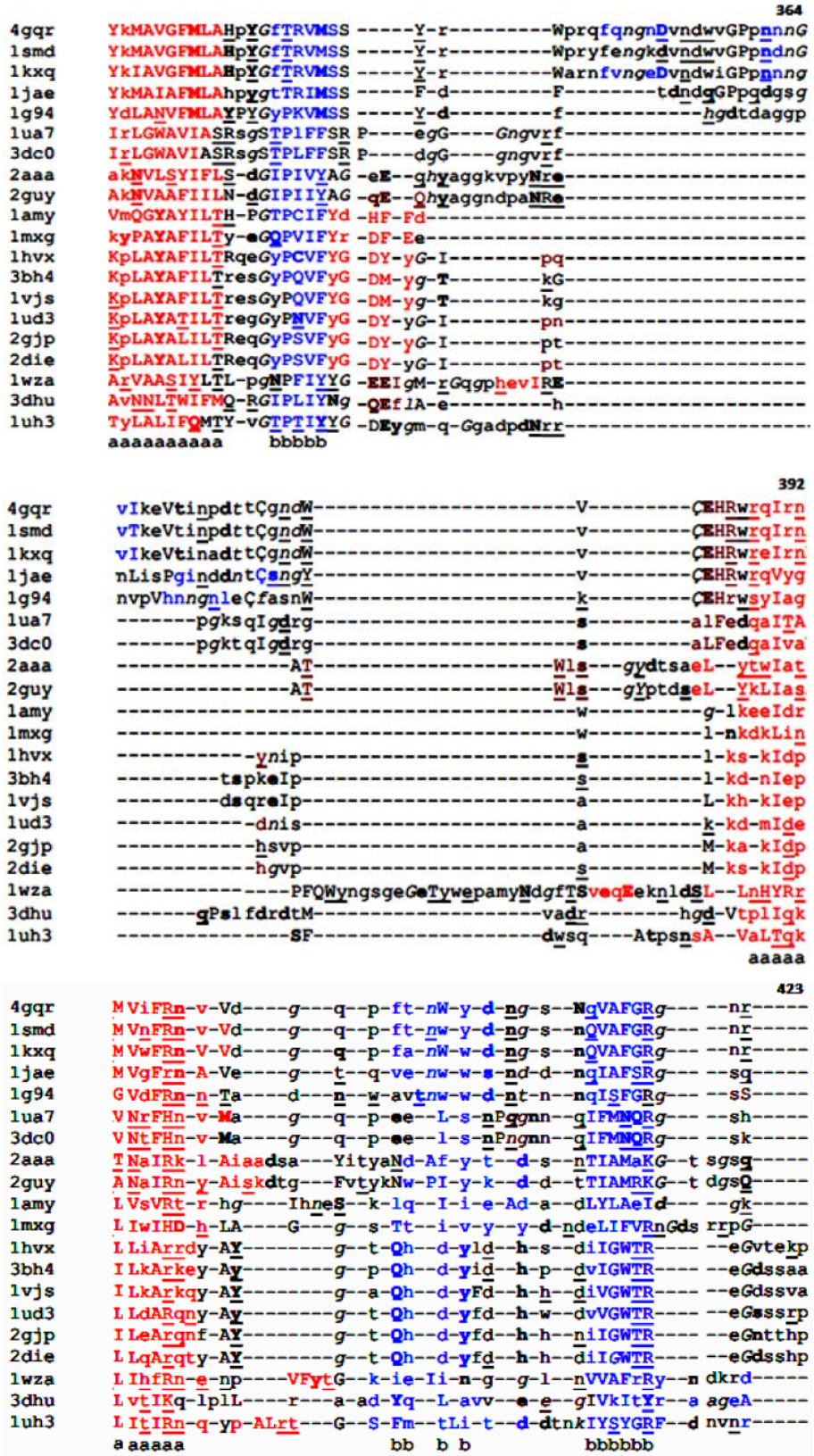

151
4gqr
1smd
1kxq
1jae
1g94
1ua7
3dc0
2aaa
2guy
1amy
1mxg
1hvx
3bh4
1vjs
1ud3
2gjp
2die
1wza
3dhu
1uh3











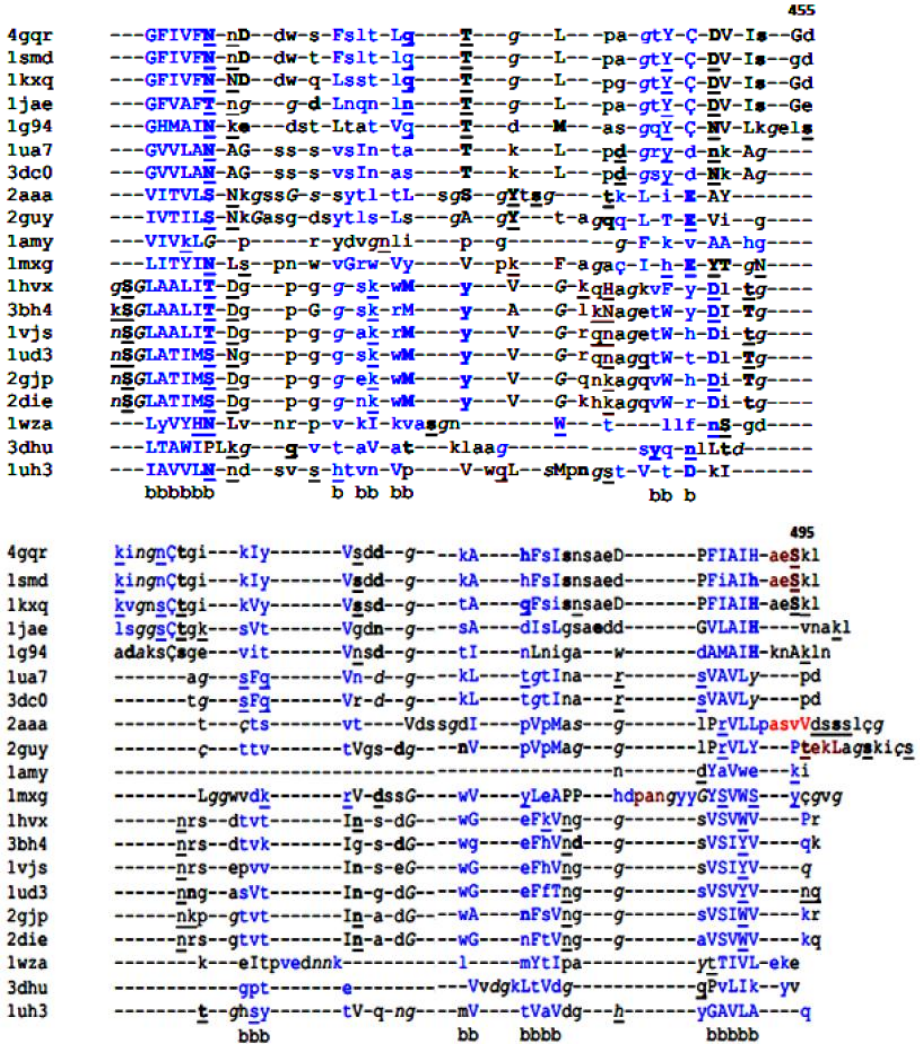


Figure A.9: Joy output representing different properties like secondary structures (red for α-helix, blue for β-strand and maroon for 3_{10} helix) along with other features like solvent accessibility (lowercase) and inaccessibility (uppercase), hydrogen bond to main chain amide (bold) or main chain carbonyl (underline).

A.4 Conclusions

A comprehensive structure and sequence based analyses was carried out to gain insights into the evolution of α-amylases using all known representative and experimentally determined X-ray crystal structures. As anticipated, the catalytic triad residues Asp197, Glu233 and Asp300 are identical across all species in both structure and sequence alignments. The structure based alignment was used to investigate the degree of conservation or variability. The phylogenetic trees of α-amylase differed from the NCBI reference tree. For example, the α-amylase of *P. halopanktis* of bacterial origin was promoted close to eukaryotic class and the α-amylase from *H. vulgare* of the eukaryotic class got demoted to the bacterial clade in the α- amylase phylogenetic tree. These variations were explained based on the regions of insertions and conserved regions in their respective profiles. We demonstrate that the Cl^- binding site is responsible for grouping the *P. halopanktis* with *H. sapiens*, *S. scrofa* and *A. oryzae.* Analysis of the Ca^{2+} binding site indicated that the cation binding may not be a necessary criterion for enzyme activity, but the cluster of negative charges from the side chain and main chain of amino acids is stabilised by positively charged cation such as Ca^{2+} and Na^+. Our analysis provides a detailed sequence and structural insights into the evolution of α-amylases that have not been studied before.

(1994). Schistosomes, liver flukes and Helicobacter pylori. IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. Lyon, 7-14 June 1994. IARC Monogr Eval Carcinog Risks Hum *61*, 1-241.

Aghajari, N., Feller, G., Gerday, C., and Haser, R. (1998a). Crystal structures of the psychrophilic alpha-amylase from Alteromonas haloplanctis in its native form and complexed with an inhibitor. Protein Sci *7*, 564-572.

Aghajari, N., Feller, G., Gerday, C., and Haser, R. (1998b). Structures of the psychrophilic Alteromonas haloplanctis alpha-amylase give insights into cold adaptation at a molecular level. Structure *6*, 1503-1516.

Ahmad, I., and Rao, D.N. (1996a). Chemistry and biology of DNA methyltransferases. Crit Rev Biochem Mol Biol *31*, 361-380.

Ahmad, I., and Rao, D.N. (1996b). Functional analysis of conserved motifs in EcoP15I DNA methyltransferase. J Mol Biol *259*, 229-240.

Algood, H.M., Gallo-Romero, J., Wilson, K.T., Peek, R.M., Jr., and Cover, T.L. (2007). Host response to Helicobacter pylori infection before initiation of the adaptive immune response. FEMS Immunol Med Microbiol *51*, 577-586.

Alison McCurdy, L.J., David A. Stauffer, Dennis A. Dougherty (1992). Biomimetic catalysis of SN2 reactions through cation–π interactions: the role of polarizability in catalysis J Am Chem Soc *114*, 10314-10321.

Allen, L.A., Schlesinger, L.S., and Kang, B. (2000). Virulent strains of Helicobacter pylori demonstrate delayed phagocytosis and stimulate homotypic phagosome fusion in macrophages. J Exp Med *191*, 115-128.

Alm, R.A., Ling, L.S., Moir, D.T., King, B.L., Brown, E.D., Doig, P.C., Smith, D.R., Noonan, B., Guild, B.C., deJonge, B.L*., et al.* (1999). Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen Helicobacter pylori. Nature *397*, 176-180.

Alm, R.A., and Trust, T.J. (1999). Analysis of the genetic diversity of Helicobacter pylori: the tale of two genomes. Journal of molecular medicine *77*, 834-846.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J Mol Biol *215*, 403-410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res *25*, 3389-3402.

Ando, T., Peek, R.M., Pride, D., Levine, S.M., Takata, T., Lee, Y.C., Kusugami, K., van der Ende, A., Kuipers, E.J., Kusters, J.G.*, et al.* (2002). Polymorphisms of Helicobacter pylori HP0638 reflect geographic origin and correlate with cagA status. Journal of clinical microbiology *40*, 239-246.

Argos, P., Hanei, M., Wilson, J.M., and Kelley, W.N. (1983). A possible nucleotide-binding domain in the tertiary fold of phosphoribosyltransferases. The Journal of biological chemistry *258*, 6450-6457.

Arnold, I.C., Dehzad, N., Reuter, S., Martin, H., Becher, B., Taube, C., and Muller, A. (2011). Helicobacter pylori infection prevents allergic asthma in mouse models through the induction of regulatory T cells. J Clin Invest *121*, 3088-3093.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T.*, et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet *25*, 25-29.

Atherton, J.C. (2006). The pathogenesis of Helicobacter pylori-induced gastro-duodenal diseases. Annu Rev Pathol *1*, 63-96.

Atherton, J.C., Cao, P., Peek, R.M., Jr., Tummuru, M.K., Blaser, M.J., and Cover, T.L. (1995). Mosaicism in vacuolating cytotoxin alleles of Helicobacter pylori. Association of specific vacA types with cytotoxin production and peptic ulceration. J Biol Chem *270*, 17771-17777.

Baker, D.J., Kan, J.L., and Smith, S.S. (1988). Recognition of structural perturbations in DNA by human DNA(cytosine-5)methyltransferase. Gene *74*, 207-210.

Baker, P.J., Hraba, T., Taylor, C.E., Stashak, P.W., Fauntleroy, M.B., Zahringer, U., Takayama, K., Sievert, T.R., Hronowski, X., Cotter, R.J.*, et al.* (1994). Molecular structures that influence the immunomodulatory properties of the lipid A and inner core region oligosaccharides of bacterial lipopolysaccharides. Infect Immun *62*, 2257-2269.

Bayle, D., Wangler, S., Weitzenegger, T., Steinhilber, W., Volz, J., Przybylski, M., Schafer, K.P., Sachs, G., and Melchers, K. (1998). Properties of the P-type ATPases encoded by the copAP operons of Helicobacter pylori and Helicobacter felis. J Bacteriol *180*, 317-329.

Baylin, S.B., and Herman, J.G. (2000). DNA hypermethylation in tumorigenesis: epigenetics joins genetics. Trends Genet *16*, 168-174.

Baylin, S.B., and Jones, P.A. (2011). A decade of exploring the cancer epigenome - biological and translational implications. Nature reviews Cancer *11*, 726-734.

Behnsen, J., Jellbauer, S., Wong, C.P., Edwards, R.A., George, M.D., Ouyang, W., and Raffatellu, M. (2014). The cytokine IL-22 promotes pathogen colonization by suppressing related commensal bacteria. Immunity *40*, 262-273.

Berendsen, H.J.C., van der Spoel, D., and van Drunen, R. (1994). GROMACS: A message-passing parallel molecular dynamics implementation. Computer physics communications *91*, 43 -56.

Bhattacharya, A., Wunderlich, Z., Monleon, D., Tejero, R., and Montelione, G.T. (2008). Assessing model accuracy using the homology modeling automatically software. Proteins *70*, 105-118.

Bickle, T.A., and Kruger, D.H. (1993). Biology of DNA restriction. Microbiol Rev *57*, 434-450.

Blanchard, T.G., and Czinn, S.J. (1998). Review article: Immunological determinants that may affect the Helicobacter pylori cancer risk. Aliment Pharmacol Ther *12 Suppl 1*, 83-90.

Boeckmann, B., Blatter, M.C., Famiglietti, L., Hinz, U., Lane, L., Roechert, B., and Bairoch, A. (2005). Protein variety and functional diversity: Swiss-Prot annotation in its biological context. C R Biol *328*, 882-899.

Boel, E., Brady, L., Brzozowski, A.M., Derewenda, Z., Dodson, G.G., Jensen, V.J., Petersen, S.B., Swift, H., Thim, L., and Woldike, H.F. (1990). Calcium binding in alpha-amylases: an X-ray diffraction study at 2.1-A resolution of two enzymes from Aspergillus. Biochemistry *29*, 6244-6249.

Boneca, I.G., de Reuse, H., Epinat, J.C., Pupin, M., Labigne, A., and Moszer, I. (2003). A revised annotation and comparative analysis of Helicobacter pylori genomes. Nucleic Acids Res *31*, 1704-1714.

Boren, T., Falk, P., Roth, K.A., Larson, G., and Normark, S. (1993). Attachment of Helicobacter pylori to human gastric epithelium mediated by blood group antigens. Science *262*, 1892-1895.

Bork, P., Dandekar, T., Diaz-Lazcoz, Y., Eisenhaber, F., Huynen, M., and Yuan, Y. (1998). Predicting function: from genes to genomes and back. J Mol Biol *283*, 707-725.

Bowie, J.U., Luthy, R., and Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. Science *253*, 164-170.

Brayer, G.D., Luo, Y., and Withers, S.G. (1995). The structure of human pancreatic alpha-amylase at 1.8 A resolution and comparisons with related enzymes. Protein Sci *4*, 1730-1742.

Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., and Karplus, M. (1983). CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem *4*, 187-217.

Buisson, G., Duee, E., Haser, R., and Payan, F. (1987). Three dimensional structure of porcine pancreatic alpha-amylase at 2.9 A resolution. Role of calcium in structure and activity. Embo J *6*, 3909-3916.

Bujnicki, J.M. (2002). Sequence permutations in the molecular evolution of DNA methyltransferases. BMC Evol Biol *2*, 3.

Bujnicki, J.M., and Radlinska, M. (1999). Molecular evolution of DNA-(cytosine-N4) methyltransferases: evidence for their polyphyletic origin. Nucleic Acids Res *27*, 4501-4509.

Bujnicki, J.M., Radlinska, M., Zaleski, P., and Piekarowicz, A. (2001). Cloning of the Haemophilus influenzae Dam methyltransferase and analysis of its relationship to the Dam methyltransferase encoded by the HP1 phage. Acta Biochim Pol *48*, 969-983.

Bussi, G., Donadio, D., and Parrinello, M. (2007). Canonical sampling through velocity rescaling. J Chem Phys *126*.

Camorlinga-Ponce, M., Perez-Perez, G., Gonzalez-Valencia, G., Mendoza, I., Penaloza-Espinosa, R., Ramos, I., Kersulyte, D., Reyes-Leon, A., Romo, C., Granados, J.*, et al.* (2011). Helicobacter pylori genotyping from American indigenous groups shows novel Amerindian vacA and cagA alleles and Asian, African and European admixture. PLoS One *6*, e27212.

Cao, P., Lee, K.J., Blaser, M.J., and Cover, T.L. (2005). Analysis of hopQ alleles in East Asian and Western strains of Helicobacter pylori. FEMS Microbiol Lett *251*, 37-43.

Cao, X., Tsukamoto, T., Seki, T., Tanaka, H., Morimura, S., Cao, L., Mizoshita, T., Ban, H., Toyoda, T., Maeda, H.*, et al.* (2008). 4-Vinyl-2,6-dimethoxyphenol (canolol) suppresses oxidative stress and gastric carcinogenesis in Helicobacter pylori-infected carcinogen-treated Mongolian gerbils. Int J Cancer *122*, 1445-1454.

Carbon, S., Ireland, A., Mungall, C.J., Shu, S., Marshall, B., and Lewis, S. (2009). AmiGO: online access to ontology and annotation data. Bioinformatics *25*, 288-289.

Casadesus, J., and Low, D. (2006). Epigenetic gene regulation in the bacterial world. Microbiol Mol Biol Rev *70*, 830-856.

Case, D.A., Cheatham, T.E., 3rd, Darden, T., Gohlke, H., Luo, R., Merz, K.M., Jr., Onufriev, A., Simmerling, C., Wang, B., and Woods, R.J. (2005). The Amber biomolecular simulation programs. J Comput Chem *26*, 1668-1688.

Celli, J.P., Turner, B.S., Afdhal, N.H., Keates, S., Ghiran, I., Kelly, C.P., Ewoldt, R.H., McKinley, G.H., So, P., Erramilli, S*., et al.* (2009). Helicobacter pylori moves through mucus by reducing mucin viscoelasticity. Proc Natl Acad Sci U S A *106*, 14321-14326.

Censini, S., Lange, C., Xiang, Z., Crabtree, J.E., Ghiara, P., Borodovsky, M., Rappuoli, R., and Covacci, A. (1996). cag, a pathogenicity island of Helicobacter pylori, encodes type I-specific and disease-associated virulence factors. Proc Natl Acad Sci U S A *93*, 14648-14653.

Chen, L., MacMillan, A.M., Chang, W., Ezaz-Nikpay, K., Lane, W.S., and Verdine, G.L. (1991). Direct identification of the active-site nucleophile in a DNA (cytosine-5)-methyltransferase. Biochemistry *30*, 11018-11025.

Chen, L., Macmillan, A.M., and Verdine, G.L. (1993a). Mutational Separation of DNA-Binding from Catalysis in a DNA Cytosine Methyltransferase. J Am Chem Soc *115*, 5318-5319.

Chen, L., MacMillan, A.M., and Verdine, G.L. (1993b). Mutational separation of DNA binding from catalysis in a DNA cytosine methyltransferase. J Am Chem Soc *115* 5318–5319.

Cheng, X. (1995). Structure and function of DNA methyltransferases. Annu Rev Biophys Biomol Struct *24*, 293-318.

Cheng, X., Kumar, S., Klimasauskas, S., and Roberts, R.J. (1993a). Crystal structure of the HhaI DNA methyltransferase. Cold Spring Harb Symp Quant Biol *58*, 331-338.

Cheng, X., Kumar, S., Posfai, J., Pflugrath, J.W., and Roberts, R.J. (1993b). Crystal structure of the HhaI DNA methyltransferase complexed with S-adenosyl-L-methionine. Cell *74*, 299-307.

Chow, C.S., Lamichhane, T.N., and Mahto, S.K. (2007). Expanding the nucleotide repertoire of the ribosome with post-transcriptional modifications. ACS chemical biology *2*, 610-619.

Clark, W.T., and Radivojac, P. (2011). Analysis of protein function and its prediction from amino acid sequence. Proteins *79*, 2086-2096.

Contreras, M., Thiberge, J.M., Mandrand-Berthelot, M.A., and Labigne, A. (2003). Characterization of the roles of NikR, a nickel-responsive pleiotropic autoregulator of Helicobacter pylori. Mol Microbiol *49*, 947-963.

Cooksley, C., Jenks, P.J., Green, A., Cockayne, A., Logan, R.P., and Hardie, K.R. (2003). NapA protects Helicobacter pylori from oxidative stress damage, and its production is influenced by the ferric uptake regulator. J Med Microbiol *52*, 461-469.

Correa, P. (1992). Human gastric carcinogenesis: a multistep and multifactorial process--First American Cancer Society Award Lecture on Cancer Epidemiology and Prevention. Cancer Res *52*, 6735-6740.

Correa, P., and Piazuelo, M.B. (2008). Natural history of Helicobacter pylori infection. Dig Liver Dis *40*, 490-496.

Covacci, A., Censini, S., Bugnoli, M., Petracca, R., Burroni, D., Macchia, G., Massone, A., Papini, E., Xiang, Z., Figura, N.*, et al.* (1993). Molecular characterization of the 128-kDa immunodominant antigen of Helicobacter pylori associated with cytotoxicity and duodenal ulcer. Proc Natl Acad Sci U S A *90*, 5791-5795.

Cover, T.L., and Blaser, M.J. (1992). Purification and characterization of the vacuolating toxin from Helicobacter pylori. J Biol Chem *267*, 10570-10575.

Cover, T.L., and Blaser, M.J. (1996). Helicobacter pylori infection, a paradigm for chronic mucosal inflammation: pathogenesis and implications for eradication and prevention. Advances in internal medicine *41*, 85-117.

Cover, T.L., and Blaser, M.J. (2009). Helicobacter pylori in health and disease. Gastroenterology *136*, 1863-1873.

Curnow, A.W., Ibba, M., and Soll, D. (1996). tRNA-dependent asparagine formation. Nature *382*, 589-590.

D'Amico, S., Gerday, C., and Feller, G. (2000). Structural similarities and evolutionary relationships in chloride-dependent alpha-amylases. Gene *253*, 95-105.

Darden, T., York, D., and Pedersen, L. (1993). Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems. J Chem Phys *98*, 10089-10092.

Darii, M.V., Cherepanova, N.A., Subach, O.M., Kirsanova, O.V., Rasko, T., Slaska-Kiss, K., Kiss, A., Deville-Bonne, D., Reboud-Ravaux, M., and Gromova, E.S. (2009). Mutational analysis of the CG recognizing DNA methyltransferase SssI: Insight into enzyme-DNA interactions. Bba-Proteins Proteom *1794*, 1654-1662.

Decatur, W.A., and Fournier, M.J. (2002). rRNA modifications and ribosome function. Trends in biochemical sciences *27*, 344-351.

Devi, S.M., Ahmed, I., Francalacci, P., Hussain, M.A., Akhter, Y., Alvi, A., Sechi, L.A., Megraud, F., and Ahmed, N. (2007). Ancestral European roots of Helicobacter pylori in India. BMC Genomics *8*, 184.

Douillard, F.P., Ryan, K.A., Lane, M.C., Caly, D.L., Moore, S.A., Penn, C.W., Hinds, J., and O'Toole, P.W. (2010). The HP0256 gene product is involved in motility and cell envelope architecture of Helicobacter pylori. BMC Microbiol *10*, 106.

Dundas, J., Ouyang, Z., Tseng, J., Binkowski, A., Turpaz, Y., and Liang, J. (2006). CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. Nucleic Acids Res *34*, W116-118.

Dunn, B.E., Vakil, N.B., Schneider, B.G., Miller, M.M., Zitzer, J.B., Peutz, T., and Phadnis, S.H. (1997). Localization of Helicobacter pylori urease and heat shock protein in human gastric biopsies. Infect Immun *65*, 1181-1188.

Dutta, S., Burkhardt, K., Swaminathan, G.J., Kosada, T., Henrick, K., Nakamura, H., and Berman, H.M. (2008). Data deposition and annotation at the worldwide protein data bank. Methods in molecular biology *426*, 81-101.

Dutta, S., Burkhardt, K., Young, J., Swaminathan, G.J., Matsuura, T., Henrick, K., Nakamura, H., and Berman, H.M. (2009). Data deposition and annotation at the worldwide protein data bank. Mol Biotechnol *42*, 1-13.

Eramian, D., Shen, M.Y., Devos, D., Melo, F., Sali, A., and Marti-Renom, M.A. (2006). A composite score for predicting errors in protein structure models. Protein Sci *15*, 1653-1666.

Estabrook, R.A., Lipson, R., Hopkins, B., and Reich, N. (2004). The coupling of tight DNA binding and base flipping: identification of a conserved structural motif in base flipping enzymes. J Biol Chem *279*, 31419-31428.

Feller, G., Bussy, O., Houssier, C., and Gerday, C. (1996). Structural and functional aspects of chloride binding to Alteromonas haloplanctis alpha-amylase. J Biol Chem *271*, 23836-23841.

Ferey-Roux, G., Perrier, J., Forest, E., Marchis-Mouren, G., Puigserver, A., and Santimone, M. (1998). The human pancreatic alpha-amylase isoforms: isolation, structural studies and kinetics of inhibition by acarbose. Biochim Biophys Acta *1388*, 10-20.

Fischer, W., Breithaupt, U., Kern, B., Smith, S.I., Spicher, C., and Haas, R. (2014). A comprehensive analysis of Helicobacter pylori plasticity zones reveals that they are integrating conjugative elements with intermediate integration specificity. BMC Genomics *15*, 310.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M.*, et al.* (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science *269*, 496-512.

Foloppe, N., and MacKerell, A. (2000). All-atom empirical force field for nucleic acids: I. Parameter optimization based on small molecule and condensed phase macromolecular target data. J Comput Chem *21*, 86–104.

Francis, K., and Gadda, G. (2006). Probing the chemical steps of nitroalkane oxidation catalyzed by 2-nitropropane dioxygenase with solvent viscosity, pH, and substrate kinetic isotope effects. Biochemistry *45*, 13889-13898.

Friedberg, I. (2006). Automated protein function prediction--the genomic challenge. Brief Bioinform *7*, 225-242.

Friedrich, T., Roth, M., Helm-Kruse, S., and Jeltsch, A. (1998). Functional mapping of the EcoRV DNA methyltransferase by random mutagenesis and screening for catalytically inactive mutants. Biol Chem *379*, 475-480.

Furuta, Y., Namba-Fukuyo, H., Shibata, T.F., Nishiyama, T., Shigenobu, S., Suzuki, Y., Sugano, S., Hasebe, M., and Kobayashi, I. (2014). Methylome diversification through changes in DNA methyltransferase sequence specificity. PLoS Genet *10*, e1004272.

Gabbara, S., Sheluho, D., and Bhagwat, A.S. (1995). Cytosine Methyltransferase from Escherichia-Coli in Which Active-Site Cysteine Is Replaced with Serine Is Partially Active. Biochemistry *34*, 8914-8923.

Gaddy, J.A., Radin, J.N., Loh, J.T., Zhang, F., Washington, M.K., Peek, R.M., Jr., Algood, H.M., and Cover, T.L. (2013). High dietary salt intake exacerbates Helicobacter pylori-induced gastric carcinogenesis. Infect Immun *81*, 2258-2267.

Geourjon, C., Combet, C., Blanchet, C., and Deleage, G. (2001). Identification of related proteins with weak sequence identity using secondary structure information. Protein Sci *10*, 788-797.

Gherardini, P.F., Ausiello, G., Russell, R.B., and Helmer-Citterich, M. (2010). Modular architecture of nucleotide-binding pockets. Nucleic Acids Res *38*, 3809-3816.

Goedecke, K., Pignot, M., Goody, R.S., Scheidig, A.J., and Weinhold, E. (2001). Structure of the N6-adenine DNA methyltransferase M.TaqI in complex with DNA and a cofactor analog. Nat Struct Biol *8*, 121-125.

Gong, W., O'Gara, M., Blumenthal, R.M., and Cheng, X. (1997a). Structure of pvu II DNA-(cytosine N4) methyltransferase, an example of domain permutation and protein fold assignment. Nucleic Acids Res *25*, 2702-2715.

Gong, W.M., OGara, M., Blumenthal, R.M., and Cheng, X.D. (1997b). Structure of PvuII DNA (cytosine N4) methyltransferase, an example of domain permutation and protein fold assignment. Nucleic Acids Res *25*, 2702-2715.

Goodwin, C.S., McCulloch, R.K., Armstrong, J.A., and Wee, S.H. (1985). Unusual cellular fatty acids and distinctive ultrastructure in a new spiral bacterium (Campylobacter pyloridis) from the human gastric mucosa. J Med Microbiol *19*, 257-267.

Gorlatova, N., Tchorzewski, M., Kurihara, T., Soda, K., and Esaki, N. (1998). Purification, characterization, and mechanism of a flavin mononucleotide-dependent 2-nitropropane dioxygenase from Neurospora crassa. Appl Environ Microbiol *64*, 1029-1033.

Graham, D.Y. (1991). Helicobacter pylori: its epidemiology and its role in duodenal ulcer disease. J Gastroenterol Hepatol *6*, 105-113.

Ha, J.Y., Min, J.Y., Lee, S.K., Kim, H.S., Kim do, J., Kim, K.H., Lee, H.H., Kim, H.K., Yoon, H.J., and Suh, S.W. (2006). Crystal structure of 2-nitropropane dioxygenase complexed with FMN and substrate. Identification of the catalytic base. The Journal of biological chemistry *281*, 18660-18667.

Hansson, L.E., Nyren, O., Hsing, A.W., Bergstrom, R., Josefsson, S., Chow, W.H., Fraumeni, J.F., Jr., and Adami, H.O. (1996). The risk of stomach cancer in patients with gastric or duodenal ulcer disease. N Engl J Med *335*, 242-249.

Heintschel von Heinegg, E., Nalik, H.P., and Schmid, E.N. (1993). Characterisation of a Helicobacter pylori phage (HP1). J Med Microbiol *38*, 245-249.

Heithoff, D.M., Sinsheimer, R.L., Low, D.A., and Mahan, M.J. (1999). An essential role for DNA adenine methylation in bacterial virulence. Science *284*, 967-970.

Hellmig, S., Hampe, J., and Schreiber, S. (2003). Helicobacter pylori infection in Africa and Europe: enigma of host genetics. Gut *52*, 1799.

Henrissat, B., and Bairoch, A. (1996). Updating the sequence-based classification of glycosyl hydrolases. Biochem J *316 (Pt 2)*, 695-696.

Hess, B. (2009). GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. Abstr Pap Am Chem S *237*.

Hess, B., Bekker, H., Berendsen, H.J.C., and Fraaije, J.G.E.M. (1997). LINCS: A linear constraint solver for molecular simulations. J Comput Chem *18*, 1463-1472.

Holm, L., and Rosenstrom, P. (2010). Dali server: conservation mapping in 3D. Nucleic Acids Res *38*, W545-549.

Holz, B., Dank, N., Eickhoff, J.E., Lipps, G., Krauss, G., and Weinhold, E. (1999). Identification of the binding site for the extrahelical target base in N6-adenine DNA methyltransferases by photo-cross-linking with duplex oligodeoxyribonucleotides containing 5-iodouracil at the target position. J Biol Chem *274*, 15066-15072.

Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A., and Simmerling, C. (2006). Comparison of multiple Amber force fields and development of improved protein backbone parameters. Proteins *65*, 712-725.

Hove-Jensen, B., Harlow, K.W., King, C.J., and Switzer, R.L. (1986). Phosphoribosylpyrophosphate synthetase of Escherichia coli. Properties of the purified enzyme and primary structure of the prs gene. The Journal of biological chemistry *261*, 6765-6771.

Hughes, N.J., Chalk, P.A., Clayton, C.L., and Kelly, D.J. (1995). Identification of carboxylation enzymes and characterization of a novel four-subunit pyruvate:flavodoxin oxidoreductase from Helicobacter pylori. J Bacteriol *177*, 3953-3959.

Hunt, R.H. (1996). The role of Helicobacter pylori in pathogenesis: the spectrum of clinical outcomes. Scandinavian journal of gastroenterology Supplement *220*, 3-9.

Huttenhower, C., Hibbs, M., Myers, C., and Troyanskaya, O.G. (2006). A scalable method for integration and functional analysis of multiple microarray datasets. Bioinformatics *22*, 2890-2897.

Ilinkin, I., Ye, J., and Janardan, R. (2010). Multiple structure alignment and consensus identification for proteins. BMC Bioinformatics *11*, 71.

Ingrosso, D., Fowler, A.V., Bleibaum, J., and Clarke, S. (1989). Sequence of the D-aspartyl/L-isoaspartyl protein methyltransferase from human erythrocytes. Common sequence motifs for protein, DNA, RNA, and small molecule S-adenosylmethionine-dependent methyltransferases. J Biol Chem *264*, 20131-20139.

Israel, D.A., Salama, N., Arnold, C.N., Moss, S.F., Ando, T., Wirth, H.P., Tham, K.T., Camorlinga, M., Blaser, M.J., Falkow, S*., et al.* (2001). Helicobacter pylori strain-specific differences in

genetic content, identified by microarray, influence host inflammatory responses. The Journal of clinical investigation *107*, 611-620.

Janecek, S. (1997). alpha-Amylase family: molecular biology and evolution. Prog Biophys Mol Biol *67*, 67-97.

Jeltsch, A. (1999). Circular permutations in the molecular evolution of DNA methyltransferases. J Mol Evol *49*, 161-164.

Jeltsch, A. (2001). The cytosine N4-methyltransferase M.PvuII also modifies adenine residues. Biol Chem *382*, 707-710.

Jeltsch, A. (2002). Beyond Watson and Crick: DNA methylation and molecular enzymology of DNA methyltransferases. Chembiochem *3*, 274-293.

Jeltsch, A., Christ, F., Fatemi, M., and Roth, M. (1999a). On the substrate specificity of DNA methyltransferases. adenine-N6 DNA methyltransferases also modify cytosine residues at position N4. J Biol Chem *274*, 19538-19544.

Jeltsch, A., Roth, M., and Friedrich, T. (1999b). Mutational analysis of target base flipping by the EcoRV adenine-N6 DNA methyltransferase. J Mol Biol *285*, 1121-1130.

Jenab, M., Riboli, E., Ferrari, P., Sabate, J., Slimani, N., Norat, T., Friesen, M., Tjonneland, A., Olsen, A., Overvad, K.*, et al.* (2006). Plasma and dietary vitamin C levels and risk of gastric cancer in the European Prospective Investigation into Cancer and Nutrition (EPIC-EURGAST). Carcinogenesis *27*, 2250-2257.

Jones, D.T. (1999). GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol *287*, 797-815.

Jones, P.A., and Laird, P.W. (1999). Cancer epigenetics comes of age. Nat Genet *21*, 163-167.

Jonsson, A.B., Nyberg, G., and Normark, S. (1991). Phase variation of gonococcal pili by frameshift mutation in pilC, a novel gene for pilus assembly. Embo J *10*, 477-488.

Jorgensen, W.L., Maxwell, D.S., and Tirado-Rives, J. (1996). Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. J Am Chem Soc *118*, 11225–11236.

Jorgensen, W.L., and Tirado-Rives, J. (2005). Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. Proc Natl Acad Sci U S A *102*, 6665-6670.

Kadziola, A., Abe, J., Svensson, B., and Haser, R. (1994). Crystal and molecular structure of barley alpha-amylase. J Mol Biol *239*, 104-121.

Kagan, R.M., and Clarke, S. (1994). Widespread occurrence of three sequence motifs in diverse S-adenosylmethionine-dependent methyltransferases suggests a common structure for these enzymes. Arch Biochem Biophys *310*, 417-427.

Kato, S., Tsukamoto, T., Mizoshita, T., Tanaka, H., Kumagai, T., Ota, H., Katsuyama, T., Asaka, M., and Tatematsu, M. (2006). High salt diets dose-dependently promote gastric chemical carcinogenesis in Helicobacter pylori-infected Mongolian gerbils associated with a shift in mucin production from glandular to surface mucous cells. Int J Cancer *119*, 1558-1566.

Katsuya, Y., Mezaki, Y., Kubota, M., and Matsuura, Y. (1998). Three-dimensional structure of Pseudomonas isoamylase at 2.2 A resolution. J Mol Biol *281*, 885-897.

Kawabata-Shoda, E., Charvat, H., Ikeda, A., Inoue, M., Sawada, N., Iwasaki, M., Sasazuki, S., Shimazu, T., Yamaji, T., Kimura, H.*, et al.* (2015). Trends in cancer prognosis in a population-based cohort survey: can recent advances in cancer therapy affect the prognosis? Cancer Epidemiol *39*, 97-103.

Kelley, L.A., and Sternberg, M.J. (2009). Protein structure prediction on the Web: a case study using the Phyre server. Nat Protoc *4*, 363-371.

Kikuchi, S., Wada, O., Nakajima, T., Nishi, T., Kobayashi, O., Konishi, T., and Inaba, Y. (1995). Serum anti-Helicobacter pylori antibody and gastric carcinoma among young adults. Research Group on Prevention of Gastric Carcinoma among Young Adults. Cancer *75*, 2789-2793.

Kim, D.E., Chivian, D., and Baker, D. (2004). Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res *32*, W526-531.

Klimasauskas, S., Kumar, S., Roberts, R.J., and Cheng, X. (1994). HhaI methyltransferase flips its target base out of the DNA helix. Cell *76*, 357-369.

Klimasauskas, S., Timinskas, A., Menkevicius, S., Butkiene, D., Butkus, V., and Janulaitis, A. (1989). Sequence motifs characteristic of DNA[cytosine-N4]methyltransferases: similarity to adenine and cytosine-C5 DNA-methylases. Nucleic Acids Res *17*, 9823-9832.

Kodaman, N., Pazos, A., Schneider, B.G., Piazuelo, M.B., Mera, R., Sobota, R.S., Sicinschi, L.A., Shaffer, C.L., Romero-Gallo, J., de Sablet, T.*, et al.* (2014). Human and Helicobacter pylori coevolution shapes the risk of gastric disease. Proc Natl Acad Sci U S A *111*, 1455-1460.

Kong, H., Lin, L.F., Porter, N., Stickel, S., Byrd, D., Posfai, J., and Roberts, R.J. (2000). Functional analysis of putative restriction-modification system genes in the Helicobacter pylori J99 genome. Nucleic Acids Res *28*, 3216-3223.

Kong, H., and Smith, C.L. (1997). Substrate DNA and cofactor regulate the activities of a multi-functional restriction-modification enzyme, BcgI. Nucleic Acids Res *25*, 3687-3692.

Kovall, R.A., and Matthews, B.W. (1999). Type II restriction endonucleases: structural, functional and evolutionary relationships. Curr Opin Chem Biol *3*, 578-583.

Kraft, C., Stack, A., Josenhans, C., Niehus, E., Dietrich, G., Correa, P., Fox, J.G., Falush, D., and Suerbaum, S. (2006). Genomic changes during chronic Helicobacter pylori infection. J Bacteriol *188*, 249-254.

Krebes, J., Morgan, R.D., Bunk, B., Sproer, C., Luong, K., Parusel, R., Anton, B.P., Konig, C., Josenhans, C., Overmann, J*., et al.* (2014). The complex methylome of the human gastric pathogen Helicobacter pylori. Nucleic Acids Res *42*, 2415-2432.

Kukimoto-Niino, M., Shibata, R., Murayama, K., Hamana, H., Nishimoto, M., Bessho, Y., Terada, T., Shirouzu, M., Kuramitsu, S., and Yokoyama, S. (2005). Crystal structure of a predicted phosphoribosyltransferase (TT1426) from Thermus thermophilus HB8 at 2.01 A resolution. Protein science : a publication of the Protein Society *14*, 823-827.

Kumar, P., Kailasam, S., Chakraborty, S., and Bansal, M. (2014). MolBridge: a program for identifying nonbonded interactions in small molecules and biomolecular structures. J Appl Crystallogr *47*, 1772-1776.

Kumar, S., Cheng, X., Klimasauskas, S., Mi, S., Posfai, J., Roberts, R.J., and Wilson, G.G. (1994). The DNA (cytosine-5) methyltransferases. Nucleic Acids Res *22*, 1-10.

Kumar, S., Horton, J.R., Jones, G.D., Walker, R.T., Roberts, R.J., and Cheng, X. (1997). DNA containing 4'-thio-2'-deoxycytidine inhibits methylation by HhaI methyltransferase. Nucleic Acids Res *25*, 2773-2783.

Kuriki, T., Takata, H., Okada, S., and Imanaka, T. (1991). Analysis of the active center of Bacillus stearothermophilus neopullulanase. J Bacteriol *173*, 6147-6152.

Labahn, J., Granzin, J., Schluckebier, G., Robinson, D.P., Jack, W.E., Schildkraut, I., and Saenger, W. (1994). Three-dimensional structure of the adenine-specific DNA methyltransferase M.Taq I in complex with the cofactor S-adenosylmethionine. Proc Natl Acad Sci U S A *91*, 10957-10961.

Labigne, A., and de Reuse, H. (1996). Determinants of Helicobacter pylori pathogenicity. Infectious agents and disease *5*, 191-202.

Lacueva, J., Gallego, J., and Diaz-Gonzalez, J.A. (2010). Updating controversies on the multidisciplinary management of gastric cancer. Clin Transl Oncol *12*, 677-685.

Lahue, R.S., and Modrich, P. (1988). Methyl-directed DNA mismatch repair in Escherichia coli. Mutat Res *198*, 37-43.

Laskowski R A, M.M.W., Moss D S & Thornton J M (1993). PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Cryst *26*.

Laskowski, R.A., Watson, J.D., and Thornton, J.M. (2005). Protein function prediction using local 3D templates. J Mol Biol *351*, 614-626.

Lau, E.Y., and Bruice, T.C. (1999). Active site dynamics of the HhaI methyltransferase: insights from computer simulation. J Mol Biol *293*, 9-18.

Lauren, P. (1965). The Two Histological Main Types of Gastric Carcinoma: Diffuse and So-Called Intestinal-Type Carcinoma. An Attempt at a Histo-Clinical Classification. Acta Pathol Microbiol Scand *64*, 31-49.

Lauster, R., Trautner, T.A., and Noyer-Weidner, M. (1989). Cytosine-specific type II DNA methyltransferases. A conserved enzyme core with variable target-recognizing domains. J Mol Biol *206*, 305-312.

Lee, D., Redfern, O., and Orengo, C. (2007). Predicting protein function from sequence and structure. Nat Rev Mol Cell Biol *8*, 995-1005.

Lehoux, I.E., and Mitra, B. (1999). (S)-Mandelate dehydrogenase from Pseudomonas putida: mutations of the catalytic base histidine-274 and chemical rescue of activity. Biochemistry *38*, 9948-9955.

Lertsethtakarn, P., Ottemann, K.M., and Hendrixson, D.R. (2011). Motility and chemotaxis in Campylobacter and Helicobacter. Annu Rev Microbiol *65*, 389-410.

Li, C., Begum, A., Numao, S., Park, K.H., Withers, S.G., and Brayer, G.D. (2005). Acarbose rearrangement mechanism implied by the kinetic and structural analysis of human pancreatic alpha-amylase in complex with analogues and their elongated counterparts. Biochemistry *44*, 3347-3357.

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics *22*, 1658-1659.

Lindahl, E., Hess, B., and van der Spoel, D. (2001). GROMACS 3.0: a package for molecular simulation and trajectory analysis. Journal of Molecular Model *7*, 306–317.

Linz, B., Windsor, H.M., McGraw, J.J., Hansen, L.M., Gajewski, J.P., Tomsho, L.P., Hake, C.M., Solnick, J.V., Schuster, S.C., and Marshall, B.J. (2014). A mutation burst during the acute phase of Helicobacter pylori infection in humans and rhesus macaques. Nat Commun *5*, 4165.

Lobley, A., Sadowski, M.I., and Jones, D.T. (2009). pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination. Bioinformatics *25*, 1761-1767.

Lu, X.J., and Olson, W.K. (2003). 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. Nucleic Acids Res *31*, 5108-5121.

Luthy, R., Bowie, J.U., and Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. Nature *356*, 83-85.

MacGregor, E.A. (1988). Alpha-amylase structure and activity. J Protein Chem *7*, 399-415.

Macnab, R.M. (2004). Type III flagellar protein export and flagellar assembly. Biochim Biophys Acta *1694*, 207-217.

Madhusoodanan, U.K., and Rao, D.N. (2010). Diversity of DNA methyltransferases that recognize asymmetric target sequences. Crit Rev Biochem Mol Biol *45*, 125-145.

Malone, T., Blumenthal, R.M., and Cheng, X. (1995). Structure-guided analysis reveals nine sequence motifs conserved among DNA amino-methyltransferases, and suggests a catalytic mechanism for these enzymes. J Mol Biol *253*, 618-632.

Mandal, R.S., and Das, S. (2014). In silico approach towards identification of potential inhibitors of Helicobacter pylori DapE. J Biomol Struct Dyn, 1-14.

Marcus, E.A., and Scott, D.R. (2001). Cell lysis is responsible for the appearance of extracellular urease in Helicobacter pylori. Helicobacter *6*, 93-99.

Marti-Renom, M.A., Stuart, A.C., Fiser, A., Sanchez, R., Melo, F., and Sali, A. (2000). Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct *29*, 291-325.

Matin, A., Zychlinsky, E., Keyhan, M., and Sachs, G. (1996). Capacity of Helicobacter pylori to generate ionic gradients at low pH is similar to that of bacteria which grow under strongly acidic conditions. Infect Immun *64*, 1434-1436.

Matsuura, K., Ogawa, M., Kosaki, G., Minamiura, N., and Tamamoto, T. (1978). alpha-Amylase from human pancreatic juice as an electrophoretically pure isozyme. J Biochem *83*, 329-332.

Matsuura, Y., Kusunoki, M., Harada, W., and Kakudo, M. (1984). Structure and possible catalytic residues of Taka-amylase A. J Biochem *95*, 697-702.

McClain, M.S., Shaffer, C.L., Israel, D.A., Peek, R.M., Jr., and Cover, T.L. (2009). Genome sequence analysis of Helicobacter pylori strains associated with gastric ulceration and gastric cancer. BMC Genomics *10*, 3.

McColl, K.E. (1997). What remaining questions regarding Helicobacter pylori and associated diseases should be addressed by future research? View from Europe. Gastroenterology *113*, S158-162.

McGuffin, L.J., Bryson, K., and Jones, D.T. (2000). The PSIPRED protein structure prediction server. Bioinformatics *16*, 404-405.

Melchers, K., Weitzenegger, T., Buhmann, A., Steinhilber, W., Sachs, G., and Schafer, K.P. (1996). Cloning and membrane topology of a P type ATPase from Helicobacter pylori. J Biol Chem *271*, 446-457.

Mendz, G.L., and Hazell, S.L. (1995). Aminoacid utilization by Helicobacter pylori. Int J Biochem Cell Biol *27*, 1085-1093.

Messer, W., and Noyer-Weidner, M. (1988). Timing and targeting: the biological functions of Dam methylation in E. coli. Cell *54*, 735-737.

Meurer, L.N., and Bower, D.J. (2002). Management of Helicobacter pylori infection. Am Fam Physician *65*, 1327-1336.

Miremadi, A., Oestergaard, M.Z., Pharoah, P.D., and Caldas, C. (2007). Cancer genetics of epigenetic genes. Human molecular genetics *16 Spec No 1*, R28-49.

Misra, V., Pandey, R., Misra, S.P., and Dwivedi, M. (2014). Helicobacter pylori and gastric cancer: Indian enigma. World J Gastroenterol *20*, 1503-1509.

Mizuguchi, K., Deane, C.M., Blundell, T.L., Johnson, M.S., and Overington, J.P. (1998). JOY: protein sequence-structure representation and analysis. Bioinformatics *14*, 617-623.

Mizuno, C.S., Chittiboyina, A.G., Kurtz, T.W., Pershadsingh, H.A., and Avery, M.A. (2008). Type 2 diabetes and oral antihyperglycemic drugs. Curr Med Chem *15*, 61-74.

Mobley, D.L., Bayly, C.I., Cooper, M.D., Shirts, M.R., and Dill, K.A. (2009). Small molecule hydration free energies in explicit solvent: An extensive test of fixed-charge atomistic simulations. J Chem Theory Comput *5*, 350-358.

Mobley, D.L., Bayly, C.I., Cooper, M.D., Shirts, M.R., and Dill, K.A. (2015). Correction to Small Molecule Hydration Free Energies in Explicit Solvent: An Extensive Test of Fixed-Charge Atomistic Simulations. J Chem Theory Comput *11*, 1347.

Moodley, Y., Linz, B., Bond, R.P., Nieuwoudt, M., Soodyall, H., Schlebusch, C.M., Bernhoft, S., Hale, J., Suerbaum, S., Mugisha, L.*, et al.* (2012). Age of the association between Helicobacter pylori and man. PLoS Pathog *8*, e1002693.

Moran, A.P. (1996). The role of lipopolysaccharide in Helicobacter pylori pathogenesis. Aliment Pharmacol Ther *10 Suppl 1*, 39-50.

Moxon, E.R., Lenski, R.E., and Rainey, P.B. (1998). Adaptive evolution of highly mutable loci in pathogenic bacteria. Perspect Biol Med *42*, 154-155.

Mueller, A., O'Rourke, J., Chu, P., Chu, A., Dixon, M.F., Bouley, D.M., Lee, A., and Falkow, S. (2005). The role of antigenic drive and tumor-infiltrating accessory cells in the pathogenesis of helicobacter-induced mucosa-associated lymphoid tissue lymphoma. Am J Pathol *167*, 797-812.

Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol *247*, 536-540.

Nabieva, E., Jim, K., Agarwal, A., Chazelle, B., and Singh, M. (2005). Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. Bioinformatics *21 Suppl 1*, i302-310.

Nagashima, T., Tada, S., Kitamoto, K., Gomi, K., Kumagai, C., and Toda, H. (1992). Site-directed mutagenesis of catalytic active-site residues of Taka-amylase A. Biosci Biotechnol Biochem *56*, 207-210.

Nakamura, S., Yao, T., Aoyagi, K., Iida, M., Fujishima, M., and Tsuneyoshi, M. (1997). Helicobacter pylori and primary gastric lymphoma. A histopathologic and immunohistochemical analysis of 237 patients. Cancer *79*, 3-11.

Neely, R.K., and Roberts, R.J. (2008). The BsaHI restriction-modification system: cloning, sequencing and analysis of conserved motifs. BMC Mol Biol *9*, 48.

Nishide, T., Emi, M., Nakamura, Y., and Matsubara, K. (1984). Corrected sequences of cDNAs for human salivary and pancreatic alpha-amylases [corrected]. Gene *28*, 263-270.

Nishide, T., Nakamura, Y., Emi, M., Yamamoto, T., Ogawa, M., Mori, T., and Matsubara, K. (1986). Primary structure of human salivary alpha-amylase gene. Gene *41*, 299-304.

Nishizawa, T., and Suzuki, H. (2015). Gastric Carcinogenesis and Underlying Molecular Mechanisms: Helicobacter pylori and Novel Targeted Therapy. Biomed Res Int *2015*, 794378.

Nonaka, T., Fujihashi, M., Kita, A., Hagihara, H., Ozaki, K., Ito, S., and Miki, K. (2003). Crystal structure of calcium-free alpha-amylase from Bacillus sp. strain KSM-K38 (AmyK38) and its sodium ion binding sites. J Biol Chem *278*, 24818-24824.

Norvell, J.C., and Berg, J.M. (2007). Update on the protein structure initiative. Structure *15*, 1519-1522.

Noto, J.M., Gaddy, J.A., Lee, J.Y., Piazuelo, M.B., Friedman, D.B., Colvin, D.C., Romero-Gallo, J., Suarez, G., Loh, J., Slaughter, J.C*., et al.* (2013). Iron deficiency accelerates Helicobacter pylori-induced carcinogenesis in rodents and humans. J Clin Invest *123*, 479-492.

O'Gara, M., Horton, J.R., Roberts, R.J., and Cheng, X. (1998). Structures of HhaI methyltransferase complexed with substrates containing mismatches at the target base. Nat Struct Biol *5*, 872-877.

O'Gara, M., Klimasauskas, S., Roberts, R.J., and Cheng, X. (1996). Enzymatic C5-cytosine methylation of DNA: mechanistic implications of new crystal structures for HhaL methyltransferase-DNA-AdoHcy complexes. J Mol Biol *261*, 634-645.

Odenbreit, S., Till, M., and Haas, R. (1996). Optimized BlaM-transposon shuttle mutagenesis of Helicobacter pylori allows the identification of novel genetic loci involved in bacterial virulence. Mol Microbiol *20*, 361-373.

OGara, M., Roberts, R.J., and Cheng, X.D. (1996). A structural basis for the preferential binding of hemimethylated DNA by HhaI DNA methyltransferase. J Mol Biol *263*, 597-606.

Oh, J.D., Kling-Backhed, H., Giannakis, M., Xu, J., Fulton, R.S., Fulton, L.A., Cordum, H.S., Wang, C., Elliott, G., Edwards, J*., et al.* (2006). The complete genome sequence of a chronic atrophic gastritis Helicobacter pylori strain: evolution during disease progression. Proc Natl Acad Sci U S A *103*, 9999-10004.

Oostenbrink, C., Villa, A., Mark, A.E., and van Gunsteren, W.F. (2004). A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. J Comput Chem *25*, 1656-1676.

Pace, C.N., Grimsley, G.R., Thomson, J.A., and Barnett, B.J. (1988). Conformational stability and activity of ribonuclease T1 with zero, one, and two intact disulfide bonds. J Biol Chem *263*, 11820-11825.

Palframan, S.L., Kwok, T., and Gabriel, K. (2012). Vacuolating cytotoxin A (VacA), a key toxin for Helicobacter pylori pathogenesis. Front Cell Infect Microbiol *2*, 92.

Parkin, D.M., Bray, F.I., and Devesa, S.S. (2001). Cancer burden in the year 2000. The global picture. European journal of cancer *37 Suppl 8*, S4-66.

Parrinello, M., and Rahman, A. (1981). Polymorphic transitions in single crystals: A new molecular dynamics method. Journal of Applied Physics *52*, 7182-7190.

Parsonnet, J. (1996). Helicobacter pylori in the stomach--a paradox unmasked. N Engl J Med *335*, 278-280.

Parsonnet, J., Friedman, G.D., Orentreich, N., and Vogelman, H. (1997). Risk for gastric cancer in people with CagA positive or CagA negative Helicobacter pylori infection. Gut *40*, 297-301.

Parsonnet, J., Friedman, G.D., Vandersteen, D.P., Chang, Y., Vogelman, J.H., Orentreich, N., and Sibley, R.K. (1991). Helicobacter pylori infection and the risk of gastric carcinoma. N Engl J Med *325*, 1127-1131.

Paulsen, M., and Ferguson-Smith, A.C. (2001). DNA methylation in genomic imprinting, development, and disease. J Pathol *195*, 97-110.

Pearson, W.R. (2013). An introduction to sequence similarity ("homology") searching. Curr Protoc Bioinformatics *Chapter 3*, Unit3 1.

Peek, R.M., Jr. (2005). Pathogenesis of Helicobacter pylori infection. Springer seminars in immunopathology *27*, 197-215.

Perez, A., Luque, F.J., and Orozco, M. (2007). Dynamics of B-DNA on the microsecond time scale. J Am Chem Soc *129*, 14739-14745.

Peterson, S.N., Bailey, C.C., Jensen, J.S., Borre, M.B., King, E.S., Bott, K.F., and Hutchison, C.A., 3rd (1995). Characterization of repetitive DNA in the Mycoplasma genitalium genome: possible role in the generation of antigenic variation. Proc Natl Acad Sci U S A *92*, 11829-11833.

Peterson, W.L. (1991). Helicobacter pylori and peptic ulcer disease. N Engl J Med *324*, 1043-1048.

Pieper, U., Webb, B.M., Barkan, D.T., Schneidman-Duhovny, D., Schlessinger, A., Braberg, H., Yang, Z., Meng, E.C., Pettersen, E.F., Huang, C.C*., et al.* (2011). ModBase, a database of annotated comparative protein structure models, and associated resources. Nucleic Acids Res *39*, D465-474.

Plant, A.R., Clemens, R.M., Morgan, H.W., and Daniel, R.M. (1987). Active-site- and substrate-specificity of Thermoanaerobium Tok6-B1 pullulanase. Biochem J *246*, 537-541.

Pogolotti, A.L., Jr., Ono, A., Subramaniam, R., and Santi, D.V. (1988). On the mechanism of DNA-adenine methylase. J Biol Chem *263*, 7461-7464.

Posfai, J., Bhagwat, A.S., Posfai, G., and Roberts, R.J. (1989). Predictive motifs derived from cytosine methyltransferases. Nucleic Acids Res *17*, 2421-2435.

Pra, D., Rech Franke, S.I., Pegas Henriques, J.A., and Fenech, M. (2009). A possible link between iron deficiency and gastrointestinal carcinogenesis. Nutr Cancer *61*, 415-426.

Pues, H., Bleimling, N., Holz, B., Wolcke, J., and Weinhold, E. (1999). Functional roles of the conserved aromatic amino acid residues at position 108 (motif IV) and position 196 (motif VIII) in base flipping and catalysis by the N6-adenine DNA methyltransferase from Thermus aquaticus. Biochemistry *38*, 1426-1434.

Qian, J., Luscombe, N.M., and Gerstein, M. (2001). Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. Journal of molecular biology *313*, 673-681.

Ramachandran, G.N., Ramakrishnan, C., and Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. J Mol Biol *7*, 95-99.

Reinisch, K.M., Chen, L., Verdine, G.L., and Lipscomb, W.N. (1995). The crystal structure of HaeIII methyltransferase convalently complexed to DNA: an extrahelical cytosine and rearranged base pairing. Cell *82*, 143-153.

Resende, T., Correia, D.M., Rocha, M., and Rocha, I. (2013). Re-annotation of the genome sequence of Helicobacter pylori 26695. J Integr Bioinform *10*, 233.

Rhodes, D., Schwabe, J.W., Chapman, L., and Fairall, L. (1996). Towards an understanding of protein-DNA recognition. Philos Trans R Soc Lond B Biol Sci *351*, 501-509.

Roberts, D., Hoopes, B.C., McClure, W.R., and Kleckner, N. (1985). IS10 transposition is regulated by DNA adenine methylation. Cell *43*, 117-130.

Roberts, R.J. (1990). Restriction enzymes and their isoschizomers. Nucleic Acids Res *18 Suppl*, 2331-2365.

Roberts, R.J. (1995). On base flipping. Cell *82*, 9-12.

Roberts, R.J., and Cheng, X. (1998). Base flipping. Annual review of biochemistry *67*, 181-198.

Roberts, R.J., Vincze, T., Posfai, J., and Macelis, D. (2015). REBASE-a database for DNA restriction and modification: enzymes, genes and genomes. Nucleic Acids Res *43*, D298-D299.

Robertson, K.D., and Wolffe, A.P. (2000). DNA methylation in health and disease. Nat Rev Genet *1*, 11-19.

Rost, B., Liu, J., Nair, R., Wrzeszczynski, K.O., and Ofran, Y. (2003). Automatic prediction of protein function. Cellular and molecular life sciences : CMLS *60*, 2637-2650.

Roth, M., Helm-Kruse, S., Friedrich, T., and Jeltsch, A. (1998). Functional roles of conserved amino acid residues in DNA methyltransferases investigated by site-directed mutagenesis of the EcoRV adenine-N6-methyltransferase. J Biol Chem *273*, 17333-17342.

Roth, M., and Jeltsch, A. (2001). Changing the target base specificity of the EcoRV DNA methyltransferase by rational de novo protein-design. Nucleic Acids Res *29*, 3137-3144.

Roy, A., Yang, J., and Zhang, Y. (2012). COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. Nucleic Acids Res *40*, W471-477.

Sali, A., and Blundell, T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol *234*, 779-815.

Santi, D.V., Norment, A., and Garrett, C.E. (1984). Covalent bond formation between a DNA-cytosine methyltransferase and DNA containing 5-azacytosine. Proc Natl Acad Sci U S A *81*, 6993-6997.

Sassone-Corsi, M., and Raffatellu, M. (2013). A hydrogen boost for salmonella. Cell Host Microbe *14*, 603-604.

Sawan, C., Vaissiere, T., Murr, R., and Herceg, Z. (2008). Epigenetic drivers and genetic passengers on the road to cancer. Mutation research *642*, 1-13.

Scavetta, R.D., Thomas, C.B., Walsh, M.A., Szegedi, S., Joachimiak, A., Gumport, R.I., and Churchill, M.E. (2000). Structure of RsrI methyltransferase, a member of the N6-adenine beta class of DNA methyltransferases. Nucleic Acids Res *28*, 3950-3961.

Schluckebier, G., Kozak, M., Bleimling, N., Weinhold, E., and Saenger, W. (1997). Differential binding of S-adenosylmethionine S-adenosylhomocysteine and Sinefungin to the adenine-specific DNA methyltransferase M.TaqI. J Mol Biol *265*, 56-67.

Schluckebier, G., Labahn, J., Granzin, J., and Saenger, W. (1998). M.TaqI: possible catalysis via cation-pi interactions in N-specific DNA methyltransferases. Biol Chem *379*, 389-400.

Schluckebier, G., Labahn, J., Granzin, J., Schildkraut, I., and Saenger, W. (1995a). A model for DNA binding and enzyme action derived from crystallographic studies of the TaqI N6-adenine-methyltransferase. Gene *157*, 131-134.

Schluckebier, G., O'Gara, M., Saenger, W., and Cheng, X. (1995b). Universal catalytic domain structure of AdoMet-dependent methyltransferases. J Mol Biol *247*, 16-20.

Schramm, V.L., and Grubmeyer, C. (2004). Phosphoribosyltransferase mechanisms and roles in nucleic acid metabolism. Progress in nucleic acid research and molecular biology *78*, 261-304.

Schreiber, S., Konradt, M., Groll, C., Scheid, P., Hanauer, G., Werling, H.O., Josenhans, C., and Suerbaum, S. (2004). The spatial orientation of Helicobacter pylori in the gastric mucus. Proc Natl Acad Sci U S A *101*, 5024-5029.

Schubert, H.L., Blumenthal, R.M., and Cheng, X. (2003). Many paths to methyltransfer: a chronicle of convergence. Trends Biochem Sci *28*, 329-335.

Scott, D.R., Marcus, E.A., Wen, Y., Oh, J., and Sachs, G. (2007). Gene expression in vivo shows that Helicobacter pylori colonizes an acidic niche on the gastric surface. Proc Natl Acad Sci U S A *104*, 7235-7240.

Scott , W.R.P., Hünenberger , P.H., Tironi , I.G., Mark , A.E., Billeter , S.R., Fennen , J., Torda , A.E., Huber , T., Krüger , P., and van Gunsteren, W.F. (1999). The GROMOS Biomolecular Simulation Program Package. The Journal of physical chemistry A *103* 3596–3607.

Sharan, R., Ulitsky, I., and Shamir, R. (2007). Network-based prediction of protein function. Mol Syst Biol *3*, 88.

Sheh, A., Chaturvedi, R., Merrell, D.S., Correa, P., Wilson, K.T., and Fox, J.G. (2013). Phylogeographic origin of Helicobacter pylori determines host-adaptive responses upon coculture with gastric epithelial cells. Infect Immun *81*, 2468-2477.

Shen, M.Y., and Sali, A. (2006). Statistical potential for assessment and prediction of protein structures. Protein Sci *15*, 2507-2524.

Shi, J., Blundell, T.L., and Mizuguchi, K. (2001). FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. Journal of molecular biology *310*, 243-257.

Shieh, F.K., and Reich, N.O. (2007). AdoMet-dependent methyl-transfer: Glu119 is essential for DNA C5-cytosine methyltransferase M.HhaI. J Mol Biol *373*, 1157-1168.

Shieh, F.K., Youngblood, B., and Reich, N.O. (2006). The role of Arg165 towards base flipping, base stabilization and catalysis in M.HhaI. J Mol Biol *362*, 516-527.

Shimamura, J., Fridhandler, L., and Berk, J.E. (1976). Unusual isomaylase in cancer-associated hyperamylasemia. Cancer *38*, 2121-2126.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J*., et al.* (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol *7*, 539.

Singh, S., Guttula, P.K., and Guruprasad, L. (2014). Structure based annotation of Helicobacter pylori strain 26695 proteome. PLoS One *9*, e115020.

Skoglund, A., Bjorkholm, B., Nilsson, C., Andersson, A.F., Jernberg, C., Schirwitz, K., Enroth, C., Krabbe, M., and Engstrand, L. (2007). Functional analysis of the M.HpyAIV DNA methyltransferase of Helicobacter pylori. J Bacteriol *189*, 8914-8921.

Smith, G.R. (2012). How RecBCD enzyme and Chi promote DNA break repair and recombination: a molecular biologist's view. Microbiol Mol Biol Rev *76*, 217-228.

Smith, H.O., Annau, T.M., and Chandrasegaran, S. (1990). Finding sequence motifs in groups of functionally related proteins. Proc Natl Acad Sci U S A *87*, 826-830.

Smith, J.L. (1995). Enzymes of nucleotide synthesis. Current opinion in structural biology *5*, 752-757.

Sokolov, A., and Ben-Hur, A. (2010). Hierarchical classification of gene ontology terms using the GOstruct method. J Bioinform Comput Biol *8*, 357-376.

Sousa da Silva, A.W., and Vranken, W.F. (2012). ACPYPE - AnteChamber PYthon Parser interfacE. BMC Res Notes *5*, 367.

Specht, M., Schatzle, S., Graumann, P.L., and Waidner, B. (2011). Helicobacter pylori possesses four coiled-coil-rich proteins that form extended filamentous structures and control cell shape and motility. J Bacteriol *193*, 4523-4530.

Stephens, C., Reisenauer, A., Wright, R., and Shapiro, L. (1996). A cell cycle-regulated bacterial DNA methyltransferase is essential for viability. Proc Natl Acad Sci U S A *93*, 1210-1214.

Stiefel, D.J., and Keller, P.J. (1973). Preparation and some properties of human pancreatic amylase including a comparison with human parotid amylase. Biochim Biophys Acta *302*, 345-361.

Strom, M.S., Nunn, D.N., and Lory, S. (1994). Posttranslational processing of type IV prepilin and homologs by PilD of Pseudomonas aeruginosa. Methods Enzymol *235*, 527-540.

Sudo, K., and Kanno, T. (1976). Properties of the amylase produced in carcinoma of the lung. Clin Chim Acta *73*, 1-12.

Sue, S., Shibata, W., and Maeda, S. (2015). Helicobacter pylori-Induced Signaling Pathways Contribute to Intestinal Metaplasia and Gastric Carcinogenesis. Biomed Res Int *2015*, 737621.

Suerbaum, S. (1995). The complex flagella of gastric Helicobacter species. Trends Microbiol *3*, 168-170; discussion 170-161.

Suerbaum, S., and Josenhans, C. (2007). Helicobacter pylori evolution and phenotypic diversification in a changing host. Nat Rev Microbiol *5*, 441-452.

Suerbaum, S., Smith, J.M., Bapumia, K., Morelli, G., Smith, N.H., Kunstmann, E., Dyrek, I., and Achtman, M. (1998). Free recombination within Helicobacter pylori. Proc Natl Acad Sci U S A *95*, 12619-12624.

Sugisaki, H., Yamamoto, K., and Takanami, M. (1991). The HgaI restriction-modification system contains two cytosine methylase genes responsible for modification of different DNA strands. J Biol Chem *266*, 13952-13957.

Sundrud, M.S., Torres, V.J., Unutmaz, D., and Cover, T.L. (2004). Inhibition of primary human T cell proliferation by Helicobacter pylori vacuolating toxin (VacA) is independent of VacA effects on IL-2 secretion. Proc Natl Acad Sci U S A *101*, 7727-7732.

Sycuro, L.K., Pincus, Z., Gutierrez, K.D., Biboy, J., Stern, C.A., Vollmer, W., and Salama, N.R. (2010). Peptidoglycan crosslinking relaxation promotes Helicobacter pylori's helical shape and stomach colonization. Cell *141*, 822-833.

Szegedi, S.S., and Gumport, R.I. (2000). DNA binding properties in vivo and target recognition domain sequence alignment analyses of wild-type and mutant RsrI [N6-adenine] DNA methyltransferases. Nucleic Acids Res *28*, 3972-3981.

Takahashi, N., Naito, Y., Handa, N., and Kobayashi, I. (2002). A DNA methyltransferase can protect the genome from postdisturbance attack by a restriction-modification gene complex. J Bacteriol *184*, 6100-6108.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol *28*, 2731-2739.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol Biol Evol *30*, 2725-2729.

Tham, K.T., Peek, R.M., Jr., Atherton, J.C., Cover, T.L., Perez-Perez, G.I., Shyr, Y., and Blaser, M.J. (2001). Helicobacter pylori genotypes, host factors, and gastric mucosal histopathology in peptic ulcer disease. Hum Pathol *32*, 264-273.

Tomb, J.F., White, O., Kerlavage, A.R., Clayton, R.A., Sutton, G.G., Fleischmann, R.D., Ketchum, K.A., Klenk, H.P., Gill, S., Dougherty, B.A*., et al.* (1997). The complete genome sequence of the gastric pathogen Helicobacter pylori. Nature *388*, 539-547.

Tran, P.H., Korszun, Z.R., Cerritelli, S., Springhorn, S.S., and Lacks, S.A. (1998). Crystal structure of the DpnM DNA adenine methyltransferase from the DpnII restriction system of streptococcus pneumoniae bound to S-adenosylmethionine. Structure *6*, 1563-1575.

Tummuru, M.K., Cover, T.L., and Blaser, M.J. (1993). Cloning and expression of a high-molecular-mass major antigen of Helicobacter pylori: evidence of linkage to cytotoxin production. Infect Immun *61*, 1799-1809.

Van der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A.E., and Berendsen, H.J.C. (2005). GROMACS: Fast, flexible, and free. J Comput Chem *26*, 1701-1718.

van der Woude, M., Hale, W.B., and Low, D.A. (1998). Formation of DNA methylation patterns: nonmethylated GATC sequences in gut and pap operons. J Bacteriol *180*, 5913-5920.

van Dijk, M., van Dijk, A.D., Hsu, V., Boelens, R., and Bonvin, A.M. (2006). Information-driven protein-DNA docking using HADDOCK: it is a matter of flexibility. Nucleic Acids Res *34*, 3317-3325.

Vidgren, J., Svensson, L.A., and Liljas, A. (1994). Crystal structure of catechol O-methyltransferase. Nature *368*, 354-358.

Vilkaitis, G., Dong, A., Weinhold, E., Cheng, X., and Klimasauskas, S. (2000). Functional roles of the conserved threonine 250 in the target recognition domain of HhaI DNA methyltransferase. J Biol Chem *275*, 38722-38730.

Vitkute, J., Stankevicius, K., Tamulaitiene, G., Maneliene, Z., Timinskas, A., Berg, D.E., and Janulaitis, A. (2001). Specificities of eleven different DNA methyltransferases of Helicobacter pylori strain 26695. J Bacteriol *183*, 443-450.

Vogiatzi, P., Cassone, M., Luzzi, I., Lucchetti, C., Otvos, L., Jr., and Giordano, A. (2007). Helicobacter pylori as a class I carcinogen: physiopathology and management strategies. J Cell Biochem *102*, 264-273.

Wang, J., Wang, W., Kollman, P.A., and Case, D.A. (2006a). Automatic atom type and bond type perception in molecular mechanical calculations. J Mol Graph Model *25*, 247-260.

Wang, J., Wolf, R.M., Caldwell, J.W., Kollman, P.A., and Case, D.A. (2004). Development and testing of a general amber force field. J Comput Chem *25*, 1157-1174.

Wang, J.M., Wang, W., Kollman, P.A., and Case, D.A. (2006b). Automatic atom type and bond type perception in molecular mechanical calculations. J Mol Graph Model *25*, 247-260.

Wang, S.Y., Shen, X.Y., Wu, C.Y., Pan, F., Shen, Y.Y., Sheng, H.H., Chen, X.M., and Gao, H.J. (2009). Analysis of whole genomic expression profiles of Helicobacter pylori related chronic atrophic gastritis with IL-1B-31CC/-511TT genotypes. J Dig Dis *10*, 99-106.

Warren JR, M.B. (1983). Unidentified curved bacilli on gastric epithelium in active chronic gastritis. Lancet *1*, 1273-1275.

Wells, J.A. (1991). Systematic mutational analyses of protein-protein interfaces. Methods Enzymol *202*, 390-411.

Wierenga, R.K., Terpstra, P., and Hol, W.G. (1986). Prediction of the occurrence of the ADP-binding beta alpha beta-fold in proteins, using an amino acid sequence fingerprint. J Mol Biol *187*, 101-107.

Wilson, G.G. (1991). Organization of restriction-modification systems. Nucleic Acids Res *19*, 2539-2566.

Wilson, G.G. (1992). Amino acid sequence arrangements of DNA-methyltransferases. Methods in enzymology *216*, 259-279.

Wilson, G.G., and Murray, N.E. (1991). Restriction and modification systems. Annu Rev Genet *25*, 585-627.

Wu, J.C., and Santi, D.V. (1987). Kinetic and catalytic mechanism of HhaI methyltransferase. J Biol Chem *262*, 4778-4786.

Xu, S., Xiao, J., Posfai, J., Maunus, R., and Benner, J., 2nd (1997). Cloning of the BssHII restriction-modification system in Escherichia coli : BssHII methyltransferase contains circularly permuted cytosine-5 methyltransferase motifs. Nucleic Acids Res *25*, 3991-3994.

Xue, H., Liu, J., Lin, B., Wang, Z., Sun, J., and Huang, G. (2012). A meta-analysis of interleukin-8 -251 promoter polymorphism associated with gastric cancer risk. PLoS One *7*, e28083.

Yang, J.M., and Tung, C.H. (2006). Protein structure database search and evolutionary classification. Nucleic Acids Res *34*, 3646-3659.

Zhang, X., and Bruice, T.C. (2006). The mechanism of M.HhaI DNA C5 cytosine methyltransferase enzyme: a quantum mechanics/molecular mechanics approach. Proc Natl Acad Sci U S A *103*, 6148-6153.

Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction. BMC Bioinformatics *9*, 40.

Zhang, Y.W., Eom, S.Y., Yim, D.H., Song, Y.J., Yun, H.Y., Park, J.S., Youn, S.J., Kim, B.S., Kim, Y.D., and Kim, H. (2013). Evaluation of the relationship between dietary factors, CagA-positive Helicobacter pylori infection, and RUNX3 promoter hypermethylation in gastric cancer tissue. World J Gastroenterol *19*, 1778-1787.

List of Publications

- Structure and dynamics of *H. pylori* 98-10 C5 cytosine specific DNA methyltransferase in complex with S-adenosyl-L-methionine and DNA. **S. Singh**, K. Tanneeru and L. Guruprasad. ***Mol. BioSyst.,*** 2016, DOI:10.1039/C6MB00306K.

- Structure Based Annotation of Helicobacter pylori Strain 26695 Proteome. **S. Singh** P.K. Guttula and L. Guruprasad (2014). ***PLoS One*** 9, e115020.

- Structure and Sequence Based Analysis of Alpha-Amylase Evolution. **S. Singh** and L. Guruprasad (2014). ***Protein and peptide letters*** 21, 948-956.

- Relation between thermal resistance and flexibility of *Saccharomyces cerevisiae* mitochondrial Hsp70 co-chaperone Mge1 determined by a gain of function mutation. A. Marada*, S. Karri*, **S. Singh***, P.K. Allu, Y. Boggula, T. Krishnamoorthy, L. Guruprasad and N.B.V. Sepuri. (***Under Revision***).
  *Authors contributed equally.

- Structure and dynamics of N6 adenine specific DNA methyltransferases from *H. pylori* 98-10: in complex with S-adenosyl-L-methionine and DNA. **S. Singh** and L. Guruprasad. (***Manuscript Communicated***).

Plagiarism Report