The Large Scale Bioinformatics Analysis of Auxiliary Activity Family 9 Enzymes

A mini-thesis submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE OF RHODES UNIVERSITY

by

Coursework / Thesis

in

Bioinformatics and Computational Molecular Biology

In the Department of Biochemistry, Microbiology & Biotechnology

Faculty of Science

by

Vuyani Moses

February 2014

ABSTRACT

Biofuels have been proposed to be a suitable replacement to the already depleting fossil fuels. The complex structures of plant biomasses present a challenge the production of biofuels due to recalcitrance. The complex cellulose structure and hydrogen bonding between repeat units of cellulose is believed to be a major contributor to the recalcitrance of cellulose. Fungal organisms come equipped with various oxidative enzymes involved in degradation of plant biomass. The exact mechanism of cellulose degradation remains elusive. The GH61 is a group of proteins which are PMOs. GH61 sequences where previously described as endoglucanases due to weak endoglucanase activity. These enzymes were later found not possess any enzyme activity of their own however they could enhance the activity of other cellulose degrading enzymes. As a result reclassification of these enzymes as AA9 has been implemented. AA9 proteins have been reported to share structural homology with the bacterial AA10 group of enzymes. Based on cleavage products that are produced when AA9 proteins interact with cellulose, AA9 proteins have been grouped into three types. To date the exact mechanism and the sequence and structural basis for differentiating between the various AA9 types remains unknown. Using various bionformatic techniques sequence and structural elements were identified for distinguishing between the AA9 types. A large dataset of sequences was obtained from the Pfam database from UNIPROT entries. Due to high divergence of AA9 sequences, a smaller dataset with the more divergent sequences removed was created. The inclusion of the reference sequences to the data set was done to observe which sequences belong to a certain type. Phylogenetic analysis was able to group AA9 proteins into three distinct groups. MSA and motif analysis revealed that the N-Terminus of these proteins is mostly responsible for type specificity. Structural analysis of AA9 PDB structures and homology models allowed the effect of physicochemical properties to be gauged structurally. The presence of 310 helices and aromatic residues the surface of AA9 sequences is an observation which still warrants further investigation.

DECLARATION

I, **Vuyani Moses**, declare that this thesis submitted to Rhodes University is a master piece of my own and has not been submitted elsewhere for a degree.

Signature……

Date……

ACKNOWLEDGMENTS

I would like to take this opportunity to acknowledge the people who made conducting this research possible:

- My supervisor Prof. Özlem Tastan Bishop for her advice, guidance and patience
- My co-supervisor Prof. Brett Pletschke for his invaluable expertise and guidance
- To my colleagues at the Rhodes Bioinformatics Research group
- To the National Research Foundation and Rhodes University for funding

Table of Contents

LIST OF TABLES

LIST OF FIGURES

SCRIPTS

LIST OF WEB SERVERS

BLAST: www.ncbi.nlm.nih.gov/BLAST/

CAZy: www.cazy.org

MAFFT: http://www.mafft.cbrc.jp/alignment/server/index.html

MEME: http://www.meme.nbcr.net/meme/cgi-bin/meme.cgi

MetaMQAPII: http://www.genesilico.pl/toolkit/unimod?method=MeaMQAPII

PDBsum: http://www.ebi.ac.uk/pdsum/

Promals3D: http://www.prodata.swmed.edu/Promals3d/Promals3d.php

PROTPARAM TOOLS: http://web.expasy.protparam/

Pfam: http://pfam.sanger.ac.uk

RAMPAGE: http://mordred.bioc.cam.ac.uk/~rapper/rampage.php

ACRONYM1S

3D Three Dimensional

AA9 auxiliary activity 9

AA10 auxiliary activity 10

BLAST Basic Local Alignment Tool

BLASTP Basic Local Alignment Tool for Protein sequences

CBM33 Carbohydrate binding module family 33 family

CBM Carbohydrate binding module

DOPE Discrete Optimized Protein Energy

GDT_TS Global Distance Test Total Score

GH61 glycoside hydrolase family 61

HMM Hidden Markov Models

JTT Jones Taylor Thornton

Kd Kyte and Doolittle

MSA Multiple sequence alignment

MAFFT Multiple sequence Alignment based on Fast Fourier Transform

MEGA Molecular Evolutionary Genetic Analysis

MQAP Model Quality Assessment Program

MSA Multiple Sequence Alignment

MUSCLE MUltiple Sequence Comparison by Log – Expectation

NCBI National Centre for Biotechnology Information

NMR Nuclear Magnetic Resonance

PDB Protein Data Bank

PMO polysaccharide monooxygenases

Promals PROfile Multiple Alignment with Local Structure

KS-test Kolmogorov-Smirnov test

TYPOGRAPHICAL CONVENTIONS

Sequence that where modeled where referred to as:

An9 - Aspergilus_niger_9

Gg9 - Gloromera_graminic_9

Hr1 - Hypocrea_rufa_1

Pt17_Pyrenophora_teres_17

Tt2 – Thievela_terestis_2

SYMBOLS USED

α -Alpha

β -Beta

ϕ -phi

ψ -ps

LIST OF AMINO ACIDS

Name 3 letter code 1 letter code

Alanine Ala A

Arginine Arg R

Asparagine Asn N

Aspartic acid Asp D

Cysteine Cys C

Glutamine Gln Q

Glutamic acid Glu E

Glycine Gly G

Histidine His H

Isoleucine Ile I

Leucine Leu L

Lysine Lys K

Methionine Met M

Phenylalanine Phe F

Proline Pro P

Serine Ser S

Threonine Thr T

Tryptophan Trp W

Tyrosine Tyr Y

Valine Val V

CHAPTER ONE

1.1 **Cellulose breakdown**

Proteomic and transcriptomic analyses of cellulolytic fungi has resulted in the identification of various oxidative enzymes involved in the degradation of plant biomass. The exact mechanism of how these enzymes actually degrade cellulose still remains unknown. There are three proposed hypotheses which could explain the breakdown of cellulose (Phillips et al., 2011).

The first hypothesis involves hydrolytic chemistry, where cellulases cleave the glycosidic bonds on glucan through acid base catalysis (Phillips et al., 2011). Cellulose genes are abundant in various groups of microorganisms. These genes are then expressed as hydrolytic proteins that are responsible for the complete breakdown of cellulose producing free glucose molecules (Merino et al., 2007: White et al., 1981: Lee et al., 1980). The breakdown of cellulose is achieved through the action of three enzymes. These enzymes include endo-β-glucanases, exo-β-glucanase (which are also known as cellobiohydrolases) and β-glucosidases. These three enzymes act synergistically to degrade cellulose (Jeoh et al., 2002). β-1,4-glycosidic bonds are hydrolyzed randomly through the activity of endo-β-glucanases to decrease the length of the cellulose chain (McCarthy et al., 2003), then exo-β-glucanases remove cellobiose from the resulting cellulose chain (Baker et al., 1998), and finally through the activity of β-glucosidases, cellobiose is hydrolyzed to form glucose (Nunoura et al., 1996). The second hypothesis involves Fenton chemistry, where hydroxyl radicals, which are produced through the reduction of extracellular metal ions, randomly oxidize cellulose (Phillips et al., 2011). The fenton chemistry hypothesis is supported from the observation made on a group of enzymes called celebiose dehydrogenases (CDHs) the observation that CDH proteins generate hydroxyl radicals formed through the reduction of iron complex which is found in the extracellular environment (Henrissat, 1991). The third hypothesis is the enzymatic oxidation of cellulose through the combination of oxidases and oxidoreductases resulting in direct oxidative enzyme-catalysis. With this step glycosidic bonds can be cleaved without having to remove a glucan chain from crystalline cellulose (Phillips et al., 2011).

1 Literature review

1.1 Background

Fossil fuels have long been utilised as one of the most important energy resources. Due to the impending depletion and as well as the environmental effects of fossil fuels, the need to identify alternative fuel sources has arisen (Bhattarai et al., 2011). There have been numerous proposed alternative sources of energy amongst which biofuels are one of the most promising. What makes biofuels so important is their ability to replace the already depleting nonrenewable fossil fuel (Al-Zuhair, 2001). Due to various factors, José et al (2009) predicts an increase in biofuel demand within the coming years. The demand in biofuel production has resulted in the development of thermochemical processes and biochemical processes for the production of biofuels (Mu et al., 2010). Production of biofuels through biochemical conversion is considered to be more advantageous since the carbohydrate structure is preserved during the process (Tramper, 1996), while the production of biofuels through thermochemical processes results in the carbohydrate structure being destroyed (Verma et al., 2012).

A major drawback in the production of biofuels is the complex structures of plant biomass. These complex structures often result in difficulties in the degradation of plant biomass. Numerous factors are believed to contribute to the complex structure of plant biomass however one of major contributors to the recalcitrance of biomass is the structural arrangement of cellulose (Himmel et al., 2007).

1.2 The structure of cellulose

The structure of cellulose is comprised of two ring anhydroglucose rings which form the cellulose repeat unit $(C_6H_{10}O_5)_n$ (Figure 1.1). Within each repeat unit, C1 from of one the glucose rings is linked covalently with a C4 of a following ring, resulting in a β 1-4 glycosidic bond. The repeat units are then organized in linear manner adopting a flat ribbon like conformation. The number of repeat units (n) in cellulose typically ranges between 10000 and 15000. This range is determined

by the source material of the cellulose. Cellulose typically adopts a linear conformation due to the presence of the hydrogen bonds which form between hydroxyl group and oxygen of the β 1-4 linked residue (Moon et al, 2011).

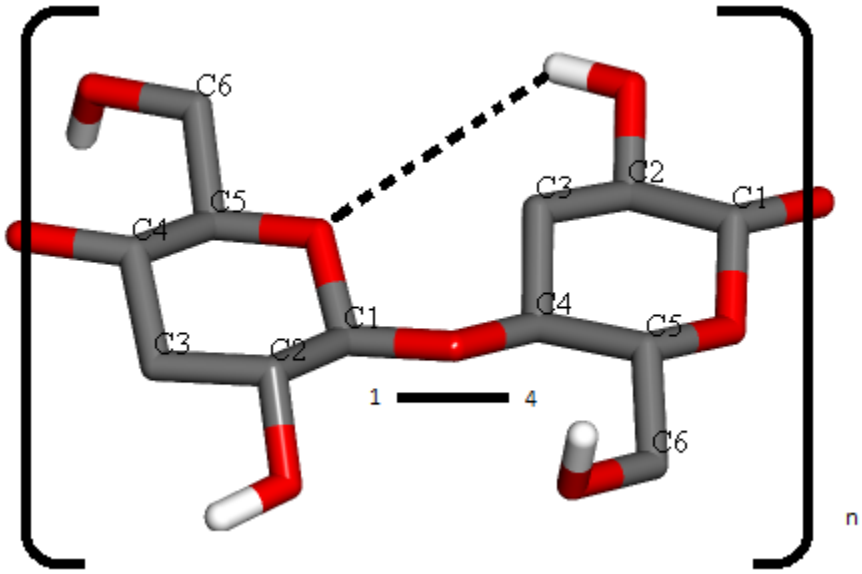


Figure 1.1: Graphical representation of the cellulose repeat unit. Carbon atoms are displayed in grey, oxygen atoms are displayed in red and hydrogen are displayed in white. The dashed line represents hydrogen bonding between the two β 1-4 linked residues. Image rendered using Discovery Studio 3.1. Adapted from Moon et al (2011).

Cellulose forms a linear homopolymer consisting of two glucose residues which adopt a D configuration. The cellulose polymer ends are different chemically. One end of the cellulose polymer consists of a D-glucopyranose unit which possesses a free anomeric carbon. The second end has a D-glucopyranose which has an anomeric carbon atom that is responsible for forming the glycosidic linkage (Perez et al., 2013).

1.3 Cellulose breakdown

Proteomic and transcriptomic of analyses of cellulolytic fungi has resulted in the identification of various oxidative enzymes involved in the degradation of plant biomass. The exact mechanism of

how these enzymes actually degrade cellulose still remains unknown. There are three proposed hypotheses which could explain the breakdown of cellulose (Phillips et al., 2011).

The first hypothesis involves hydrolytic chemistry, where cellulases cleave the glycosidic bonds on glucan through acid base catalysis (Phillips et al., 2011). Cellulose genes are abundant in various groups of microorganisms. These genes are then expressed as hydrolytic proteins that are responsible for the complete breakdown of cellulose producing free glucose molecules (Merino et al., 2007: White et al., 1981: Lee et al., 1980). The breakdown of cellulose is achieved through the action of three enzymes. These enzymes include endo-β-glucanases, exo-β-glucanase (which are also known as cellobiohydrolases) and β-glucosidases. These three enzymes act synergistically to degrade cellulose (Jeoh et al., 2002). β-1,4-glycosidic bonds are hydrolyzed randomly through the activity of endo-β-glucanases to decrease the length of the cellulose chain (McCarthy et al., 2003), then exo-β-glucanases remove cellobiose from the resulting cellulose chain (Baker et al., 1998), and finally through the activity of β-glucosidases, cellobiose is hydrolyzed to form glucose (Nunoura et al., 1996). The second hypothesis involves Fenton chemistry, where hydroxyl radicals, which are produced through the reduction of extracellular metal ions, randomly oxidize cellulose (Phillips et al., 2011). The fenton chemistry hypothesis is supported from the observation made on a group of enzymes called celebiose dehydrogenases (CDHs) the observation that CDH proteins generate hydroxyl radicals formed through the reduction of iron complex which is found in the extracellular environment (Henrissat, 1991). The third hypothesis is the enzymatic oxidation of cellulose through the combination of oxidases and oxidoreductases resulting in direct oxidative enzyme-catalysis. With this step glycosidic bonds can be cleaved without having to remove a glucan chain from crystalline cellulose (Phillips et al., 2011).

1.4 **Glycoside hydrolase family 61**

The glycoside hydrolase family 61 (GH61) is a group of proteins which are polysaccharide monooxygenases (PMOs). GH61 sequences where previously described as endoglucanases due to weak endoglucanase activity previously detected by Karlsson et al. (2001). However recent studies have demonstrated that GH61 proteins are not typical glycosyl hydrolases (Karkehabadi et al., 2008: Harris et al., 2010). As a result under the carbohydrate-active enzymes database (CAZy;

www.cazy.org) classification, GH61 proteins have been reclassified as auxiliary activity 9 (AA9) proteins. AA9 proteins are believed to interact directly with cellulose, allowing other enzymes to access it for degradation (Vaaje-Kolstad et al., 2010). AA9 proteins are characterized as copper dependent PMOs which contain a distinct conserved motif of copper coordinating histidine residues (Levasseur et al., 2013) without removing the glucan chain from the surface of cellulose. GH61 proteins are capable of oxidatively cleaving glycosidic bonds on the cellulose (Beeson et al., 2012).

1.4.1 **AA9 types**

PMOs including AA9 proteins are known to oxidize the C1 carbon of the glucose ring structure but may also be able to oxidize the C4 and C6 carbons of the glucose ring (Li et al., 2012). The C1 and C4 cleavage products come in the form of aldonolactone or 4-ketoaldose, respectively (Beeson et al., 2012; Phillips et al., 2011). As a result, PMOs are grouped into three types based on the sequence similarity as well as the cleavage product. The first type: Type-1 PMOs cleave the C1 carbon of pyranose residues producing aldonolactone. The second type: Type-2 PMOs cleave the C4 carbon of pyranose residues producing 4-ketoaldose. The third type: Type-3 PMOs do not display any specificity in the production of either aldonolactone or 4-ketoaldose (Beeson et al., 2012; Phillips et al., 2011; Quinlan et al., 2011).

1.4.2 **AA9 similarities with AA10**

AA9 proteins are believed to perform a crucial role in the initial stage of cellulose. The initial step of cellulose breakdown is believed to involve the disruption of the cellulose structure. The disrupted structure provides access to traditional cellulases, resulting in cellulose degradation (Quinlam et al., 2011; Vaaje-Kolstad et al., al 2010). CPB21 is an enzyme which belongs to the Carbohydrate family 33 (CBM33), the CBM33 family has been reclassified as AA10. The CPB21 enzyme has no apparent activity, but it was demonstrated that enzymatic hydrolysis of Chitin is greatly increased when this enzyme is present in concert with other cellulases. AA9 proteins are structurally very similar to AA10 as they have a similar conserved metal binding motif in the substrate binding site (Aachmann et al., 2012).

Harris et al. (2010) conducted a study further describing similarities between AA9 and AA10 proteins. In the conducted study, it was shown that the metal binding site residues in AA9 proteins share similarities to those previously described in work done by Vaaje-Kolsaad et al. (2005) on CBM33 proteins.The active site coordinates a copper ion using the amino terminus and Nδ a of bidentate N-Terminus histidine. Due to post translational modifications, the N-Terminus histidine is methylated at the Nε, however the purpose of this modification remains unknown. The other site belongs to a second histidine residue which coordinates copper with its Nε (Beeson et al., 2012). In addition to the two histidine residues, AA9s have a tyrosine and a glutamine that contributes to metal coordination through interaction with a water molecule which in turn, interacts with the copper ion (Aachmann et al., 2012). Site directed mutagenisis of these residues in both AA10 and AA9 eliminated activity.

The similarities between AA10 and AA9 are not only limited to the active site residues, figures 1.2A and 1.2B shows that AA10 and AA9 sequences also possess a similar beta-sandwich fold. It has been proposed that due to the similarities between AA9 and AA10 at both structural and functional level, these two groups of enzymes share the same evolutionary origin. This observation therefore suggests that both AA9 and AA10 originate from a previously unknown superfamily of proteins (Harris et al., 2010).

1.4.3 **AA9 features**

A role in molecular recognition has been suggested for AA9. This is based on the observation made on the immunoglobulin heavy chain variable domain (figure 1. 2C) and fibronectin III (Figure 1.2D) protein. These two structures share the same beta-sandwich fold and the loop regions contain regions which are responsible for molecular recognition (Li et al., 2012). The immunoglobulin heavy chain variable domain contains highly variable regions indicated in figure 1.2C which are responsible for interacting with an antigen (Li et al., 2012). The fibronectin III protein contains a conserved RGD motif located in one of the loops of the beta sandwich fold and is believed to be responsible for the interaction with αv integrins (Takahashi et al., 2007). GH61 proteins contain a highly variable loop region designated loop 2, which is located on the planar surface of the molecule which is in a similar conformation to both the immunoglobulin heavy chain variable domain and fibronectin III, which may suggest a role in molecular recognition.

There are observable differences between AA9 and AA10 structures. All AA9 sequences contain a metal ion within the active site of the enzyme, whereas AA10 proteins do not all possess a metal ion in the active site. On both AA10 and AA9 proteins the catalytic site comprises of aromatic and polar residues which are believed to bind chitin and cellulose, respectively. The flat surface of the LPMOs catalytic site is most suited for cleaving glycosidic bonds without actually decrystallizing the overall cellulose structure. Whereas the endoglucanases have a catalytic cleft instead of a flat surface that is believed to be responsible to cleave more accessible glycosidic bonds on the cellulose substrate (Wu et al., 2013). AA9 enzymes are secreted with CDHs to act as electron donors which supports a hypothesis which describes the breakdown of Cellulose-involving the use of hydroxyl reticules in the specific breakdown of cellulose degrading proteins (Eastwood et al., 2011: Goodell et al., 1997). The co-expression of these enzymes may also indicate a possible genetic link between AA9 and of CDHs. It was discovered that the transcription of AA9 genes in *Heterobasidion annonsum sensu lato* complex can be divided into two groups: AA9 proteins that are upregulated in the presence of certain substrates and AA9 proteins whose expression is not affected by substrates (Yakolev et al., 2012)

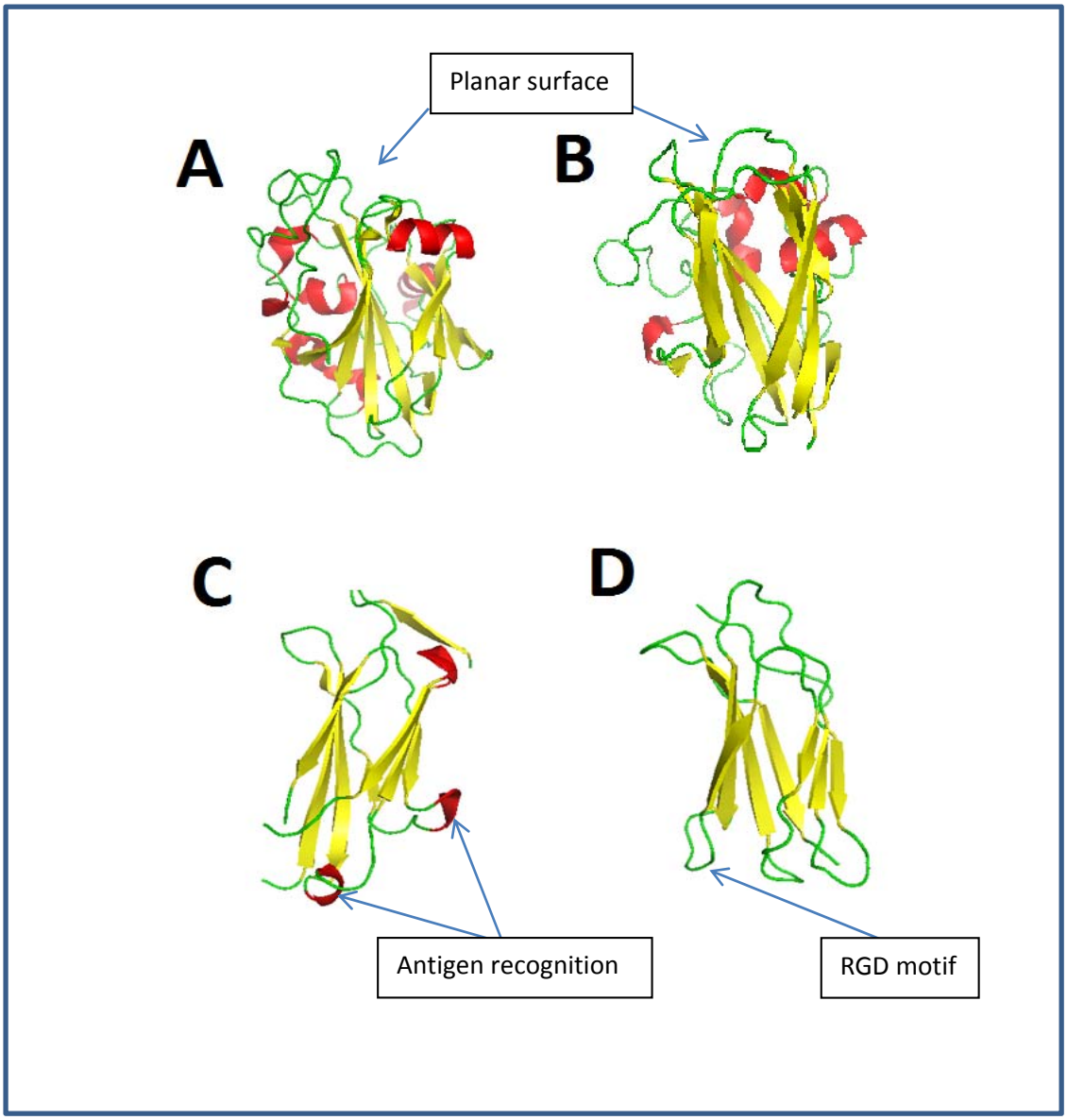


Figure 1.2: overall fold displayed by A: AA9 (PDB ID: 2VTC), B: AA10 (PDB ID: 2VTC), C: Immunogobulin Heavy chain variable domain (PDB ID: 2NY1) and D: Fibronectin III (PDB ID: 1FNF). Structures are coloured based on their secondary structure, alpha helices are displayed in red, beta) sheets are displayed in yellow and random coil is displayed in green.Adapted from Li et al. (2012).

1.4.4 **Domain organization**

The Pfam database currently reports 827 AA9 sequences (Pfam ID: PF03443) present within the database and these sequences are distributed amongst 87 fungal species and 1 plant sequence from maize. The AA9 sequences display 12 distinct architectures as seen in Figure 1.3.

From the 12 architectures described in the Pfam database a large proportion (626 sequences) of AA9 protein sequences share the architecture of a single AA9 domain without the presence other

domains. The second most prominent architecture is that of the AA9 protein with a carbohydrate binding motif (CBM) from the carbohydrate binding motif family 1 which represents 168 of the Pfam database sequences. There are 10 other architectures with distinctly varying domain organisations. However, these other architectures are represented to a lesser extent.

1.4.5 **The carbohydrate binding domain**

A large proportion of AA9 protein sequences in Pfam database were found to associate with varying numbers of CBMs. The glycosidic bonds of polysaccharides often present the challenge of accessibility to the active site glycoside hydrolases. As a consequence, some AA9 proteins are associated with carbohydrate binding modules (CBMs). These CBMs are responsible for enhancing the accessibility of glycosidic bonds by associating the enzyme with the substrate. The structural analysis of 22 CBM families has shown that the CBMs from different families are similar in terms of structure (Shoseyov et al., 2006). CBMs use their hydrophobic surface which binds cellulos. Substrate disruption by CBMs was first observed in *C. fimi* endoglucanase which showed the substrate disruption without any detectable hydrolytic activity (Din et al., 1991). Other roles believed to be played by CBMs include: (i) a proximity effect and (ii) a targeting function.

Comparison of the structures of CBM1 and AA9 showed that these two are similar in terms the polar aromatic residues found on the substrate binding surface. The polar aromatic residues on CBM1 have a different spatial distribution to those on the AA9 protein because of the size differences between the two. This observation suggests that AA9 bind cellulose in a similar manner to CBM1. The number of polar residues on AA9 tends to vary, which is believed to affect substrate binding and product formation (Li et al., 2012).

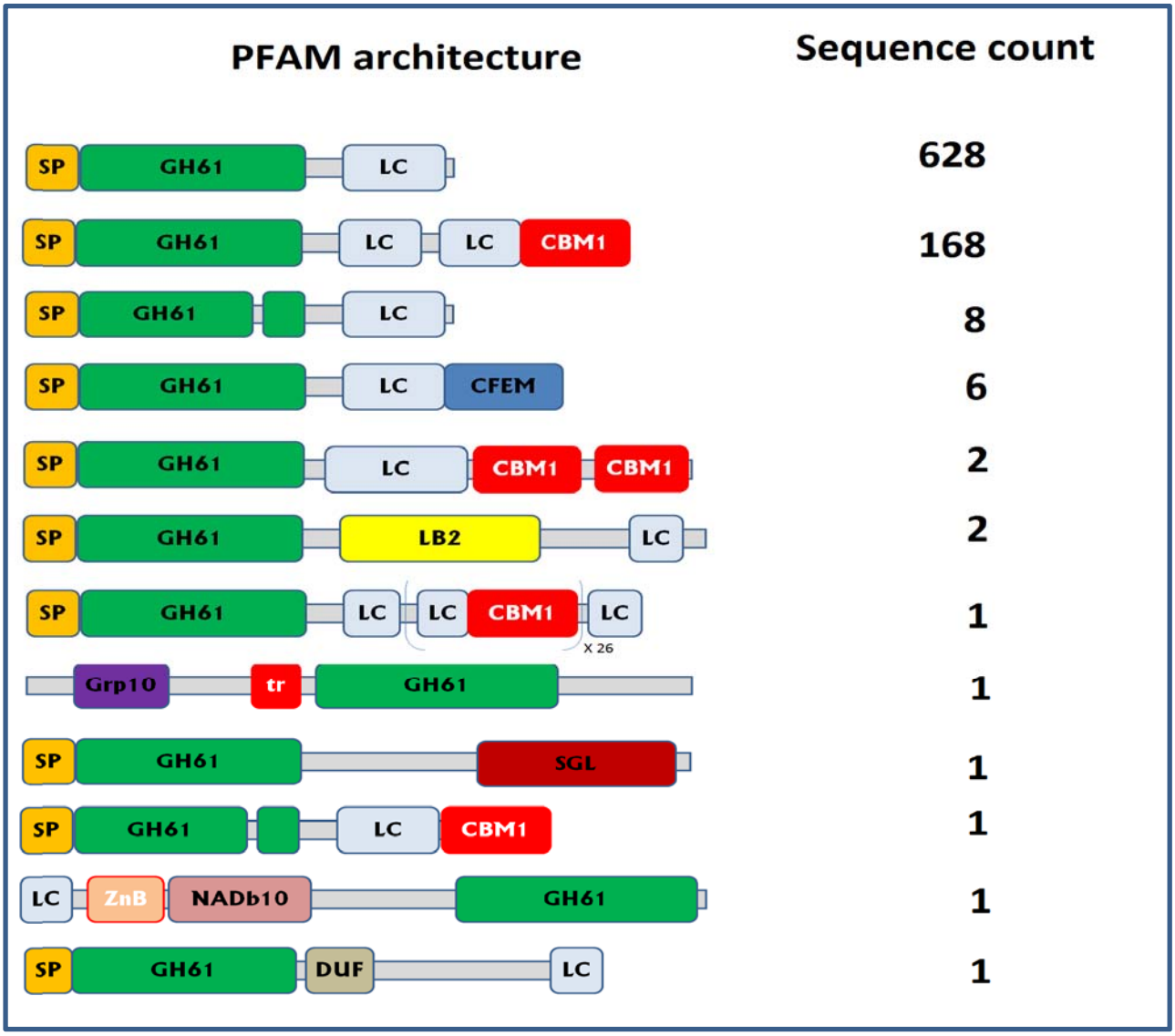


Figure 1.3: Representation of the 12 domain architectures of AA9 protein sequence as described in the Pfam database (http://pfam.sanger.ac.uk). The number of sequences for each architecture is displayed. The AA9 proteins form distinct architectures which consist of varying modular organisations as well as distinct regions which include: A Signal peptide(SP) region, the glycoside hydrolase 61 (AA9) functional domain (Pfam ID: PF03443), low complexity(LC) regions, the carbohydrate binding family 1(CBM1) module (Pfam ID: PF00734), the CFEM domain (Pfam ID: PF05730), the beta lactamase superfamily (LB2) domain (Pfam ID: PF12706), the chaperonin 10Kd (Cpn10) subunit(Pfam ID: PF00166), a transmembrane region (tr), a SMP-30/Gluconolaconase/LRE-like (SGL) region (Pfam ID: PF08450), The Zinc binding (ZnB) domain (Pfam ID: PF13695), the NADH(P)-binding (Nad10) domain (Pfam ID: PF13460) and a domain of unknown function (DUF) (Pfam ID: PF07142).

1.5 **The problem**

It is known that there are three AA9 types because of cleavage products each AA9 protein forms when interacting with cellulose. There are currently numerous AA9 structures deposited in the PDB database. Even with the presence of these structures the sequence and structural basis of distinguishing between the various AA9 types remains unknown. AA9 Sequences are highly divergent and yet their functionality is conserved. Therefore the basis for this conservation needs to be elucidated and the structural and the consequences of this divergence needs to be investigated.

1.6 **The hypothesis**

Structural and sequence analysis of AA9 protein sequences will provide insights in the understanding off cellulose degradation enhancing enzymes.

1.7 **Aim**

To identify unique type specific AA9 features at both sequence and structural levels.

1.8 **Objectives**

- Retrieval of AA9 protein sequences and crystal structures from various databases.
- Identification of sequence markers of the three AA9 types.
- Alignment of AA9 protein sequences with the marker sequences.
- Motif discovery on selected AA9 protein sequences with the marker sequences.
- Phylogenetic analysis of AA9 sequences with their specific AA9 markers.
- Physicochemical property analysis of various specific AA9 types.
- Mapping of unique type specific AA9 features.

CHAPTER TWO

2 Sequence analysis

AA9 proteins are mostly expressed within members of fungal kingdom, with only two known cases were these proteins are found in the plant kingdom. AA9 proteins are highly diverse and abundant in fungal genomes as many fungal organisms encode more than one AA9 protein. These diverse AA9 protein form three separate groups based on the type of product that they produce. This chapter mainly focuses on the identification of sequence feature that allow distinction between the various AA9 types. To identify these sequence features tools such as multiple sequence alignment (MSA), motif analysis, phylogenetic analysis and sequence feature analysis.

2.1 Introduction

The 3D structure of proteins often determines the function performed by that particular protein. The primary structure of amino acids within the protein determines the overall 3D arrangement of a protein which in turn determines the function of that protein. As a result it is important to understand sequence features of proteins to gain better understanding of their biological roles.

The analysis of these sequence features usually begins with the identification of relevant sequence data. This sequence information can be obtained from various biological databases, which store a wide variety of biological information. Biological databases are divided into three categories: primary databases which contain data straight from the scientific community which may include redundancy of data, secondary databases which contain manually curated data, and specialized databases which contain which only contain a specific type of data. Biological databases are an important resource for retrieving sequence information.

2.2 Database searches

In biological databases access to the data is important. As a result tools such as the Basic local Alignment Search Tool (BLAST) have been developed. BLAST employs a heuristic word approach to identify short stretches of characters in the query sequence. Once these words have been identified BLAST searches the database for the occurrence of these words. A substitution matrix is used to score the matching words. If a word match is above a certain threshold, a pairwise alignment is conducted to extend the matching words. When the threshold drops below a certain point, extension is terminated (Altschul et al., 1990).

2.3 Specialized databases -The Pfam database

The Pfam database is a collection of protein domains for a specific protein family. The Pfam Database is manually curated with over 10000 entries. The HMMER3 suite (http://hmmr.janelia.org) determines the similarity between proteins. Each entry contains a protein alignment as well as a hidden Markov model. The hidden Markov model is used to search sequence databases for homologs. There are two types of Pfam families in the database, manually curated Pfam-A and automatically generated Pfam-B. Pfam-A families are more accurate and these families are generated in a four step process. The first step involves constructing a multiple sequence alignment called the seed alignment. The seed alignment is then used to construct a profile hidden Markov model (HMM). The HMM is searched against the UniprotKB database to identify sequences which are identical to the protein domain which are then added to the final alignment (Punta et al., 2012).

2.4 Multiple sequences alignment

Multiple sequence alignment (MSA) of proteins sequences can be regarded as crucial for many biological applications (pei , 2007). If one is given a set of sequences MSA will permit the identification as well as the visualization of patterns of conservation of sequences, this is achieved by organizing homologue positions of different sequences in columns. The presence of sequence similarity amongst sequences usually suggests that proteins have diverged from a common ancestor or have similar functions. The presence of conserved amino acids in within a group of proteins can also be regarded as a strong indication of preserved 3D structure amongst the group

of proteins (Do, 2005). It has been observed that the quality of a particular alignment is greatly influenced by the presence of sequences with a sequence identity less than 30% which falls below the twilight zone which results in the accuracy of alignment dropping considerably (Do, 2005). A problem that arises with the construction of an alignment is defining an objective function that will assess the quality of an alignment as well as the creating an algorithm that will find optimal alignment based on the objective function (Do, 2005). For the alignment of just 2 sequences evaluation is carried out by the addition of match or mismatch scores for aligned pair positions and gap penalties are affined for unaligned amino acids (Do, 2005). Pair hidden Markov models (HHM) are utilized to act as an alternative of the formulation of the sequence alignment problem (Do, 2005). There are numerous alignment programs which utilize different approaches to obtain alignments. Currently there is no clear way to say one program is superior to another it is therefore important to evaluate different programs and identify which produces the most preferred alignment. The alignment programs used in this study are explained below.

2.4.1 **MAFFT**

Multiple sequence alignment through MAFFT uses the fast Fourier transform (FFT). This because amino acids may be substituted for different ones given that the physiochemical properties, particularly polarity and size are similar. The correlation($c(k)$) between two amino sequences is calculated. The next step would be to find homologous regions. If two sequences are aligned and are found to have homologous regions, then there will be $c(k)$ peaks ,some peaks corresponding to these homologous regions. However FFT only allows the positional lag of k of a homologous region and not the position of the region. A sliding window analysis is carried out to determine the positions. A window size of 30 is used and the degree of local homologies is calculated for each of the first 20 highest peaks in c(k). A dividing homology matrix is constructed in order to obtain an alignment of the sequences and segments in the sequences must be arranged consistently the main aim of this step is to ensure optimal arrangement of the homologous regions (Katoh et al., 2008).

2.4.2 **Promals3D**

The first stage for Promals3D is the alignment of similar sequences making use of a scoring function of weighted sum of pairs BLOSUM62 scores. This is then followed by the second

alignment stage were one sequence is selected and represents the target sequence. PSI-BLAST searches are then subjected to the Target sequences to find additional homologs from UNIREF90 database and to PSIPRED in order to predict secondary structures. The pair of representatives is hen subjected to a hidden Markov model profile-profile alignment together with the predicted secondary structures this is done to obtain posterior probabilities of the residue matches. These probabilities serve as sequence-based constraints that are used to derive a probabilistic consistency scoring function. The representative target sequences are progressively aligned using such a consistency scoring function, and the pre-aligned groups obtained in the first stage are merged into the alignment of representatives to form the final multiple alignment of all sequences (Pei et al., 2008).

2.5 **Motif analysis**

Sequences obtained from biological databases may also be analyzed for conserved motifs. A motif, with respect to a protein sequence can be defined in four ways. The first describes conserved residues which may evolve independently from the rest of the other residues in a particular protein. These conserved motifs may perform specialized functions within the protein. Secondly, other conserved motifs may describe regions in the protein which are crucial for protein structure. These motifs may possess structural constraints which may govern the formation of secondary structural elements. Thirdly, conserved motif may be indicative of regions in the protein that serve as signal peptides, sorting signals or trans-membrane regions. Fourth and finally conserved may present regions in proteins that can be used to distinguish a particular group of proteins against another. Discovery of such motifs in a group of proteins often implies evolutionary relatedness in that group (Bork et al., 1996).

2.5.1 **MEME**

To identify motifs on protein sequences numerous programs may be used such as MEME. MEME is based the MM algorithm which takes a dataset of sequences that are not aligned and then estimates, for probabilistic model, the parameters which could have been used to create the dataset. A two component finite mixture model best describes the probabilistic method mentioned above. The first component finds a set of similar smaller sequences within the dataset. These smaller sequences are called motifs and these motifs have a fixed width. The second component of the

probabilistic method is the Background which essentially describes all other positions in the dataset. The MM model is fitted to a dataset by estimating the frequency of the specific motifs occurrence in the dataset. The MM algorithm is capable of identifying motifs on sequences with more than one occurrence of that motif. The MM algorithm is also not fazed by the sequences with zero occurrence of a particular motif. This then means that sequences containing none zero or multiple motifs can be modeled (Bailey et al., 2009).

2.5.2 **MAST**

MAST jobs can be performed in parallel with MEME jobs (Bailey et al., 2009). MAST is used to evaluating the probability that two motifs are different from each other. This is achieved through computing the pairwise correlations between each pair of motifs. The maximum, found by trying all alignments of the two motifs, is the sum of Pearson's correlation coefficients for aligned columns divided by the width of the shortest motif in the pair (Bailey and Gribskov, 1998). Pairs of motifs with higher correlations are too similar and so the two motifs are the same. The server returns a pairwise similarity table where correlations above the similarity threshold are shown in red text.

2.6 **Phylogenetic analysis**

Even though AA9 proteins are quite diverse they are known to have three types therefore the evaluation of the evolutionary relationship is important. The evolutionary relationship which occurs in these proteins will be identified through phylogenetic analysis. Phylogenetic analysis will allow for the determination of groups which occur amongst AA9 proteins as well as determining the evolutionary history amongst these proteins.

2.6.1 **Maximum likelihood**

The computer program MEGA5 (Tamura et al., 2011) is used to infer evolutionary relationships in proteins. The evolutionary relationship can be estimated using the maximum likelihood method (ML). The ML method utilizes a likelihood function to find a tree that best describes the phylogenetic relationship. Therefore maximum likelihood methods find the most probable tree that would most likely describe the evolution of the observed data (Felstein, 1981).

2.7 Physico-chemical property analysis

Numerous studies have concluded that the physico-chemical properties of proteins play a crucial role in the functionality of a protein. The physicochemical properties displayed by a particular protein are determined by the amino acids present on the protein. A protein's functional properties are in turn determined by the protein's physicochemical properties. As a result it is important to fully understand the physic-chemical properties of a protein.

2.7.1 Aromaticity of proteins and hydrophobicity

Using Python programming various physicochemical properties of protein sequences can be investigated these include, but not limited to aromaticity and hydrophobicity. The Python module ProtParam comes equipped with the functionality of elucidating physicochemical properties of protein sequences. The aromaticity of a protein can be elucidated using the method previously developed by Lobry, 1994, which computes the relative frequencies of aromatic residues (phenylalanine, tryptophan, histidine and tyrosine. The ProtParam module is also capable of computing the hydrophobicity of a protein. The hydrophobicity of a protein is computed through the use of protein scales which assign a value to each amino acid. Even though there are numerous others, the most commonly used scale is Kyte Doolittle (kd) scale (Kyte et al., 1982). The kd scale is shown in table 2.1 below

2.8 Aims of the chapter

The main emphasis of the chapter is to analyze sequence features presented by AA9 proteins which can be used to distinguish between the various AA9 types. To identify these sequence elements, the techniques mentioned in this chapter will be employed. Results obtained from the various analyses will be scrutinized for their ability to provide significant information in terms of distinguishing between the three AA9 types as well as gaining better insights into these enzymes.

Table 2.1 Kyte and Doolittle hydrophobicity scale

Residue Type	Kd-Hydrophobicity
Ile	4.5
Val	4.2
Leu	3.8
Phe	2.8
Cys	2.5
Met	1.9
Ala	1.8
Gly	-0.4
Thr	-0.7
Ser	-0.8
Trp	-0.9
Tyr	-1.3
Pro	-1.6
His	-3.2
Glu	-3.5
Gln	-3.5
Asp	-3.5
Asn	-3.5
Lys	-3.9
Arg	-4.5

2.9 Methodology

2.9.1 Sequence retrieval

The Pfam database (http://pfam.sanger.ac.uk) was used to obtain all UniProt sequences which contained the AA9 domain (Pfam id: PF03443). There were a total of 827 AA9 sequences which belonged to 87 fungal and one plant organisms. Fungal organisms generally encode multiple copies of the AA9 protein in their genomes. As a result there were multiple fungal organisms with multiple AA9 sequences. These sequences where respectively stored for all 88 organisms. All AA9 sequences for each respective organism where individually named from one to the total number of sequences within that organism. Once all AA9 where obtained the sequences where inspected to assess the dataset. Using the Clean_redunt.py script redundant sequences were removed. The sequences were combined and aligned using MAFFT (http://toolkit.tuebingen.mpg.de/mafft/). Short fragments and highly divergent sequences where found and these sequences where removed

from the dataset. The number of sequences in the final dataset was 160 sequences. The accession numbers of the initial dataset can be found in Supplementary Data 1. There are currently a number of AA9 protein sequences which are known AA9 to be specific AA9 types, these sequences where described by Li et al., 2012 (Table 2.2). The sequences in Table 2.2 were used as reference sequences in the analyses to be carried out.

Table 2.2: Sequences already identified as AA9 types

Sequence type	Accession number and PDB id
Type 1	NCU08760
	3EJA
	NCU03328
	NCU02344
	NCU00836
Type 2	NCU01050
	NCU02240
	NCU02916
Type 3	NCU07898
	3ZUD
	2VTC
	NCU05969
	NCU07760

2.9.2 **Multiple sequence alignment**

MAFFT was used to align the original sequences so that problematic sequences could be removed. Promals3D (http://prodata.swmed.edu/Promals3d/Promals3d.php) were used to align the 160 AA9 sequences together with the reference sequences. The alignments where inspected and it was found that the presence of other domains made analysis of alignments difficult. As The AA9 domain with the signal peptide region was extracted using Jalview. Raw sequences were extracted using the gap_remover.py script. The sequences were then re-aligned using Promals3D. The resulting alignment possessed problematic regions. As a result Jalview was used to manually adjust the

problematic regions. The addition of the reference sequences into the alignment allowed for the identification of sequence features specific to each AA9 types which can then be used to classify unknown AA9 Proteins.

2.9.3 **Motif analysis**

The webserver MEME (http://www.meme.nbcr.net/meme/cgi-bin/meme.cgi) was used to identify globally conserved motifs in AA9 sequences. The dataset containing full length AA9 sequences as well as the references were uploaded to the MEME webserver. MEME was instructed to identify motifs with a minimum and maximum width of 5 and 50 respectively. The number of motifs to be identified was varied from 5, 10 and 15. The optimum number of sites for each motif was not specified. The MEME output was parsed through MAST to evaluate the similarities of these motifs.

2.9.4 **Polygenetic analysis**

The evolutionary relationship of the AA9 proteins in the dataset was evaluated using MEGA5. MEGA5 was used to evaluate the best amino acid substitution models for conducting phylogenetic analysis. Once the best models were evaluated, phylogenetic analysis was carried out. Maximum Likelihood tree was constructed based on the Whelan and Goldman + Freq. model since it was found to be the best model. The initial tree for heuristic search was obtained by applying the neighbor joining method to a matrix of pairwise distances estimated using a JTT model. The confidence of the inner node topology was evaluated using 500 bootstrap replicates. The same tree was constructed using the 100% site The Promals3D alignment generated during the alignment stage was used. The AA9 domain was extracted from alignment before the analysis.

2.9.5 **Physicochemical property analysis**

Once all sequences were separated into their types the analysis of the aromaticity of each type was evaluated. The script Physicochemical_properties.py was used to evaluate the aromaticity.

2.10 **Results and discussion**

2.10.1 **Physicochemical property analysis**

Aromatic residues have long been implicated in in the importance of AA9 activity. As a result the aromaticity of all three AA9 types was identified using the ProtParmam module. The aromaticity values can be viewed in the supplementary material Table A.2

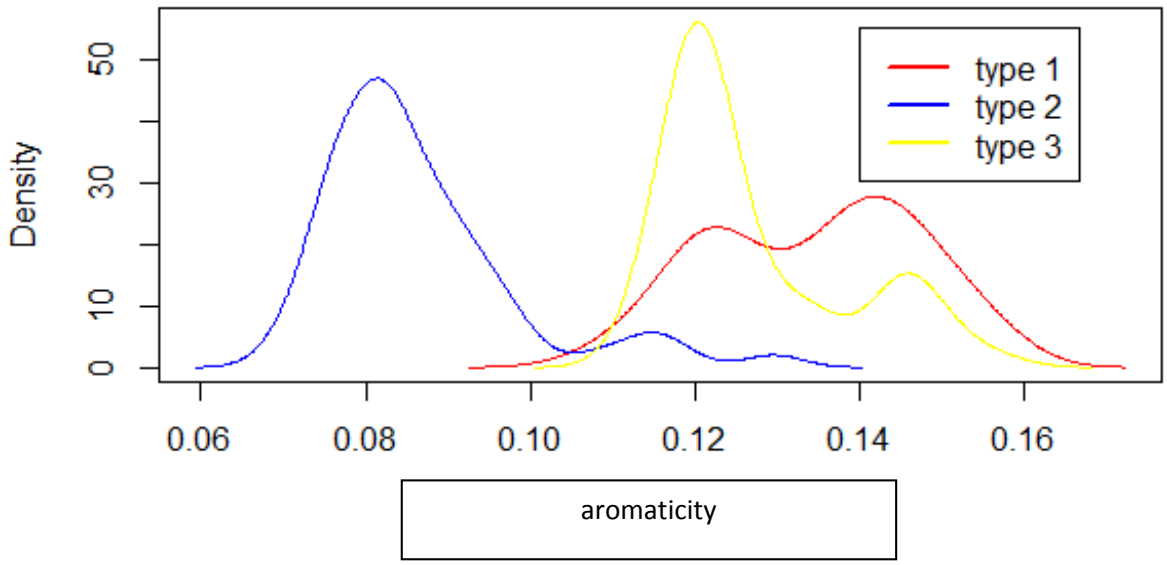


Figure 2.1: Density distribution of the aromaticity for type 1, 2 and 3 protein sequences

Density distribution of all the AA9 types is displayed in figure 2.1. The density of type 2 sequences is mostly concentrated at the lower aromaticty range with an average of 0.08635783 while type 1 and 3 had similar distributions with averages of 0.1345318 and 0.1268923. Using The Kolmogorov-Smirnov test (KS-test) it was found that the Cumulative frequency distribution of type 2 aromaticity was, at 5% level of significance, less than that of type 1 and type 3 sequences. At the 5 % level of significance there was insufficient evidence to suggest that type 1 and 3 sequences had different aromaticity distributions.

2.10.2 **Multiple sequence alignment**

The Promals3D alignment revealed high sequence variability in the C-Terminus of aligned AA9 proteins and as a consequence the AA9 domains with the signal peptides were extracted and re-aligned using Promals3D (Figure 2.2 A and 2.2B) . The resulting alignment revealed distinct groups within the alignment which contained specific features. These features where correlated

with specific features present on the reference sequences to determine if the AA9 proteins group according to type.

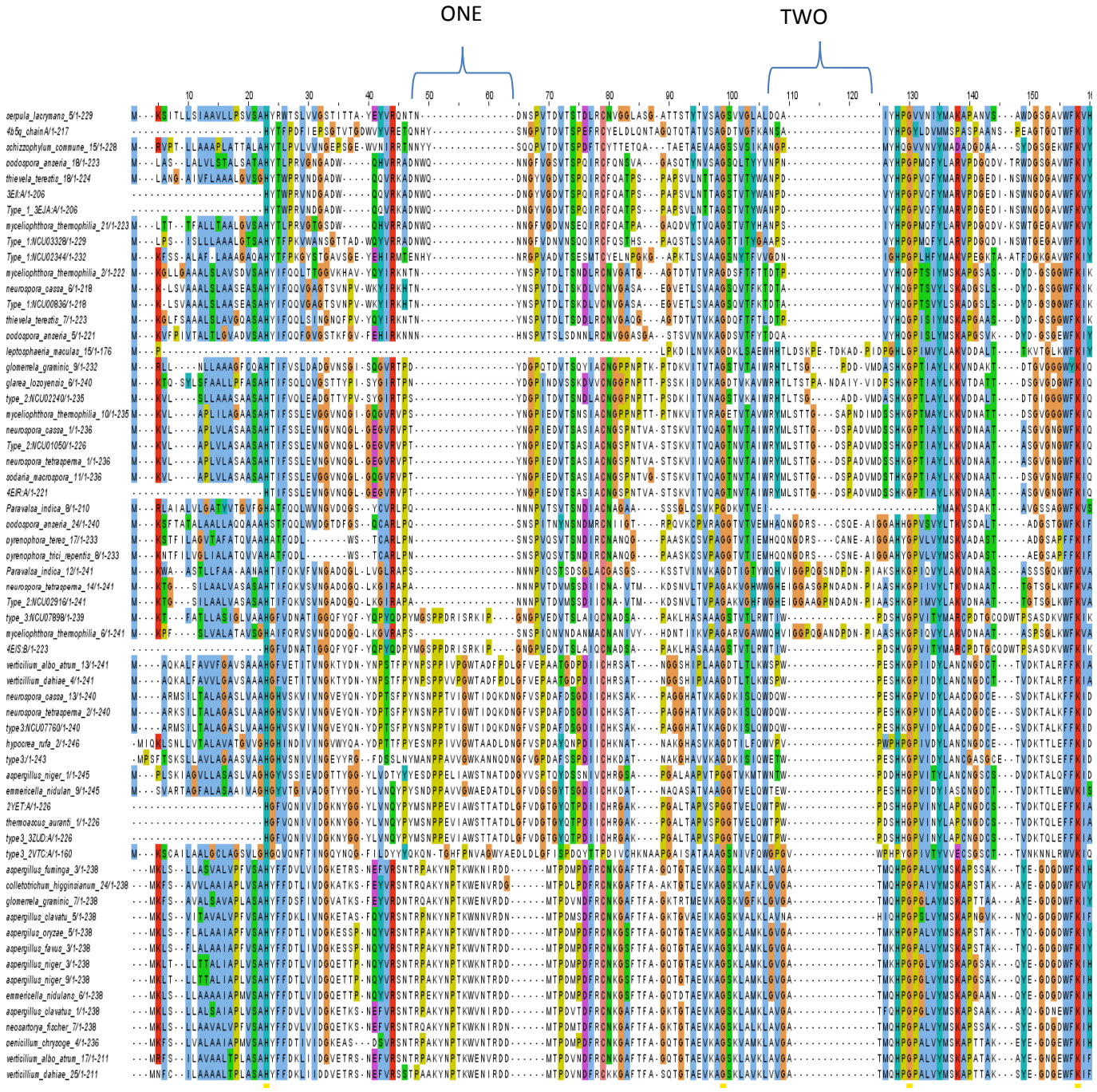


Figure 2.2A: Promals3D alignment of AA9 domain. Only the N Terminus is displayed major insertions are illustrated (ONE and TWO). Sequences which were similar to the type 1 reference sequences are highlighted in yellow, sequences similar to type 2 reference sequences are highlighted in red, sequences that were similar to type 3 reference sequences are highlighted in blue and sequences that were not similar to any of the reference sequences are highlighted in green**.**

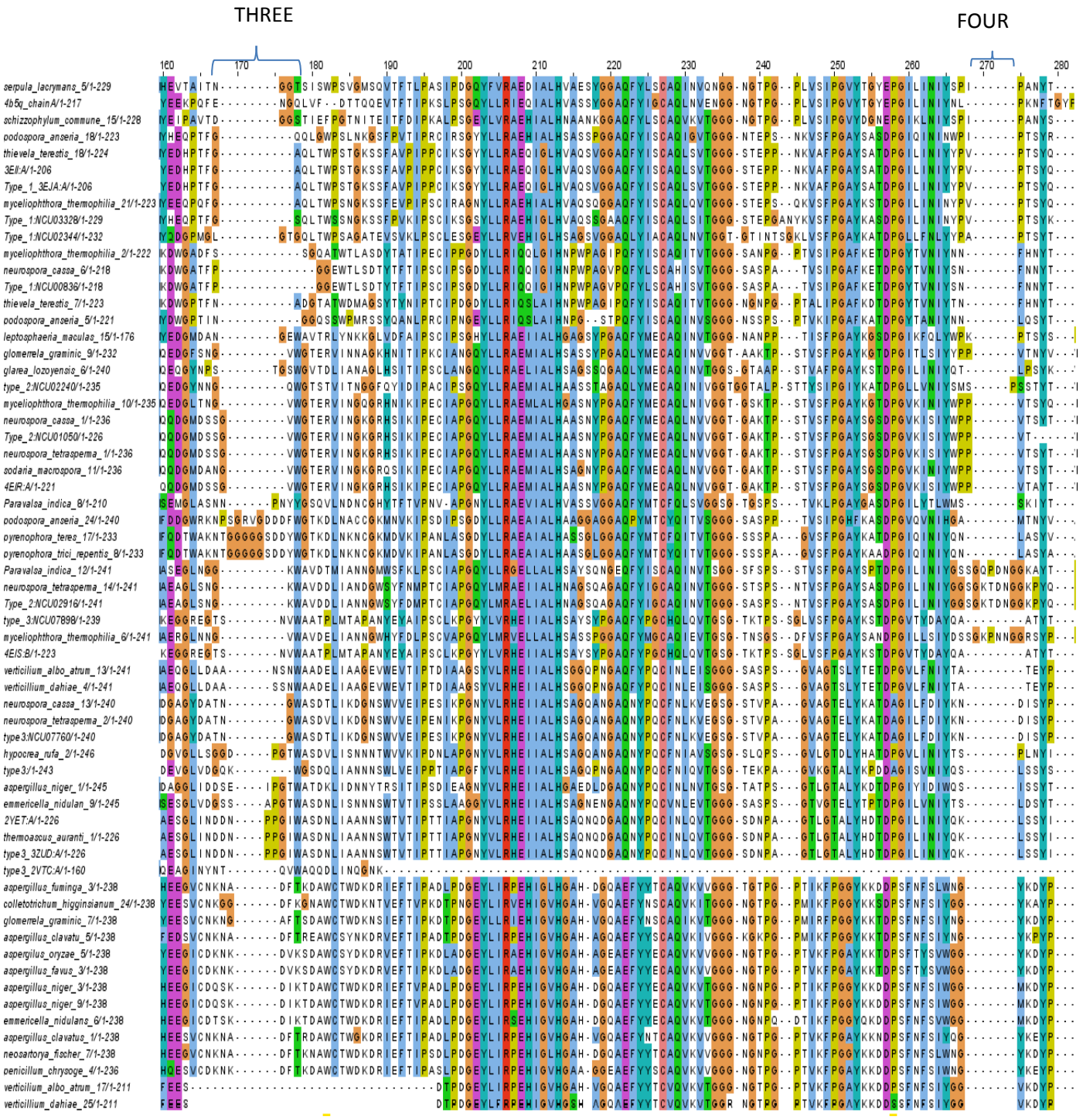


Figure 2.2B: Promals3D alignment of AA9 domain. Only the N Terminus is displayed major insertions are illustrated (THREE and FOUR). Sequences which were similar to the type 1 reference sequences are highlighted in yellow, sequences similar to type 2 reference sequences are highlighted in red, sequences that were similar to type 3 reference sequences are highlighted in blue and sequences that were not similar to any of the reference sequences are highlighted in green.

Aligned AA9 protein sequences where found to form distinct groups within the alignment. These AA9 groups can easily be identified based on the presence or absence of inserts ONE and TWO

(Figure 2.2 A). There were four observed groups the first of which were sequences which only had insert ONE. Reference sequences for type 1 were found to be associated with this group. The second group of sequences observed contained insert two which included all type two reference sequences. The third group of sequence observed did not contain either insert ONE or TWO and this group sequences was associated with type 3 reference sequences. A fourth group, which contained a relatively smaller insert ONE and lacked insert TWO, was observed. The N-Terminus (Figure 2.2B) of the aligned AA9 protein sequences was more conserved when compared to the C-Terminus. There was observable variability in this region, even though this variability did not appear to have an impact on the grouping of these proteins.

2.10.3 **Phylogenetic analysis**

To investigate the evolutionary relationship which occurs between various AA9 types phylogenetic analysis was carried out using MEGA5. The best amino acid models for conducting the phylogenetic investigation of AA9 are displayed in table 2.3. The amino acid substitution models were also identified these are shown in table A5. The observed clustering of the sequences with the data set with the reference sequences was used to determine which sequences were of a specific type. The results of the analysis are displayed in figure 2.3.

Table 2.3. Maximum Likelihood fits of 48 different amino acid substitution model using 95% site coverage

Model	Parameters	BIC	AICc	*lnL*
WAG+G+I	323	52315.445	49645.130	-24495.969
WAG+G	322	52324.623	49662.553	-24505.702
WAG+G+I+F	342	52439.266	49612.326	-24460.129
WAG+G+F	341	52448.580	49629.882	-24469.931
rtREV+G+I	323	52735.063	50064.748	-24705.777
rtREV+G+I+F	342	52901.519	50074.579	-24691.255
rtREV+G+F	341	52908.923	50090.225	-24700.102
rtREV+G	322	52938.188	50276.118	-24812.485
Dayhoff+G+I	323	52989.606	50319.291	-24833.049
Dayhoff+G	322	53001.463	50339.393	-24844.122

The data in Table 2.3 shows the 10 best amino acid substitution models identified for the dataset. The analysis was carried out at 95 % site coverage. The top 3 models were used to conduct phylogenetic analysis. The best tree was chosen based on significant bootstrap values and the best tree is shown in figure 2.3.

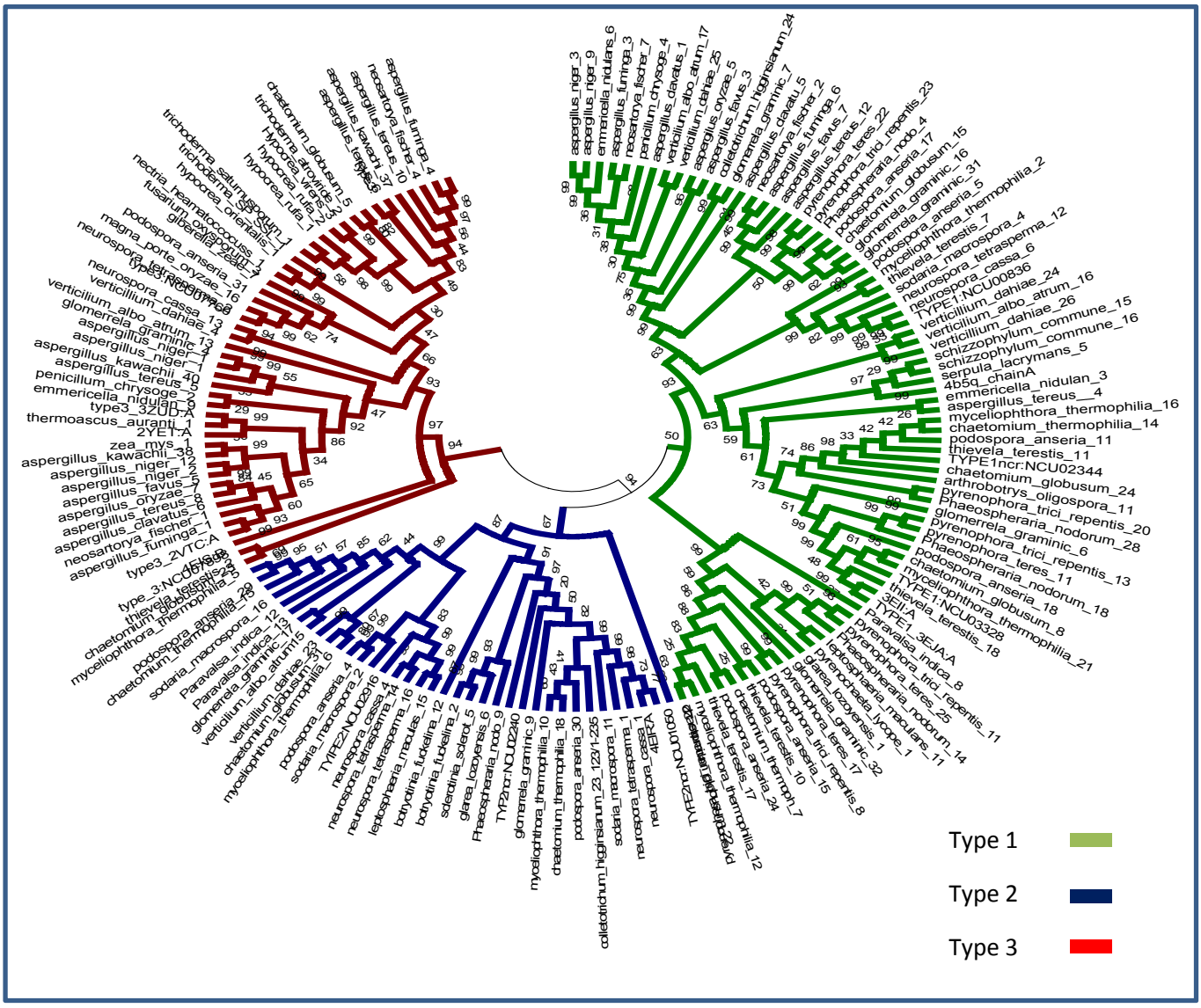


Figure 2.3: Molecular Phylogenetic analysis by Maximum Likelihood method of AA9 proteins at 95% site coverage. Type 1 sequences are highlighted in red. Type 2 sequences are highlighted in red. Type 3 sequences highlighted in green and an unknown group of sequences shown in pink dashed lines.

Phylogenetic analysis revealed that AA9 sequences separate out into three distinct clusters based on their types. This is observation was made by identifying where the reference sequences would

occur in the phylogenetic tree. Figure 2.3 reveals that type 3 forms a separate mono phyletic group which is a sister group to both type 2 and type 1. After diverging from each other type 1 and type 2 can be seen forming separate monophyletic groups with respect to each other. The tree displayed in figure reveals the formation of three distinct groups which are based type. The tree was constructed using 95% site coverage and this produced a tree with favorable bootstrap values. When the same tree was constructed using 100% site coverage **(**Figure A10), a similar clustering of sequences was observed. However the bootstrap values for this tree were lower than those observed figure 2.3.

2.10.4 **Motif analysis**

Motif analysis was conducted investigate the presence of sequence motifs displayed by AA9 protein dataset. The motifs analysis was carried to investigate sequence features that are specific to certain AA9 types. The webserver MEME was used to identify motifs. 20 motifs where identified (Table A1) in the dataset and in the reference sequences. The organization as well as the presence and absence of certain motifs appeared to play an important role in AA9 proteins. To clearly show the motifs presented by AA9 proteins, the three separate protein clusters observed the phylogenetic tree constructed in figure 2.3 were separated into 3 different datasets. Each of the datasets was re-aligned and smaller phylogenetic trees were constructed. The results are shown in figure 2.4 (Type 1 data set), figure 2.5 (Type 2 dataset) and (Type 3 dataset).

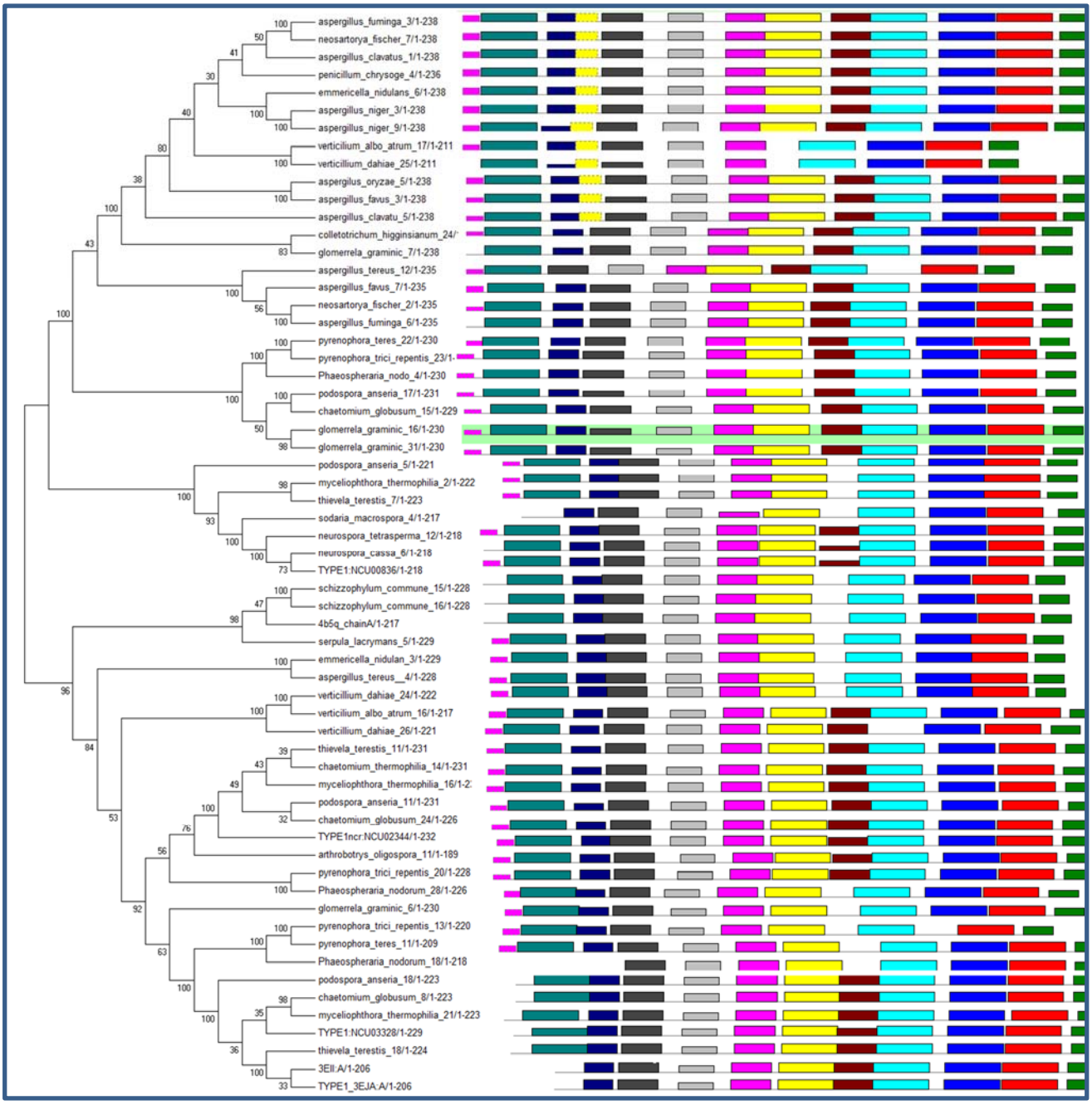


Figure 2.4: Phylogenetic tree illustrating the motif organization present on Type 1 protein sequences.

The motifs colours correspond to MEME colour coding**.**

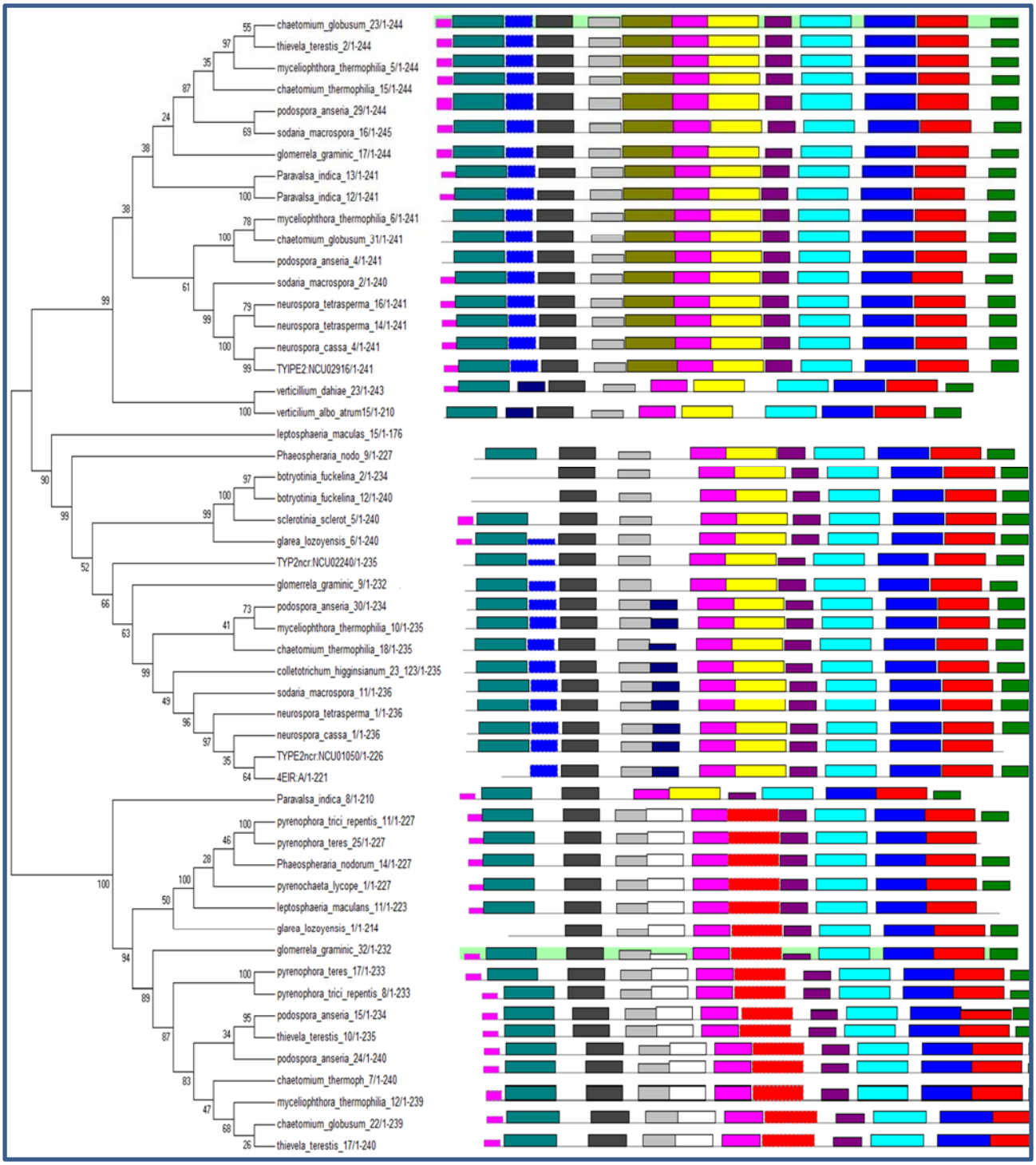


Figure 2.5: Phylogenetic tree showing the various motif organizations displayed by type 2 AA9 proteins. The motifs colours correspond to MEME colour coding.

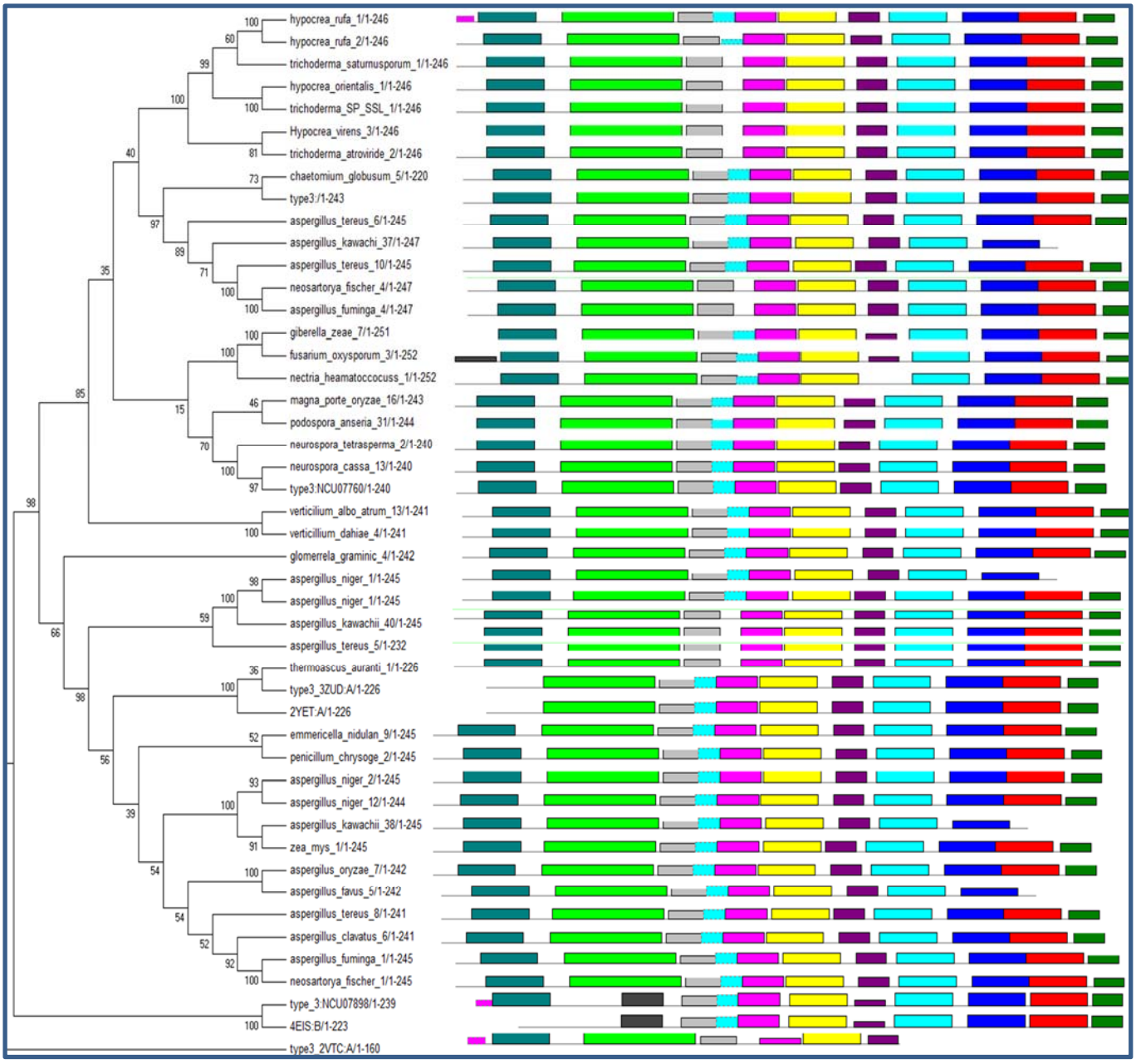


Figure 2.6 Phylogenetic trees showing the various motif organizations displayed by type 3 AA9 proteins. The motifs colours correspond to MEME colour coding.

Phylogenetic analysis of separate types was carried out and the result of this analysis is displayed in figure 2.4, 2.5 and 2.6 for type 1, 2 and 3 respectively. It was found that for each of the 3 respective types, there were variants observed. The variants where identified through motif analysis. The absence and presence of certain motifs was found used to identify variants in each

AA9 type. The use of the reference sequences was useful in the identification of AA9 Clustering however motif analysis was able to detect certain sequence differences at certain positions. Figure 2.5 revealed that the type 1 sequences formed two variants. These variants were identified by the presence of motif 20 which is exclusive to type 1 sequence but is not found in all of them.

Table 2.4: Type 1 determining motifs

	Motif		
Variant	13	14	20
1	*+	*+	
2	*+	*+	*

Table 2.4 illustrates the motifs that determine type 1 AA9 sequences. The motifs found to determine type 1 AA9 sequences were motifs13, 14 and 20. Motifs 13 and 14 were both found to be absent in some of the sequences given that one of them is present on the sequences. Motif 20 was only specific group of sequences and so this group of sequence was regarded as a variant.

Table 2.5: Type 2 determining motifs

	Motifs						
Variant	**5**	**11**	**12**	**13**	**15**	**17**	**18**
1	*	*		*		*	
2		*			*	*+	*
3	*	*	*			*+	

Motifs which determine type two sequences. Three type 2 variants where identified through motif analysis. These type 2 variants were characterized by the presence of certain motifs shown in table 2.5. The first type two variant was found to have motif 5, 11, 13 and 17. The second type two variant was found to have motif 11, 15, 17 and 18. For variant 2 motif 17 was found to be a type two determining motif however its presence is not required since it was found to be absent in some variant two sequences. The third variant observed for type 2 sequences were motifs 5, 11, 12 and 17. Similar to variant 2, variant one was not two specific to the presence of motif 17. It is important to observe that even though motif 5 is a common motif in all AA9 types, it is can be replaced with motif 18 in type 2 sequences. Motif 12, 13 and 15 are exclusively type 2 motifs which are located in relatively the similar positions in type 2 sequences.

Table 2.6: Type 3 determining motifs

	Motif		
Variant	**6**	**11**	**16**
1	*		*
2	*	*	*

Type three sequences displayed the presence of two subtypes shown in table 2.6. The first type three variant contained motif 6 and 16. The second type 2 variant also possessed motif 6 and 16 however the additional motif 11 was also present.

The characterization of AA9 sequences by Motif analysis revealed that all full length AA9 proteins analyzed possess any one of the motif arrangements displayed in figures 2.2, 2.3 and 2.4. Motifs 5, 11, 12, 13, 14, 15, 16, 17 and 18 were found to be important for differentiating between the various AA9 types and the newly discovered subtypes. The remaining motifs were found in all sequences analyzed suggesting that these motifs do not play any role in distinguishing between the various AA9 types.

2.10.4.1 **Similarity and motif comparison**

MEME was able to identify numerous motifs on the AA9 proteins which allowed for differentiating between the AA9 types. The motifs which are type specific were found to occur in overlapping regions. In order to observe how similar these overlapping motifs are to each other MAST was used. The result of the analysis is shown in Table 2.7.

Table 2.7: Pearson correlation coefficients of the discovered motifs computed using MAST

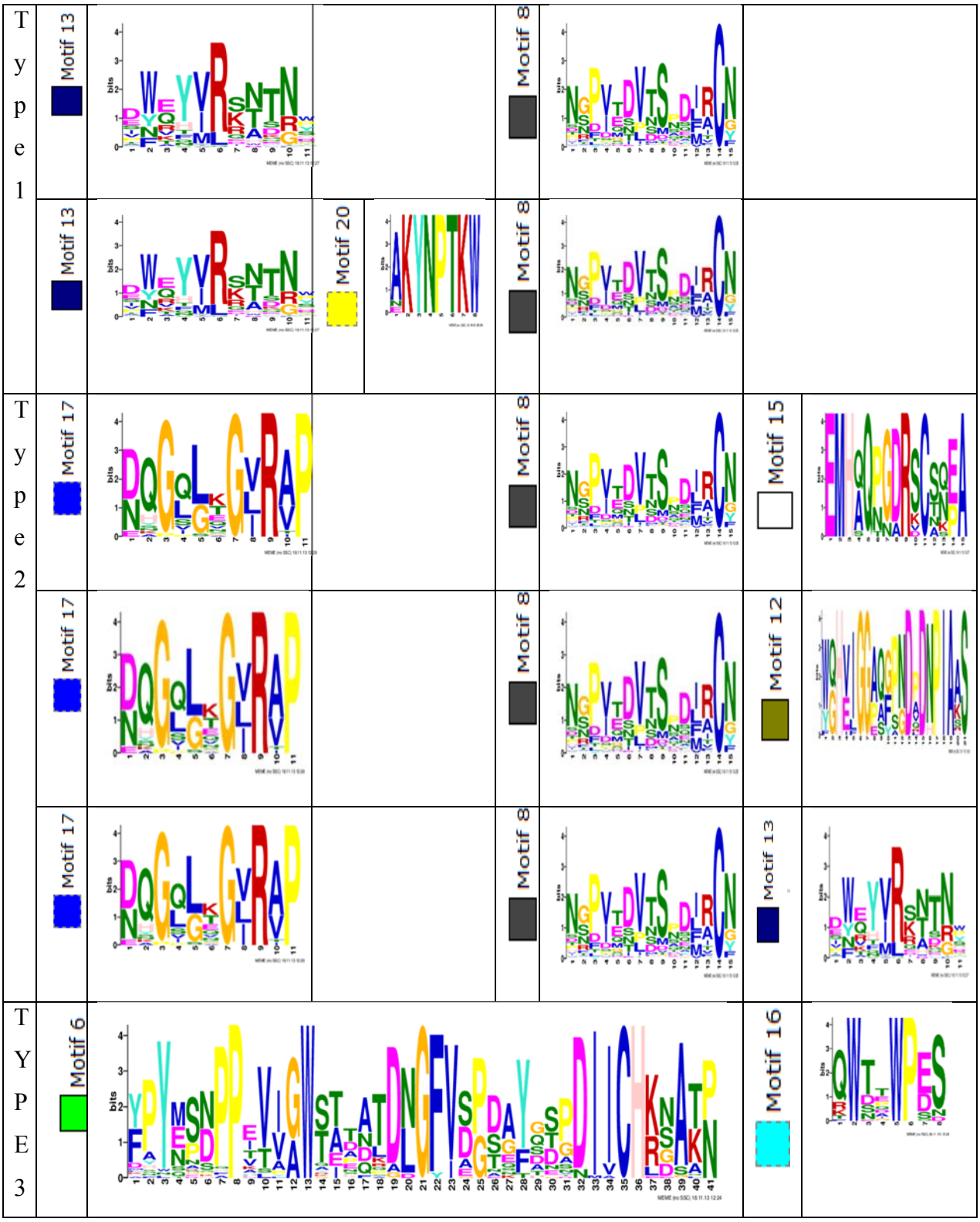
Motif	Width	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	21	-	0.14	0.19	0.19	0.2	0.22	0.16	0.23	0.29	0.31	0.25	0.12	0.22	0.15	0.18	0.35	0.24	0.15	0.33	0.15
2	21	0.14	-	0.17	0.34	0.2	0.22	0.17	0.18	0.21	0.31	0.29	0.16	0.2	0.19	0.21	0.19	0.2	0.23	0.28	0.2
3	21	0.19	0.17	-	0.2	0.19	0.23	0.19	0.15	0.23	0.24	0.28	0.19	0.22	0.21	0.25	0.23	0.27	0.18	0.28	0.23
4	15	0.19	0.34	0.2	-	0.19	0.28	0.25	0.19	0.17	0.19	0.26	0.15	0.25	0.14	0.15	0.25	0.23	0.18	0.27	0.22
5	21	0.2	0.2	0.19	0.19	-	0.31	0.17	0.17	0.19	0.23	0.23	0.11	0.26	0.3	0.15	0.28	0.3	0.48	0.21	0.22
6	41	0.22	0.22	0.23	0.28	0.31	-	0.29	0.32	0.35	0.26	0.3	0.19	0.22	0.26	0.24	0.31	0.32	0.21	0.18	0.24
7	21	0.16	0.17	0.19	0.25	0.17	0.29	-	0.29	0.17	0.25	0.28	0.12	0.21	0.18	0.14	0.24	0.27	0.22	0.47	0.15
8	15	0.23	0.18	0.15	0.19	0.17	0.32	0.29	-	0.21	0.23	0.26	0.22	0.16	0.18	0.22	0.27	0.25	0.18	0.2	0.16
9	11	0.29	0.21	0.23	0.17	0.19	0.35	0.17	0.21	-	0.34	0.16	0.17	0.13	0.22	0.17	0.16	0.2	0.22	0.13	0.2
10	13	0.31	0.31	0.24	0.19	0.23	0.26	0.25	0.23	0.34	-	0.24	0.15	0.2	0.36	0.19	0.19	0.28	0.23	0.32	0.13
11	11	0.25	0.29	0.28	0.26	0.23	0.3	0.28	0.26	0.16	0.24	-	0.23	0.29	0.23	0.28	0.22	0.15	0.29	0.14	0.12
12	21	0.12	0.16	0.19	0.15	0.11	0.19	0.12	0.22	0.17	0.15	0.23	-	0.26	0.24	0.25	0.34	0.26	0.16	0.15	0.23
13	11	0.22	0.2	0.22	0.25	0.26	0.22	0.21	0.16	0.13	0.2	0.29	0.26	-	0.3	0.23	0.29	0.15	0.22	0.19	0.14
14	15	0.15	0.19	0.21	0.14	0.3	0.26	0.18	0.18	0.22	0.36	0.23	0.24	0.3	-	0.19	0.41	0.19	0.21	0.18	0.26
15	15	0.18	0.21	0.25	0.15	0.15	0.24	0.14	0.22	0.17	0.19	0.28	0.25	0.23	0.19	-	0.27	0.26	0.2	0.1	0.2
16	8	0.35	0.19	0.23	0.25	0.28	0.31	0.24	0.27	0.16	0.19	0.22	0.34	0.29	0.41	0.27	-	0.15	0.28	0.1	0.2
17	11	0.24	0.2	0.27	0.23	0.3	0.32	0.27	0.25	0.2	0.28	0.15	0.26	0.15	0.19	0.26	0.15	-	0.28	0.22	0.09
18	21	0.15	0.23	0.18	0.18	0.48	0.21	0.22	0.18	0.22	0.23	0.29	0.16	0.22	0.21	0.2	0.28	0.28	-	0.25	0.22
19	6	0.33	0.28	0.28	0.27	0.21	0.18	0.47	0.2	0.13	0.32	0.14	0.15	0.19	0.18	0.1	0.1	0.22	0.25	-	0.22
20	8	0.15	0.2	0.23	0.22	0.22	0.24	0.15	0.16	0.2	0.13	0.12	0.23	0.14	0.26	0.2	0.2	0.09	0.22		

The motifs which appear to determine AA9 types where investigated for their similarity with each other using MAST. To check how similar the overlapping motifs are, a Pearson coefficient motif similarity matrix (Pictrokavoski, 1996) was generated using MAST. All motifs analyzed produced Pearson coefficients that where below 0.6 which then suggests that the motifs identified where different from each other.

2.10.4.2 *Characterization of overlapping motifs*

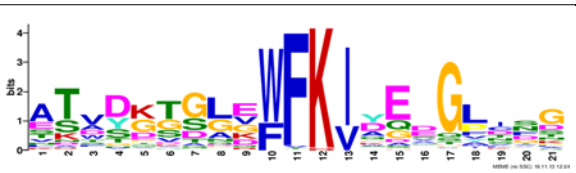
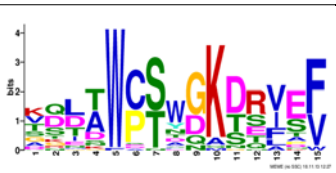
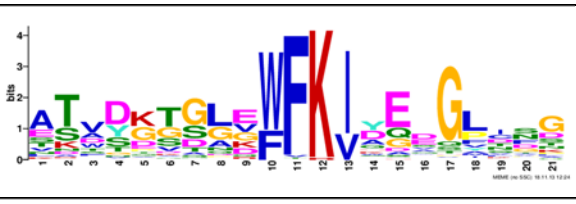
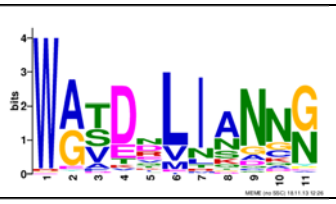
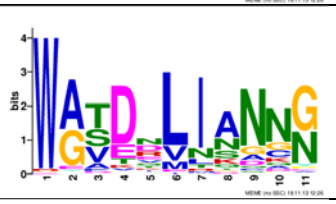
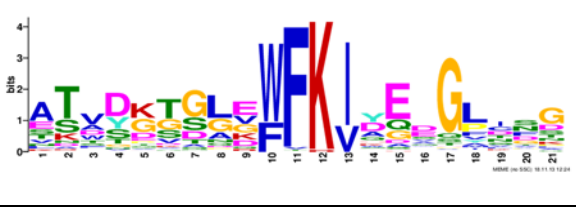
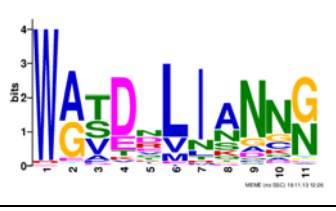
Motifs which were found to occur in relatively the same positions where further characterized in terms of the composition of residues (Table 2.7). This was carried out to observe any significant changes in residue composition which was brought on by the difference in motifs on the AA9 sequences.

Table 2.8: Comparison of the C-terminus type specific motifs


Type 1
Motif 13
Motif 8
Motif 13
Motif 20
Motif 8
Type 2
Motif 17
Motif 8
Motif 15
Motif 17
Motif 8
Motif 12
Motif 17
Motif 8
Motif 13
TYPE 3
Motif 6
Motif 16

Based on motif analysis the N-Terminus Terminus region of AA9 Sequences is important for the type specificity. Table 2.8illustrates the relative positions of the N-Terminus type determining motifs columns show the motifs which occur in relatively the same positios.in is apparent from table 2.7 that the N-Terminus of all AA9 proteins changes in composition and this appear to have an impact on type specificity motif 13 according to table 2.8 is found in both type 1 variants and a single type 2 variant. However this motif is not found in the same relative position in the sequences of both these types as shown in figure 2.5 and 26.

Table 2.9: Comparison of the N-terminus motifs type specific motifs

TYPE 1	Motif 5		Motif 14	
TYPE 2	Motif 5		Motif 11	
	Motif 18		Motif 11	
TYPE 3	Motif 5		Motif 11 +	

The C-Terminus of AA9 protein was found to have variation in AA9 proteins however, this variation was not as extensive as the variation observed in the N-Terminus according to the data in Table 2.8 and 2.9.

2.11 The identification of variation in different AA9 types by multiple sequence alignment

The individual type 1, type 2 and type 3 sequences where realigned using Promals3D to observe distinct sequence features which are specific to AA9 types. MEME Motifs logos were mapped out to the individual alignments to aid in the visualization of sequence features. The results of the analysis are shown in figure 2.7, 2.

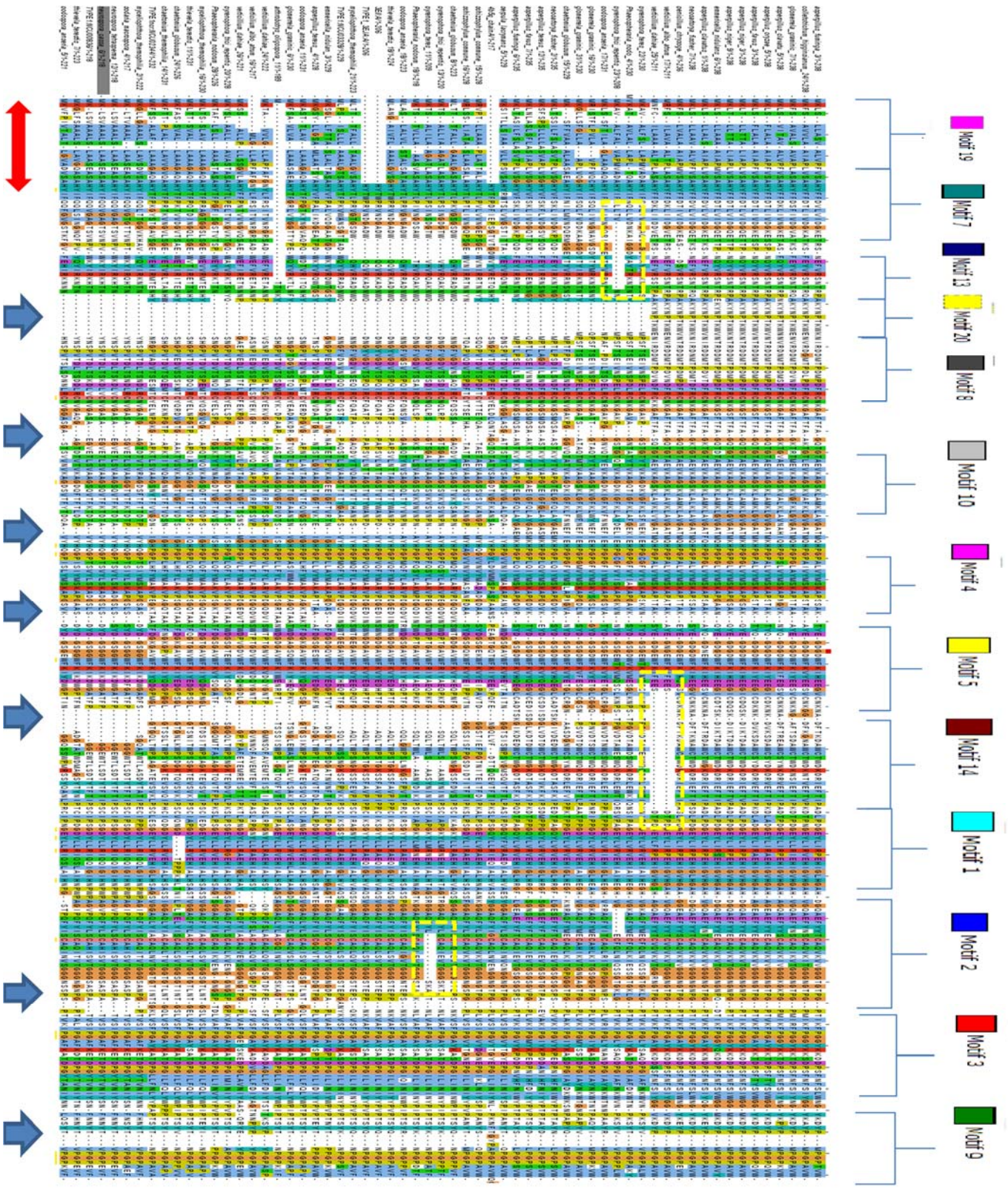


Figure 2.7: Promals3D alignment of type 1 protein sequences. Motifs are indicated.

Figure 2.8: Promal3D alignment of type 2 protein sequences. Motifs are shown.

Figure 2.9: Promals3D alignment of type 3 protein sequences. Motifs are shown**.**

Figures 2.7, 2.8, and 2.9 all reveal that AA9 sequences are fairly well conserved. However there were regions of variability which were observed in all three types. Most AA9 type contains a cleavage signal peptide indicated by the red arrow. This region is represented by motif 5 and is highly variable. The blue arrows indicate regions which contain inserts or highly variable regions. All AA9 types contain inserts and variable regions which are distributed through the entire span of the respective types. Out of the 3 types, type 3 sequences were found to be the most conserved sequences since these sequences contain the least number of inserts and variable regions. The highly conserved 50–KYNTWP-56 insert was observed in 13 type 1 sequences in figure 2.7. This insert was found to correspond with motif 20. The yellow dashed line in the alignments indicates deletions in these sequences. Often the observed deletions results in the absence of part or even complete motifs in a sequence.

2.12 **Conclusion**

The sequence analysis conducted in this chapter was able reveal numerous sequence features on AA9 proteins. It was observed that type 1 AA9 sequences have a distribution of aromatic residues that significantly differs from that of type 2 and 3 proteins. Type 2 and 3 sequences had comparable aromaticity data distributions. MSA was able to reveal that AA9 sequences are highly variable especially in the N-Terminus of the sequences. This variability was based on the observation of inserts in this region and was found that the presence or absence of certain inserts plays a role in type specificity. Motif analysis of the N and C-Terminus of AA9 (table 2.8 and 2.9 respectively) supports the results observed for MSA in figure 2.2A and 2.2B which reveal that the AA9 sequences are more conserved in the N-Terminus as opposed to the C-Terminus. Phylogenetic analysis of AA9 sequence revealed the formation of three distinct groupings of AA9 sequences. Through the use of references it was discovered that the observed grouping of AA9 sequences was based on types. Motif analysis was also capable of distinguishing between the various AA9 types

as well as identifying variants which occur amongst AA9 proteins. Motif analysis revealed that it was only specific regions on AA9 proteins that are responsible for type specificity. The motifs identified were also found to offer a means of identifying type variants. It was also found that deletion present on certain AA9 sequences results in the absence of certain motifs.

CHAPTER 3

3 Structural analysis

It has been shown that protein structure and function are more conserved than the sequence. As such, it is possible to have different proteins sharing low sequence similarity and identity but having similar functions. Protein structures are often determined by experimental means such as X-ray crystallography or nuclear magnetic resonance (NMR). However both these approaches tend to be time consuming and expensive. To better understand the function of a protein it is often necessary to understand the structural properties of the proteins. This then requires a viable 3D representative model (Marti-Renom et al., 2000; Fiser, 2004; Misura and Baker, 2005; Petrey and Honig, 2005; Misura et al., 2006). When a NMR or X-ray structure is not available computational techniques such as comparative (homology modeling) can be used.

3.1 Comparative modeling

Comparative modeling is computer based technique for predicting the 3D structure of a query sequence based on the alignment of this query sequence with another protein with a known structure (template). Comparative modeling has four steps which include: fold assignment or template identification, aligning the query sequence with the template, model building and model evaluation (Cavasotto et al., 2009).

3.1.1 Template Identification

The identification of a suitable template is a crucial step in homology modeling because the accuracy of the model generated will depend on the template used. Therefore the identification of a suitable template from a suitable database such as the PDB is important. Tools such as HHpred can be used to identify a suitable template for a particular protein. HHpred searches databases to identify a template based on a query which can be in the form of a single sequence or a multiple sequence alignment. HHpred initially generates an alignment of homologous sequences to the query sequence using PSI-BLAST. The homologous sequences are obtained from the NCBI non-

redundant database. HHpred allows the user to specify input parameters such as the E-value threshold, PSI-BLAST iterations, the minimum sequence identity and the minimum number of PSI-BLAST matches. When a PSI-BLAST match is obtained it is annotated with a predicted secondary structure by PSIPRED (Söding et al., 2005). The next step involves the generation of a profile hidden Markov model (HMM) from the multiple sequence alignment. The profile HMM is a statistical representation of the alignment, each column in the profile contains the probability of each of the 20 amino acids (Söding et al., 2005. Factors to consider when choosing a template to use for the homology modeling process are the quality of the template and how much coverage do the target sequence and the template structure have with each other (di Luccio and Koehl, 2011).

3.1.2 **Template query alignment methods**

There are numerous ways to align the query sequence with the template. The first of which is pair wise alignment. Given that sequence share a high enough sequence identity, usually in the range of 30 to 40 % and above, pairwise alignment can be used. This method is regarded as being very quick however this method fails for alignment of template query combinations of low sequence identity. The accuracy of the model building process heavily relies on the accuracy of the alignment as a result almost all errors which are obtained from the alignment step will be carried forward to the final model (Sanchez and Sali, 1997).

3.1.3 **Homology modeling using spatial restraints**

MODELLER is a comparative modeling program that models proteins by satisfying of spatial restraints. Modeler generates many constraints on the query sequence structure using the alignment and the template structure to guide the process. To create the restraints, it is assumed that the distances corresponding between the aligned residues of the template and query are similar. Stereo chemical restraints on bond lengths, bond angles, dihedral angles, and non-bonded atom-atom contacts are used to supplement the previously mentioned homology derived restraints. A model is generated by minimizing all the deviation which occur from the restraints. Another important consideration in terms of comparative modeling is loop modeling. Loop modeling is crucial because even if models may share a high sequence identity the loop regions will tend to show variability while the core regions will be more relatively conserved. Loops are often responsible for specificity of protein interaction such as ligand binding (Eswar, N. et al., 2007).

3.1.3.1 **Model evaluation**

It is important to evaluate the accuracy of the model generated because the amount of information that can be extracted from the model will depend on how accurate the generated model is. There are numerous methods which can be used to evaluate protein model quality. To evaluate models scoring functions can be used. These scoring functions are based on the observed properties of amino acids of already known protein structures models can also be evaluated by comparing structural feature of a generated model to those observed in actual experimentally generated 3D structures (Xiang, 2006). Model assessment tools are explained below.

3.1.3.1.1 Normalized DOPE Score

Native proteins, under the native conditions, are known to possess the lowest free energy of all states. The discrete optimized energy (DOPE) score is a statistical potential to measure of protein stability. The DOPE score is distance dependent and has been derived from a sample of native structures. However one cannot compare two different proteins together using the DOPE score. The DOPE score can be normalized to allow for the comparison of two different structures (Z-DOPE) (Shen & Šali, 2006).

3.1.3.1.2 Rosetta Energy Score

The Rosetta energy score is a method for modeling which relies on thermodynamics principles specifically and the way that large biomolecules will arrange themselves structurally when in equilibrium will tend to favor the structures which possess the lowest free energy. Free energy for a given structural arrangement is obtained using a mathematical function that takes into account factors such as hydrogen bonding and electrostatics (Lazaridis and Karplus 1999).

3.1.3.1.3 RMSD and the GDT Total Score

Root Mean Square Deviation (RMSD) is a measure that is widely used to evaluate protein structures. The RMSD does however come with a few drawbacks. RMSD is believed to greatly underestimate the quality of a model because it can find that most of the structure is accurately predicted but the predicted parts are not located were they should be. Another problem with the RMSD value is that it cannot accurate lately evaluate models which have a high distance between each other As a result the GDT TS (GDT Total Score) has been proposed as a more accurate

measure. The GDT TS is used to identify residues which deviate from each other (specifically the alpha carbon) under a specified cut off value (Li et al., 2011).

3.1.3.1.4 MetaMQAPII

MetaMQAPII is a meta server that combines the scores from other servers which VERIFY3D, ProSA (Wiederstein & Sippl, 2007), ANOLEA, BALA-SNAPP (Krishnamoorthy & Tropsha, 2003), TUNE (Lin, May & Taylor, 2002), REFINER (Boniecki et al., 2003) and PROQRES (Wallner & Elofsson, 2006) These servers have been created to detect local inaccuracies in X-ray structures and computational models which display more severe errors. (Pawlowski et al., 2008). MetaMQAPII evaluates individual residues of a structure. This is achieved through first placing it into one of 315. a unique linear regression for each of the 315 groups is created , model has been developed to determine the RMSD of a residue within that group from its location in the native structure, depending on how well it was scored by the combination of eight different quality assessment servers (Pawlowski et al., 2008). The Meta server takes as input a structural file. After completing the analysis the output is in the form three files. The first file is a log file describing which servers were executed. The second file contains the raw scores of the servers. The third file contains a PDB file that has the beta factors column replaced by MQAP score. The PDB file can be visualized using visualization programs such PyMol.

3.1.3.1.5 RAMPAGE

The Cα is the most locus to consider when evaluating the distortion of the covalent geometry in protein structures. This is based on the observation that the Cα joins the side chain to the back bone. In the case where a misfit occurs in the junction between the side chain and the back bone the result is the wrong local minimum is obtained and during model refinement the Cα geometry is distorted a compromise. As a result there are three major components for geometrical structure validation these include backbone conformation, side chain conformation and Cα geometry.
RAMPAGE is a program that makes use of Ramachandran diagram plots to show ϕ versus ψ Angles of the backbone structure of a protein backbone (Lovell et al., 2008).

3.2 Aims of the chapter

The main aim of the chapter was to investigate structural features possessed by AA9 proteins, which are a consequence of sequence variability observed in this group of enzymes. AA9 structures were collected and analyzed in with the methods described in chapter 2 and these structures were grouped into their respective types. The grouped structures were then analyzed structurally to investigate distinct features.

3.3 Methodology

3.3.1 Structure Retrieval

All currently available structures where obtained from the PDB based on the analysis conducted in chapter 2, it was now possible to group these data according to types (Table 3.1).

Table 3.1: AA9 PDB structures grouped by type

AA9 types	PDB structures
1	4B5Q
	3EII
	3EJA
2	4EIR
3	2YET
	2VTC
	3ZUD
	4EIS

3.3.2 Homology Modelling

HHpred was used to identify the best template for modeling the selected AA9 variantss. The type 1 variant an9, was selected for homology modelling and the best template for its modeling was found to be 4B5Q. Phylogenetic analysis grouped the 4BQ5 structure among type 1 sequences as a result it was considered acceptable to model the an9 type 1 protein. The 4B5Q structure was

found to have the highest sequence identity (39%) with an9. The 4B5Q structure seemed to result in the least gaps which and no disruption of secondary structures was observed unlike the other possible template alignment. Phylogenetic analysis revealed that the crystal structure 4EIR was the only type 2 crystal structure. As a result, HHpred determined that this template was the most suited for modelling type 2 proteins. HHpred revealed that the 4EIR structure had a sequence identity of 71%, 40% and 45% with gg9, pt17 and tt2 respectively. The best crystal structure for modeling type 3 protein hr1 was determined to be 3ZUD with a sequence identity of 56%. The 3ZUD crystal structure was the most for modeling type 3 proteins because it had complete query coverage and was determined phylogenetic analysis to be a type 3 protein. The PDB files of the structures (shown in Table 3.1) where obtained and edited using the Python scripting language. Single chains (chain A) were extracted from all the PDB files removing all other chains and hetero atoms. Extraction of chain A was carried out using the script Clean_pdb.py. To generate alignment for homology modelling, the fasta sequences of the PDB files were extracted using the script Respdb.py. Once PDB structure sequences were obtained they were aligned with their respective sequences. Homology modeling was carried out using Modeller (Eswar et al., 2008; Shen & Sali, 2006) using the Modelling.py script. 100 hundred models were generated for each modeling round.

3.3.3 **Model Validation**

The quality of the generated homology models was evaluated using global and local techniques through the calculating several parameters. The RMSD, Rosetta energy and the N-DOPE Z score were used to rank the models the calculation of these values can be found in table 3.2. The RMSD, Rosetta energy and the N-DOPE Z score were computed using the Scoring.py script. Rampage was used to generate Ramachandran of the homology models. Homology models were submitted to MetaMQAPII server (https://genesilico.pl/toolkit/unimod?method=MetaMQAPII), the resulting PDB files were visualized in PyMol. A script (Scoring.py) previously developed by Mathys Kroon was used to compute the DOPE score and RMSD. The ROSETTA energy calculation functionality was also added to the script resulting in the script Scoring.py which can be found in the supplementary data.

3.4 **Results and Discussion**

3.4.1 **Homology modeling**

Homology models were generated for an9, gg9, pt17, tt2 and hr1. In order to search a large enough sample space for a correct model, 100 models were generated for each of these sequences. Using the Rosetta and DOPE scores the top three models were selected for each sequence for validation.

Table 3.2: The best homology models generated for AA9 variants based on DOPE score and Rosetta energy

Type	DOPE score	CA RMSD	Rosetta energy	Model
Type 1				
An9	-0.76136	16.69448	1478.728	aspergillus_niger_9.B99990023.pdb
	-0.77137	16.68421	1484.187	aspergillus_niger_9.B99990008.pdb
	-0.7803	16.6558	1491.847	aspergillus_niger_9.B99990019.pdb
Type 2				
Gg9	-1.62715	6.730763	528.9251	glomerrela_graminic_9.B99990093.pdb
	-1.61744	6.727348	670.7561	glomerrela_graminic_9.B99990090.pdb
	-1.61521	6.727826	712.1386	glomerrela_graminic_9.B99990030.pdb
Pt2	-0.7288	16.11672	1051.177	pyrenophora_teres_17.B99990035.pdb
	-0.73487	16.14643	1084.793	pyrenophora_teres_17.B99990052.pdb
	-0.75428	16.21822	1178.857	pyrenophora_teres_17.B99990064.pdb
Tt2	-1.18807	4.993337	1210.106	thievela_terestis_2.B99990089.pdb
	-1.13134	5.173565	1239.368	thievela_terestis_2.B99990022.pdb
	-1.26898	5.141533	1272.788	thievela_terestis_2.B99990078.pdb
Type 3				
Hr1	-1.53592	2.788537	847.5774	hypocrea_rufa_1.B99990022.pdb
	-1.52168	2.790164	879.3846	hypocrea_rufa_1.B99990037.pdb
	-1.50144	2.787762	884.4931	hypocrea_rufa_1.B99990049.pdb

Self-models were generated for all three templates used. 100 models were created for 3ZUD, 4EIR and 4B5Q. These models were ranked by using their DOPE score and Rosetta energy. Top the three models were selected for each of the templates used (Table A5). These selected models were then validated to observe errors which are the result of the homology modelling procedure.

3.4.2 **Model evaluation**

RAMPAGE and MetaMQAPII were used to evaluate the models in Table 3.2. The criteria for selecting the best models involved selecting models with the lowest DOPE score and Rosetta energy. This is because models Positive normalized DOPE scores are regarded to be poor models, while models whose scores fall below -0.5 are near native (Shen & Šali, 2006). The top 3 models in group were then submitted to RAMPAGE and MetaMQAPII and the models displaying the most optimal results are shown below.

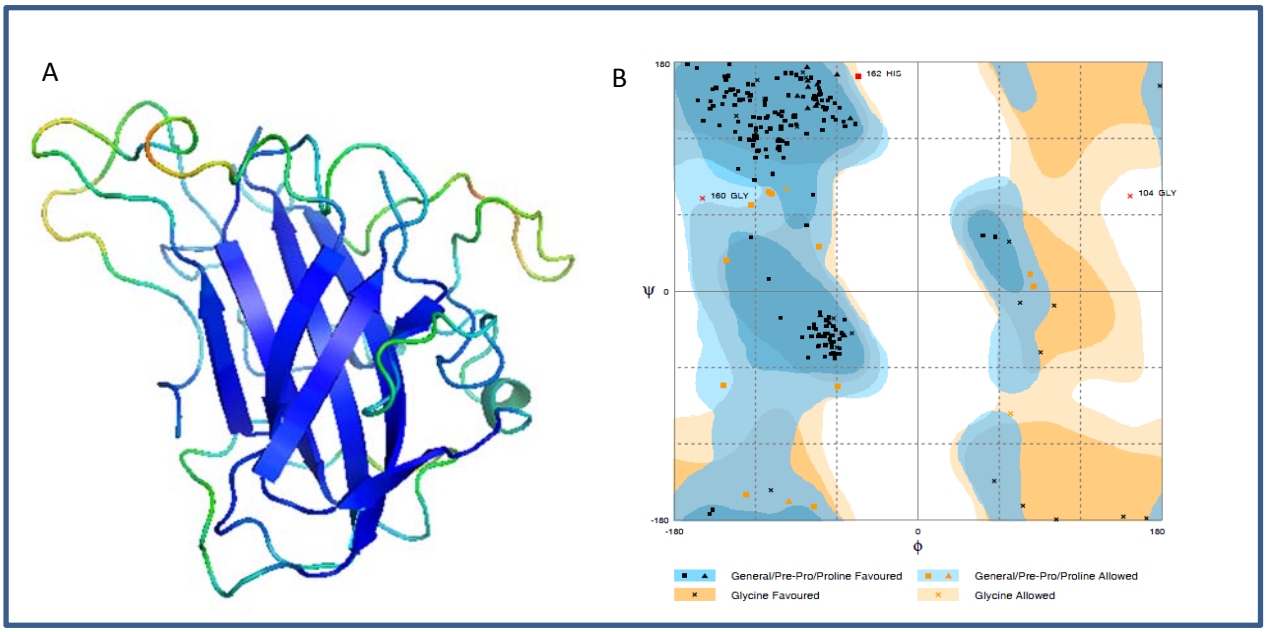


Figure 3.1: MetaMQAPII and RAMPAGE validation results for aspergilus_niger_9.B99990019.pdb. A. MetaMQAP output an9 model, The image is coloured by quality in a spectrum of blue to red where blue is highly scored residues and red is poorly scored and erroneous residues. B. RAMPAGE Ramachandran plots the an99 homology model.

aspergilus_niger_9.B99990019.pdb was found to be the best model. MetaMQAPII did not report any major region of deviation in the structure. All residues were identified as being stable (figure 3.1 A). Rampage detected 95.3% of residues are found in the Number of residues in favored region, 4.2% of residues were found in of residues in allowed region and 0.5 % residues were found in the outlier region. The aspergilus_niger_9.B99990019.pdb model was selected and used for further analysis (Figure 3.1 B).

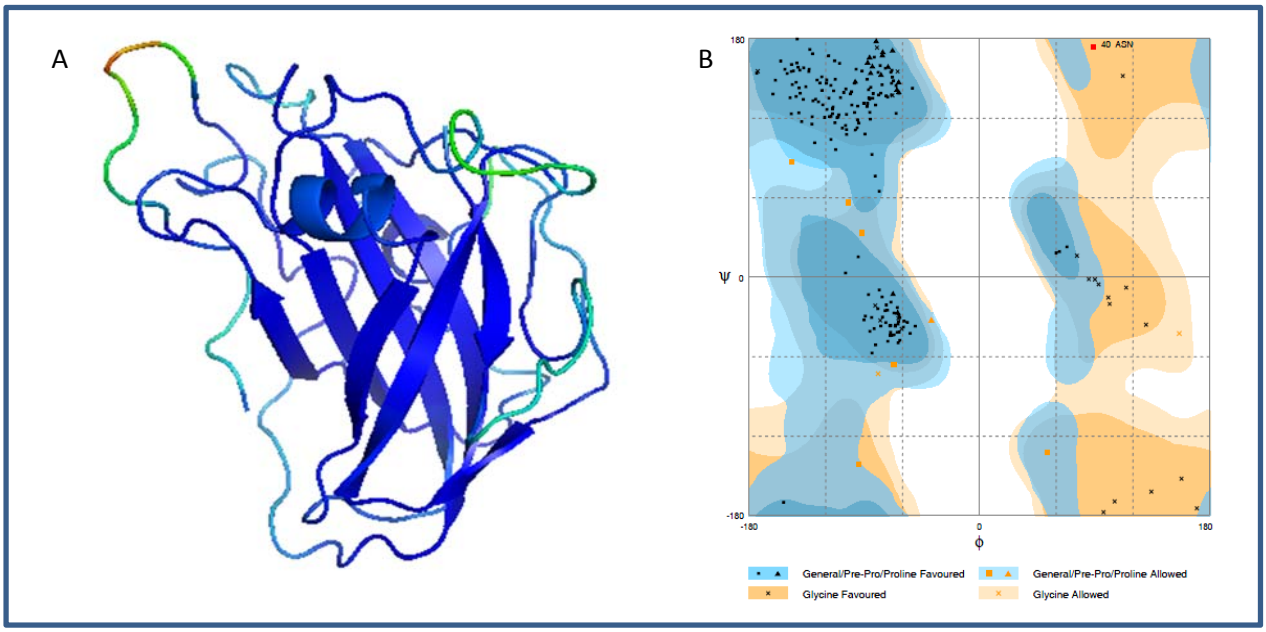


Figure 3.2: MetaMQAPII and RAMPAGE validation results for Glomerrela_graminic_9.B99990030. A. MetaMQAP output gg9 model, the image is coloured by quality in a spectrum of blue to red where blue is highly scored residues and red is poorly scored and erroneous residues. B. RAMPAGE Ramachandran plots the gg9 homology model.

Glomerrela_graminic_9.B99990030.pdb was found to be the best model. MetaMQAPII did not report any major region of deviation in the structure. All residues were identified as being stable Figure 3.2 A) Rampage detected 95.3% of residues are found in the Number of residues in favored region, 4.2% of residues were found in of residues in allowed region and 0.5 % residues were found in the outlier region. The glomerrela_graminic_9.B99990030.pdb model was selected and used for further analysis.

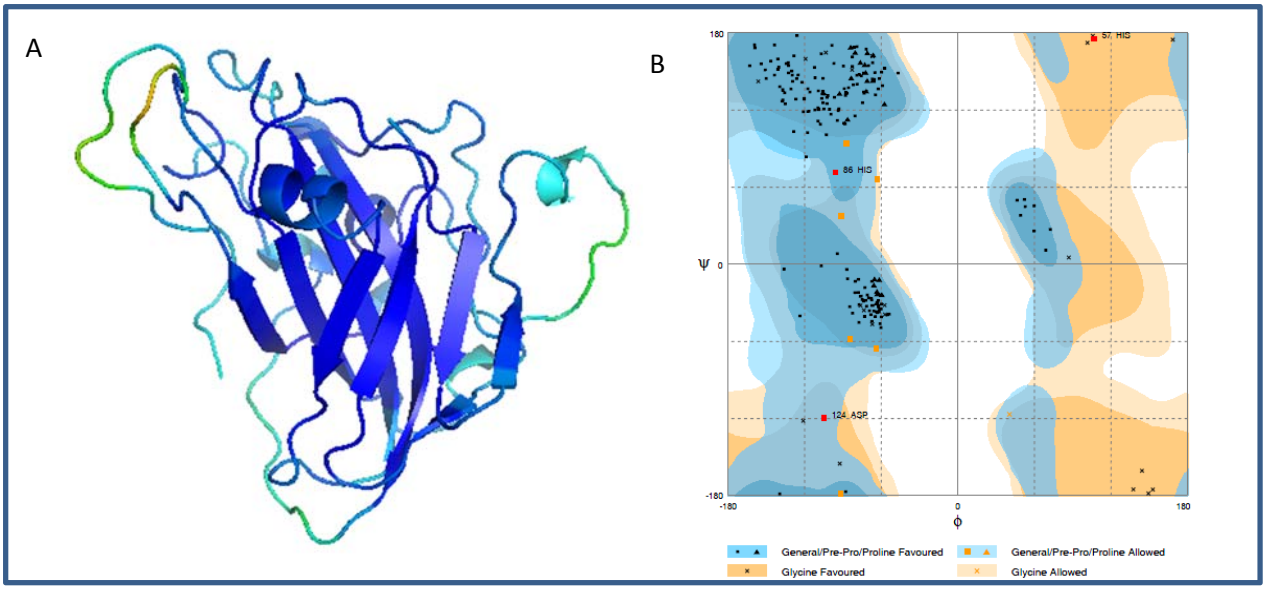


Figure 3.3: MetaMQAPII and RAMPAGE validation results for Hypocrea_rufa_1.B99990037. A. MetaMQAP output hr1 model, the image is coloured by quality in a spectrum of blue to red where blue is highly scored residues and red is poorly scored and erroneous residues. B. RAMPAGE Ramachandran plots the hr1 homology model.

Hypocrea_rufa_1.B99990037.pdb was found to be the best model for the type 3 hr1 sequence. MetaMQAPII did not report any major region of deviation in the structure. All residues were identified as being stable (figure 3.3A). Rampage detected 95.5% of residues are found in the Number of residues in favored region, 3.1% of residues were found in of residues in allowed region and 1.3% residues were found in the outlier region. The hypocrea_rufa_9.B99990037.pdb model was selected and used for further analysis. (Figure 3.3 B).

Pyrenophora_teres_17.B99990035.pdb was found to be the best model. MetaMQAPII did not report any major region of deviation in the structure. All residues were identified as being stable (figure 3.4A). Rampage detected 95.3% of residues are found in the Number of residues in favored region, 4.2% of residues were found in of residues in allowed region and 0.5 % residues were found in the outlier region. The Pyrenophora_teres_17.B99990035.pdb model was selected and used for further analysis (Figure 3.4 B).

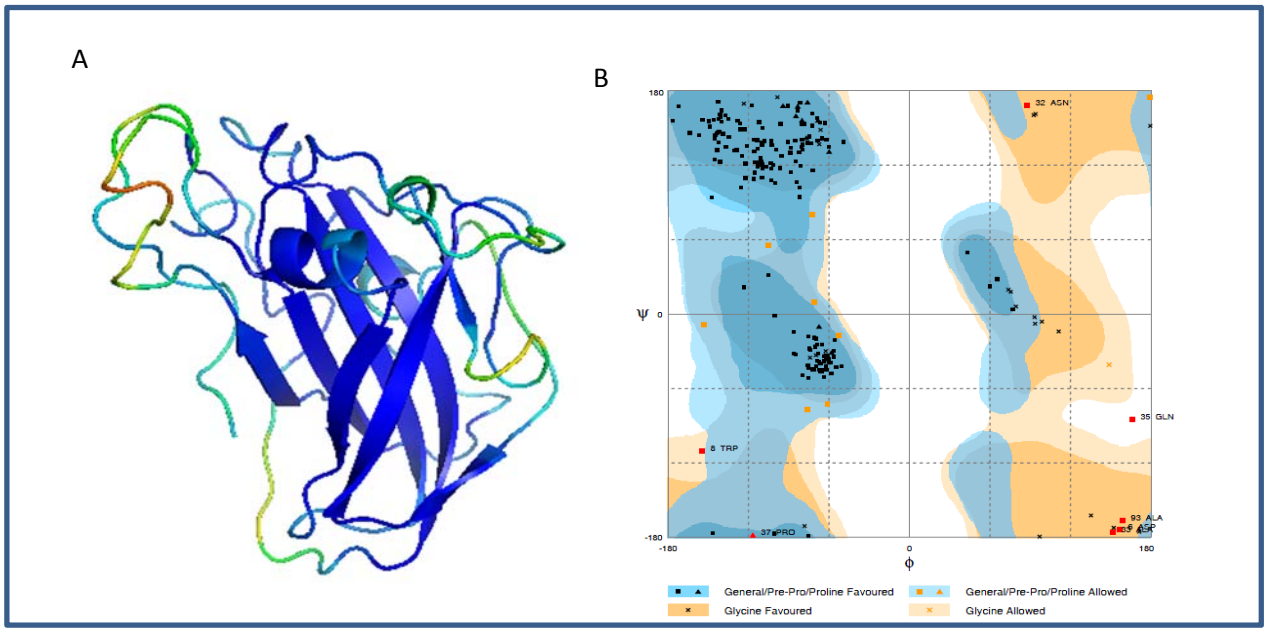


Figure 3.4: MetaMQAPII and RAMPAGE validation results for Pyrenophora_teres_17. B99990035. A. MetaMQAP output pt17 model, the image is coloured by quality in a spectrum of blue to red where

blue is highly scored residues and red is poorly scored and erroneous residues. B. RAMPAGE Ramachandran plots the pt9 homology model.

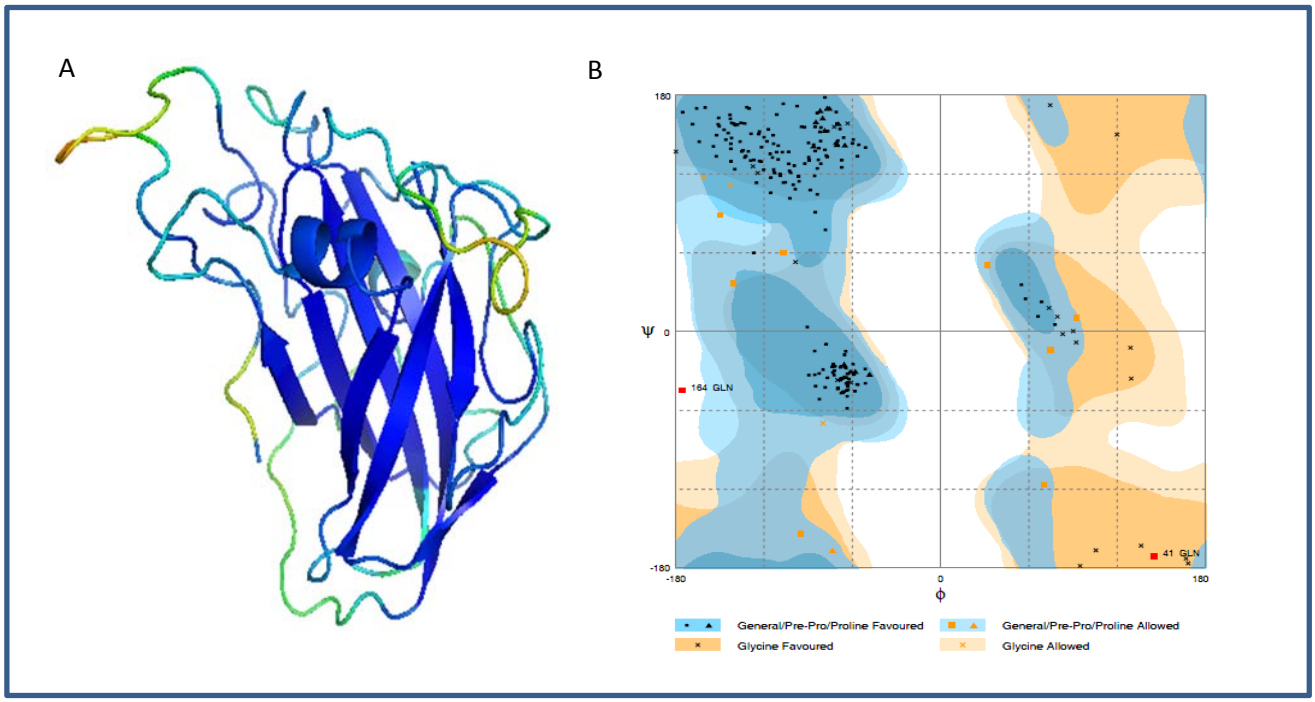


Figure 3.5: MetaMQAPII and RAMPAGE validation results Thievela_terestis_2.B99990022.
A. MetaMQAP output tt2 model, the image is coloured by quality in a spectrum of blue to red where blue is highly scored residues and red is poorly scored and erroneous residues. B. RAMPAGE Ramachandran plots the tt2 homology model.

Thievela_terestis_2.B99990022.pdb was found to be the best model. MetaMQAPII did not report any major region of deviation in the structure. All residues were identified as being stable (figure 3.4 A). Rampage detected 93.8% of residues are found in the Number of residues in favored region, 5.3% of residues were found in of residues in allowed region and 0.9%residues were found in the outlier region. The Thievela_terestis_2.B99990022.pdb model was selected and used for further analysis.

Model evaluation revealed that the best generated models were: aspergilus_niger_9.B99990019. pdb, Hypocrea_rufa_1.B99990037.pdb, Pyrenophora_teres_17. B99990035.pdb and Thievela_terestis_2.B99990022.pdb.

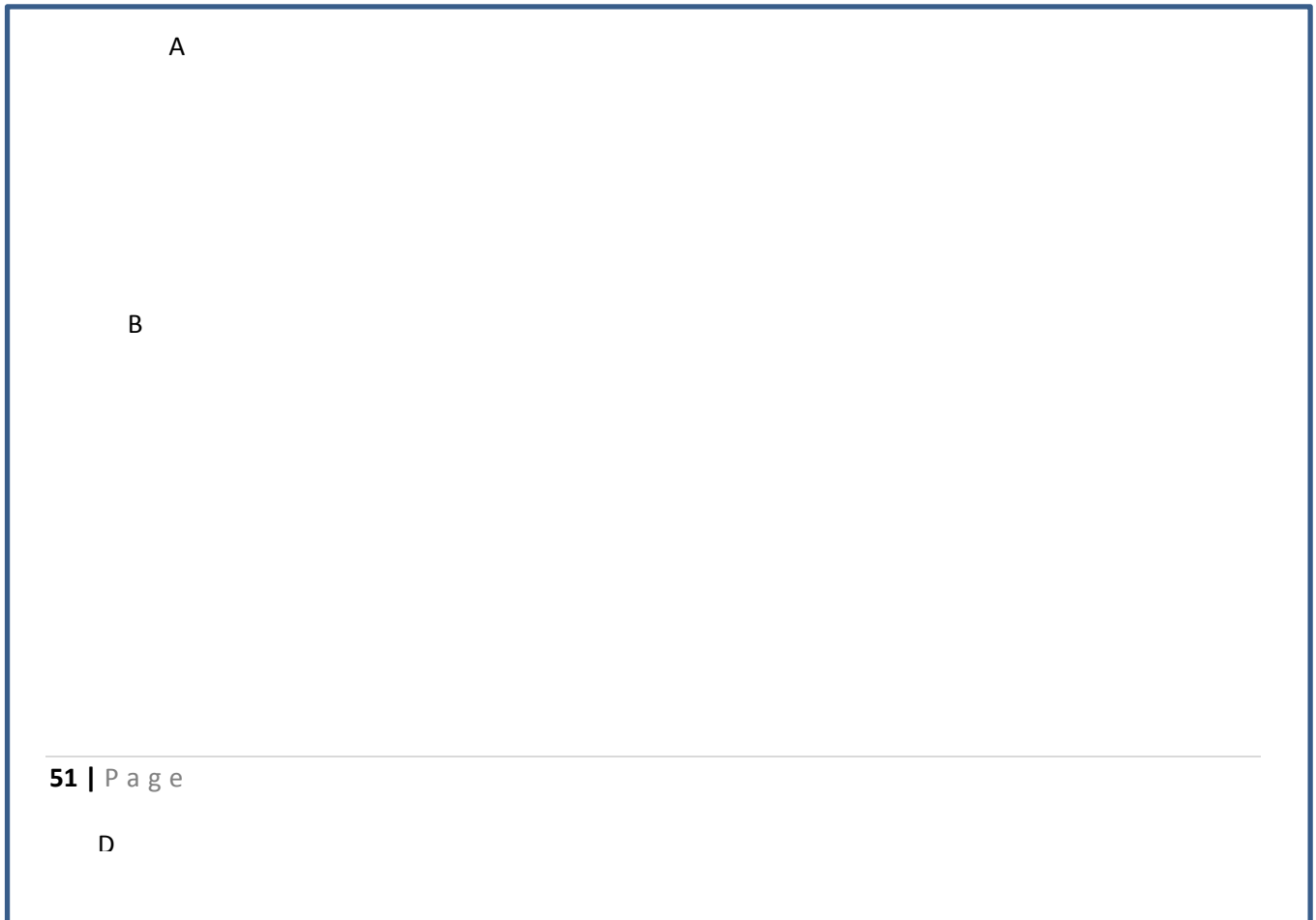
3.4.3 **Physiochemical property analysis-Hydrophobicity**

The effect of motif variation of on overall surface AA9 proteins was investigated through the identification of hydrophobic regions on AA9 protein sequences. The analysis involved taking all

the PDB structures and generated models and extracting their amino acid sequences. The sequences where then analyzed for hydrophobicity using the kd scale. Hydrophobicity plots were obtained and this data was correlated to the motif data by mapping the motifs identified in Chapter 2 onto their respective structures. The results of the analysis are shown in this section.

3.4.3.1 ***Hydrophobicity***

Type 1



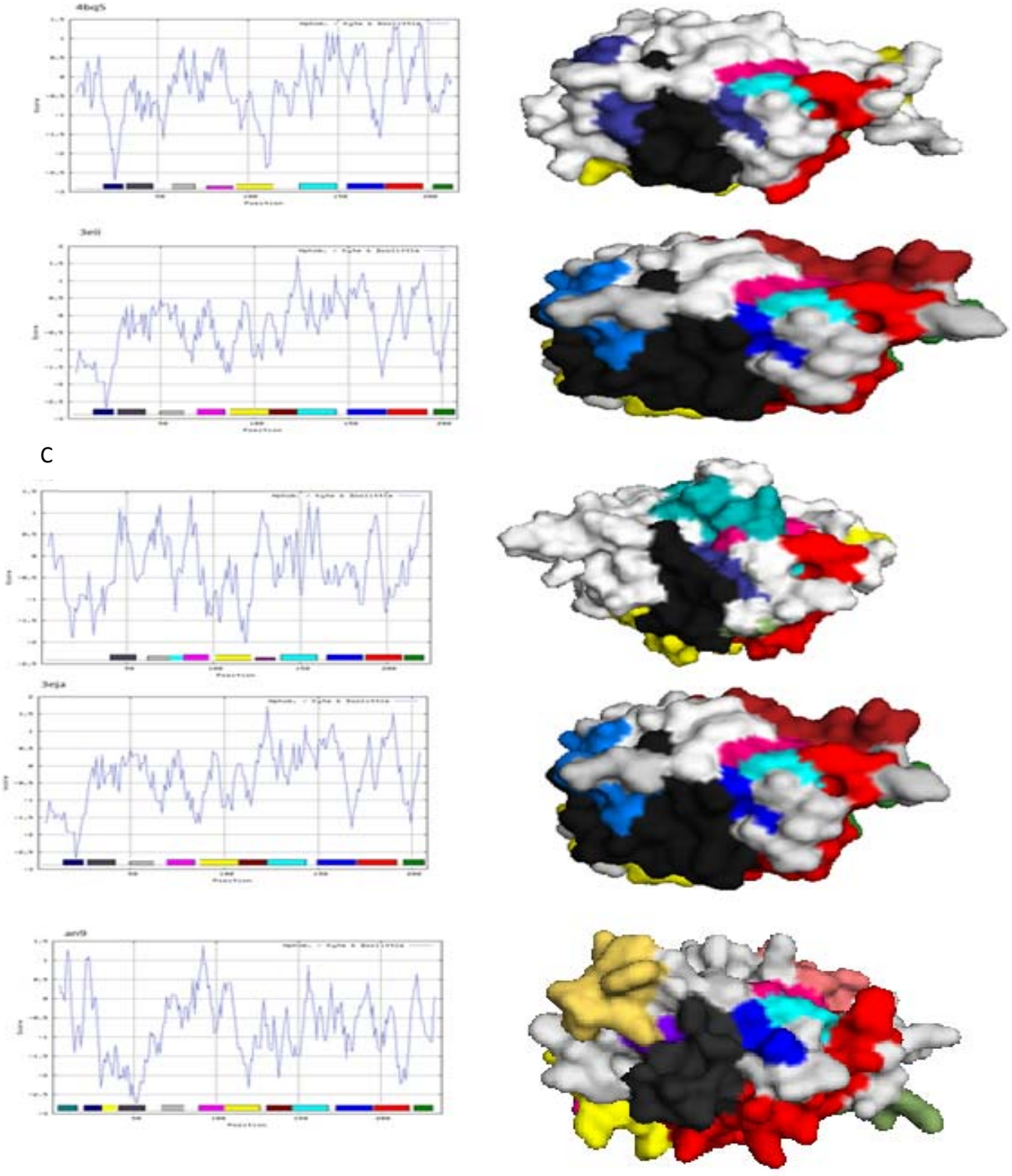


Figure 3.6: the hydrophobicity plots of type 1 AA9 structures based on the KD scale. A=45BQ, B=3EII, C=4EIS, 3EJA, E=An9.

Type 2

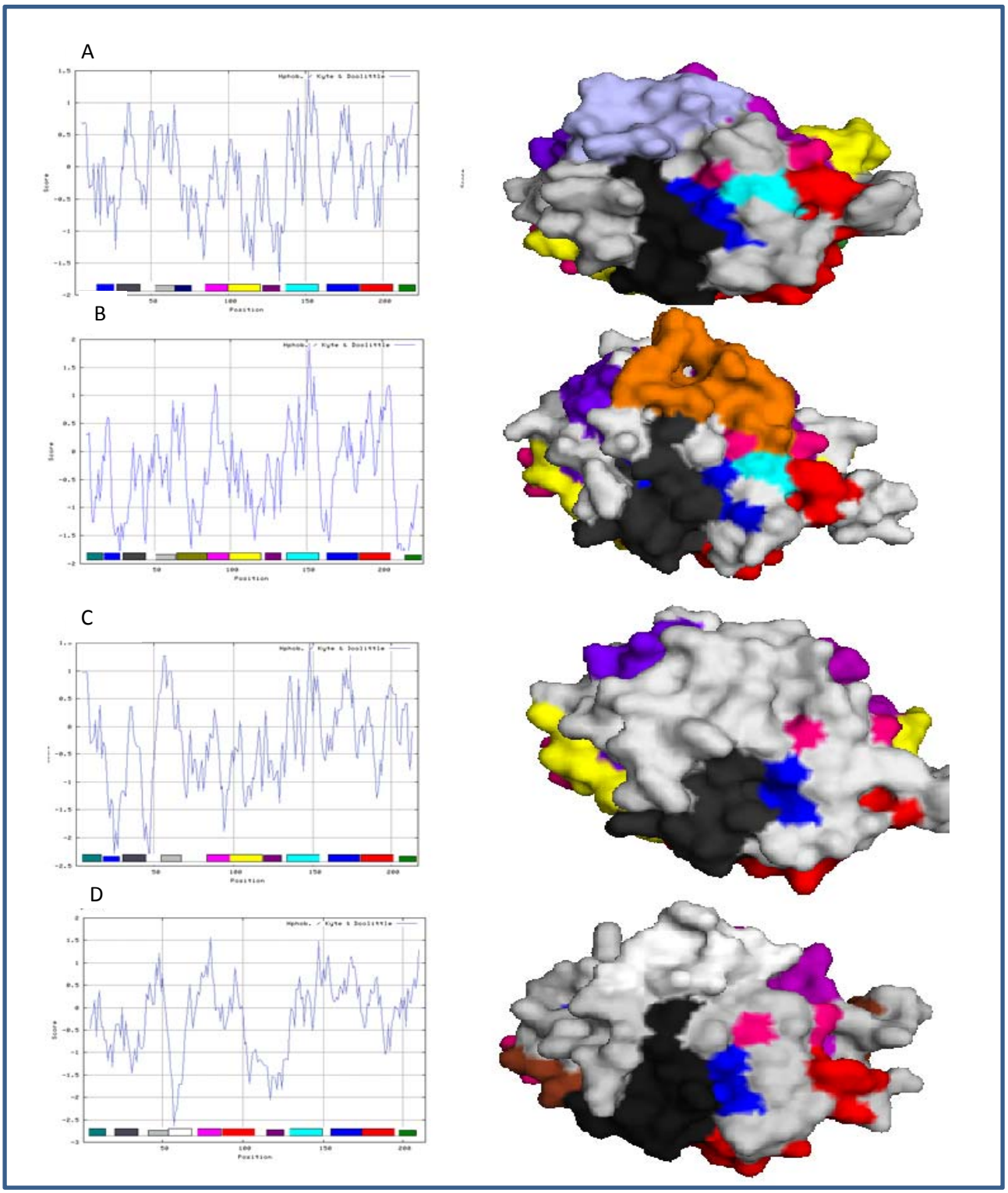


Figure 3.7: The hydrophobicity plots of type 2 AA9 structures based on the KD scale. A=4EIR, B=Tt2, C=Gg9, D=Pt17.

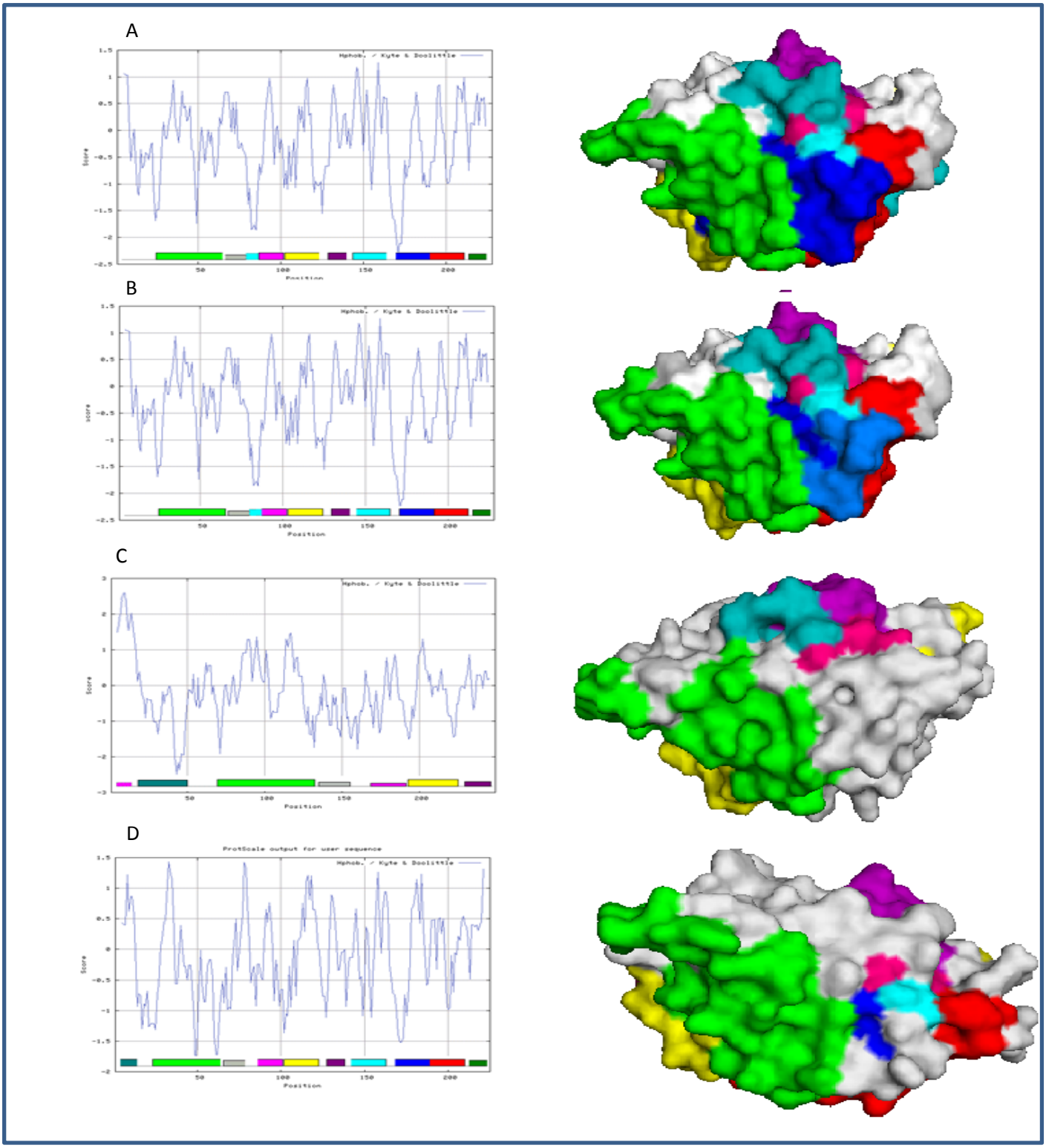


Figure 3.8: The hydrophobicity plots of type 3 AA9 structures based on the Kd scale. A=3ZUD, B=2YET, C=2VTC, D=Hr1.

Figure 3.7 shows the type 1 hydrophocity plots for all templates and homology models. the surface diagram of their respective structures is shown. On the surface diagrams the structures have been mapped with the sequence motifs identified by MEME. It can be seen from this figure that the type specific motifs 13, 14 and 20 are hydrophilic motifs. Visualisation of the surface diagrams of motif 13, 14 and 20 reveals that these motifs are surface exposed. The same analysis was conducted for type 2 sequences and the results are displayed in figure 3.7. The results show that motifs 5 ,11,12, 15,17 and 18 are hydrophilic and can be found protruding on the surface of type 2 structures. Type 3 sequences also displayed the same features as the other AA9 types. The motifs responsible for type 3 specificity, which are motifs 6, 11 and 16 were found to be hydrophilic and just like the other two AA9 types they were found on the surface.

Findings from figures 3.7, 3.8, 3.9 suggests that motifs 5, 11, 12, 13, 14 ,15 ,16,17 and 18 which were found to be type specific are hydrophilic motifs. This observation is supported by the fact that these motifs are mostly found located on the surface off their respective AA9 molecule. In particular these motifs are located on the proposed active site surface.

3.4.4 **Motif structural analysis**

To assess the effect that motifs 5, 11, 12, 13, 14, 15, 16, 17 and 18 have on the overall structure of AA9 proteins PyMol was used to map out these motifs on their respective structures (Figure 3.10). The analysis revealed that these motifs were primarily located on the flat surface active site of the enzyme.

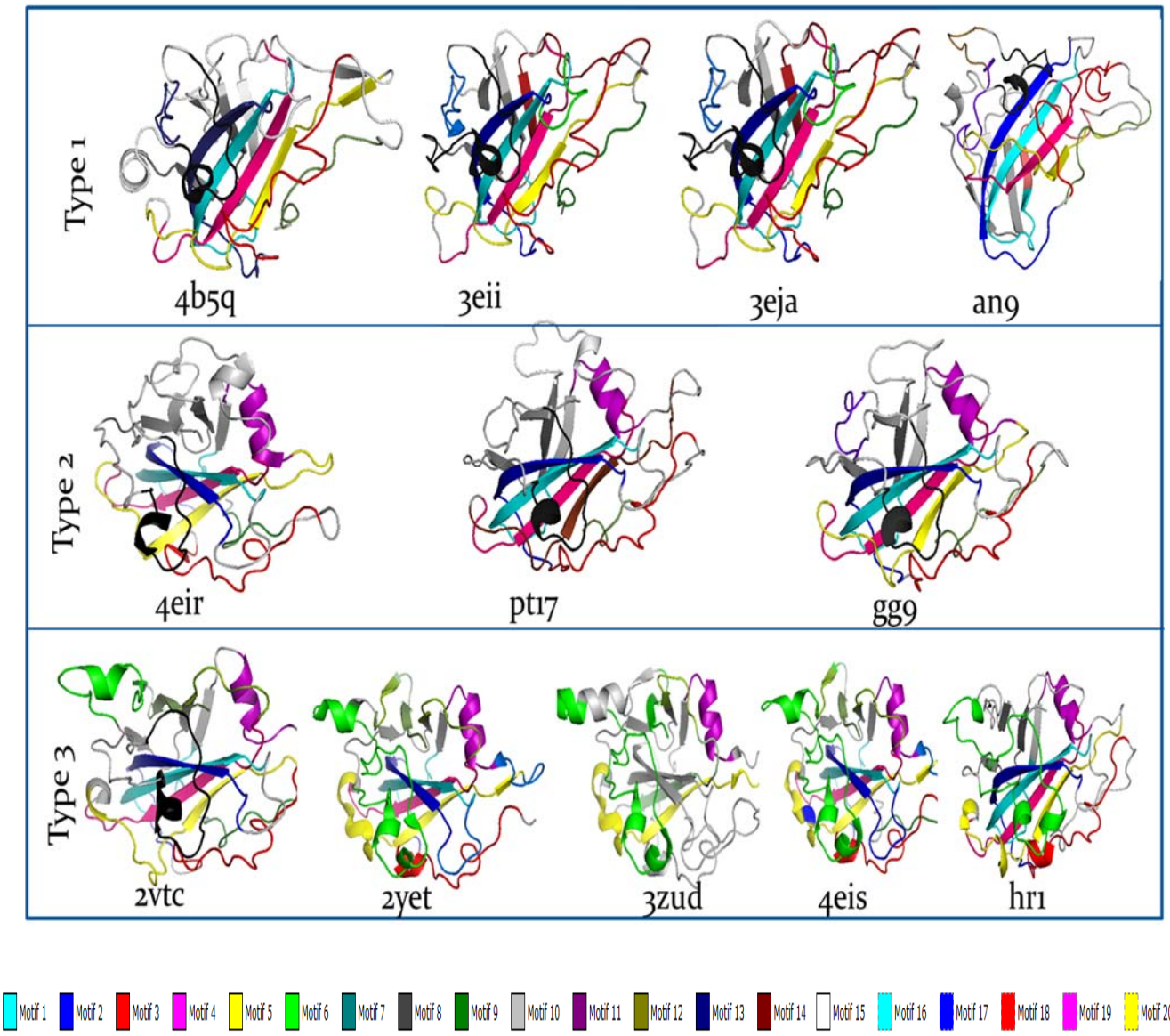


Figure 3.9: Aerial view of the AA9 flat surface active site with MEME motifs indicated with their respective colors.

Figure 3.9 illustrates the to view of AA9 protein variants that were identified through the course of the conducted study it is apparent from Figure 3.9 that the overall beta sandwich arrangement is conserved at both sequence and structural level. This is based on the observation that beta sandwich fold is almost always composed of motif 2, 6 and 16. The only exception to this is the type 2 variant pt17 which has motif 5 replaced with motif 18. Conservation was however not only limited to the β sandwich regions because it was also observed in motifs which corresponds to a loop region on the surface off AA9 structures which can be observed in all AA9 variants. Super

imposing all the AA9 variants (Figure 3.10) revealed that most of the variability that is observed mostly occurs in the loop regions on the proposed active site surface. This observation suggests that the loop regions play a significant role in type specificity.

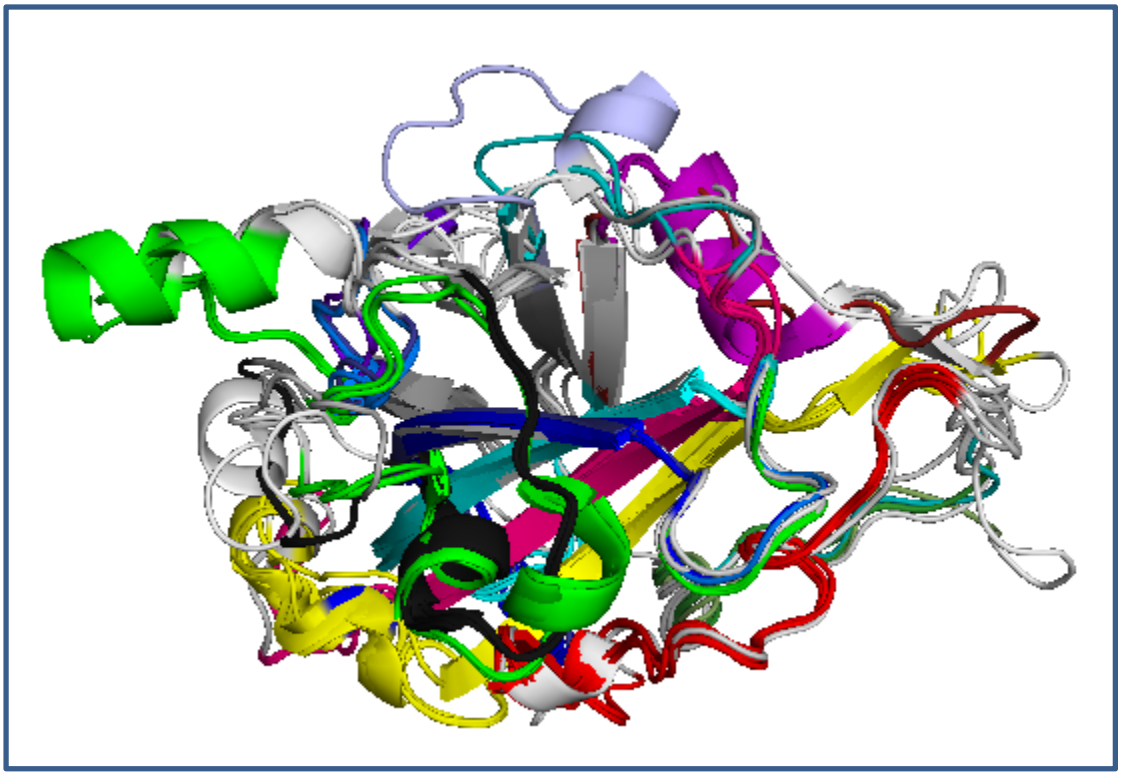


Figure 3.10: Super imposed AA9 variants. The motifs are coloured using MEME colour coding.

3.4.5 **Aromatic residue distribution**

After evaluating the distribution aromaticity amongst various AA9 types, it was observed that AA9 possess varying amounts of aromatic residues. To observe the role that these residues have on the residues play on type specificity and the effect of these aromatic residues on AA9 structures in general.

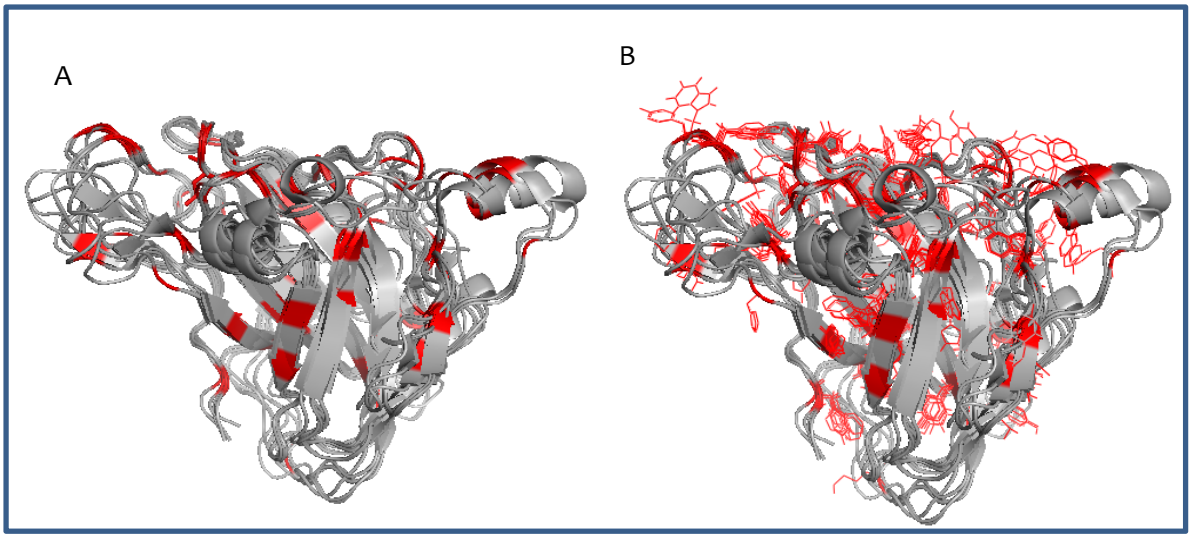


Figure 3.11: Diagrammatic representation of aromatic residues on the superimposed AA9 structures. Aromatic residues are coloured in red (A) and shown in line representation in B.

Figure 3.11 shows that in all the superimposed AA9 structures the aromatic residue positions are conserved. As a result there was no clear observable difference in aromatic residue positions within the various AA9 types. Figure 3.11 B revealed that even though the aromatic residues seem to be distributed thought the structure of AA9 proteins, they are on only surface exposed in the active site region. This suggests that the aromatic residues may play a role in substrate interaction and in turn type specify.

3.4.6 **AA9 structural features**

The structure listed in table where submitted to the PDBsum webserver to observe any unique features displayed AA9 crystal structures. A summary of findings is displayed in this section.

3.4.6.1 ***310 helices***

PDBsum was able to detect the presence of 310 helices on a group of AA9 structures. 310 helix are different from the classical α-helix. This is based on the hydrogen-bonding pattern that they form. In α helices hydrogen bonds are made between residues ***i*** and ***i*** + 4. The case of 310 helices, hydrogen bonds are formed between residues ***i*** and ***i*** + 3. Changes in the backbone dihedrals (ϕ and ψ) and of less than 15 degr**ees** distinguish these two helical conformations. 310 helices are not a prominent feature in proteins as compared to their α-helix counterparts. The biological function of 310 remains unknown however the leading belief is that these structures are intermediate structures that occur when α helices unfolds (Millhauser, 1999). There were three helical structures observed for the type 1 PDB structure 4EIS (Table 3.3).

Table 3.3: 310 helices observed on the type 1 crystal structure 4EIS

Start	End	Type	Number of residues
Ala49	Cys52	alpha	4
Cys98	Asp100	310	3
Thr128	Met131	310	4

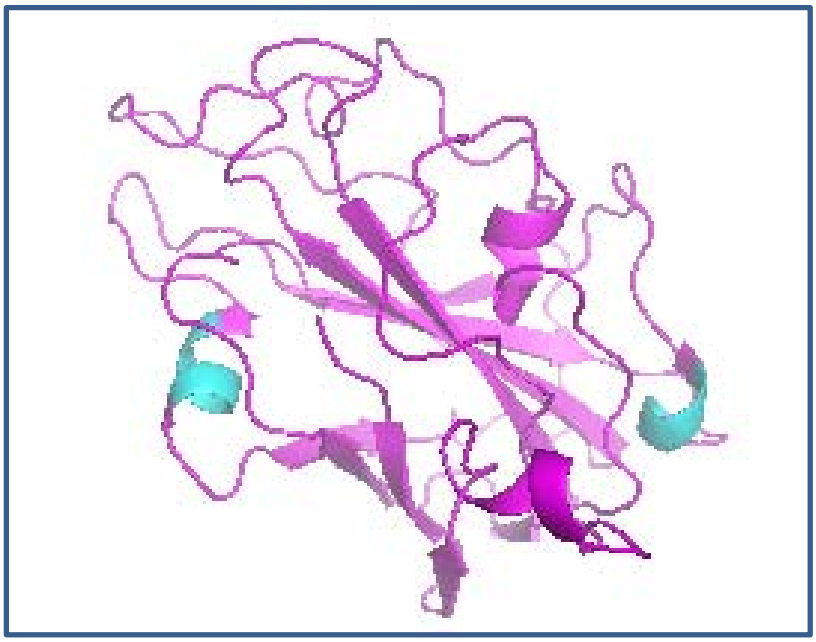


Figure 3.12: 4EIS crystal structure showing the location of the 310 helices. The 310 in helices are shown in blue while the rest of the molecule is displayed in purple

Two of the helical structures were found to be 310 helices. The position of the 310 alpha helices on the 4EIS crystal structure reveals that these helices are located on the surface accessible area of the AA9 structure.

The type 2 crystal structure 4EIR was found to possess 7 helical structures table 3.4.

Table 3.4: 310 helices observed on the type 1 crystal structure 4EIR

Start	End	Type	Number of residues
Tyr22	Met25	alpha	4
Gly46	Gly48	310	3
Asp53	Cys56	Alpha	4
Cys101	Thr103	310	3
Lys106	Gln108	310	3
Ala130	Ala136	Alpha	7
Gly196	Ala198	310	3

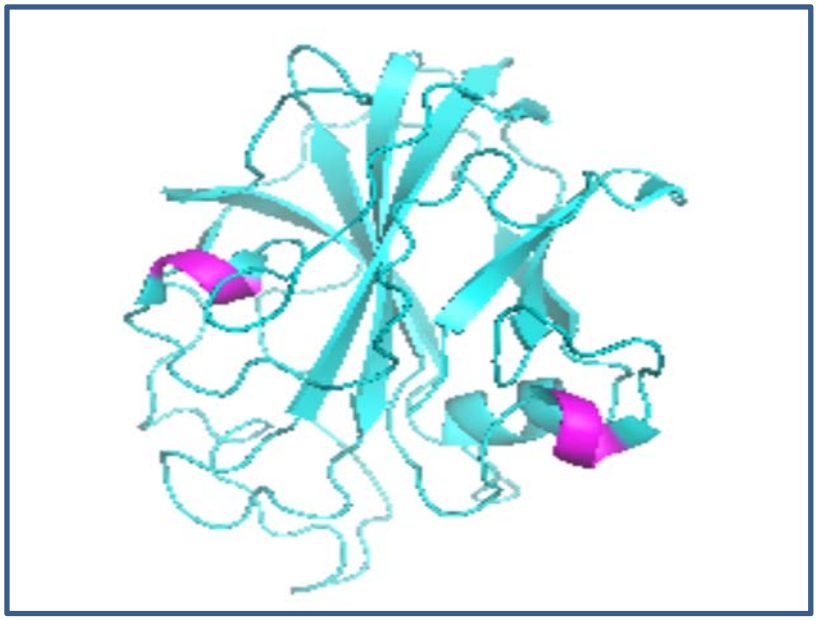


Figure 3.13: 4EIR crystal structure showing the location of the 310 helices. 310 helices are coloured in purple while rest of the molecule is shown in blue.

Four out of 7 of these helical structures were found to be 310 helices. These 310 helices were then mapped on the 4EIS crystal structures to observe their position on the structure (Figure 3.13).The visualization of these 310 helices on the structure of 4EIS reveals that these helices are surface exposed, located on opposite ends of the enzyme structure.

The type 3 crystal structure 2YET was found to possess 7 helical structures and it was found that 4 of these helices were 310 helices (table 3.5)

Table 3.5: 310 helices observed on the type 1 crystal structure 2YET

Start	End	Type	Number of residues
Leu19	Thr28	alpha	10
Pro49	Gln51	310	3
Asp56	Cys59	alpha	4
Cys105	Thr107	310	3
Lys110	Asn112	310	3
Ala131	Asn137	alpha	7
Ala198	Gln200	310	3

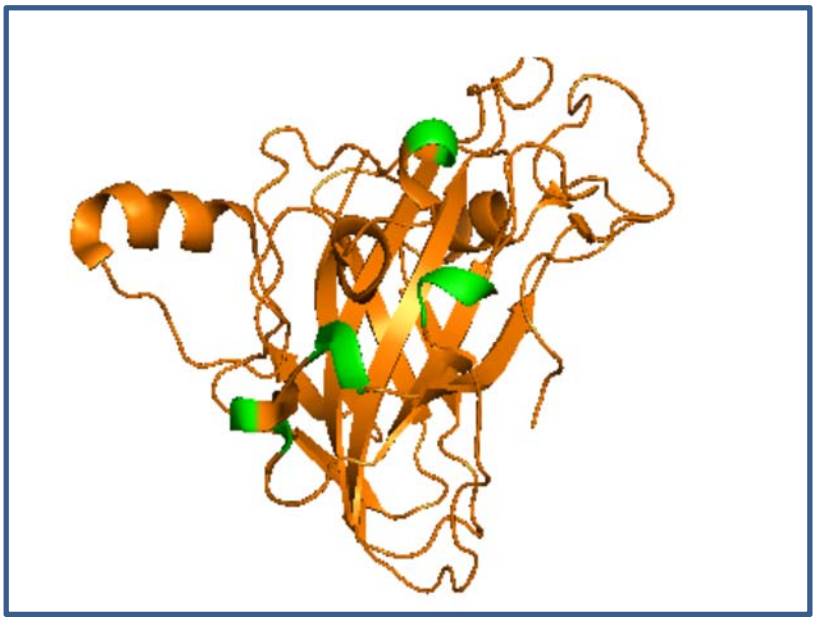


Figure 3.14: 2YET Crystal structure showing the location of the 310 helices. 310 helices are coloured in green and the rest of the molecule is shown in orange.

The 310 helix residues observed in table 3.5 were mapped out on the structure of 2YET to observe the localization of these structures on the protein (figure 3.14). I was found that the 310 residues on the structure are found on the outside surface of the enzyme where they form a ridge like organization on the side of the AA9 protein surface.

The type 3 crystal structure 2VTC was found to possess 7 helical structures and it was found that 4 of these helices were 310 helices (table 3.5)

Table 3.6: 310 helices observed on the type 1 crystal structure 2VTC.

Start	End	Type	Number of residues
Leu19	Thr28	alpha	10
Pro49	Gln51	310	3
Asp56	Cys59	alpha	4
Cys105	Thr107	310	3
Lys110	Asn112	310	3
Ala131	Asn137	alpha	7
Ala198	Gln200	310	3

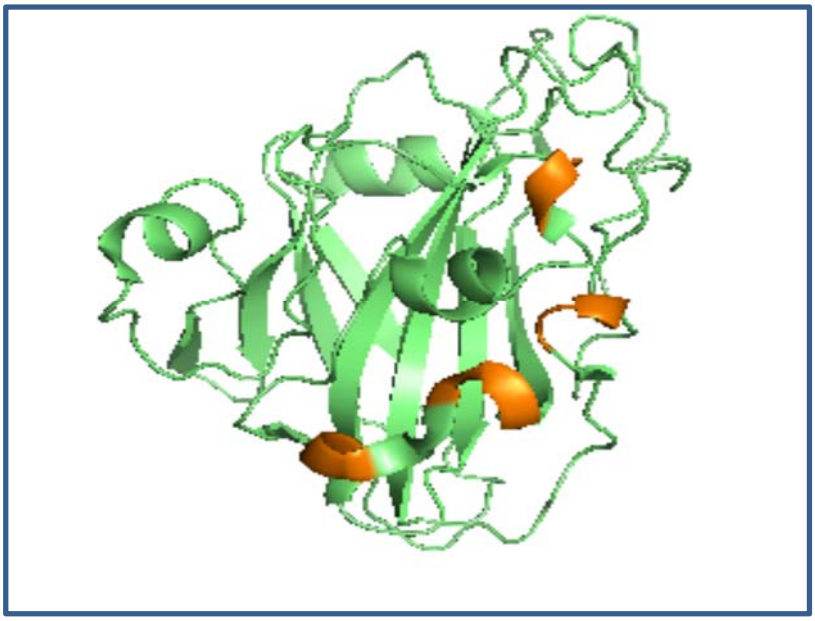


Figure 3.15: 2VTC Crystal structure showing the location of the 310 helices. 310 helices are coloured in orange.

The 310 helix residues observed in table 3.5 were mapped out on the structure of 2VTC to observe the localization of these structures on the protein (figure 3.15). Much like the 2YET structure (figure 3.15), the 310 residues on the 2VTC structure were also found on the outside surface of the enzyme forming a ridge like organization on the side of the AA9 protein surface.310 observed in every AA9 type however other crystal structures such as 4B5Q, 3ZUD, 3EII and 3EJA. 310 helices appear to be a common occurrence in AA9 proteins. 310 helices are not type specific however there are also not found in every AA9.w

3.5 **Conclusions**

Viable homology models where generated for an9, gg9, pt17, tt2 and hr1. With these models at hand, all the determined variants had a representative model to conduct analysis. Structural features to distinguish between the various AA9 types have been identified. Sequence variability was observed which severely affects the loop regions. As a result we observe that the proposed active site of AA9 is largely affected by the presences or absence of certain motifs which results in a complete change in residue composition of the active. This observed variability could play a role in type specificity. The presence and absence of motifs was also observed to play a role in subtyping. An example of such an observation is in the case of type 1 sequences which have motif 20 which is in the form of an insert (figure 2.5). The presence of these motifs was shown to structurally alter the active site surface of type 1 AA9 proteins by inserting 6 residues consisting of mainly aromatic residues. Sub typing was not only limited to active site region but also the β-sandwich fold. This was observed for type 2 variants which show the substitution of the globally conserved motif 5 with motif 18. This swapping off motifs did not appear to affect the site. The motifs which were found responsible for AA9 typing where mostly found to be hydrophilic and surface exposed. These motifs were also found to be primarily located at the active site. This may suggest that these motifs play a role in substrate targeting and orientation which a feature which long been proposed as the reason why there are numerous AA9 types. Analysis of the aromatic residues on the structure of AA9 proteins revealed that the active site has numerous protruding aromatic residues. The presence of aromatic residues on the active site may suggest that residues play a role in substrate binding. The fact that 310 helices are found in all AA9 types but not in all AA9 models warrants more investigation.

CHAPTER 4

During the course of the study, sequence and structural elements have been identified for the distinction of the three AA9 types. The original intention of the study was to carry out the analysis of AA9 proteins on a large scale using all the available Pfam sequences. However the high divergence amongst AA9 proteins proved to be a problem that resulted in the use of a smaller dataset of sequences. Even with the use of this smaller dataset, it was still possible to complete the objectives specified in Chapter one and therefore completing the main aim of this project.

Obtaining the original sequences (Supplementary Data 1) was relatively straight forward as the Pfam database offers a straight forward method for obtaining protein sequences (Finn et al., 2008). Prior to this study a sequence and structural basis for distinguishing between various types had not been determined. As a result the sequences previously published by li et al (2012) where included in the dataset. These sequences were already determined experimentally to be one of the 3 AA9 types. These sequences where then used as references sequences in subsequent analysis.

Multiple sequence alignment was able to identify distinct groups of sequences which could be characterised by the presence or absence of inserts in certain positions. These inserts were found to be located on the N-Terminus of AA9 proteins. It was determined that that type one sequences lacked both inserts I and II however a few AA9 protein sequences were found to have a short residue insertion corresponding with insert 1. The type 2 sequences lacked only insert I while type 3 sequences did not lack any inserts. This observation suggested that the N-Terminus of these proteins is responsible for type specificity, a notion which is supported by motif analysis which found that the N-Terminus motifs are variable as compared to the more conserved C-Terminus.

Phylogenetic analysis was able to significantly group the AA9 sequences based on type at 95% site coverage. This grouping of was crucial as it allowed the analysis of individual AA9 types

to identify unique features in a specific type. This grouping of sequences also made the process of homology modelling easier since it enabled the elucidation determination of the types that AA9 crystal structures belong to (Table 3.1). The phylogenetic analysis of individual types as well motif analysis allowed for the discovery of variants which occur in in each type. The types where characterized by the presence or absence of certain motif such as in the case of the type one sequences, the presence of motif 20 results in the presence of a shorter insert the insert I position.

Physiochemical property analysis revealed that the individual types can be distinguished using aromaticity as it was found that the distribution of aromaticity values for the individual types is significantly different from each other. The aromatic residues were found to be mostly situated on the active site of the AA9 enzymes. The localization of these aromatic residues on the surface of AA9 enzymes may implicate them in substrate interaction. Homology modelling using suitable templates was used to generate homology models for the different AA9 variation that was observed to assess its impact on the protein structure. Evaluation of these models and the template self-models revealed that errors generated in the modelling process are most likely a consequence of the modelling process. The newly generated models as well as the crystal structures were used to evaluate the effect that motifs have on the hydrophobicity of the accessible surface of AA9 structures. The hydrophobicity did not reveal any features that allow for differentiating between AA9 types. However, it was found that hydrophobicity of the motifs that determine type specificity are primary located on the surface of the AA9 protein structure and these motifs are hydrophilic according to the Kd scale.

PDBsum identified 310 helices in all three AA9 types even though they were not present on all the available structures. These helices were found on the surface of AA9 structures. The biological function of these 310 helical structures is yet to be elucidated but it is commonly believed that these helices are an intermediary state between random coil and a fully formed alpha helix. In conclusion the study was successful in determining a means of distinguishing between the various AA9 types. Unique structural features were also identified on these enzymes which warrant the need for further investigation through the use of techniques such as molecular dynamics as well as docking.

References

Aachmann, F.L., Sorlie, M., Skjak-Braek, G., Eijsink, V.G. & Vaaje-Kolstad, G. 2012, "NMR structure of a lytic polysaccharide monooxygenase provides insight into copper binding, protein dynamics, and substrate interactions", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 109, no. 46, pp. 18779-18784.

Al-Zuhair, S., Ramachandran, K. B., Farid, M., Kheireddine, Aroua, M., Vadlani, P., Ramakrishnan, S., and GardossI, L., Enzymes in Biofuels Production.

Baker, J. O., C. I. Ehrman, W. S. Adney, S. R. Thomas, and M. E. Himmel.1998. Hydrolysis of cellulose using ternary mixtures of purified celluloses.Appl. Biochem. Biotechnol. 70–72:395–403.

Bailey, T.L., Williams, N., Misleh, C., Li W.W., 2006. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic acids research*, 34(Web Server issue), pp.W369-73.

Beeson, W.T., Phillips, C.M., Cate, J.H. & Marletta, M.A. 2012, "Oxidative cleavage of cellulose by fungal copper-dependent polysaccharide monooxygenases", *Journal of the American Chemical Society,* vol. 134, no. 2, pp. 890-892.

Bhattarai, K., Stalick, W.M., McKay, S., Geme, G. & Bhattarai, N. 2011, "Biofuel: an alternative to fossil fuel for alleviating world energy and economic crises", *Journal of environmental science and health.Part A, Toxic/hazardous substances & environmental engineering,* vol. 46, no. 12, pp. 1424-1442.

Bork, P. & Koonin, E.V. 1996, "Protein sequence motifs", *Current opinion in structural biology,* vol. 6, no. 3, pp. 366-376.

Cavasotto, C.N. & Phatak, S.S., 2009. Homology modeling in drug discovery: current trends and applications. *Drug discovery today*, 14(13-14), pp.676-83.

Din, N., Gilkes, N.R., Tekant, B., Miller, R.C., Warren, R.A.J. & Kilburn, D.G. 1991, *Non–Hydrolytic Disruption of Cellulose Fibres by the Binding Domain of a Bacterial Cellulase*.

Do, C.B. et al., 2005. ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome research*, 15(2), pp.330-40.

Eastwood, D.C., Floudas, D., Binder, M., Majcherczyk, A., Schneider, P., Aerts, A., Asiegbu, F.O., Baker, S.E., Barry, K., Bendiksby, M., Blumentritt, M., Coutinho, P.M., Cullen, D., de Vries, R.P., Gathman, A., Goodell, B., Henrissat, B., Ihrmark, K., Kauserud, H., Kohler, A., LaButti, K., Lapidus, A., Lavin, J.L., Lee, Y.H., Lindquist, E., Lilly, W., Lucas, S., Morin, E., Murat, C., Oguiza, J.A., Park, J., Pisabarro, A.G., Riley, R., Rosling, A., Salamov, A., Schmidt, O., Schmutz, J., Skrede, I., Stenlid, J., Wiebenga, A., Xie, X., Kues, U., Hibbett, D.S., Hoffmeister, D., Hogberg, N., Martin,

F., Grigoriev, I.V. & Watkinson, S.C. 2011, "The plant cell wall-decomposing machinery underlies the functional diversity of forest fungi", *Science (New York, N.Y.),* vol. 333, no. 6043, pp. 762-765.

Eswar, N. et al., 2007. Comparative protein structure modeling using MODELLER. *Current protocols in protein science / editorial board, John E. Coligan ... [et al.]*, Chapter 2, p.Unit 2.9.

Goodell, B., Jellison, J., Liu, J., Daniel, G., Paszczynski, A., Fekete, F., Krishnamurthy, S., Jun, L. & Xu, G. 1997, "Low molecular weight chelators and phenolic compounds isolated from wood decay fungi and their role in the fungal biodegradation of wood", *Journal of Biotechnology; Low Molecular Weight Compounds in Lignin Degradation,* vol. 53, no. 2, pp. 133-162.

Harris, P.V., Welner, D., McFarland, K.C., Re, E., Navarro Poulsen, J., Brown, K., Salbo, R., Ding, H., Vlasenko, E., Merino, S., Xu, F., Cherry, J., Larsen, S. & Lo Leggio, L. 2010, "Stimulation of Lignocellulosic Biomass Hydrolysis by Proteins of Glycoside Hydrolase Family 61: Structure and Function of a Large, Enigmatic Family", *Biochemistry,* vol. 49, no. 15, pp. 3305-3316.

Henrissat, B. 1991. A classification of glycosyl hydrolases based on amino acid sequence similarities. Biochem. J. 280:309–316

Himmel, M.E., Ding, S.Y., Johnson, D.K., Adney, W.S., Nimlos, M.R., Brady, J.W. & Foust, T.D. 2007, "Biomass recalcitrance: engineering plants and enzymes for biofuels production", *Science (New York, N.Y.),* vol. 315, no. 5813, pp. 804-807.

Hodgman, T.C. 1989, "The elucidation of protein function by sequence motif analysis", *Computer applications in the biosciences : CABIOS,* vol. 5, no. 1, pp. 1-13.

Horn, S.J., Vaaje-Kolstad, G., Westereng, B. & Eijsink, V.G. 2012, "Novel enzymes for the degradation of cellulose", *Biotechnology for biofuels,* vol. 5, no. 1, pp. 45-6834-5-45.

Jeoh, T., D. B. Wilson, and L. P. Walker. 2002. Cooperative and competitivebinding in synergistic mixtures of Thermobifida fusca cellulases Cel5A,Cel6B, and Cel9A. Biotechnol. Prog. 18:760–769..

Karkehabadi, S., Hansson, H., Kim, S., Piens, K., Mitchinson, C. & Sandgren, M. 2008, "The first structure of a glycoside hydrolase family 61 member, Cel61B from Hypocrea jecorina, at 1.6 A resolution", *Journal of Molecular Biology,* vol. 383, no. 1, pp. 144-154.

Karlsson, J., Saloheimo, M., Siika-Aho, M., Tenkanen, M., Penttila, M. & Tjerneld, F. 2001, "Homologous expression and characterization of Cel61A (EG IV) of Trichoderma reesei", *European journal of biochemistry / FEBS,* vol. 268, no. 24, pp. 6498-6507

Katoh, K. & Toh, H., 2008. Recent developments in the MAFFT multiple sequence alignment program. *Briefings in bioinformatics*, 9(4), pp.286-98.

Lazaridis, T. & Karplus, M., 1999. Effective energy function for proteins in solution. *Proteins*, 35(2), pp.133-52.

Levasseur, A., Drula, E., Lombard, V., Coutinho, P.M. & Henrissat, B. 2013, "Expansion of the enzymatic repertoire of the CAZy database to integrate auxiliary redox enzymes", *Biotechnology for biofuels,* vol. 6, no. 1, pp. 41-6834-6-41.

Li, X., Beeson, W.T.,4th, Phillips, C.M., Marletta, M.A. & Cate, J.H. 2012, "Structural basis for substrate targeting and catalysis by fungal polysaccharide monooxygenases", *Structure (London, England : 1993),* vol. 20, no. 6, pp. 1051-1061.

Li, S.C. Bu, D., Xu, J., Li, M., 2011. Finding nearly optimal GDT scores. *Journal of computational biology : a journal of computational molecular cell biology*, 18(5), pp.693-704

di Luccio, E. & Koehl, P. (2011). A Quality Metric for Homology Modelling: The HFactor.Bioinformatics, 12, 48.

Lobry, J.R., Gautier, C. (1994) Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 Escherichia coli chromosome-encoded genes. Nucleic Acids Research, 22:3174-3180

Lovell, S.C.; Davis, I.W.; Arendall, W.B.; De Bakker, P.I.W.; Word, J.M.; Prisant, M.G.; Richardson, J.S.; Richardson, D.C. (2003). "Structure validation by Cα geometry: ϕ,ψ and Cβ deviation". Proteins: Structure, Function, and Genetics 50 (3): 437–50.

Merino, S.T. & Cherry, J. 2007, "Progress and challenges in enzyme development for biomass utilization", *Advances in Biochemical Engineering/Biotechnology,* vol. 108, pp. 95-120.

Moon, R.J., Martini, A., Nairn, J., Simonsen, J. & Youngblood, J. 2011, "Cellulose nanomaterials review: structure, properties and nanocomposites", *Chemical Society Reviews,* vol. 40, no. 7, pp. 3941-3994.

Mu, D., Seager, T., Rao, P.S. & Zhao, F. 2010, "Comparative life cycle assessment of lignocellulosic ethanol production: biochemical versus thermochemical conversion", *Environmental management,* vol. 46, no. 4, pp. 565-578.

Millhauser, G.L., 1999. r and 3 10 : The Split Personality of Polypeptide Helices. , pp.1027-1033.

Nunoura, N., K. Ohdan, T. Yano, K. Yamamoto, and H. Kumagai. 1996.Purification and characterization of beta-D-glucosidase (beta-D-fucosidase) from Bifidobacterium breve clb acclimatedto cellobiose. Biosci. Biotechnol. Biochem. 60:188–193. Pawlowski, M. et al., 2008. MetaMQAP: a meta-server for the quality assessment of protein models. *BMC bioinformatics*, 9, p.403.

Pei, J., 2008. Multiple protein sequence alignment. *Current opinion in structural biology*, 18(3), pp.382-6.

Pei, J., Kim, B.-H. & Grishin, N.V., 2008. PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic acids research*, 36(7), pp.2295-300.

Pei, J., Tang, M. & Grishin, N.V., 2008. PROMALS3D web server for accurate multiple protein sequence and structure alignments. *Nucleic acids research*, 36(Web Server issue), pp.W30-4.

Perez, S. & Samain, D. 2010, "Structure and engineering of celluloses", *Advances in Carbohydrate Chemistry and Biochemistry,* vol. 64, pp. 25-116.

Phillips, C.M., Beeson, W.T., Cate, J.H. & Marletta, M.A. 2011, "Cellobiose dehydrogenase and a copper-dependent polysaccharide monooxygenase potentiate cellulose degradation by Neurospora crassa", *ACS chemical biology,* vol. 6, no. 12, pp. 1399-1406.

Quinlan, R.J., Sweeney, M.D., Lo Leggio, L., Otten, H., Poulsen, J.C., Johansen, K.S., Krogh, K.B., Jorgensen, C.I., Tovborg, M., Anthonsen, A., Tryfona, T., Walter, C.P., Dupree, P., Xu, F., Davies, G.J. & Walton, P.H. 2011, "Insights into the oxidative degradation of cellulose by a copper metalloenzyme that exploits biomass components", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 108, no. 37, pp. 15079-15084.

Rohl, C. a, 2005. Protein structure estimation from minimal restraints using Rosetta. *Methods in enzymology*, 394, pp.244-60.

Shen, M.-yi & Sali, A., 2006. Statistical potential for assessment and prediction of protein structures. , pp.2507-2524.

Sánchez, R. & Sali, a, 1997. Advances in comparative protein-structure modelling. *Current opinion in structural biology*, 7(2), pp.206-14.

Söding, J., Biegert, A. & Lupas, A.N., 2005. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic acids research*, 33(Web Server issue), pp.W244-8.

Takahashi, S., Leiss, M., Moser, M., Ohashi, T., Kitao, T., Heckmann, D., Pfeifer, A., Kessler, H., Takagi, J., Erickson, H.P. & Fassler, R. 2007, "The RGD motif in fibronectin is essential for development but dispensable for fibril assembly", *The Journal of cell biology,* vol. 178, no. 1, pp. 167-178.

Tramper, J. 1996, "Chemical versus biochemical conversion: when and how to use biocatalysts", *Biotechnology and bioengineering,* vol. 52, no. 2, pp. 290-295.

Uchiyama, T. et al., 2001. Roles of the exposed aromatic residues in crystalline chitin hydrolysis by chitinase A from Serratia marcescens 2170. *The Journal of biological chemistry*, 276(44), pp.41343-9.

Vaaje-Kolstad , G., Houston, D.R., Riemen, A.H., Eijsink, V.G. & van Aalten, D.M. 2005, "Crystal structure and binding properties of the Serratia marcescens chitin-binding protein CBP21", *The Journal of biological chemistry,* vol. 280, no. 12, pp. 11313-11319.

Vaaje-Kolstad, G., Westereng, B., Horn, S.J., Liu, Z., Zhai, H., Sorlie, M. & Eijsink, V.G. 2010, "An oxidative enzyme boosting the enzymatic conversion of recalcitrant polysaccharides", *Science (New York, N.Y.),* vol. 330, no. 6001, pp. 219-222.

Vieira-Pires, R.S. & Morais-Cabral, J.H., 2010. 3(10) Helices in Channels and Other Membrane Proteins. *The Journal of general physiology*, 136(6), pp.585-92.

Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ (2009) Jalview Version 2-a multiple sequence alignment editor and analysis workbench. Bioinformatics 25: 1189-1191

White, A.R. & Brown, R.M. 1981, "Enzymatic hydrolysis of cellulose: Visual characterization of the process", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 78, no. 2, pp. 1047-1051.

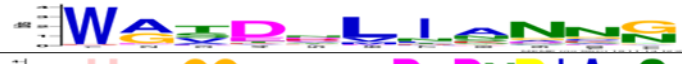
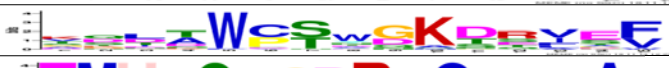
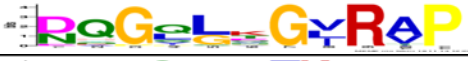
Wu, M., Beckham, G.T., Larsson, A.M., Ishida, T., Kim, S., Payne, C.M., Himmel, M.E., Crowley, M.F., Horn, S.J., Westereng, B., Igarashi, K., Samejima, M., Stahlberg, J., Eijsink, V.G. & Sandgren, M. 2013, "Crystal structure and computational characterization of the lytic polysaccharide monooxygenase GH61D from the Basidiomycota fungus Phanerochaete chrysosporium", *The Journal of biological chemistry,* vol. 288, no. 18, pp. 12828-12839.

Yakovlev, I., Vaaje-Kolstad, G., Hietala, A.M., Stefanczyk, E., Solheim, H. & Fossdal, C.G. 2012, "Substrate-specific transcription of the enigmatic GH61 family of the pathogenic white-rot fungus Heterobasidion irregulare during growth on lignocellulose", *Applied Microbiology and Biotechnology,* vol. 95, no. 4, pp. 979-990.

Appendices

- **Motif analysis**

Table A1: Web logos of the MEME identified motifs.

Motif	E-value	Sites		Web logo
1	4.5e-230	159		
2	6.0e-1888	157		
3	4.6e-1702	159		
4	2.4e-1230	159		
5	8.5e-1076	144		
6	4.6e-1053	44		
7	6.2e-1038	148		
8	6.7e-674	113		
9	5.0e-653	153		
10	1.8e-555	159		
11	1.9e-367	99		
12	1.0e-214	19		
13	3.9e-193	62		
14	2.0e-191	40		
15	1.1e-142	15		
16	4.4e-137	37		
17	5.1e-104	28		
18	1.4e-090	15		
19	2.6e-072	84		
20	1.1e-050	13		

- Scripts

Clean_redunt.py

```
from itertools import groupby


file1=open("all_sequences.fa", 'r+') #input file


if __name__ == '__main__':
    string_in="./original/"+fasta.rstrip("\n ") #input redundant file
    string_out= "./cleaned/"+fasta.rstrip("\n") #outtput noredundant file

    ishead = lambda x: x.startswith('>')
    all_seqs = set()
    with open(string_in) as handle:
    with open(string_out, 'w') as outhandle:
        head = None
        for h, lines in groupby(handle, ishead):
            if h:
                head = lines.next()
            else:
                seq = ''.join(lines)
                if seq not in all_seqs:
                    all_seqs.add(seq)
                    outhandle.write('%s\n%s\n' % (head, seq)) # writing output file
```

Gap_remover.py

```
import re
string = open('promals_selection1.fa').read()
new_str = re.sub('[-]', '', string)
open('selection1.fa', 'w').write(new_str)
```

Physicochemical_properties.py

```
from Bio.SeqUtils import ProtParam
from Bio.SeqUtils import ProtParamData
import matplotlib.pyplot as plt
import pylab as p

handle = open("group_all_domains_sequences.fa", "r")

hydrophobicity= []
aromaticity = []
amino_percent =[]

for rec in SeqIO.parse(handle, "fasta"):
    seq = str(rec.seq)
    X = ProtParam.ProteinAnalysis(seq)
    hydrophobicity = hydrophobicity + [X.protein_scale(ProtParamData.kd, 9, 0.4)]
    aromaticity = aromaticity + [X.aromaticity()]
    amino_percent =amino_percent + [X.get_amino_acids_percent()]
prop1=open("aromaticity.txt",'w')
prop2=open("hydrophobicity.txt",'w')
handle = open("group_all_domains_sequences.fa", "r")
names=[]
for j in handle:
    if j[0]=='>':
        names=names+[j[1:].rstrip("\n").rstrip("\r")]
properties=''
menu="yuvani"
while menu !="":
    menu=raw_input("*****************************"+"\n*"+"  what do yo want to do ?   *"+"\n*                          *"+"\n*"+"\t"+"1-aromaticity       *"+"\n*

    if menu==str(1):
        all_or_org=raw_input("Do you want to analyze all (1) or individually (2)"+"\n"+"  :")
        if all_or_org==str(1):
            fig = p.figure()
            ax = fig.add_subplot(1,1,1)
            x = len(names)
            y = aromaticity
            ax.bar(range(x),y)
            p.show()
        if all_or_org==str(2):

            organism=raw_input("Enter organism number : "+"\n")
            names[int(organism)]
            print "The Aromaticity of ",names[int(organism)][:-6],"is",str(aromaticity[int(organism)])
    if menu==str(2):
        choice=raw_input("Select organism you want to analyse : "+"\n")

        fig = p.figure()
        ax = fig.add_subplot(1,1,1)
        x = len(hydrophobicity[int(choice)])
        y = hydrophobicity[int(choice)]
        ax.bar(range(x),y)
        p.show()

    if menu==str(3):
        choose=raw_input("Select organism you want to analyse : "+"\n")
        i=amino_percent[int(choose)]
        plt.bar(range(len(i)), i.values(), align='center')
        plt.xticks(range(len(i)), i.keys())
        plt.draw()
        plt.show()
```

Clean_pdb.py

```
file1=open("3ZUD.pdb", 'r+')

lines=[]
CL=[]
for i in file1:
    lines=lines+[i]
file1.close()
for line in lines:
    if line[0:4]=="ATOM":
        if line[21]=="A":
            CL.append(line)
    if line[0:3]=="TER":
        if line[21]=="A":
            CL.append(line)

with open("cleaned.pdb","w") as f:
    f.write("".join(CL)) #writing output file
    print "File SAVED."

```

Respdb.py

```
from modeller import *
env = environ()
code = '3ZUD'
mdl = model(env, file=code)
aln = alignment(env)
aln.append_model(mdl, align_codes=code)
aln.write(file=code+'.seq')
```

Modelling.py

```
# modeling by the automodel class
from modeller import *# Load standard Modeller classes
from modeller.automodel import *# Load the automodel class
log.verbose()
env = environ()# request verbose output
        # create a new MODELLER environment to build this model in
        # directories for input atom files
env.io.atom_files_directory = "."
a = automodel(env,
alnfile = '3ZUD_control.pir',# alignment filename
knowns= '3ZUD',# codes of the templates
sequence = '3ZUD_sequence')# code of the target
a.starting_model= 1# index of the first model
a.ending_model = 100# index of the last model
            # (determines how many models to calculate)
a.md_level=refine.slow


# Thorough MD optimization:
a.md_level = refine.very_slow


a.make() # do the actual homology modeling
```

Scoring.py

```
import sys
# This script computes dope scores for protein structure,it sorts them in ascendiorder and
# outputs the best three (with highest dope score)
# Example for: model.assess_normalized_dope()
from modeller import *
from modeller.scripts import complete_pdb
from rosetta import *
init()
env = environ()
env.libs.topology.read(file='$(LIB)/top_heav.lib')
env.libs.parameters.read(file='$(LIB)/par.lib')
# directories for input atom files
env.io.atom_files_directory = '.'
# Read a model previously generated by Modeller's automodel class
files = sys.argv[1]
f1 = open (files)
filename = ""
f2 = open ("scores.txt","w")
f2.write("Z-SCORE"+"\t"*2+"CA RMSD"+"\t"*2+"Rosetta"+"\t"*3+"pdb file"+"\n")
for line in f1:
    if(len(line)>1):
    filename = str.strip(line)
    mdl = complete_pdb(env, filename)
    zscore = mdl.assess_normalized_dope()
    #rosseta and  RMSD
    pose1=Pose()
    pose2=Pose()
    pose_from_pdb(pose1, filename)
    pose_from_pdb(pose2, "template.pdb")
    scorefxn = create_score_function("standard")
    ros=scorefxn(pose1)
    rmsd=CA_rmsd(pose1,pose2)
    f2.write(str(zscore)+"\t"+str(rmsd)+"\t"+str(ros)+"\t"+filename+"\n")
f2.close()
#Script originally written by Matthys Kroon and adapted by vuyani moses
```

Type-1 sequence	Aromaticity
chaetomium_globusum_8	0.1435
3eja_chainA_p001	0.1538
3EII:A_	0.1546
TYPE1_3EJA:A	0.1553
thievela_terestis_18	0.1473
myceliophthora_thermophilia_21	0.1435
TYPE1:NCU03328_	0.1485
podospora_anseria_18	0.1480
pyrenophora_trici_repentis_13	0.1227
pyrenophora_teres_11	0.1292
Phaeospheraria_nodorum_18	0.1239
emmericella_nidulan_3	0.1223
aspergillus_tereus_4	0.1316
podospora_anseria_11	0.1212
glomerrela_graminic_6	0.1391
arthrobotrys_oligospora_11	0.1217
pyrenophora_trici_repentis_20	0.1272
Phaeospheraria_nodorum_28	0.1372
myceliophthora_thermophilia_16	0.1217
thievela_terestis_11	0.1212
chaetomium_globusum_24	0.1195
chaetomium_thermophilia_14	0.1342
TYPE1ncr:NCU02344	0.1379
schizzophylum_commune_15	0.1096
schizzophylum_commune_16	0.1184
serpula_lacrymans_5	0.1135
neurospora_tetrasperma_12	0.1396
neurospora_cassa_6	0.1429
TYPE1NCU00836	0.1422
thievela_terestis_7	0.1422
podospora_anseria_5	0.1422
aspergillus_fuminga_3	0.1519
aspergillus_favus_3	0.1525
penicillum_chrysoge_4	0.1447
aspergillus_niger_9	0.1296
emmericella_nidulan_6	0.1392
aspergillus_clavatus_1	0.1477
neosartorya_fischer_7	0.1519
verticilium_albo_atrum_17	0.1476
verticillium_dahiae_25	0.1524
aspergillus_clavatus_5	0.1603
cholleotitracum_hig24	0.1561

glomerrela_graminic_7	0.1603
aspergillus_fuminga_3	0.1519

Table A2: AA9 Type 1 Domain aromaticity.

Table A3: Type 2 AA9 Domain aromaticity

Type 2 sequence	Aromaticity
leptosphaeria_maculans_15	0.1364
glomerrela_graminic_9	0.1039
4EIR:A	0.0987
neurospora_cassa_1	0.0936
TYPE2ncr:NCU01050	0.0889
neurospora_tetrasperma_1	0.0936
sodaria_macrospora_11	0.0894
myceliophthora_thermophilia_10	0.0940
podospora_anseria_30	0.1116
colleotitracum_higginsianum_23	0.0983
chaetomium_thermophilia_18	0.1026
Phaeospheraria_nodorum_9	0.0973
glarea_lozoyensis_6	0.1088
TYP2ncr:NCU02240	0.0983
botryotinia_fuckelina_2	0.0944
botryotinia_fuckelina_12	0.1088
sclerotinia_sclerotiorum_5	0.1088
P_indica_8	0.1053
pyrenophora_trici_repentis_11	0.1150
leptosphaeria_maculans_11	0.1036
Phaeospheraria_nodorum_14	0.1062
pyrenochaeta_lycope_1	0.1062
pyrenophora_teres_25	0.1150
glarea_lozoyensis_1	0.0935
myceliophthora_thermophilia_12	0.1046
chaetomium_thermophilia_7	0.1004
podospora_anseria_15	0.1030
thievela_terestis_10	0.1111
podospora_anseria_24	0.1046
pyrenophora_teres_17	0.0991
pyrenophora_trici_repentis_8	0.0991
chaetomium_globusum_22	0.1008
thievela_terestis_17	0.1046
glomerrela_graminic_32	0.1039
verticillium_dahiae_23	0.1157
verticilium_albo_atrum_15	0.1388
glomerrela_graminic_17	0.1111
chaetomium_globusum_23	0.0988
podospora_anseria_29	0.1235

chaetomium_thermophilia_15	0.1276
sodaria_macrospora_16	0.1025
myceliophthora_thermophilia_5	0.1111
thievela_terestis_2	0.1111
P_indica_13	0.1000
P_indica_12	0.1042
myceliophthora_thermophilia_6	0.1083
chaetomium_globusum_31	0.1125
podospora_anseria_4	0.1292
sodaria_macrospora_2	0.1004
neurospora_tetrasperma_16	0.0958
neurospora_cassa_4	0.0958
neurospora_tetrasperma_14	0.0958
TYPE2:NCU02916	0.0958
verticillium_dahiae_24	0.1312
verticilium_albo_atrum_16	0.1435
verticilium_dahiae_26	0.1364

Table A4: Type 3 AA9 domain aromaticity

Type 3 sequence	Aromaticity
type_3ncr:NCU07898	0.1471
4EIS:B	0.1511
giberella_zeae_7	0.1240
nectria_heamatoccocus_1	0.1195
fusarium_oxysporum_3	0.1355
nectria_heamatoccocuss_1	0.1292
verticilium_albo_atrum_13	0.1250
verticillium_dahiae_4	0.1213
neurospora_cassa_13	0.1213
neurospora_tetrasperma_2	0.1213
type3:NCU07760	0.1364
magna_porte_oryzae_16	0.1235
podospora_anseria_31	0.1224
hypocrea_orientalis_1	0.1184
trichoderma_SP_SSL__1	0.1184
Hypocrea_virens_3	0.1306
trichoderma_atrovir_2	0.1224
hypocrea_rufa_1	0.1224
hypocrea_rufa_2	0.1224
trichoderma_saturnisporum_1	0.1220
aspergillus_kawachii_37	0.1184
aspergillus_tereus_6	0.1220
neosartorya_fischer_4	0.1179
aspergillus_fuminga_4	0.1189
aspergillus_tereus_10	0.1324
chaetomium_globusum_5	0.1198
type3:NCU05969	0.1189
aspergillus_fuminga_1	0.1189
neosartorya_fischer_1	0.1189
aspergillus_niger_11	0.1189
aspergillus_niger_1	0.1189
aspergillus_kawachii_40	0.1169
aspergillus_tereus_5	0.1148
emmericella_nidulan_9	0.1228
2YET:A	0.1239
type3_3ZUD:A	0.1239
penicillum_chrysoge_2	0.1148
aspergillus_niger_2	0.1189
aspergillus_niger_12	0.1152
aspergillus_kawachii_38	0.1189

zea_mys_1	0.1189
aspergillus_clavatus_6	0.1250
aspergillus_tereus_8	0.1125
aspergilus_oryzae_7	0.1203
aspergillus_favus_5	0.1203
glomerrela_graminic_4	0.1203
podospora_anseria_17	0.1435
glomerrela_graminic_16	0.1565
glomerrela_graminic_31	0.1478
chaetomium_globusum_15	0.1447
pyrenophora_teres_22	0.1354
Phaeospheraria_nodo_4	0.1310
pyrenophora_trici_repentis_23	0.1256
neosartorya_fischer_2	0.1453
aspergillus_tereus_12	0.1496
aspergillus_favus_7	0.1453
aspergillus_fuminga_6	0.1453
type3_2VTC:A	0.1447

- **Modeling**

Control evaluation

Table A5: Top 3 self models for the 3 templates used for modelling.

Template	DOPE score	CA RMSD	Rosetta energy	Model
3ZUD	-1.80784	0.149898	241.1738	3ZUD_sequence.B99990041.pdb
	-1.81324	0.129595	244.9788	3ZUD_sequence.B99990090.pdb
	-1.79227	0.130587	250.4152	3ZUD_sequence.B99990040.pdb
4B5Q	-1.90491	0.135702	167.2205	4B5Q.B99990098.pdb
	-1.90463	0.118083	177.2999	4B5Q.B99990099.pdb
	-1.8961	0.133634	118.4848	4B5Q.B99990011.pdb
4EIR	-1.4943	0.137133	92.91127	4EIR_SEQUENCE.B99990091.pdb
	-1.50292	0.13813	100.2746	4EIR_SEQUENCE.B99990098.pdb
	-1.44673	0.142631	129.4237	4EIR_SEQUENCE.B99990053.pdb

3ZUD self models

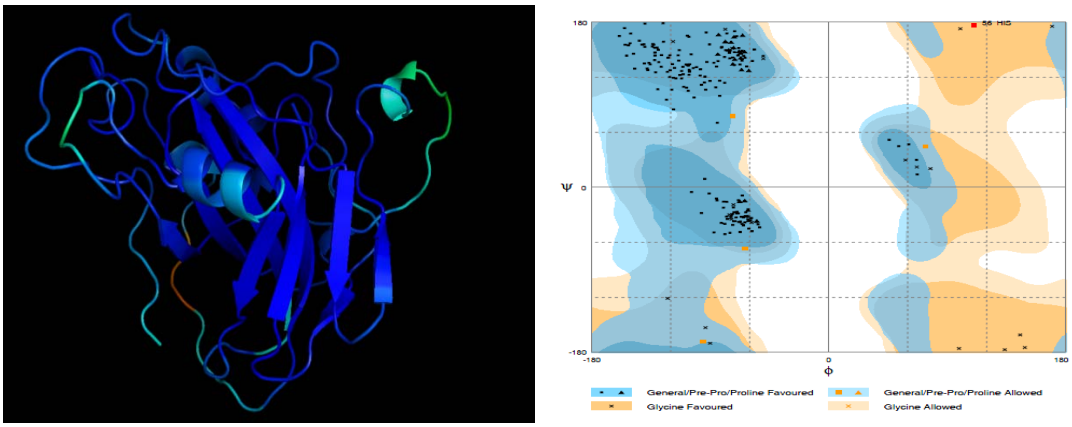


Figure A1: MetaMQAP and RAMPAGE analysis of 3ZUD_sequence.B99990040.pdb

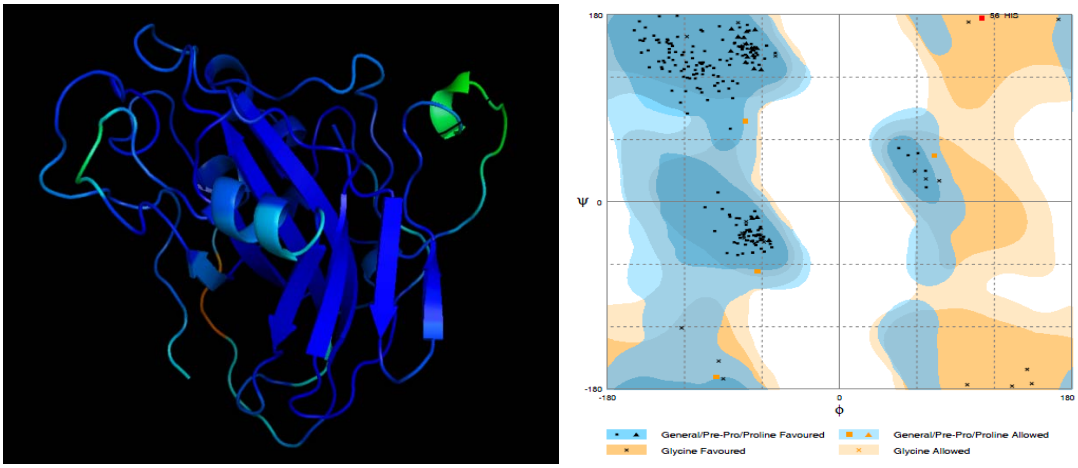


Figure A2: MetaMQAP and RAMPAGE analysis of 3ZUD_sequence.B99990041.pdb

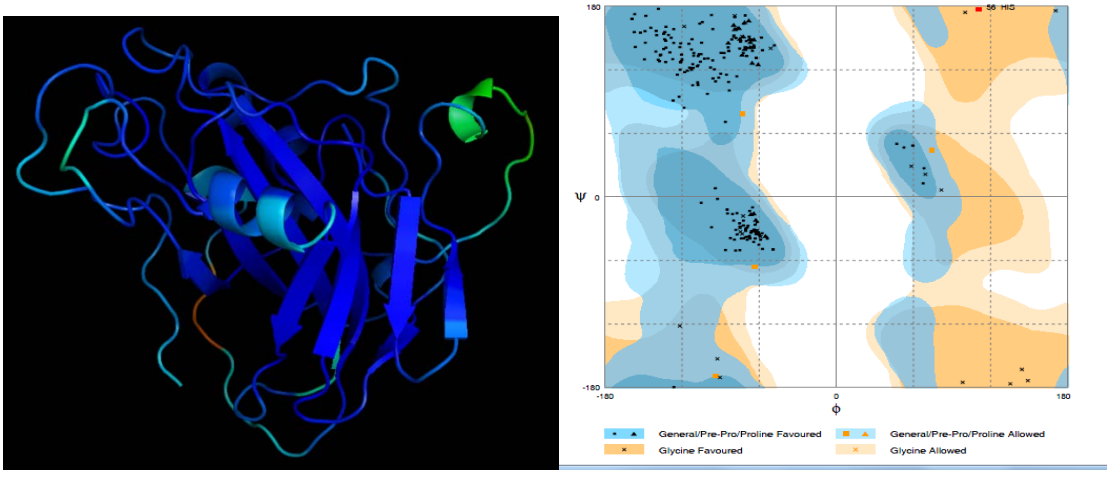


Figure A3: MetaMQAP and RAMPAGE analysis of 3ZUD_sequence.B99990090.pdb

4BQ5 self models

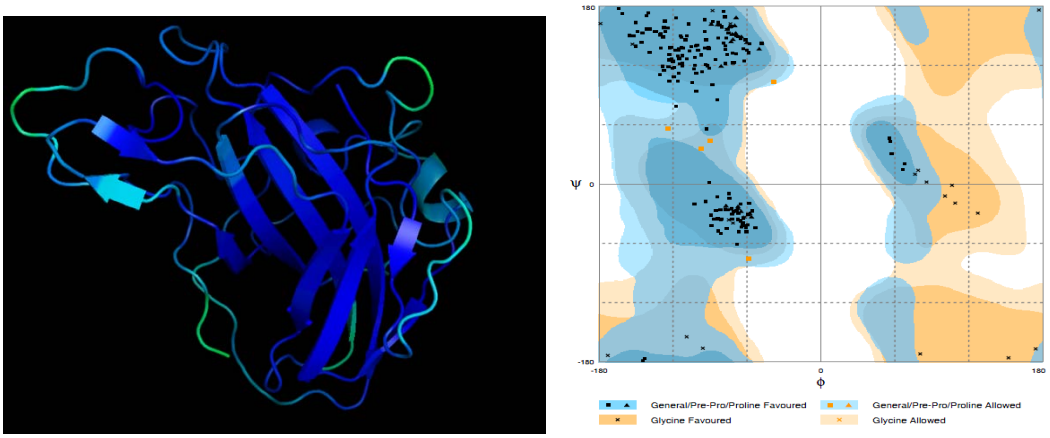


Figure A4: MetaMQAP and RAMPAGE analysis of 4B5Q.B99990011.pdb

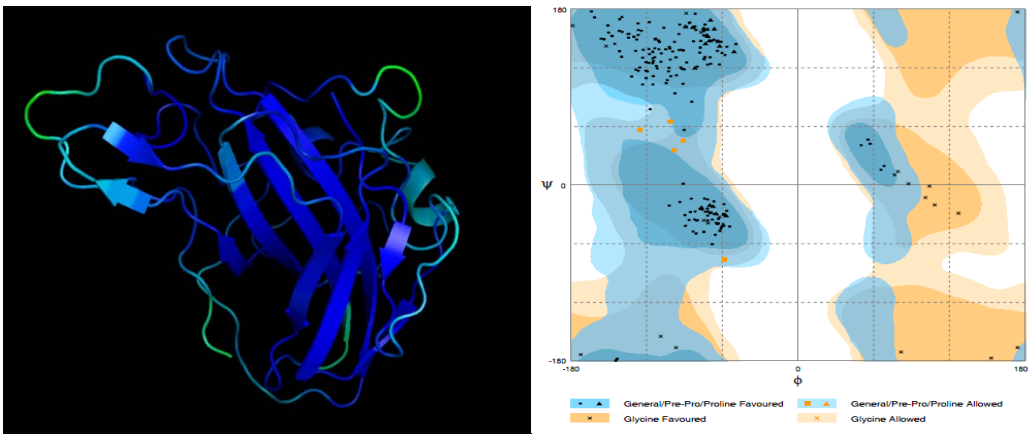


Figure A5: MetaMQAP and RAMPAGE analysis of 4B5Q.B99990098.pdb

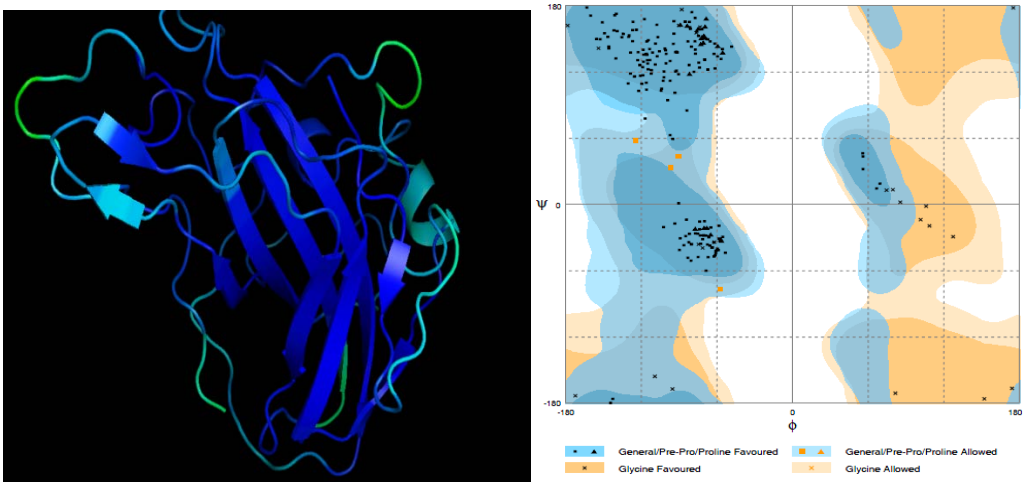


Figure A6: MetaMQAP and RAMPAGE analysis of 4B5Q.B99990099.pdb

- 4EIR self models

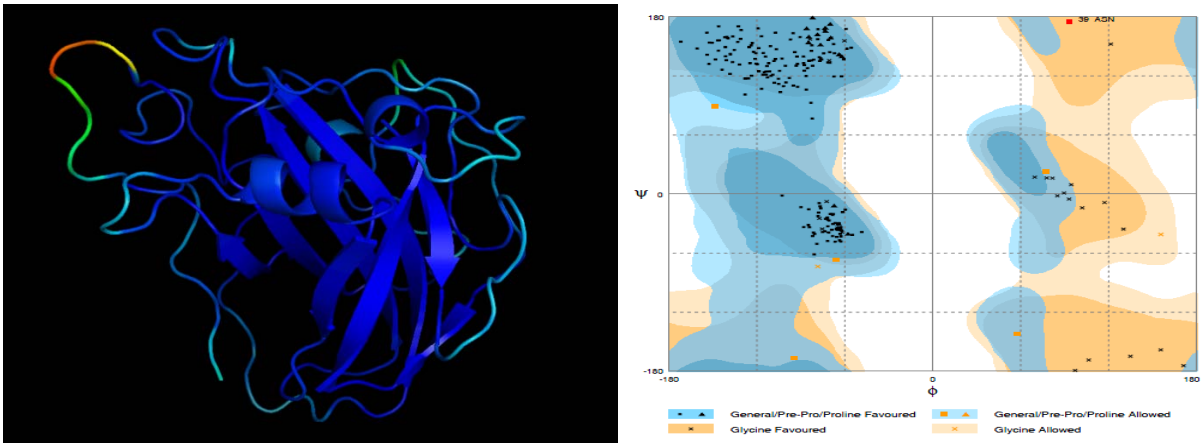


Figure A7: MetaMQAP and RAMPAGE analysis of 4EIR_SEQUENCE.B99990053.pdb

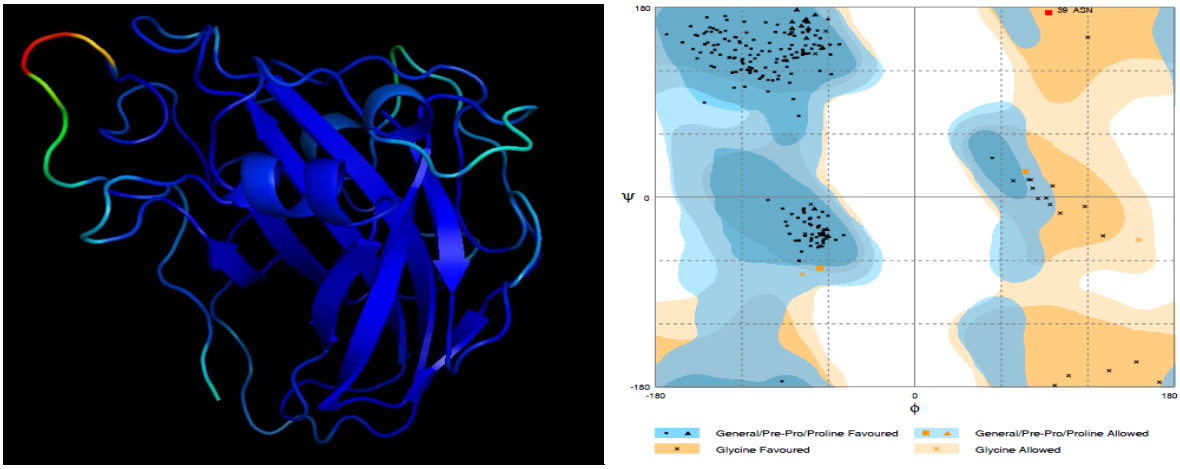


Figure A8: MetaMQAP and RAMPAGE analysis of 4EIR_SEQUENCE.B99990091.pdb

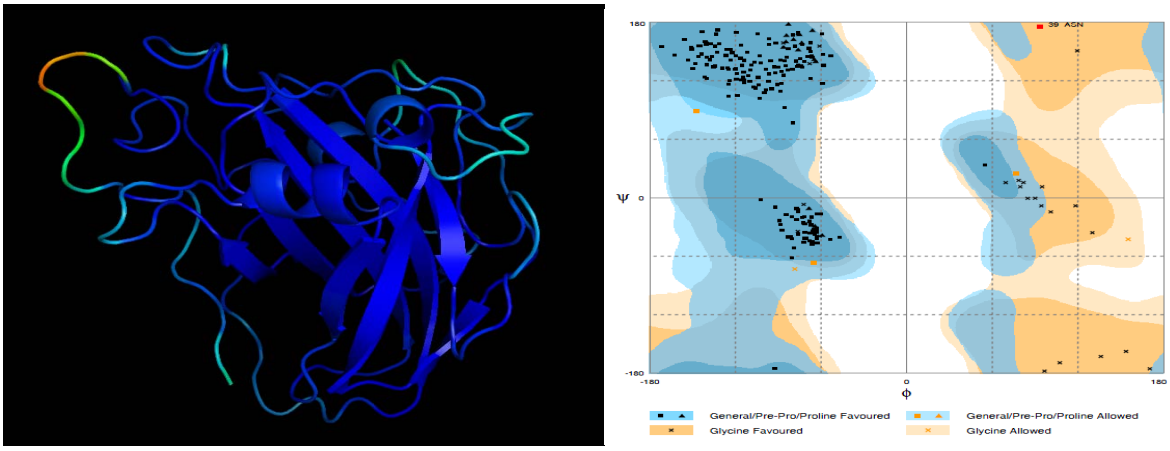


Figure A9: MetaMQAP and RAMPAGE analysis of 4EIR_SEQUENCE.B99990098.pdb

- **Phylogenetic analysis**

Table A6: Maximum Likelihood fits of 48 different amino acid substitution models at 100% Site coverage

Model	Parameters	BIC	AICc	*lnL*
WAG+G	322	10574.998	8496.329	-3905.458
WAG+G+I	323	10583.582	8498.594	-3905.458
rtREV+G	322	10616.681	8538.011	-3926.300
rtREV+G+I	323	10625.265	8540.277	-3926.300
Dayhoff+G	322	10724.361	8645.692	-3980.140
JTT+G	322	10726.092	8647.422	-3981.005
Dayhoff+G+I	323	10732.945	8647.957	-3980.140
JTT+G+I	323	10734.676	8649.688	-3981.005
WAG+G+F	341	10793.747	8595.179	-3933.284
WAG	321	10794.285	8721.935	-4019.394
WAG+G+I+F	342	10802.331	8597.462	-3933.284
WAG+I	322	10802.869	8724.199	-4019.394

Figure A10: Molecular Phylogenetic analysis by Maximum Likelihood method of AA9 proteins at 100% site coverage.