The genetic diversity of the CbpA gene in *Streptococcus pneumoniae*

Steven Okinyi Sewe, Bsc

Reg. No I56/70951/2007

Centre for Biotechnology and Bioinformatics

College of Biological and Physical Sciences

University of Nairobi

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

Master of Science in Bioinformatics

July 2015

Declaration and approval

I declare that this thesis is my own work, and to the best of my knowledge has never been submitted as proposed work of study or examined for the award of degree in any university.

Steven Okinyi Sewe

Registration Number: I56/70951/2007

Signature; …………………………… Date; …………………………

Thesis Approval

This thesis has been submitted with our approval as university supervisors.

Dr. George Obiero, Ph.D.

Centre for Biotechnology and Bioinformatics

University of Nairobi

Signature; …………………………….. Date; ………………………………

Dr. Benard Kulohoma, Ph.D.

Centre for Biotechnology and Bioinformatics

University of Nairobi

Signature; ………………………………. Date; ………………………………

Acknowledgements

I would like to thank the almighty God for the patience and strength he gave me to painstakingly gather all the information that was needed and put it all together. I sincerely appreciate my supervisors, Dr. Benard Kulohoma and Dr. George Obiero for their support. It would not have been possible without their frequent inputs. I thank my family for support and understanding when I was engrossed in the research work. Last but not least, many thanks to friends and colleagues who were always there to help.

Dedication

To my family, for their prayers and support. Your encouragement saw me through.

Abstract

Streptococcus pneumoniae (pneumococcus) is a common human pathogen responsible for morbidity and mortality worldwide. It causes mild to life-threatening, inflammatory diseases such as otitis media, pneumonia, sepsis and meningitis. The prevention and management of pneumococcal infections has been very challenging. Over time there has been increasing drug resistance of pneumococcus strains against the available antibiotics. Moreover, the pneumococcal vaccines currently available in the market do not offer broad coverage against the more than 90 serotypes currently identified. If novel treatment and preventative strategies are not adopted soon, then pneumococcal disease will continue to devastate human populations, especially in the developing countries where it causes the most damage. This study comprised of 213 fully annotated complete genome sequences of *S. pneumoniae* downloaded from GenBank. Amino acid and nucleic acid sequences of Choline binding protein A (CbpA) were successfully extracted from 211 genomes (99%) for study of genetic variation and identification of possible conserved, immunogenic regions eligible for novel vaccine targets. Multiple SequenCe alignment by Log-Expectation (MUSCLE) was used for alignment and the phylogenetic trees and heat maps created by PhyML and R respectively. The CbpA locus was found to be highly polymorphic at both the nucleic acid and amino acid sequences. However, RaptorX server showed that 83.3% of the pneumococcal protein domains predicted were the conserved modular teichoic acid phosphorylcholine esterase Pce (2bib:A) and the CbpA R2 domain (1w9r:A). Using Transmembrane Hidden Markov Model (TMHMM) server and VaxiJen v 2.0 server, these conserved domains were shown to be extracellularly located and immunogenic. The high variability observed in CbpA suggests its importance as a natural target for host defense and essential element for the colonization of different niches within the host. Further evaluation of 2bib:A and 1w9r:A conserved regions would be required to design novel, efficacious, serotype- independent CbpA- fusion protein vaccine candidate.

Table of contents

List of figures

List of abbreviations

PPV - Pneumococcal polysaccharide vaccine

PCV - Protein Conjugate Vaccine

GSK - Glaxosmithkline

GAVI - Global Alliance for Vaccination and Immunization

IPD - Invasive Pneumococcal Disease

LAC - Latin American and The Carribean

CI - Confidence interval

PspA - Pneumococcus surface protein A

CbpA - Choline binding protein A

PsaA - Pneumococcal surface adhesin A

PspC - Pneumococcus surface protein C

SpsA - Streptococcus pneumoniae secretory immunoglobulin A binding protein A

PbcA - C3-binding protein A

Hic - Factor H- binding inhibitor of complement

hpIgR - Human polymeric immunoglobulin receptor

NCBI - National Centre for Biotechnology Information

MUSCLE - Multiple Sequence Comparison by Log-Expectation

GTR - General Time Reversible

HIV - Human Immunodeficiency Virus

CSF - Cerebrospinal Fluid

PD - Pneumococcal Disease

Perl - Practical extraction and reporting language

CGE - Centre for Genomic Epidemiology

MLST - Multi-Locus Sequence Typing

ST - Sequence Typing

TMHMM - Trans-Membrane Hidden Markov Model

INDELS - Insertions and Deletions

DNA - Deoxyribonucleic acid

RCSB - Research Collaboration for Structural Bioinformatics

CBP - Choline Binding Protein

CBPE - Choline Binding Protein E

SLV - Single Locus Variant

TM - Transmembrane

PC - Phosphorylcholine

PAF - Platelet-Activating Factor

BBR - Blood Brain Barrier

Aa - Amino acid

CCs – Clonal Complexes

Cplx – Complex

Chapter 1

Introduction

1.1 *Streptococcus pneumoniae*

Streptococcus pneumoniae (pneumococcus) is a major cause of global health concern. It is estimated to have caused more than 800,000 deaths annually in children aged under 5 years [1]. Approximately 1.6 million worldwide deaths are reported each year as a result of pneumococcal diseases [2]. The burden of pneumococcal pneumonia is highest in developing countries [3]. In spite of the fact that young children and the elderly are most at risk of disease, all age groups including older children, adolescents, and adults may have this infectious disease [4]. The groups most at risk are children (<5 years), the elderly (>65 years), and people with immuno-compromising conditions, such as a removed spleen, HIV, and autoimmune disorders.

Pneumococcus naturally colonizes the nasopharynx thought, to act as the reservoir and source of pneumococcal transmission between individuals [5]. It can be found in about 10% of healthy adults and 40% of healthy children [6]. It may invade sterile tissue sites that include the cerebrospinal fluid (CSF), blood and the middle ear [5], and may result in life-threatening diseases such as pneumonia, meningitis, otitis media, and sepsis [7].

The emergence of multi-drug resistant strains and lack of broad coverage by existing vaccines has led to research towards the development of better preventive strategies capable of protecting all populations at risk. Currently, more than 90 pneumococcal serotypes of pneumococci exist, classified by different capsular polysaccharide structures [8].

The pneumococcal polysaccharide capsule is regarded as the major virulence determinant [9], but other antigens have also recently been identified, for example pneumococcal surface proteins. The vast majority of diseases are caused by the following serogroups, in descending order: 14, 6, 19, 18, 9, 23, 4, 1 and 15 in developed countries, but 6, 14, 8, 5, 1, 19, 9, 23, 18, 15 and 17 in developing countries [10].

The first pneumococcal polysaccharide vaccine (PPV14) was developed in 1976. It consisted of 14 polysaccharides from 14 pneumococcal serotypes. It was then replaced by 23-valent polysaccharide vaccine (PPV23; Pneumovax 23) in 1983 to offer wider protective coverage [6]. However, both vaccines failed to protect a major risk group - children (<2 years) due to their inability to induce T cell-dependent immune response in this group [1]. Consequently, PPV23 is not recommended for children under 18 months of age [11]. Pneumococcal conjugate vaccines (PCVs), were then developed by chemically bonding pneumococcal capsular polysaccharides to a carrier protein [12]. PCVs proved successful in inducing T cell - dependent immune response in children (<2 years) and are included in infant immunization schedules. Several PCVs have since been developed for example, (PCV7; Prevenar, Pfizer Inc.), (PCV10; Synflorix, GSK Biologicals), and (PCV13, Prevenar 13, Pfizer Inc.). PCVs are different in terms of serotype composition, the carrier proteins used and in the methods of conjugation applied [13]. In November 2011, Kenya, Pakistan and Madagascar under Global Alliance for Vaccination and Immunization (GAVI) initiative introduced PCV10 in their childhood immunization programs.

The current 10-valent and 13-valent formulations of the pneumococcal conjugate vaccines (PCVs) cover 70% of pneumococcal serotypes, which cause serious pneumococcal disease in children in all geographic regions. Figure1 below shows bar graphs of global distribution of invasive pneumococcal disease (IPD), number of pneumococcal diseases and deaths in children under 5 years of age due to serotypes in existing PCV formulations [14].

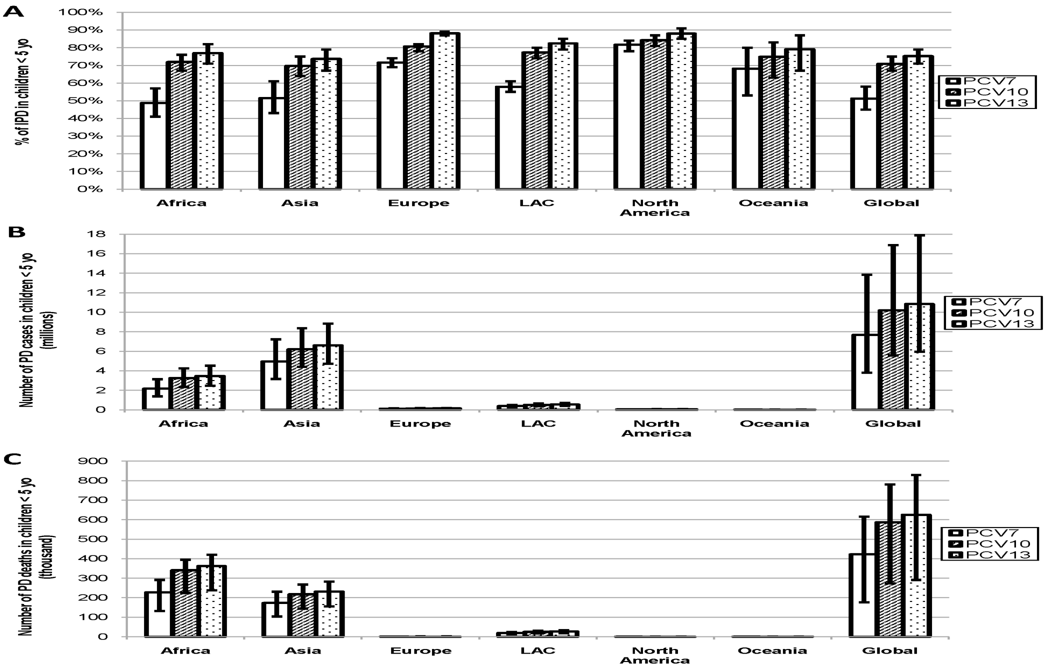


Figure 1. (A) Percentage of IPD cases, (B) Cases of PD, and (C) Mortality rate in children <5yrs of age due to serotypes in existing PCV formulations. Adopted from [14].

The serotypes included in PCV7, PCV10, and PCV13 are found in all major geographic regions in the world. These serotypes account for 55-85% of all IPD in children <5 years in each region. However, morbidity and mortality rates are very high in the developing world. This is largely because of strained vaccination programs by the governments and scarcity of the vaccines due to cost constraints. Africa, Latin American and the Caribbean (LAC) and Asia make up the developing countries (Figure 1. A, B, and C). Europe, Oceania and North American countries represent developed world. Figure 1 assumes serotype 6A/6B cross-protection, globally and by region. Error bars indicate the 95% CIs (A) or uncertainty estimates (B, C). PCV serotypes include: 4, 6B, 9V, 14, 18C, 19F, 23F. PCV10 adds serotypes: 1, 5, and 7F. PCV13 adds serotypes: 3, 6A, and 19A [14].

Despite widespread use of PCVs, reduction of pneumococcal disease is still below the expectation. This is in part, because non-vaccine serotypes may thrive in the absence of their vaccine serotype competitors causing serotype replacement [15]. Furthermore, inclusion of all serotypes in one PCV vaccine is complex and very expensive [1].

Recently, focus has shifted towards the development of protein based serotype-independent pneumococcal vaccines [16]. Recent studies show that pneumococci cells utilize proteins exposed on the surface of their cell wall to subvert host immunity or host protein function. Comprehensive knowledge of the mechanisms employed by pneumococci to achieve this is key to the development of next-generation pneumococcal vaccines [9].

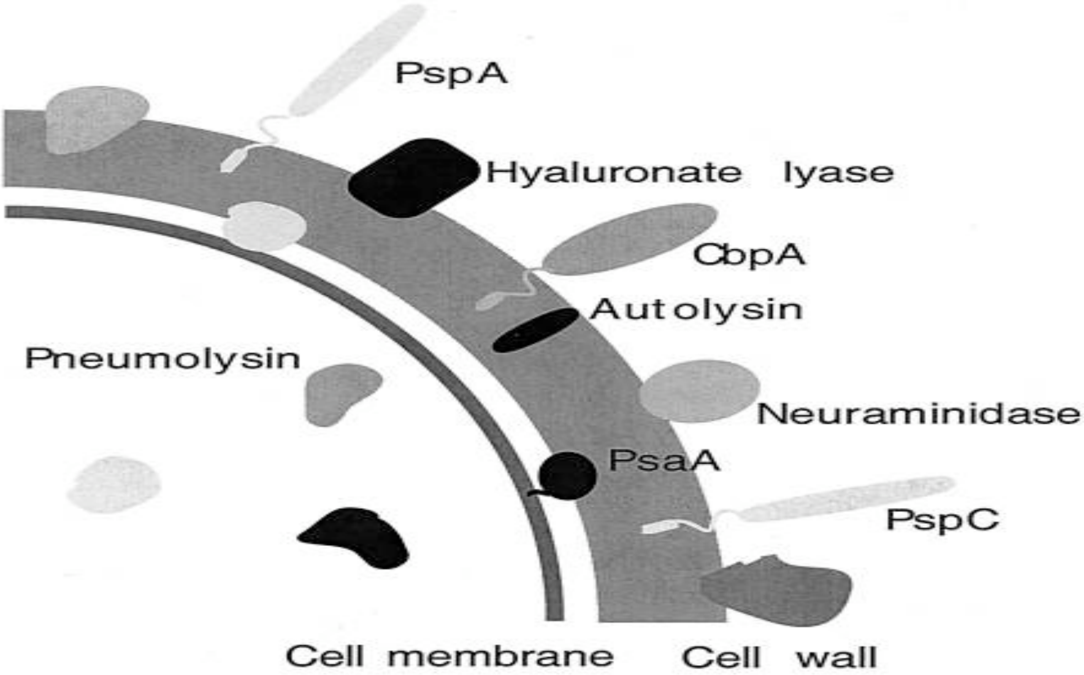


Figure 2. Schematic diagram of the virulent surface proteins of *S. pneumoniae*. Adopted from [17].

This study aims to explore the genetic diversity of choline binding protein A (CbpA), which is expressed on the surface of the cell wall of virtually all pneumococci. This will also expand the understanding on how these genetic variations influence the structural conformation of expressed proteins, which are potential candidates for the development of next-generation vaccines.

1.2 CbpA gene

CbpA gene is present in more than 75% of all *S. pneumoniae* strains [18]. The gene occurs in different allelic forms each generating a slightly different biological structure. CbpA gene is referred to by different names that include, choline-binding protein A (CbpA) [19], *S. pneumoniae* secretory IgA binding protein A (SpsA) [20], pneumococcal surface protein C (PspC) because of its strong molecular and serologic similarities to PspA [18], C3-binding protein A (PbcA) [21], and factor H-binding inhibitor of complement (Hic) [22].

It is known that the polymorphic protein plays an important role in the pathogenesis of *S. pneumoniae* in the following ways:

1. It is required for adhesion to the human nasopharyngeal epithelium and activated lung epithelial and endothelial cells [19]. Experiments have shown that without CbpA, pneumococci have reduced ability to invade nasopharyngeal cells by over 90% compared to the parent strain [23]. CbpA interacts with hpIgR to enable the adherence and invasion of the mucosal cells [16].
2. CbpA initially serves as an adhesin for colonization and then binds complement component C3 either in the epithelial cells or in the blood stream. Interaction with C3 increases adherence capability of pneumococcus [24].
3. It binds to human serum factor H to prevent pneumococcal cells from opsonisation by the components of alternative complement pathway [16].
4. CbpA appears to be preferentially expressed in the transparent phase of colony, which is associated with bacteria in the nasopharynx. The upregulation in the nasopharynx suggests the importance of CbpA in maintaining pneumococci in the ecological niche for nasal carriage [25].

Due to high polymorphism at the CbpA locus, which generates multiple phenotypes, analysis of this loci from different isolates will highlight how sequence diversity correlates with structural variation [26]. The conserved epitope regions of the CbpA protein fragments can be exploited to develop more efficacious serotype-independent vaccines [27].

1.3 Problem statement

Despite the availability of broad-spectrum antibiotics and pneumococcal vaccines, Pneumococcal diseases still pose an enormous burden worldwide. There are more than 90 different pneumococcal serotypes, complicating clinical management and broad coverage vaccination strategies. There is a requirement of novel candidate vaccine antigens that overcomes the shortcomings of the current vaccines for example inadequate serotype coverage, serotype replacement, suboptimal protection and cost constraints in developing countries [28]. Serotype-independent vaccines can solve many of the challenges. Identification of a highly immunogenic surface protein, present in most or all strains, which elicits sufficient immune response, is key to discovery of novel new-generation vaccine. CbpA is a good candidate as it is involved in the first step of the disease carriage in the nasopharynx. Furthermore, its tertiary structure is thought to be conserved in certain regions across all serotypes, offering significant cross reactivity. In this study, we analyze the genetic and structural diversity of the CbpA loci from different pneumococcal genomes, to examine its potential as a next-generation vaccine candidate.

1.4 Research Question

Does the nucleotide sequence and protein structure of the CbpA locus vary across the strains of *S. pneumoniae*?

1.5 Hypothesis

The nucleotide sequence and protein structure of the CbpA locus varies across different strains of *S. pneumoniae*. However, there are conserved CbpA domains that are sufficiently antigenic and that qualify as potential vaccine targets.

1.6 Objectives

1.6.1 General Objective

To determine how the nucleotide sequence and protein structure of the CbpA locus vary across the strains of *S. pneumoniae*.

1.6.2 Specific Objectives

1. To identify variation in the CbpA gene locus across different strains of *S. pneumoniae.*
2. To determine whether these variations significantly alter the structural conformation of the CbpA protein.

1.7 Justification

Novel next-generation vaccine targets are required to develop vaccines that provide broad coverage against pneumococcal infection. Surface-exposed proteins play important roles during the infectious process of *S. pneumoniae*. Furthermore, most of the proteins are common to essentially all pneumococcus serotypes for example, CbpA proteins. Vaccines based on such proteins could potentially offer broad, affordable protection to children and other risk groups worldwide. Currently, identified proteins only give varying levels of protection to certain groups at risk. Understanding how the variability in the CbpA locus affects CbpA protein structure, function and expression can inform its inclusion in novel optimal coverage multi-protein vaccines or single protein vaccines.

Chapter 2

Materials and methods

2.1 Datasets collection

213 complete genome sequences of *S. pneumoniae* were downloaded from the National Centre for Biotechnology Information (NCBI) database – GenBank. The fully annotated genomes were complete and were widely sampled to give a global representation, that is, the various strains were originally sampled from patients in different parts of the world. The respective accession numbers of the strains in the Genbank were used to download the genomes, which were then saved both as Genbank files and fasta files. The list of the *S. pneumoniae* accession numbers is shown in Appendix A.

The CbpA gene sequences and the corresponding amino acid sequences were then extracted from each downloaded genome and saved as fasta files. Each fully annotated genome had links to both the CbpA gene sequence and the translation (protein) sequence. See Appendix B showing the 211 CbpA proteins, their positions in the genome and lengths.

The extracted CbpA files were concatenated into 2 distinct multi-fasta files of nucleotide sequences and amino acid sequences using UNIX shell script command - “cat”. See Appendix C.

2.2 Sequence alignment

The CbpA multi-fasta files were each aligned from the command line using Multiple SequenCe alignment by Log-Expectation (MUSCLE) version 3.8.31 [29]. The aligned nucleic and protein sequence files were saved in fasta format. MUSCLE was chosen over other alignment softwares due to it's precise nature of localized sequence alignment [30].

2.3 Phylogenetic tree Construction

Maximum likelihood phylogenies were estimated using PhyML software [31]. The generated phylogenetic trees were then used to infer evolutionary relationships between CbpA genes and proteins [32]. PhyML implements the maximum likelihood approach of finding the topology and branch lengths of the best tree from the provided sequences [31]. The tree building was done under stringent parameters including GTR (General Time- Reversible) model, and 100 replicates for bootstrapping.

2.4 Heat map generation

Sequence diversity heat map was generated using the heat map function of R and Bioconductor. The packages that contain the codes to make heat map were installed from the Bioconductor website (http://www.bioconductor.org/packages). Heat maps allow visualization of the provided data by way of representing values contained in the matrix as colors in a graph [33]. Perl scripts were first used to prepare both the CbpA protein and nucleic acid dataset.

2.5 CbpA Secondary structure determination

Each CbpA amino acid sequence was submitted to the RaptorX server (http://raptorx.uchicago.edu/StructurePrediction/predict/) for 3-D structure prediction and visualization. RaptorX generated full-length sequence structures with all the possible domains. Out of all the predicted models, the best template (domain) for each sequence was chosen based on the P-value, Score and global distance test. The 3-D structure of the best domain of each CbpA sequence was downloaded and saved for further analysis. These domains were used to determine conserved regions and their role in pneumococcal pathogenicity. RaptorX was chosen for its precise secondary and tertiary protein structure prediction capabilities producing high quality structural models for many sequences with just a few templates [34].

2.6 Multi- locus sequence typing of *S. pneumoniae*

The fasta formats of the genome sequences were all submitted to the centre for genomic epidemiology (CGE) server (www.genomicepidemiology.org) for multi-locus sequence typing (MLST). The allelic profile (STs) of each submitted genome were identified and recorded along with the house keeping allele numbers. MLST is a common method for the analysis of relationships among strains of clinically relevant microbial species. The sequence type (ST) is a unique identifier (numerical allelic profile) assigned to each according to the sequence of the seven (for pneumococcus) housekeeping genes. How two strains are related to each other can be inferred from the differences between their sequence types (ST) [35]. Lineage assignment was achieved by eBURST V3 (eburst.mlst.net). eBURST is an algorithm that explores patterns of evolutionary descent by dividing MLST data set into groups of related isolates and clonal complexes. The ancestral genotype of each clonal complex is predicted and bootstrap support for the assignment computed. For all isolates in each clonal complex from the predicted ancestor, parsimonious patterns of descent are then displayed [36].

2.7 Transmembrane protein test

CbpA proteins were submitted to the TransMembrane Hidden Markov Model (TMHMM) server (http://www.cbs.dtu.dk/services/TMHMM/) to determine whether indeed CbpA is located on the surface of pneumococcus. TMHMM can predict to 97% - 98% accuracy whether a protein is transmembrane domain using hidden markov models [37].

2.8 Immunogenecity

The CbpA proteins were then tested for immunogenicity using VaxiJen v2.0 software (www.Ddg-pharmfac.net/vaxijen) - the first alignment-free bioinformatics tool for the in silico identification of antigens. The score of antigenicity for each CbpA protein was determined based on the physicochemical properties of each protein.

2.9 Tajima's D test on CbpA loci

The DNA polymorphisms across the CbpA loci was used to determine selection neutrality in all the 211 CbpA nucleic acid sequences [38].Tajima's D tests for selection by neutrality based on allele frequencies in a sample size [39].

Chapter 3

Results

3.1 *S. pneumoniae* genome and CbpA sequences

Out of the 213 genome sequences downloaded from GenBank, 211 CbpA amino acid sequences and 211 CbpA nucleic acid sequences were retrieved. This represented 99% presence and expression, which is an important fact in support of wide availability of the protein in pneumococcus. The two genomes that did not have CbpA gene and protein were genome numbers 76 and 151.

3.2 Multiple sequence alignments of CbpA

CbpA gene and protein multi-fasta input files are shown in Appendix D and Appendix E respectively below. Multiple sequence alignments of both CbpA nucleic acid sequences and CbpA amino acid sequences are shown in Appendix F and G respectively. The multiple alignments showed widespread divergence in both the datasets. INDELS were clearly observed in both the alignments. However, there were regions that are consistently conserved for all the sequences, especially in the first 100 residues located upstream.

3.3 CbpA phylogenetic trees and heat maps

Phylogenetic trees and the corresponding heat maps for both the CbpA amino acid sequences and nucleotide sequences (Figures 3 and 4) further shows the high polymorphism at both nucleic acid and amino acid level. This is consistent with previous observations that the surface protein (CbpA) is highly diverse both genetically and phenotypically. Due to constraints of space, the taxon labels and bootstrap values of both the phylogenetic trees could not be included. However, in the protein phylogeny for example, the most distant ancestral proteins were CbpA protein with the labels 121 on the top first taxon and 13 different CbpA proteins on the bottom first 13 taxa comprising of CbpAs labeled 185, 24, 100, 161, 210, 130, 79, 75, 166, 70, 94, 209, and 193. The most recently evolved CbpA gene loci are 192 and 168 represented by the furthest taxa from the root. The colour keys on the heat maps represent the residues. Similar residues across the datasets have the same color allowing visualization of conserved regions. The yellow and white patches represent insertions and deletions (INDELs) in Figures 3 and 4 respectively. The alignments of both amino acid and DNA residues are largely inconsistent across the entire sequence dataset. However, there was clear alignment of some amino acid residues and nucleic acids visible from the same colors at the same positions. This allowed for identification of broadly, conserved CbpA domains.

The statistical test of selection neutrality (Tajima's test) across the CbpA loci showed a significant signature for balancing selection (Tajima's D = 0.59). It is possible that the observed polymorphism at the CbpA locus is an evolutionary strategy of the pneumococcus to evade the host's defense mechanisms and Multi- locus sequence typing of the *S. pneumoniae* genomes adapt for colonization and invasion of different niches within the host.

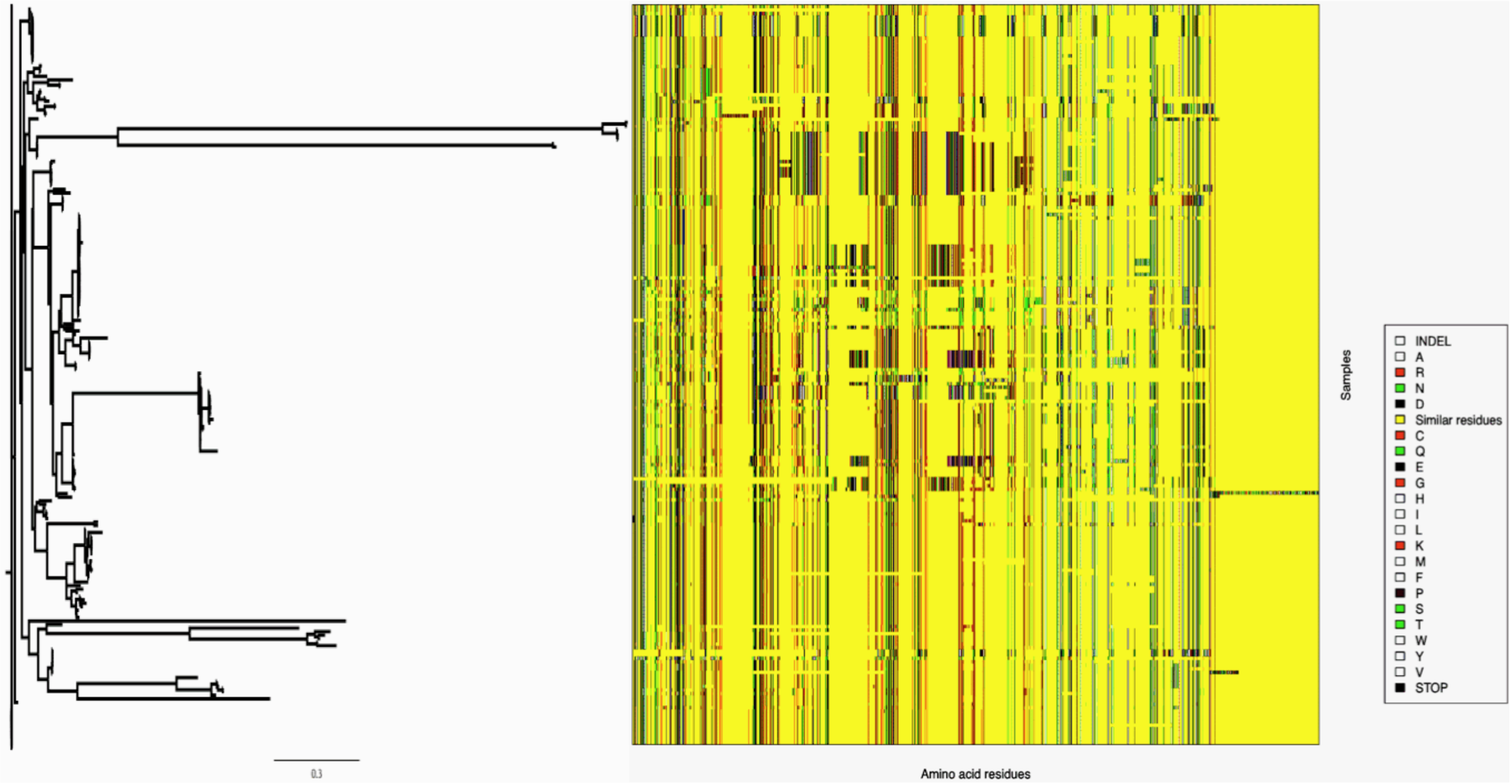


Figure 3. Phylogenetic tree of *S. pneumoniae* CbpA amino acid residues and their corresponding heat map.

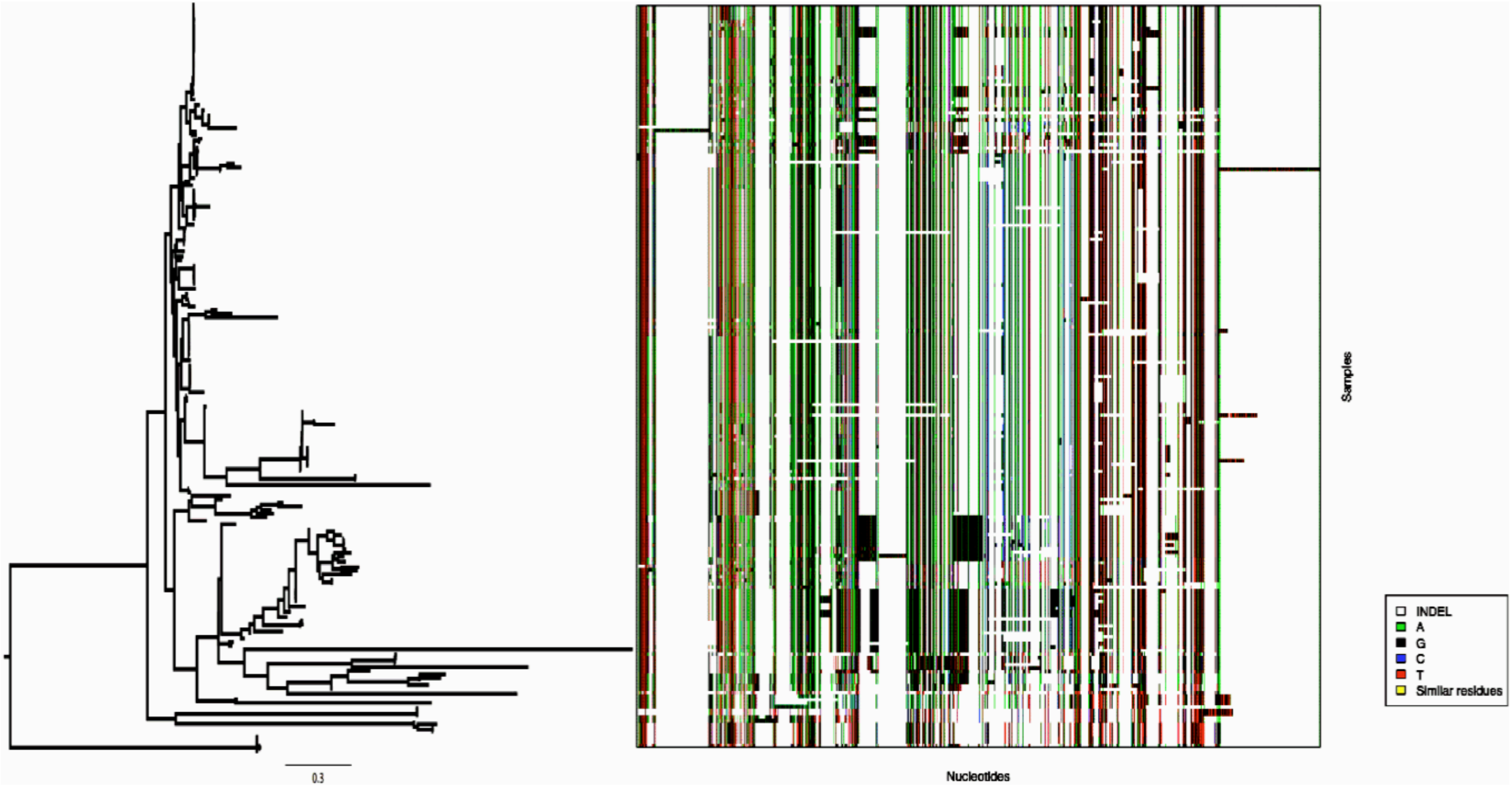


Figure 4. Phylogenetic tree of *S. pneumoniae* CbpA nucleic acid residues and their corresponding heat map.

In figures 3 and 4 above, the phylogenetic trees were constructed using PhyML (Maximum likelihood approach at 100 bootstraps) and the heat maps generated in R and bioconductor.

3.4 CbpA full sequence structures and domains

RaptorX server was used to predict the tertiary structures of the CbpA proteins and to determine whether there are conserved regions across the entire data set. Out of the 211 complete 3-D structures of each CbpA sequences returned, only the ones that predicted choline-binding proteins (CBPs) as corresponding best domains were considered in Appendix J. Most of the structures were different from each other but there were those that appeared in high frequency. The best domain template for each amino acid sequence was identified (Appendix K). The predicted best domains were then counterchecked in the Research Collaboration for Structural Bioinformatics (RCSB) website (www.rcsb.org) to filter out irrelevant domains. There were 30 different proteins predicted to have the best domains (see Appendix L), 7 of which were found to belong to the choline binding proteins (CBPs) of *Streptococcus pneumoniae*, namely:

1. 2bib:A : Modular teichoic acid phosphorylcholine esterase (CBPE)
2. 2pms:C: Lactoferrin-binding domain of pspA
3. 2vyuA: Choline binding protein F (CbpF)
4. 2m6u:A: Choline binding protein A (CbpA/CbpAN)
5. 1w9r:A: Choline binding protein A (domain R2)
6. 3hia:A: Choline binding protein A
7. 4k12:B: Choline binding protein A

It was observed that the most common domains matched previously characterized pneumococcal templates in the protein data bank (PDB) for the modular teichoic acid phosphorylcholine esterase Pce (2bib:A) and the CbpA R2 domain (1w9r:A).

CbpA R2 domain (1w9r:A) and 2bib:A made up 83.3% of the best pneumococcal protein domains, implying that the two domains are highly conserved and are relevant to CbpA from almost all known pneumococcal strains. The 3-D structures of 2bib:A and 1w9r:A are shown in figures 5 and 6 below:

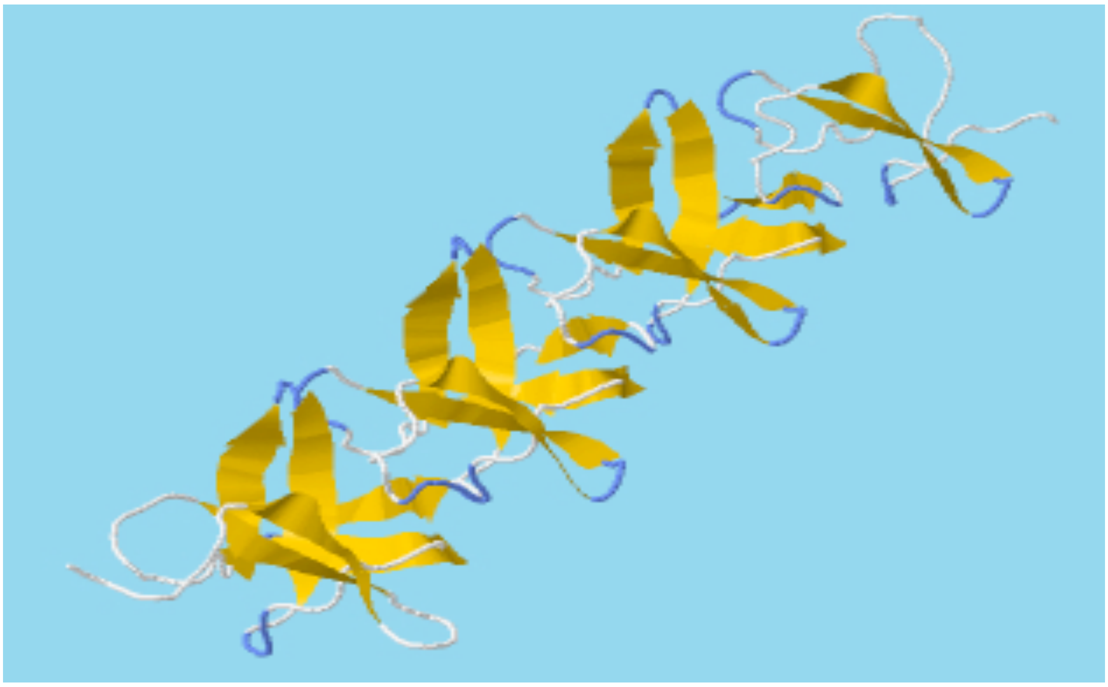


Figure 5. 3-D structure of 2bib:A domain. Adopted from raptorx.uchicago.edu.

The arrows show the direction of the beta- sheets beginning from the N- terminus to the C- terminus. The catalytic site is between the beta-sheets in the N- terminal of the structure [40].

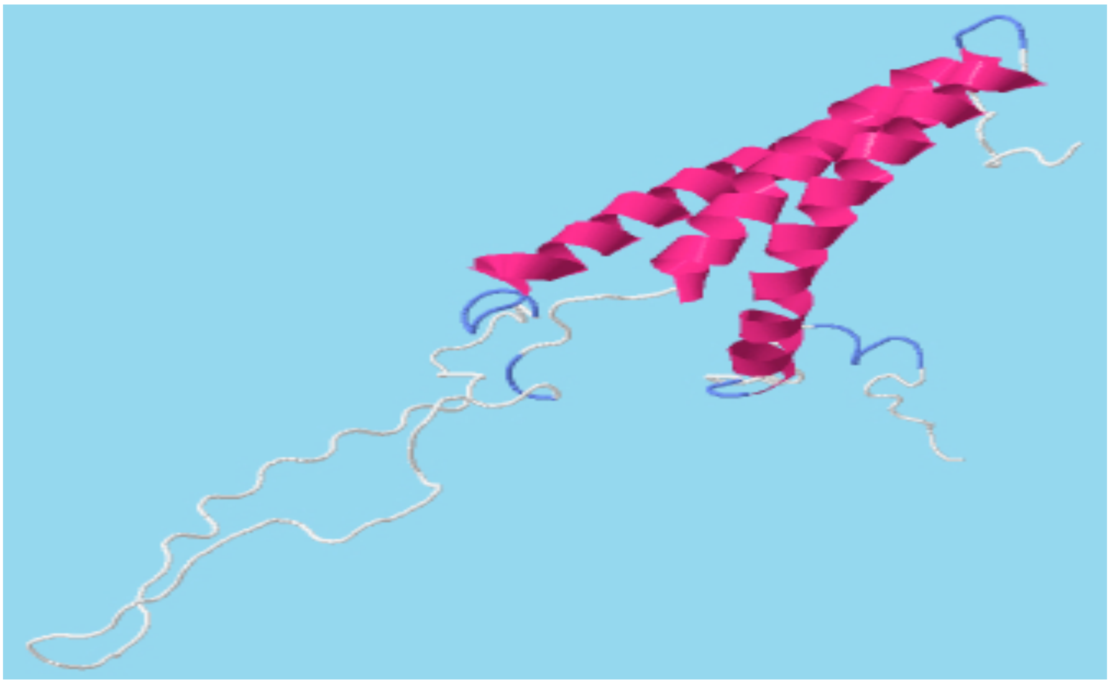


Figure 6. 3-D structure of 1w9r:A domain. Adopted from raptorx.uchicago.edu

The 3-D structure of 1w9r:A above showed 3 alpha- helix ribbons. The conserved residues within the tyrosine fork (not visible in the diagram) - next to the loop between the first and second helix are likely to bind Ig-like domains of pIgR [41].

3.5 Multi-locus sequence typing of *S. pneumoniae* genomes

Multi-locus sequence typing results (MLST) (Appendix H) of the 213 genomes submitted produced 114 (53.5%) distinct pneumococcus genotypes based on their unique sequence types (ST). This represents the genotypic diversity in the global pneumococcal data set used. ST-156, ST-180 and ST-199 were the most abundant, each having a frequency of 8 (3.8%). ST- 13, ST-81, and ST-376 had a frequency of 7 (3.3%) followed by ST-191 and ST-320 with frequency of 6(2.8%). ST- 37, ST-62, ST-236, ST-433 and ST-695 had frequency of 4 (1.9%). ST-63, ST-242, ST-384, ST-595 and ST-651 each had frequency of 3 (1.4%). ST-43, ST-53, ST-90, ST-205, ST-

271, ST-338, ST-507, ST-558, ST-1292, ST-1296, ST-1339, ST-1536, ST-2150, ST-2344, ST-2705 and ST-3039 appeared twice each (1%). The rest of the STs appeared only once. There were 8 strains of previously unidentified genotypes (3.8%) as they did not return any allele number in the pneumococcal MLST website. These non-identities can be submitted to the curator of the pneumococcal MLST database for assignment of new allele numbers.

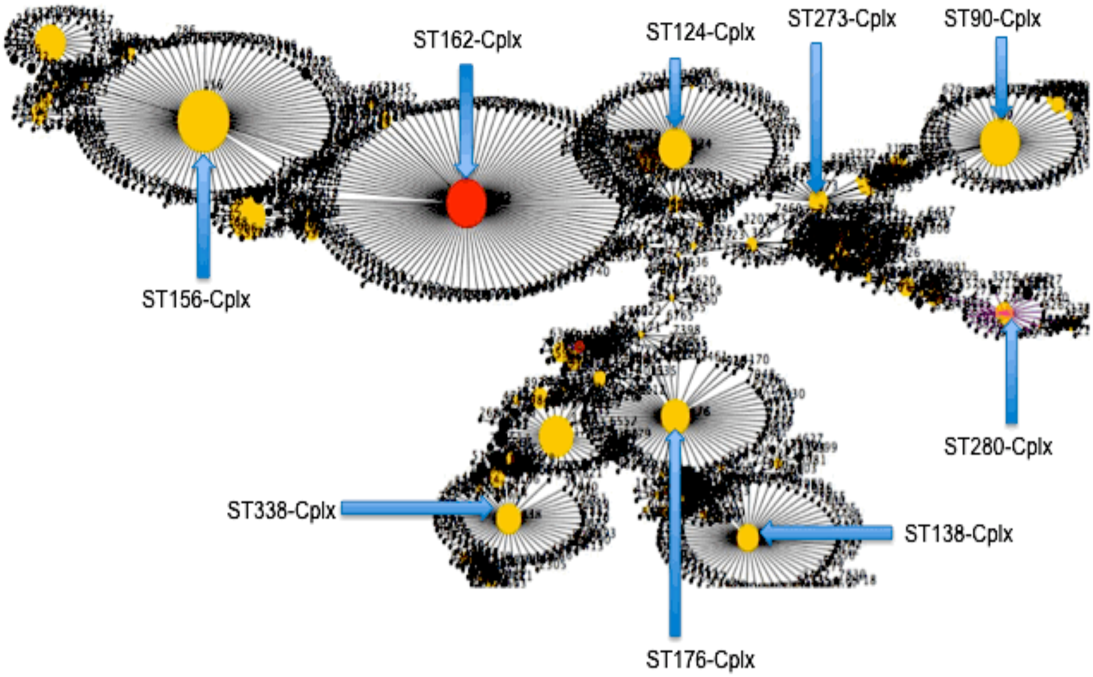


Figure 7. eBURST population clusters of *S. pneumoniae* (213 isolates) showing progressive outward growth from the founding complex (Cplx) – ST162-Cplx.

Figure 7 above shows more than 12 clonal complexes (CCs) varying in size depending on the number of STs linked to the group. The centrally placed ST in each clonal complex represents the founding genotype (the primary founder). ST162 represent the primary founder of all the clonal complexes and belongs to the largest clonal complex (ST162-Cplx). ST156 and ST124 are single locus variants (SLVs) of ST162 and have become successful and diversified to produce their own SLVs.

Therefore, ST156 and ST124 are subgroup founders of ST156-Cplx and ST124-Cplx clonal complexes. ST124 have SLVs that formed few SLVs of their own but later had descendants that were more successful and diversified into large clonal complexes such as ST273-Cplx, ST176-Cplx and others as shown in figure 7 above.

3.6 Cellular locations of CbpA best domain structures.

Membrane topology of 1w9r:A and 2bib:A were determined using Transmembrane Hidden Markov Model (TMHMM) server (version 2.0) (www.cbs.dtu.dk/services/TMHMM). Figures 8 and 9 shows the posterior probabilities plots of 1w9r:A and 2bib:A respectively. The x-axis represents the sequence length of the protein helix and the y-axis represents the posterior probabilities of protein helix locations. The posterior probabilities of the protein helix being inside (blue line), outside (pink line) or transmembranous - TM (red line) is calculated by summing up probabilities of each model state found by forward – backward algorithm [37]. In both the plots, the thick pink line is between 1 and 1.2 (the N-best prediction) indicating total extracellular localization, which would allow access of the host's immune system to both 2bib:A and 1w9r:A.

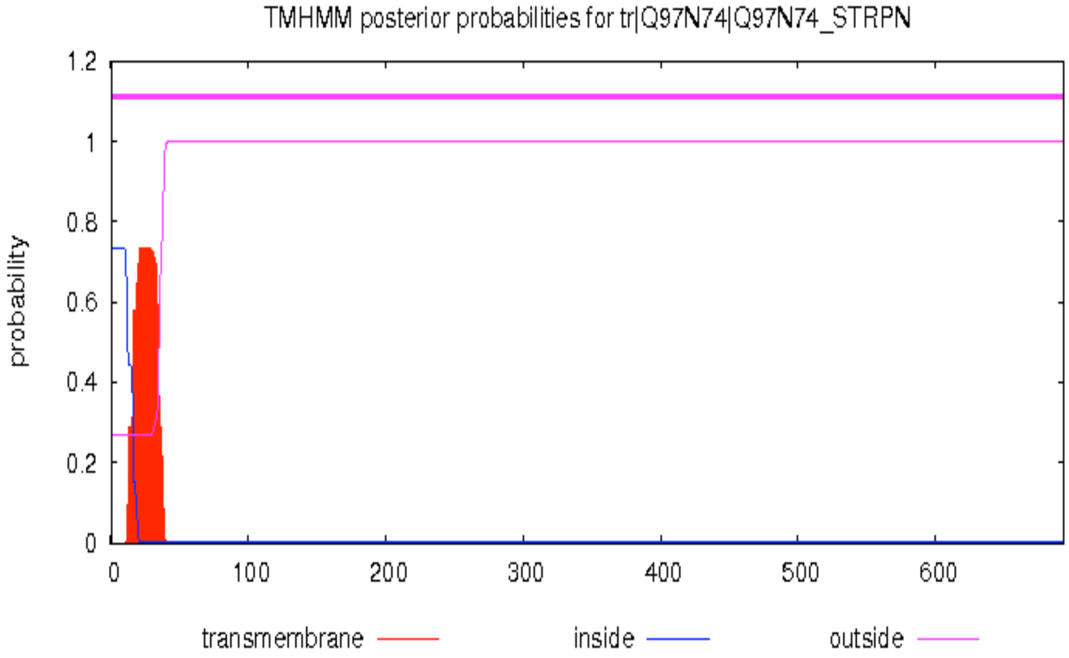


Figure 8. Cellular locations of 1w9r:A.

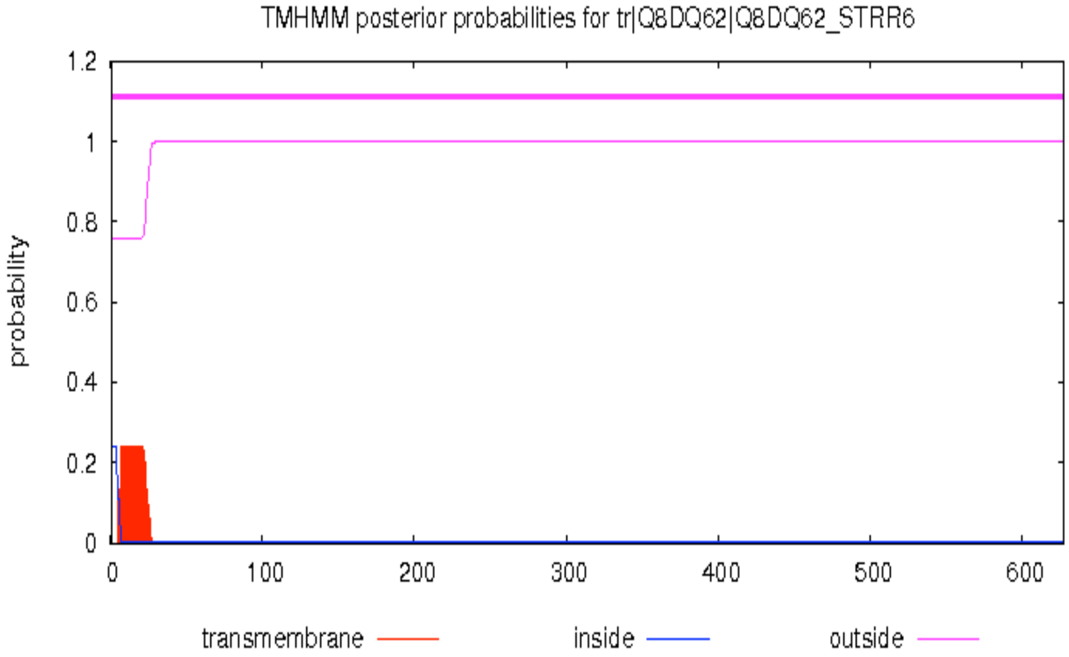


Figure 9. Cellular location of 2bib:A.

3.7 Antigenicity of CbpA protein

The CbpA amino acid sequences were analyzed using VaxiJen antigen prediction server (VaxiJen v 2.0) (www.Ddg-pharmfac.net/vaxijen) and the results returned as shown in Appendix I. The default parameters (Threshold = 0.4, ACC Output) were used against all the 211 CbpA protein sequences. Both 2bib:A and 1w9r:A domain templates passed the 0.4 threshold hence, qualifying them as probable protective CbpA epitopes.

Given the prevalence, accessibility and immunogenicity of the two domains across most of the strains of *S. pneumoniae*, the respective amino acid sequences qualify as potential candidates for a novel vaccine against most of the strains.

Chapter 4

Discussion and Conclusion

Treatment of pneumococcal infections is still a big challenge worldwide. Pneumococci continue to exhibit increasing resistance to antibiotics and current vaccine formulations are not able to fully protect global populations. Development of vaccines that would protect against over 90 serotypes is complicated and unaffordable, especially to the developing countries. Novel vaccine technologies and strategies offer hope of developing serotype-independent vaccines that are cost-effective and protective of both children and adults [42].

One such strategy involves the use of one or more pneumococcal cell surface proteins involved in pathogenesis. The suitability of the protein as candidate for vaccine depends on its prevalence across the strains, antigenicity and absence of autoimmunity induction. The latter ensures that the eventual vaccine is safe for use in humans. The antigenic subunit of the protein needs to be conserved across the strains to develop serotype-independent vaccine. However, no single pneumococcal protein has been able to elicit protection comparable to protein-conjugate vaccines (PCV) [43]. Experiments have shown that a mixture of the proteins could have significant protection such as was observed with choline-binding proteins (CBPs) in mice [44]. CbpA was presumed to be the primary protective antigen in this particular mixture.

The high variability of CbpA (also designated as PspC, SpsA, PbcA, and Hic) suggests its importance as a natural target for host defense. It has a number of biologic functions that establish colonization within different niches and modulate the host immune response. Firstly, CbpA is important in adherence and colonization of the nasopharynx, which is the carrier of pneumococcus in humans. It has been reported that *S. pneumoniae* mutants lacking CbpA were unable to colonize the nasopharynx of a mouse model nor were they able to invade and multiply in the lungs [25]. In addition, [19] demonstrated that CbpA-negative *S. pneumoniae* mutants had reduced ability to bind cytokine-activated human lung epithelial cells and endothelial cells compared to the parent strain as well as 100 fold reduction in nasopharyngeal carriage ability in an infant rat model. Furthermore, CbpA interacts with human pIgR to facilitate bloodstream and blood brain barrier invasion.

In this study, the diversity of CbpA is confirmed by the multiple differences observed in both the amino acid and nucleic acid sequence alignments, the heat maps and the phylogenetic trees. However, conserved tertiary domains do exist as has been seen in amino acid sequences. Conserved sequences can be seen at positions 11-24, 34-38, and 41-57. It is possible that the conserved regions are part of the prominent domains 2bib:A and lw9r and that they are essential in the roles of CbpA protein e.g. receptor binding.

The pneumococcal Pce phosphorylcholine esterase (2bib:A) is a key virulent factor, as it is known to modify the distribution and the content of phophorylcholine (PC) residues present in human molecules such as Platelet-activating factor (PAF) and other phospholipids. [45]. The R2 domain binds to the host's polymeric immunoglobulin receptor (PigR) facilitating pneumococcal intracellular translocation at the mucosal epithelium and BBR, allowing bloodstream and cerebrospinal fluid invasion respectively.

In conclusion, new generation serotype-independent vaccines against pneumococcal disease will be protein-based. However, the formulation will likely require multiple antigenic surface proteins. CbpA is a major candidate given its role in pathological processes in the nasopharynx, ear, lung, blood and brain [46]. In this study, it was observed that both nucleic acid sequences and amino acid sequences of CbpA were highly divergent and resulted in different structural conformations as seen in (appendix J). This could be an evolutionary strategy of pneumococcus to evade various human immune responses and also to colonize different niches. The diversity observed in CbpA protein is a big challenge but that can be surmounted by more work in identification of conserved epitopic regions as seen here with 2bib:A and 1w9r:A. Future work might require that the conserved residues of 2bib:A (modular teichoic acid phosphorylcholine esterase) and 1w9r:A (CbpA R2 domain) be studied further to understand their interactions with host immune system. This would inform on how to design the next generation CbpA- fusion vaccine and optimize safety and efficacy.

References

1. Gladstone R A, Jefferies JM, Faust SN, Clarke SC. Pneumococcal 13-valent conjugate vaccine for the prevention of invasive pneumococcal disease in children and adults. Expert Rev Vaccines. 2012;11: 889–902. doi:10.1586/erv.12.68
2. Jiang Y, Gauthier A, Annemans L, van der Linden M, Nicolas-Spony L, Bresse X. Cost-effectiveness of vaccinating adults with the 23-valent pneumococcal polysaccharide vaccine (PPV23) in Germany. Expert Rev Pharmacoecon Outcomes Res. 2012;12: 645–60. doi:10.1586/erp.12.54
3. Frolet C, Beniazza M, Roux L, Gallet B, Noirclerc-Savoye M, Vernet T, et al. New adhesin functions of surface-exposed pneumococcal proteins. BMC Microbiol. 2010;10: 190. doi:10.1186/1471-2180-10-190
4. Mitchell R, Trück J, Pollard AJ. Use of the 13-valent pneumococcal conjugate vaccine in children and adolescents aged 6 - 17 years. Expert Opin Biol Ther. 2013;13: 1451–65. doi:10.1517/14712598.2013.824419
5. Simell B, Auranen K, Käyhty H, Goldblatt D, Dagan R, O'Brien KL. The fundamental link between pneumococcal carriage and disease. Expert Rev Vaccines. 2012;11: 841–55. doi:10.1586/erv.12.53
6. Cordonnier C, Averbuch D, Maury S, Engelhard D. Pneumococcal immunization in immunocompromised hosts: where do we stand? Expert Rev Vaccines. 2014;13: 59–74. doi:10.1586/14760584.2014.859990
7. Zangeneh TT, Baracco G, Al-Tawfiq J a. Impact of conjugate pneumococcal vaccines on the changing epidemiology of pneumococcal infections. Expert Rev Vaccines. 2011;10: 345–53. doi:10.1586/erv.11.1

8. Farkouh R A, Klok RM, Postma MJ, Roberts CS, Strutton DR. Cost-effectiveness models of pneumococcal conjugate vaccines: variability and impact of modeling assumptions. Expert Rev Vaccines. 2012;11: 1235–47. doi:10.1586/erv.12.99

9. Bergmann S, Hammerschmidt S. Versatility of pneumococcal surface proteins. Microbiology. 2006;152: 295–303. doi:10.1099/mic.0.28610-0

10. Huang Y-L, Wu C-Y. Carbohydrate-based vaccines: challenges and opportunities. Expert Rev Vaccines. 2010;9: 1257–74. doi:10.1586/erv.10.120

11. O'Grady K-AF, Chang AB, Grimwood K. Vaccines for children and adults with chronic lung disease: efficacy against acute exacerbations. Expert Rev Respir Med. 2014;8: 43–55. doi:10.1586/17476348.2014.852960

12. Pace D. Glycoconjugate vaccines. Expert Opin Biol Ther. 2013;13: 11–33. doi:10.1517/14712598.2012.725718

13. Poolman JT, Peeters CC a M, van den Dobbelsteen GPJM. The history of pneumococcal conjugate vaccine development: dose selection. Expert Rev Vaccines. 2013;12: 1379–94. doi:10.1586/14760584.2013.852475

14. Johnson HL, Deloria-Knoll M, Levine OS, Stoszek SK, Freimanis Hance L, Reithinger R, et al. Systematic evaluation of serotypes causing invasive pneumococcal disease among children under five: the pneumococcal global serotype project. PLoS Med. 2010;7. doi:10.1371/journal.pmed.1000348

15. Reinert RR, Paradiso P, Fritzell B. Advances in pneumococcal vaccines: the 13-valent pneumococcal conjugate vaccine received market authorization in Europe. Expert Rev Vaccines. 2010;9: 229–36. doi:10.1586/erv.10.6

16. Tai SS. Streptococcus pneumoniae Protein Vaccine Candidates : Properties , Activities and Animal Studies. 2006; 139–153. doi:10.1080/10408410600822942

17. Jedrzejas MJ. Pneumococcal Virulence Factors : Structure and Function. 2001;65: 187–207. doi:10.1128/MMBR.65.2.187

18. Brooks-walter A, Briles DE, Susan K, Hollingshead SK. The pspC Gene of Streptococcus pneumoniae Encodes a Polymorphic Protein , PspC , Which Elicits Cross-Reactive Antibodies to PspA and Provides Immunity to Pneumococcal Bacteremia The pspC Gene of Streptococcus pneumoniae Encodes a Polymorphic Protein , Psp. 1999;

19. Rosenow C, Ryan P, Weiser JN, Johnson S, Fontan P, Ortqvist a, et al. Contribution of novel choline-binding proteins to adherence, colonization and immunogenicity of Streptococcus pneumoniae. Mol Microbiol. 1997;25: 819–29. Available: http://www.ncbi.nlm.nih.gov/pubmed/9364908

20. Hammerschmidt S, Talay SR, Brandtzaeg P, Chhatwal GS. SpsA, a novel pneumococcal surface protein with specific binding to secretory immunoglobulin A and secretory component. Mol Microbiol. 1997;25: 1113–1124. doi:10.1046/j.1365-2958.1997.5391899.x

21. Qi C, Finkel D, Hostetter MK. Novel purification scheme and functions for a C3-binding protein from Streptococcus pneumoniae. Biochemistry. 2000;39: 5450–5457. doi:10.1021/bi992157d

22. Janulczyk R, Iannelli F, Sjoholm a G, Pozzi G, Bjorck L. Hic, a novel surface protein of Streptococcus pneumoniae that interferes with complement function. J Biol Chem. 2000;275: 37257–63. doi:10.1074/jbc.M004572200

23. LeMessurier KS, Ogunniyi AD, Paton JC. Differential expression of key pneumococcal virulence genes in vivo. Microbiology. 2006;152: 305–11. doi:10.1099/mic.0.28438-0

24. Smith BL, Hostetter MK. C3 as substrate for adhesion of Streptococcus pneumoniae. J Infect Dis. 2000;182: 497–508. doi:10.1086/315722

25. Balachandran P, Brooks-walter A, Virolainen-julkunen A, Hollingshead SK, Briles DE. Role of Pneumococcal Surface Protein C in Nasopharyngeal Carriage and Pneumonia and Its Ability To Elicit Protection against Carriage of Streptococcus pneumoniae. 2002;70: 2526–2534. doi:10.1128/IAI.70.5.2526

26. Iannelli F, Oggioni MR, Pozzi G. Allelic variation in the highly polymorphic locus pspC of Streptococcus pneumoniae. Gene. 2002;284: 63–71. doi:10.1016/S0378-1119(01)00896-4

27. Briles DE, Hollingshead SK, Uab T. Patentes Pneumococcal surface protein C (PspC), epitopic regions and strain selection thereof , and uses therefor. 2006;

28. Dinleyici EC. Current status of pneumococcal vaccines: lessons to be learned and new insights. Expert Rev Vaccines. 2010;9: 1017–22. doi:10.1586/erv.10.86

29. Edgar RC. MUSCLE : a multiple sequence alignment method with reduced time and space complexity. 2004;19: 1–19. doi:10.1186/1471-2105-5-113

30. Edgar RC. MUSCLE User Guide. 2010;32: 1–17.

31. Guindon S, Gascuel O. A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. Syst Biol. 2003;52: 696–704. doi:10.1080/10635150390235520

32. Rizzo J, Rouchka EC. Review of Phylogenetic Tree Construction Review of Phylogenetic Tree Construction. 2007;

33. Schroeder MP, Gonzalez-perez A, Lopez-bigas N. Visualizing multidimensional cancer genomics data. 2013; 1–13.

34. Källberg M, Wang H, Wang S, Peng J, Wang Z, Lu H, et al. Template-based protein structure modeling using the RaptorX web server. Nat Protoc. 2012;7: 1511–22. doi:10.1038/nprot.2012.085

35. Francisco AP, Bugalho M, Ramirez M, Carriço J a. Global optimal eBURST analysis of multilocus typing data using a graphic matroid approach. BMC Bioinformatics. 2009;10: 152. doi:10.1186/1471-2105-10-152

36. Ej F, Bc L, Dm A, Wp H, Bg S. eBURST : inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data . PubMed Commons Supplemental Content Full text links. 2004;

37. Krogh a, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. 2001;305: 567–80. doi:10.1006/jmbi.2000.4315

38. Tajima F. Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. 1989;595: 585–595.

39. Korneliussen TS, Moltke I, Albrechtsen A, Nielsen R. Calculation of Tajima ' s D and other neutrality test statistics from low depth next-generation sequencing data. 2013;

40. Hermoso J A, Lagartera L, González A, Stelter M, García P, Martínez-Ripoll M, et al. Insights into pneumococcal pathogenesis from the crystal structure of the modular teichoic acid phosphorylcholine esterase Pce. Nat Struct Mol Biol. 2005;12: 533–8. doi:10.1038/nsmb940

41. Luo R, Mann B, Lewis WS, Rowe A, Heath R, Stewart ML, et al. Solution structure of choline binding protein A, the major adhesin of Streptococcus pneumoniae. EMBO J. 2005;24: 34–43. doi:10.1038/sj.emboj.7600490

42. Aviv T. 7 th INTERNATIONAL SYMPOSIUM ON PNEUMOCOCCI AND PNEUMOCOCCAL DISEASES SPECIAL REPORT. 2010;

43. Ogunniyi AD, Grabowicz M, Briles DE, Cook J, Paton JC. Development of a vaccine against invasive pneumococcal disease based on combinations of virulence proteins of Streptococcus pneumoniae. Infect Immun. 2007;75: 350–7. doi:10.1128/IAI.01103-06

44. Swiatlo E, Ware D. Novel vaccine strategies with protein antigens of Streptococcus pneumoniae. FEMS Immunol Med Microbiol. 2003;38: 1–7. doi:10.1016/S0928-8244(03)00146-9

45. Hermoso JA, Lagartera L, Gonza ANA, Garci LA, Mene M, A PG. contains a metal binuclear center that is essential for substrate binding and catalysis. 2005; 3013–3024. doi:10.1110/ps.051575005.this

46. Mann B, Thornton J, Heath R, Wade KR, Tweten RK, Gao G, et al. Broadly protective protein-based pneumococcal vaccine composed of pneumolysin toxoid-CbpA peptide recombinant fusion protein. J Infect Dis. 2014;209: 1116–25. doi:10.1093/infdis/jit502

Appendices

Appendix A. S.pneumoniae strains accession numbers in the dataset.

Seq_ID	Seq_ID	Seq_ID	Seq_ID	Seq_ID
2061376	cdc1873_00	ga14688	ga44378	ga60190
2061617	cdc3059_06	ga14798	ga44386	ga62331
2070005	cgsp14	ga16121	ga44452	ga62681
2070035	d39	ga16242	ga44500	hungary19_6
2070108	england14_9	ga16531	ga44511	jja
2070109	eu_np01	ga16833	ga47033	mlv_016
2070335	eu_np02	ga17227	ga47179	netherlands15b_37
2070425	eu_np03	ga17301	ga47210	northcarolina6a_23
2070531	eu_np04	ga17328	ga47281	np070
2070768	eu_np05	ga17371	ga47283	np112
2071004	g54	ga17457	ga47360	np127
2071247	ga02254	ga17484	ga47368	np141
2072047	ga02270	ga17545	ga47373	np170
2080076	ga02506	ga17570	ga47388	oxc141
2080913	ga02714	ga17719	ga47439	p1031
2081074	ga04175	ga17971	ga47461	r6
2081685	ga04216	ga18068	ga47502	sp11_bs70
2082170	ga04375	ga18523	ga47522	sp14_bs292
2082239	ga04672	ga19077	ga47562	sp14_bs69
2090008	ga05245	ga19101	ga47597	sp18_bs74
3063_00	ga05248	ga19451	ga47628	sp195
4027_06	ga05578	ga19690	ga47688	sp19_bs75
4075_00	ga06083	ga19923	ga47751	sp23_bs72
459_5	ga07228	ga19998	ga47760	sp3_bs71
5185_06	ga07643	ga40028	ga47778	sp6_bs73
5652_06	ga07914	ga40183	ga47794	sp9_bs68
5787_06	ga08780	ga40410	ga47901	sp_bs293
670_6b	ga08825	ga40563	ga47976	spar27
6735_05	ga11184	ga41277	ga49138	spar48
6901_05	ga11304	ga41301	ga49194	spar55
6963_05	ga11426	ga41317	ga49447	spar95
70585	ga11663	ga41410	ga49542	spn034156
7286_06	ga11856	ga41437	ga52306	spn034183
7533_05	ga13224	ga41538	ga52612	spn994039
7879_04	ga13338	ga41565	ga54354	spna45
8190_05	ga13430	ga41688	ga54644	st556
ap200	ga13455	ga43257	ga56113	sv35
bs397	ga13494	ga43264	ga56348	sv36
bs455	ga13499	ga43265	ga58581	taiwan19f_14
bs457	ga13637	ga43380	ga58771	tch8431_19a
bs458	ga13723	ga44128	ga58981	tigr4
cdc0288_04	ga13856	ga44194	ga60080	
cdc1087_00	ga14373	ga44288	ga60132	

Appendix B: CbpA proteins positions and size.

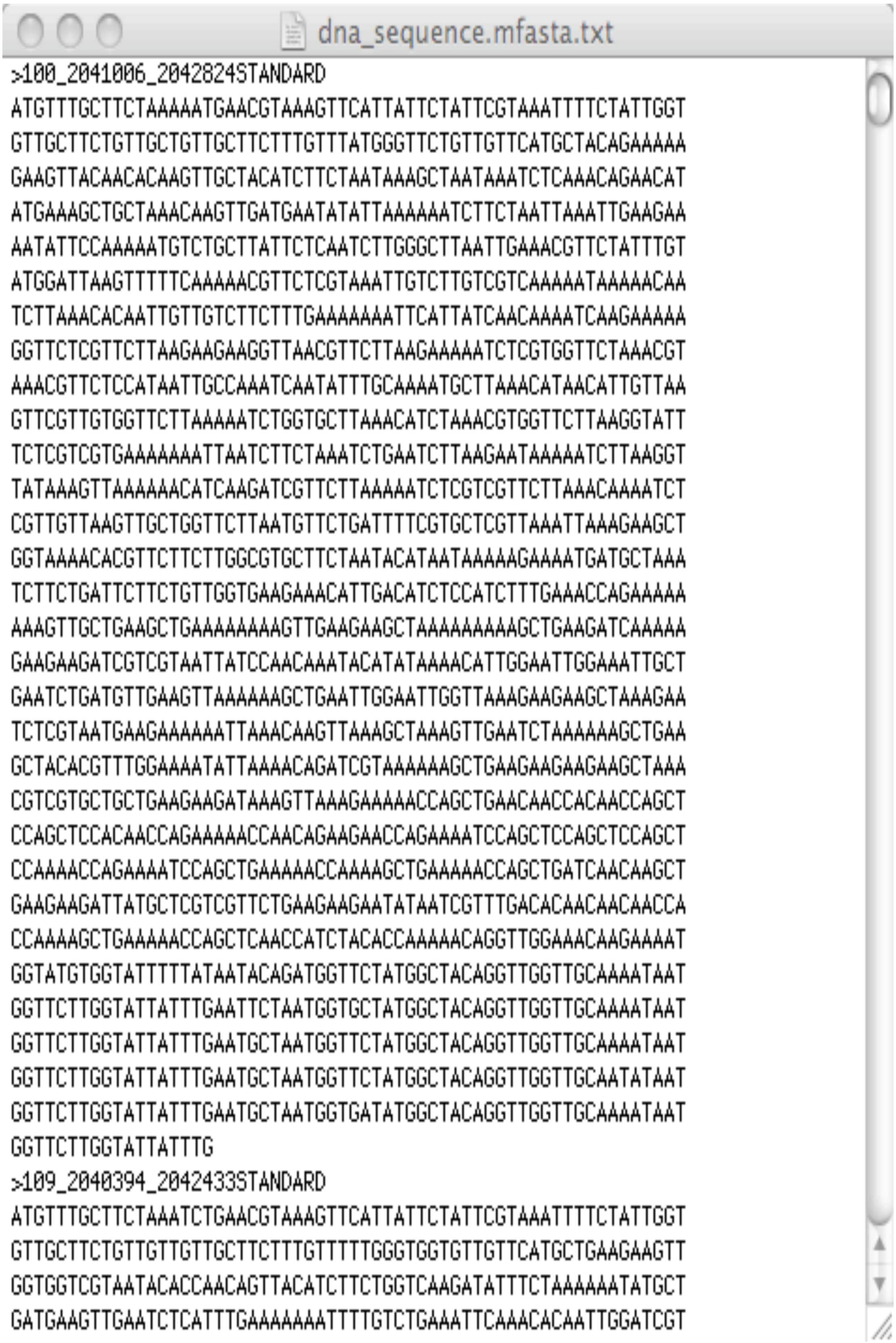
Seq_ID	Position	Size(bp)	Seq_ID	Position	Size(bp)
100	2041006-2042824	606	85	1973805_1975892	696
109	2040394-2042433	680	86	2107878_2110793	972
191	1152750-1155177	809	87	2030927_2033733	936
192	1812647-1814609	654	88	2035162_2037108	649
193	1137342-1139483	714	89	2075948_2077987	680
194	1753749_1755690	647	101	2101225_2103264	680
195	761979_764574	865	118	2136864_2138838	658
196	840685_842621	646	8	2011333_2013450	706
197	912598_915624	1009	90	2000206_2002397	731
198	19507_21562	685	91	2038229_2040255	676
199	1045743_1048207	822	92	2028999_2031037	680
19	1973814_1975840	676	93	2093146_2095196	684
10	1947220_1949708	830	94	2055093_2057234	714
1	1976482_1978977	832	95	2136261_2139175	972
200	1979749_1981368	540	96	2136261_2139175	680
201	2090416_2092497	694	97	2027649_2029673	675
202	65284_67225	647	98	1972800_1975790	997
203	2038813_2040914	701	119	2083723_2086596	958
204	1057353_1059500	716	99	2034067_2036198	711
205	1986660_1988675	672	9	1960342_1962399	686
206	1975833_1977926	698	11	1991275_1993820	849
207	2079566_2081489	641	120	2036538_2038748	737
208	2095718_2097610	631	121	2136815_2138473	553
110	495690_498359	890	122	2041042_2043781	913
209	2093389_2095530	714	123	2036922_2039591	890
20	2037096_2039829	911	124	2092457_2095258	934
210	1999751_2001892	714	125	2057134_2059100	656
211	2061822_2063861	680	126	2016531_2018068	513
212	195941_197437	499	102	2051939_2053995	686
213	2110015_2112096	694	127	2081963_2083893	644
21	2002063_2004039	659	128	2036837_2038449	538
22	2040504_2042471	656	129	2007307_2009346	680
23	2036149_2038188	680	12	1943919_1946006	696
24	2096354_2098300	649	130	2139715_2141856	714
111	2013418_2015463	682	131	2118922_2121762	947
25	2040678_2042717	680	132	2078089_2080868	927
26	2026983_2029022	680	133	2118213_2120239	676
27	2049806_2051773	656	134	2058215_2060728	838
28	2189909_2192707	933	135	1532024_1534188	722
29	2047399_2049425	676	103	2094092_2096413	774
2	1946537_1948588	684	136	2060979_2063008	677
30	2031138_2033164	676	137	2058664_2060714	684
31	2043992_2045886	632	138	2051810_2053789	660
32	2134034_2136112	693	139	1987726_1989769	681
33	2028632_2030671	680	13	1933609_1935678	690
112	2066712_2068679	656	140	2088014_2090215	734
34	2034814_2036853	680	141	2106261_2109065	935
35	2021945_2023984	680	142	2091647_2093749	701
36	2015230_2017269	680	143	2055483_2057876	798
37	2078650_2080857	736	144	1953892_1956162	757
38	1292293_1294758	822	104	1663236_1665273	679
39	167795_169354	520	145	1977301_1980231	977
3	1962555_1965321	922	146	2173286_2175400	705
40	747136_748401	422	147	2044279_2046443	722

41	1002497_1004056	520	148	1992501_1994588	696
42	1162300_1164357	680	149	2006475_2008503	676
113	2048064_2050031	656	14	1943201_1945686	829
43	756277_758322	682	150	2059811_2061850	680
44	1642022_1643963	647	152	2088769_2091576	936
45	1347390_1349357	656	153	1958395_1960611	739
46	2158221_2160383	721	154	2027460_2029499	680
47	1995044_1997149	702	105	273576_275517	647
48	2015826_2018600	925	155	2032903_2035211	770
49	2139758_2141959	734	156	2014344_2015924	527
4	1976360_1979130	924	157	2041319_2043395	692
50	2046209_2048134	642	158	2031490_2033529	680
51	2041048_2043135	696	159	2022339_2024426	696
114	2058118_2060181	688	15	1954875_1956868	665
52	2037135_2040089	985	160	2113285_2116006	907
53	2024808_2026972	722	161	2066124_2068265	714
54	2027843_2030152	770	162	2047380_2050409	1010
55	2095443_2097992	850	163	2003932_2006148	739
56	2097885_2100686	934	106	1994989_1997232	748
57	1985791_1987982	731	164	2031114_2033264	717
58	2124513_2127500	996	165	2041992_2044636	882
59	2090589_2092556	656	166	2068760_2070901	714
5	1907830_1909758	643	167	2068138_2070993	952
60	1989373_1991565	731	168	1612524_1614477	651
115	2080623_2082589	656	169	2034009_2036048	680
61	1990902_1992878	659	16	1993498_1995552	685
62	2064692_2066719	676	170	2193373_2196156	928
63	2091628_2093712	695	171	2065249_2067379	710
64	2057132_2059192	687	172	2090077_2092179	701
65	2042148_2044340	731	107	2038692_2040731	680
66	2037668_2039833	722	173	2030713_2033742	1010
67	1976352_1978414	688	174	2116957_2119106	717
68	2060715_2062796	694	175	2050671_2052835	722
69	1979875_1981962	696	176	2197008_2199110	701
6	1894817_1898268	1151	177	2069879_2071978	700
116	2116874_2118914	680	178	1005047_1007194	716
70	2059441_2061582	714	179	2034168_2037344	1059
71	2031016_2033055	680	17	1914367_1916105	580
72	2068249_2070350	701	180	2091652_2094566	972
73	2065313_2068107	932	181	2030441_2033173	911
74	2111480_2113446	656	108	819609_821570	654
75	2101458_2103598	714	182	2001373_2003565	731
77	2017714_2019843	710	183	2092876_2095790	972
78	2027782_2030534	918	184	1985740_1987785	682
79	2059902_2062043	714	185	2066421_2068562	714
7	1968733_1970063	444	186	1986096_1988288	731
117	2020019_2022106	696	187	2061360_2063099	580
80	2043269_2045308	680	188	1987544_1989649	702
81	1827867_1830475	870	189	1214300_1216207	636
82	2037677_2039653	659	18	370968_373047	693
83	2070179_2072278	700	190	1198012_1200417	802
84	2036943_2038497	518			

Appendix C: Unix shellscript for concatenation of strings.

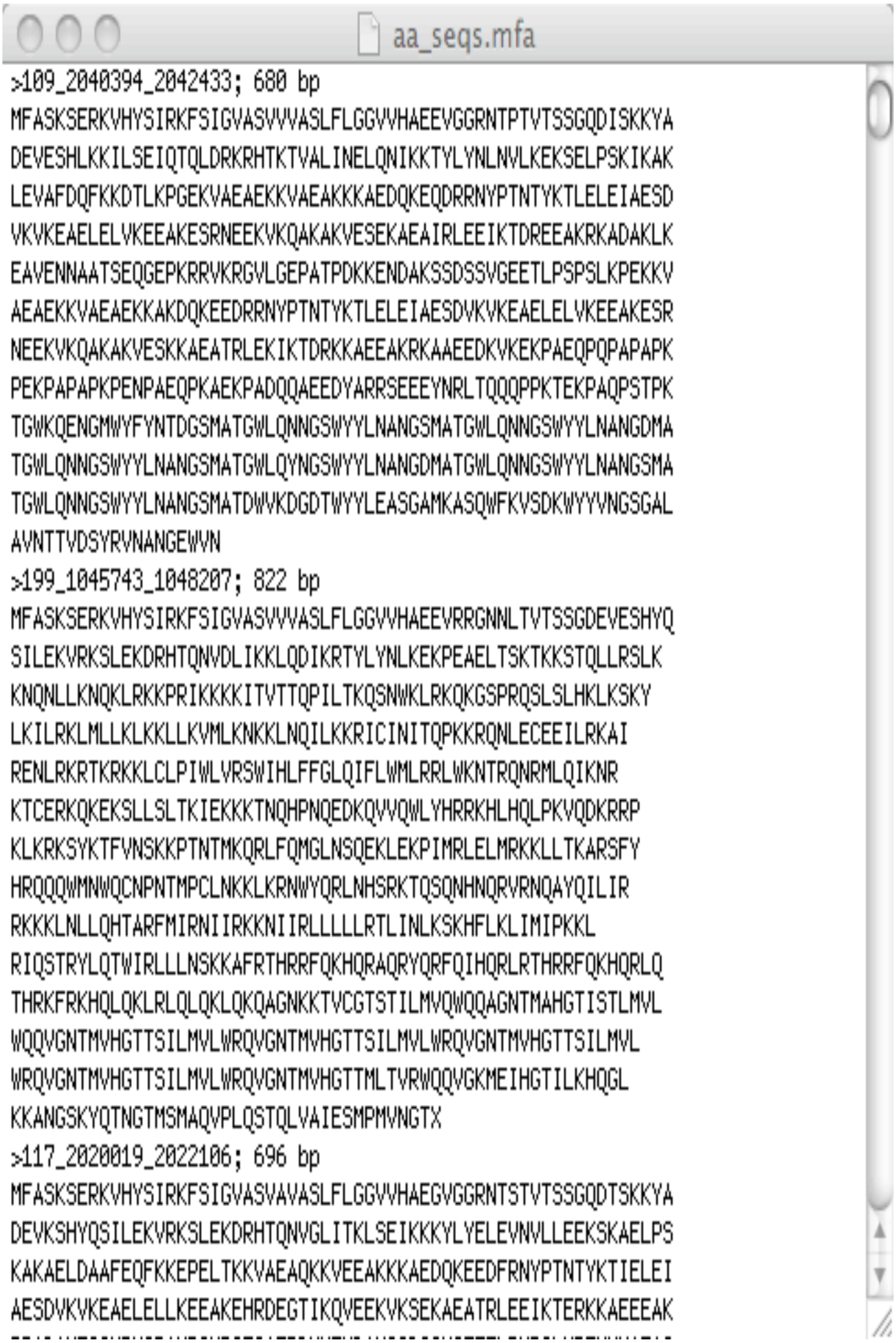
```
$ cd /Users/stevensewe/Desktop/dna_sequence_data

$ cat *.fasta > dna_sequence.mfasta
```

Appendix D. Multi-fasta file of CbpA dna sequences.



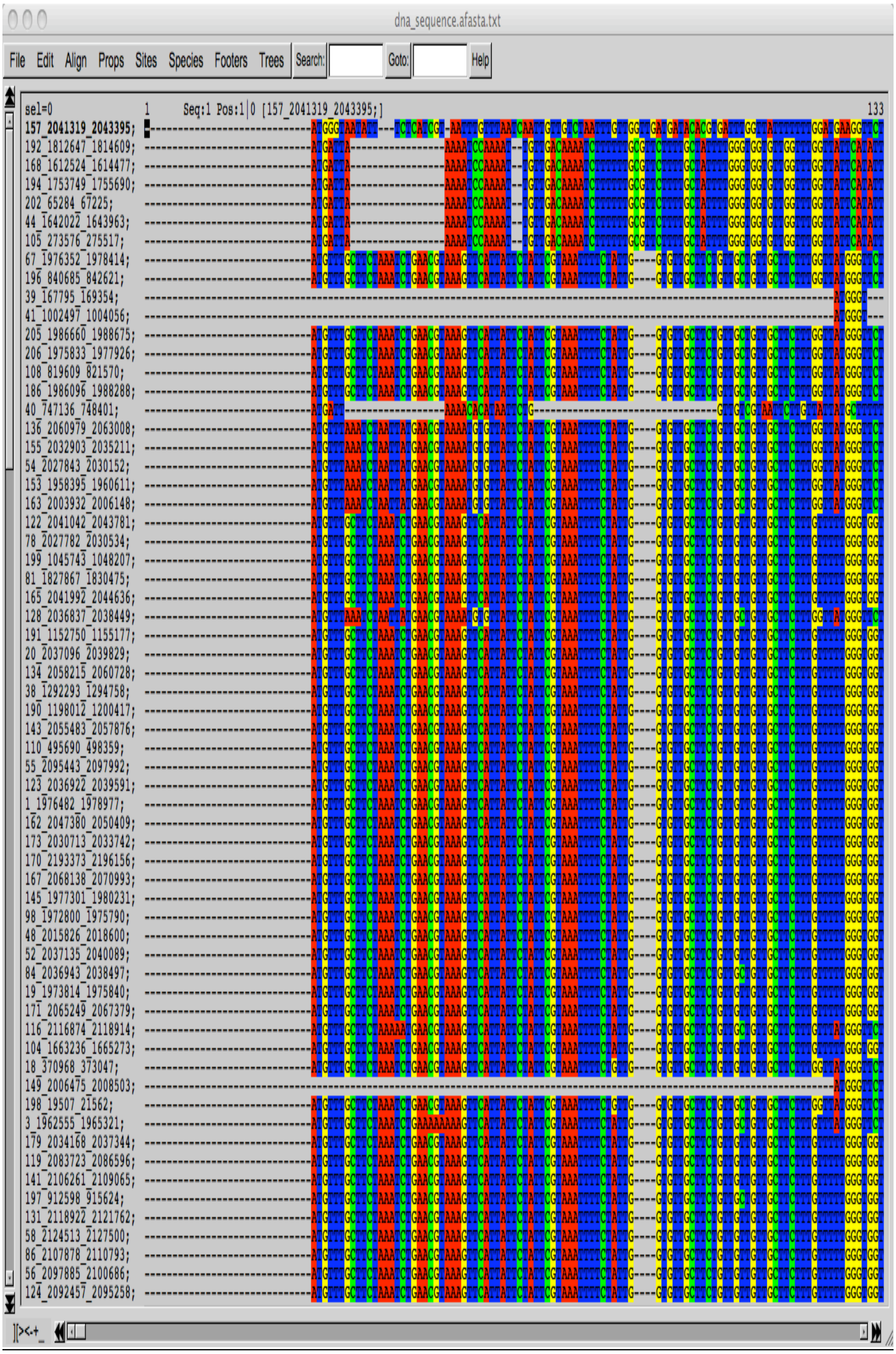
```
>100_2041006_2042824STANDARD
ATGTTTGCTTCTAAAAATGAACGTAAAGTTCATTATTCTATTCGTAAATTTTCTATTGGT
GTTGCTTCTGTTGCTGTTGCTTCTTTGTTTATGGGTTCTGTTGTTCATGCTACAGAAAAA
GAAGTTACAACACAAGTTGCTACATCTTCTAATAAAGCTAATAAATCTCAAACAGAACAT
ATGAAAGCTGCTAAACAAGTTGATGAATATATTAAAAAATCTTCTAATTAAATTGAAGAA
AATATTCCAAAAATGTCTGCTTATTCTCAATCTTGGGCTTAATTGAAACGTTCTATTTGT
ATGGATTAAGTTTTTCAAAAACGTTCTCGTAAATTGTCTTGTCGTCAAAAATAAAAACAA
TCTTAAACACAATTGTTGTCTTCTTTGAAAAAAATTCATTATCAACAAAATCAAGAAAAA
GGTTCTCGTTCTTAAGAAGAAGGTTAACGTTCTTAAGAAAAATCTCGTGGTTCTAAACGT
AAACGTTCTCCATAATTGCCAAATCAATATTTGCAAAATGCTTAAACATAACATTGTTAA
GTTCGTTGTGGTTCTTAAAAATCTGGTGCTTAAACATCTAAACGTGGTTCTTAAGGTATT
TCTCGTCGTGAAAAAAATTAATCTTCTAAATCTGAATCTTAAGAATAAAAATCTTAAGGT
TATAAAGTTAAAAAACATCAAGATCGTTCTTAAAAATCTCGTCGTTCTTAAACAAAATCT
CGTTGTTAAGTTGCTGGTTCTTAATGTTCTGATTTTCGTGCTCGTTAAATTAAAGAAGCT
GGTAAAACACGTTCTTCTTGGCGTGCTTCTAATACATAATAAAAAGAAAATGATGCTAAA
TCTTCTGATTCTTCTGTTGGTGAAGAAACATTGACATCTCCATCTTTGAAACCAGAAAAA
AAAGTTGCTGAAGCTGAAAAAAAAGTTGAAGAAGCTAAAAAAAAAGCTGAAGATCAAAAA
GAAGAAGATCGTCGTAATTATCCAACAAATACATATAAAACATTGGAATTGGAAATTGCT
GAATCTGATGTTGAAGTTAAAAAAGCTGAATTGGAATTGGTTAAAGAAGAAGCTAAAGAA
TCTCGTAATGAAGAAAAAATTAAACAAGTTAAAGCTAAAGTTGAATCTAAAAAAGCTGAA
GCTACACGTTTGGAAAATATTAAAACAGATCGTAAAAAAGCTGAAGAAGAAGAAGCTAAA
CGTCGTGCTGCTGAAGAAGATAAAGTTAAAGAAAAACCAGCTGAACAACCACAACCAGCT
CCAGCTCCACAACCAGAAAAACCAACAGAAGAACCAGAAAATCCAGCTCCAGCTCCAGCT
CCAAAACCAGAAAATCCAGCTGAAAAACCAAAAGCTGAAAAACCAGCTGATCAACAAGCT
GAAGAAGATTATGCTCGTCGTTCTGAAGAAGAATATAATCGTTTGACACAACAACAACCA
CCAAAAGCTGAAAAACCAGCTCAACCATCTACACCAAAAACAGGTTGGAAACAAGAAAAT
GGTATGTGGTATTTTTATAATACAGATGGTTCTATGGCTACAGGTTGGTTGCAAAATAAT
GGTTCTTGGTATTATTTGAATTCTAATGGTGCTATGGCTACAGGTTGGTTGCAAAATAAT
GGTTCTTGGTATTATTTGAATGCTAATGGTTCTATGGCTACAGGTTGGTTGCAAAATAAT
GGTTCTTGGTATTATTTGAATGCTAATGGTTCTATGGCTACAGGTTGGTTGCAATATAAT
GGTTCTTGGTATTATTTGAATGCTAATGGTGATATGGCTACAGGTTGGTTGCAAAATAAT
GGTTCTTGGTATTATTTG
>109_2040394_2042433STANDARD
ATGTTTGCTTCTAAATCTGAACGTAAAGTTCATTATTCTATTCGTAAATTTTCTATTGGT
GTTGCTTCTGTTGTTGTTGCTTCTTTGTTTTTGGGTGGTGTTGTTCATGCTGAAGAAGTT
GGTGGTCGTAATACACCAACAGTTACATCTTCTGGTCAAGATATTTCTAAAAAATATGCT
GATGAAGTTGAATCTCATTTGAAAAAAATTTTGTCTGAAATTCAAACACAATTGGATCGT
```

Appendix E. Multi-fasta file of CbpA protein sequences.



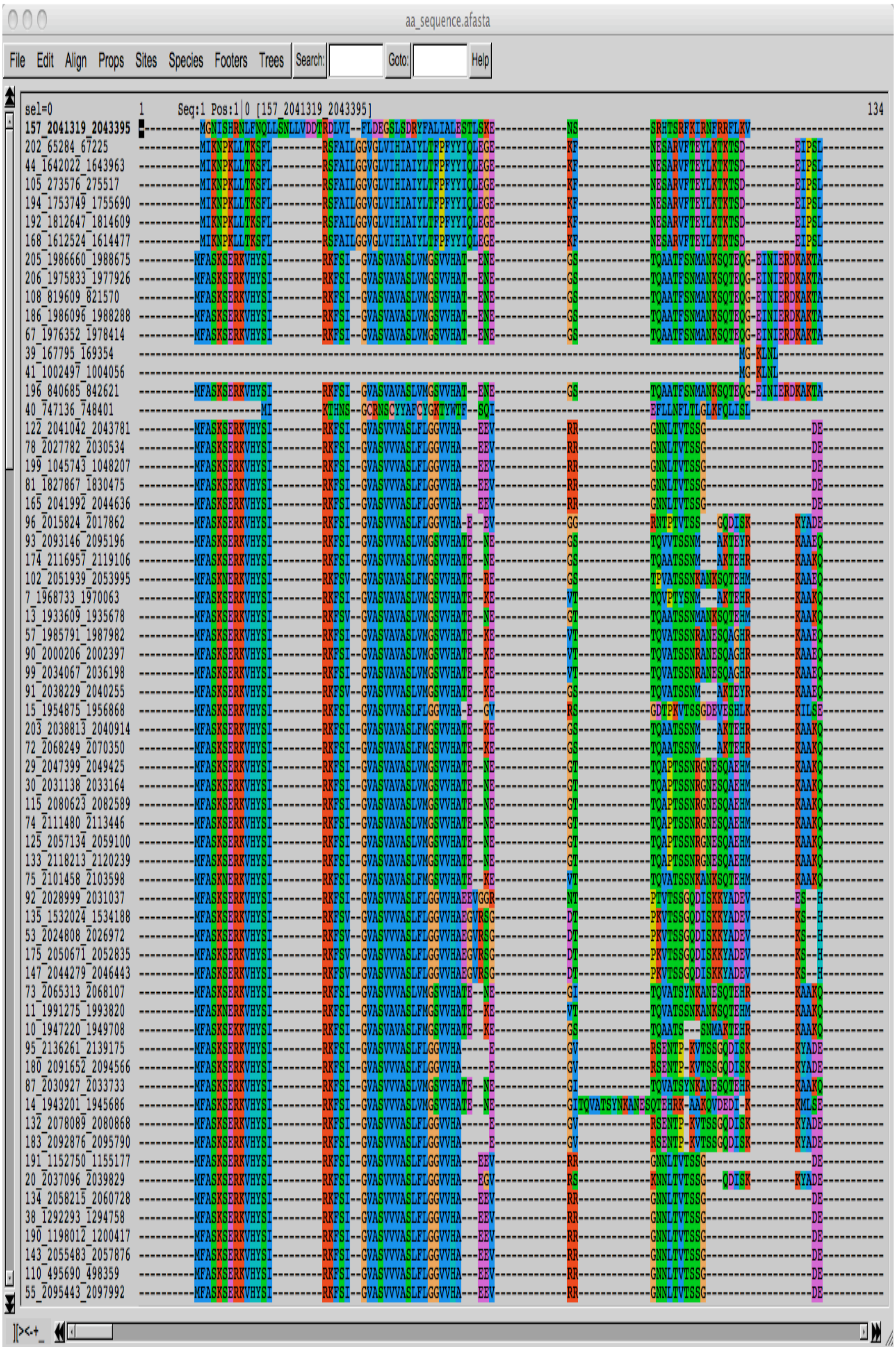
```
>109_2040394_2042433; 680 bp
MFASKSERKVHYSIRKFSIGVASVVVASLFLGGVVHAEEVGGRNTPTVTSSGQDISKKYA
DEVESHLKKILSEIQTQLDRKRHTKTVALINELQNIKKTYLYNLNVLKEKSELPSKIKAK
LEVAFDQFKKDTLKPGEKVAEAEKKVAEAKKKAEDQKEQDRRNYPTNTYKTLELEIAESD
VKVKEAELELVKEEAKESRNEEKVKQAKAKVESEKAEAIRLEEIKTDREEAKRKADAKLK
EAVENNAATSEQGEPKRRVKRGVLGEPATPDKKENDAKSSDSSVGEETLPSPSLKPEKKV
AEAEKKVAEAEKKAKDQKEEDRRNYPTNTYKTLELEIAESDVKVKEAELELVKEEAKESR
NEEKVKQAKAKVESKKAEATRLEKIKTDRKKAEEAKRKAAEEDKVKEKPAEQPQPAPAPK
PEKPAPAPKPENPAEQPKAEKPADQQAEEDYARRSEEEYNRLTQQQPPKTEKPAQPSTPK
TGWKQENGMWYFYNTDGSMATGWLQNNGSWYYLNANGSMATGWLQNNGSWYYLNANGDMA
TGWLQNNGSWYYLNANGSMATGWLQYNGSWYYLNANGDMATGWLQNNGSWYYLNANGSMA
TGWLQNNGSWYYLNANGSMATDWVKDGDTWYYLEASGAMKASQWFKVSDKWYYVNGSGAL
AVNTTVDSYRVNANGEWVN
>199_1045743_1048207; 822 bp
MFASKSERKVHYSIRKFSIGVASVVVASLFLGGVVHAEEVRRGNNLTVTSSGDEVESHYQ
SILEKVRKSLEKDRHTQNVDLIKKLQDIKRTYLYNLKEKPEAELTSKTKKSTQLLRSLK
KNQNLLKNQKLRKKPRIKKKKITVTTQPILTKQSNWKLRKQKGSPRQSLSLHKLKSKY
LKILRKLMLLKLKKLLKVMLKNKKLNQILKKRICINITQPKKRQNLECEEILRKAI
RENLRKRTKRKKLCLPIWLVRSWIHLFFGLQIFLWMLRRLWKNTRQNRMLQIKNR
KTCERKQKEKSLLSLTKIEKKKTNQHPNQEDKQVVQWLYHRRKHLHQLPKVQDKRRP
KLKRKSYKTFVNSKKPTNTMKQRLFQMGLNSQEKLEKPIMRLELMRKKLLTKARSFY
HRQQQWMNWQCNPNTMPCLNKKLKRNWYQRLNHSRKTQSQNHNQRVRNQAYQILIR
RKKKLNLLQHTARFMIRNIIRKKNIIRLLLLLRTLINLKSKHFLKLIMIPKKL
RIQSTRYLQTWIRLLLNSKKAFRTHRRFQKHQRAQRYQRFQIHQRLRTHRRFQKHQRLQ
THRKFRKHQLQKLRLQLQKLQKQAGNKKTVCGTSTILMVQWQQAGNTMAHGTISTLMVL
WQQVGNTMVHGTTSILMVLWRQVGNTMVHGTTSILMVLWRQVGNTMVHGTTSILMVL
WRQVGNTMVHGTTSILMVLWRQVGNTMVHGTTMLTVRWQQVGKMEIHGTILKHQGL
KKANGSKYQTNGTMSMAQVPLQSTQLVAIESMPMVNGTX
>117_2020019_2022106; 696 bp
MFASKSERKVHYSIRKFSIGVASVAVASLFLGGVVHAEGVGGRNTSTVTSSGQDTSKKYA
DEVKSHYQSILEKVRKSLEKDRHTQNVGLITKLSEIKKKYLYELEVNVLLEEKSKAELPS
KAKAELDAAFEQFKKEPELTKKVAEAQKKVEEAKKKAEDQKEEDFRNYPTNTYKTIELEI
AESDVKVKEAELELLKEEAKEHRDEGTIKQVEEKVKSEKAEATRLEEIKTERKKAEEEAK
```

Appendix F. Multiple sequence alignment of CbpA dna in seaview.



```
sel=0               1      Seq:1 Pos:1|0 [157_2041319_2043395;]      133
157_2041319_2043395;
192_1812647_1814609;
168_1612524_1614477;
194_1753749_1755690;
202_65284_67225;
44_1642022_1643963;
105_273576_275517;
67_1976352_1978414;
196_840685_842621;
39_167795_169354;
41_1002497_1004056;
205_1986660_1988675;
206_1975833_1977926;
108_819609_821570;
186_1986096_1988288;
40_747136_748401;
136_2060979_2063008;
155_2032903_2035211;
54_2027843_2030152;
153_1958395_1960611;
163_2003932_2006148;
122_2041042_2043781;
78_2027782_2030534;
199_1045743_1048207;
81_1827867_1830475;
165_2041992_2044636;
128_2036837_2038449;
191_1152750_1155177;
20_2037096_2039829;
134_2058215_2060728;
38_1292293_1294758;
190_1198012_1200417;
143_2055483_2057876;
110_495690_498359;
55_2095443_2097992;
123_2036922_2039591;
1_1976482_1978977;
162_2047380_2050409;
173_2030713_2033742;
170_2193373_2196156;
167_2068138_2070993;
145_1977301_1980231;
98_1972800_1975790;
48_2015826_2018600;
52_2037135_2040089;
84_2036943_2038497;
19_1973814_1975840;
171_2065249_2067379;
116_2116874_2118914;
104_1663236_1665273;
18_370968_373047;
149_2006475_2008503;
198_19507_21562;
3_1962555_1965321;
179_2034168_2037344;
119_2083723_2086596;
141_2106261_2109065;
197_912598_915624;
131_2118922_2121762;
58_2124513_2127500;
86_2107878_2110793;
56_2097885_2100686;
124_2092457_2095258;
```

Appendix G. Multiple sequence alignment of CbpA protein in seaview.



```
sel=0               1      Seq:1 Pos:1|0 [157_2041319_2043395]                                                        134
157_2041319_2043395 -----------MGNISHRNLFNQLLSNLLVDDTRDLVI--FLDEGSLSDRYFALIALESTLSKE-------------NS-------------SRHTSRFKIRNFRRFLKV------------------------
202_65284_67225     -----------MIKNPKLLTKSFL---------RSFAILGGVGLVIHIAIYLTFPFYYIQLEGE-------------KF-------------NESARVFTEYLKTKTSD---------EIPSL-----------
44_1642022_1643963  -----------MIKNPKLLTKSFL---------RSFAILGGVGLVIHIAIYLTFPFYYIQLEGE-------------KF-------------NESARVFTEYLKTKTSD---------EIPSL-----------
105_273576_275517   -----------MIKNPKLLTKSFL---------RSFAILGGVGLVIHIAIYLTFPFYYIQLEGE-------------KF-------------NESARVFTEYLKTKTSD---------EIPSL-----------
194_1753749_1755690 -----------MIKNPKLLTKSFL---------RSFAILGGVGLVIHIAIYLTFPFYYIQLEGE-------------KF-------------NESARVFTEYLKTKTSD---------EIPSL-----------
192_1812647_1814609 -----------MIKNPKLLTKSFL---------RSFAILGGVGLVIHIAIYLTFPFYYIQLEGE-------------KF-------------NESARVFTEYLKTKTSD---------EIPSL-----------
168_1612524_1614477 -----------MIKNPKLLTKSFL---------RSFAILGGVGLVIHIAIYLTFPFYYIQLEGE-------------KF-------------NESARVFTEYLKTKTSD---------EIPSL-----------
205_1986660_1988675 ----------MFASKSERKVHYSI---------RKFSI--GVASVAVASLVMGSVVHAT--ENE-------------GS-------------TQAATFSNMANKSQTEQG-EINIERDKAKTA-----------
206_1975833_1977926 ----------MFASKSERKVHYSI---------RKFSI--GVASVAVASLVMGSVVHAT--ENE-------------GS-------------TQAATFSNMANKSQTEQG-EINIERDKAKTA-----------
108_819609_821570   ----------MFASKSERKVHYSI---------RKFSI--GVASVAVASLVMGSVVHAT--ENE-------------GS-------------TQAATFSNMANKSQTEQG-EINIERDKAKTA-----------
186_1986096_1988288 ----------MFASKSERKVHYSI---------RKFSI--GVASVAVASLVMGSVVHAT--ENE-------------GS-------------TQAATFSNMANKSQTEQG-EINIERDKAKTA-----------
67_1976352_1978414  ----------MFASKSERKVHYSI---------RKFSI--GVASVAVASLVMGSVVHAT--ENE-------------GS-------------TQAATFSNMANKSQTEQG-EINIERDKAKTA-----------
39_167795_169354    ------------------------------------------------------------------------------------------------------------MG-KLNL-------------------
41_1002497_1004056  ------------------------------------------------------------------------------------------------------------MG-KLNL-------------------
196_840685_842621   ----------MFASKSERKVHYSI---------RKFSI--GVASVAVASLVMGSVVHAT--ENE-------------GS-------------TQAATFSNMANKSQTEQG-EINIERDKAKTA-----------
40_747136_748401    ----------------------MI---------KTHNS--GCRNSCYYAFCYGKTYWTF--SQI----------------------------EFLLNFLTLGLKFQLISL------------------------
122_2041042_2043781 ----------MFASKSERKVHYSI---------RKFSI--GVASVVVASLFLGGVVHA---EEV-------------RR-------------GNNLTVTSSG-------------------DE-----------
78_2027782_2030534  ----------MFASKSERKVHYSI---------RKFSI--GVASVVVASLFLGGVVHA---EEV-------------RR-------------GNNLTVTSSG-------------------DE-----------
199_1045743_1048207 ----------MFASKSERKVHYSI---------RKFSI--GVASVVVASLFLGGVVHA---EEV-------------RR-------------GNNLTVTSSG-------------------DE-----------
81_1827867_1830475  ----------MFASKSERKVHYSI---------RKFSI--GVASVVVASLFLGGVVHA---EEV-------------RR-------------GNNLTVTSSG-------------------DE-----------
165_2041992_2044636 ----------MFASKSERKVHYSI---------RKFSI--GVASVVVASLFLGGVVHA---EEV-------------RR-------------GNNLTVTSSG-------------------DE-----------
96_2015824_2017862  ----------MFASKSERKVHYSI---------RKFSI--GVASVVVASLFLGGVVHA-E--EV-------------GG-------------RNTPTVTSS---GQDISK--------KYADE-----------
93_2093146_2095196  ----------MFASKSERKVHYSI---------RKFSI--GVASVVVASLVMGSVVHATE--NE-------------GS-------------TQVVTSSNM---AKTEYR--------KAAEQ-----------
174_2116957_2119106 ----------MFASKSERKVHYSI---------RKFSI--GVASVVVASLVMGSVVHATE--NE-------------GS-------------TQAATSSNM---AKTEHR--------KAAKQ-----------
102_2051939_2053995 ----------MFASKNERKVHYSI---------RKFSV--GVASVAVASLFMGSVVHATE--RE-------------GS-------------TPVATSSNKANKSQTEHM--------KAAEQ-----------
7_1968733_1970063   ----------MFASKSERKVHYSI---------RKFSI--GVASVAVASLVMGSVVHATE--KE-------------VT-------------TQVPTYSNM---AKTEHR--------KAAKQ-----------
13_1933609_1935678  ----------MFASKSERKVHYSI---------RKFSV--GVASVAVASLVMGSVVHATE--NE-------------GT-------------TQAATSSNMANKSQTEHM--------KAAKQ-----------
57_1985791_1987982  ----------MFASKSERKVHYSI---------RKFSI--GVASVAVASLVMGSVVHATE--KE-------------VT-------------TQVATSSNRANESQAGHR--------KAAEQ-----------
90_2000206_2002397  ----------MFASKSERKVHYSI---------RKFSI--GVASVAVASLVMGSVVHATE--KE-------------VT-------------TQVATSSNRANESQAGHR--------KAAEQ-----------
99_2034067_2036198  ----------MFASKSERKVHYSI---------RKFSI--GVASVAVASLVMGSVVHATE--KE-------------VT-------------TQVATSSNRANESQAGHR--------KAAEQ-----------
91_2038229_2040255  ----------MFASKSERKVHYSI---------RKFSV--GVASVVVASLVMGSVVHATE--KE-------------GS-------------TQVATSSNM---AKTEYR--------KAAEQ-----------
15_1954875_1956868  ----------MFASKSERKVHYSI---------RKFSI--GVASVVVASLFLGGVVHA-E--GV-------------RS-------------GDTPKVTSSGDEVESHLK--------KILSE-----------
203_2038813_2040914 ----------MFASKSERKVHYSI---------RKFSI--GVASVAVASLFMGSVVHATE--KE-------------GS-------------TQAATSSNM---AKTEHR--------KAAKQ-----------
72_2068249_2070350  ----------MFASKSERKVHYSI---------RKFSI--GVASVAVASLFMGSVVHATE--KE-------------GS-------------TQAATSSNM---AKTEHR--------KAAKQ-----------
29_2047399_2049425  ----------MFASKSERKVHYSI---------RKFSI--GVASVAVASLVMGSVVHATE--NE-------------GT-------------TQAPTSSNRGNESQAEHM--------KAAKQ-----------
30_2031138_2033164  ----------MFASKSERKVHYSI---------RKFSI--GVASVAVASLVMGSVVHATE--NE-------------GT-------------TQAPTSSNRGNESQAEHM--------KAAKQ-----------
115_2080623_2082589 ----------MFASKSERKVHYSI---------RKFSI--GVASVAVASLVMGSVVHATE--NE-------------GT-------------TQAPTSSNRGNESQAEHM--------KAAKQ-----------
74_2111480_2113446  ----------MFASKSERKVHYSI---------RKFSI--GVASVAVASLVMGSVVHATE--NE-------------GT-------------TQAPTSSNRGNESQAEHM--------KAAKQ-----------
125_2057134_2059100 ----------MFASKSERKVHYSI---------RKFSI--GVASVAVASLVMGSVVHATE--NE-------------GT-------------TQAPTSSNRGNESQAEHM--------KAAKQ-----------
133_2118213_2120239 ----------MFASKSERKVHYSI---------RKFSI--GVASVAVASLVMGSVVHATE--NE-------------GT-------------TQAPTSSNRGNESQAEHM--------KAAKQ-----------
75_2101458_2103598  ----------MFASKNERKVHYSI---------RKFSI--GVASVAVASLFMGSVVHATE--KE-------------VT-------------TQVATSSNKANKSQTEHM--------KAAKQ-----------
92_2028999_2031037  ----------MFASKSERKVHYSI---------RKFSI--GVASVVVASLFLGGVVHAEEVGGR-------------NT-------------PTVTSSGQDISKKYADEV--------ES--H-----------
135_1532024_1534188 ----------MFASKSERKVHYSI---------RKFSV--GVASVVVASLFLGGVVHAEGVRSG-------------DT-------------PKVTSSGQDISKKYADEV--------KS--H-----------
53_2024808_2026972  ----------MFASKSERKVHYSI---------RKFSV--GVASVVVASLFLGGVVHAEGVRSG-------------DT-------------PKVTSSGQDISKKYADEV--------KS--H-----------
175_2050671_2052835 ----------MFASKSERKVHYSI---------RKFSV--GVASVVVASLFLGGVVHAEGVRSG-------------DT-------------PKVTSSGQDISKKYADEV--------KS--H-----------
147_2044279_2046443 ----------MFASKSERKVHYSI---------RKFSV--GVASVVVASLFLGGVVHAEGVRSG-------------DT-------------PKVTSSGQDISKKYADEV--------KS--H-----------
73_2065313_2068107  ----------MFASKSERKVHYSI---------RKFSI--GVASVVVASLVMGSVVHATE--NE-------------GI-------------TQVATSYNKANESQTEHR--------KAAKQ-----------
11_1991275_1993820  ----------MFASKNERKVHYSI---------RKFSI--GVASVAVASLFMGSVVHATE--KE-------------VT-------------TQVATSSNKANKSQTEHM--------KAAKQ-----------
10_1947220_1949708  ----------MFASKSEKKVHYSI---------RKFSI--GVASVAVASLFMGSVVHATE--KE-------------GS-------------TQAATS---SNMAKTEHR--------KAAKQ-----------
95_2136261_2139175  ----------MFASKSERKVHYSI---------RKFSI--GVASVVVASLFLGGVVHA-----E-------------GV-------------RSENTP-KVTSSGQDISK--------KYADE-----------
180_2091652_2094566 ----------MFASKSERKVHYSI---------RKFSI--GVASVVVASLFLGGVVHA-----E-------------GV-------------RSENTP-KVTSSGQDISK--------KYADE-----------
87_2030927_2033733  ----------MFASKSERKVHYSI---------RKFSI--GVASVVVASLVMGSVVHATE--NE-------------GI-------------TQVATSYNKANESQTEHR--------KAAKQ-----------
14_1943201_1945686  ----------MFASKSERKVHYSI---------RKFSI--GVASVVVASLVMGSVVHATE--NE-------------GITQVATSYNKANESQTEHRK-AAKQVDEDI-K--------KMLSE-----------
132_2078089_2080868 ----------MFASKSERKVHYSI---------RKFSI--GVASVVVASLFLGGVVHA-----E-------------GV-------------RSENTP-KVTSSGQDISK--------KYADE-----------
183_2092876_2095790 ----------MFASKSERKVHYSI---------RKFSI--GVASVVVASLFLGGVVHA-----E-------------GV-------------RSENTP-KVTSSGQDISK--------KYADE-----------
191_1152750_1155177 ----------MFASKSERKVHYSI---------RKFSI--GVASVVVASLFLGGVVHA---EEV-------------RR-------------GNNLTVTSSG-------------------DE-----------
20_2037096_2039829  ----------MFASKSERKVHYSI---------RKFSI--GVASVVVASLFLGGVVHA---EGV-------------RS-------------KNNLTVTSSG---QDISK--------KYADE-----------
134_2058215_2060728 ----------MFASKSERKVHYSI---------RKFSI--GVASVVVASLFLGGVVHA---EEV-------------RR-------------GNNLTVTSSG-------------------DE-----------
38_1292293_1294758  ----------MFASKSERKVHYSI---------RKFSI--GVASVVVASLFLGGVVHA---EEV-------------RR-------------GNNLTVTSSG-------------------DE-----------
190_1198012_1200417 ----------MFASKSERKVHYSI---------RKFSI--GVASVVVASLFLGGVVHA---EEV-------------RR-------------GNNLTVTSSG-------------------DE-----------
143_2055483_2057876 ----------MFASKSERKVHYSI---------RKFSI--GVASVVVASLFLGGVVHA---EEV-------------RR-------------GNNLTVTSSG-------------------DE-----------
110_495690_498359   ----------MFASKSERKVHYSI---------RKFSI--GVASVVVASLFLGGVVHA---EEV-------------RR-------------GNNLTVTSSG-------------------DE-----------
55_2095443_2097992  ----------MFASKSERKVHYSI---------RKFSI--GVASVVVASLFLGGVVHA---EEV-------------RR-------------GNNLTVTSSG-------------------DE-----------
```

Appendix H. MLST results.

	HOUSE KEEPING ALLELE NUMBER							
Seq. no.	**aroe**	**ddl**	**gdh**	**gki**	**recp**	**spi**	**xpt**	**ST**
1	1	17	1	4	1	18	58	433
2	15	9	8	8	18	15	1	1030
3	1	6	8	6	2	6	20	507
4	5	27	7	4	2	10	1	97
5	8	17	9	2	1	6	1	191
6	8	17	9	2	1	6	1	191
7	7	21	2	1	1	100	1	
8	7	14	11	10	16	6	493	8138
9	7	14	11	10	1	6	8	162
10	1	6	8	6	2	6	20	507
11	7	14	5	1	1	13	31	440
12	2	14	5	1	11	16	3	53
13	10	145	11	34	16	15	1	1233
14	5	6	7	4	2	10	1	1551
15	1	6	8	4	1	1	4	36
16	16	14	13	4	5	6	10	247
17	2	14	5	1	11	16	3	53
18	10	18	5	4	5	13	10	205
19	15	31	8	8	18	15	1	235
20	1	17	1	4	1	18	58	433
21	15	26	16	19	15	6	86	652
22	8	1	13	14	4	17	4	2344
23	4	78	16	19	15	55	20	635
24	4	1	4	2	4	4	1	81
25	4	1	144	19	15	6	20	2393
26	4	1	16	19	102	6	20	2588
27	8	97	13	14	4	17	77	2542
28	5	4	6	1	2	6	3	90
29	8	1	13	14	4	17	4	2344
30	8	14	13	152	4	17	4	2381
31	8	232	13	14	4	17	4	2392
32	16	33	12	9	1	41	33	289
33	4	1	16	19	67	6	202	2514
34	4	1	16	19	250	6	20	9356
35	4	1	16	19	4	6	20	2383
36	4	1	16	19	15	6	40	2347
37	2	14	5	29	12	16	3	62
38	1	8	5	4	5	5	27	13
39	1	101	5	4	5	5	3	2011
40	1	8	5	4	5	5	27	13
41	1	8	5	4	5	5	27	13
42	10	29	20	14	1	9	1	220
43	8	17	9	2	1	6	1	191
44	6	77	11	1	1	15	72	376
45	8	14	13	14	4	17	4	199
46	1	8	5	4	5	5	3	15

47	7	15	5	1	1	10	7	595
48	1	8	5	4	5	5	1	9
49	16	18	13	4	4	6	113	695
50	10	1	16	19	15	6	20	1451
51	5	1	12	29	16	9	39	2705
52	1	20	13	9	12	94	28	1292
53	203	14	13	14	4	17	4	
54	2	14	5	36	12	17	21	63
55	7	14	5	1	8	14	11	124
56	6	14	11	1	67	15	72	1339
57	7	67	6	9	2	6	1	384
58	6	77	11	1	1	15	72	376
59	8	14	13	14	4	17	4	199
60	7	67	25	9	2	6	1	2150
61	15	26	16	19	15	6	20	236
62	8	14	13	14	4	17	4	199
63	15	14	29	4	21	30	1	242
64	7	67	6	9	2	6	1	384
65	7	67	25	9	2	6	1	2150
66	8	548	13	14	4	17	4	8497
67	7	22	15	2	10	6	1	180
68	16	18	13	4	4	6	113	695
69	2	14	5	29	12	16	3	62
70	7	1	11	10	1	6	8	156
71	15	26	5	19	15	6	20	651
72	8	14	13	1	4	6	4	649
73	5	4	6	1	2	6	3	90
74	8	207	13	14	4	17	4	1936
75	4	1	4	2	4	4	1	81
76	4	26	16	19	15	6	20	271
77	5	6	9	6	5	6	1	1791
78	1	8	5	4	5	5	27	13
79	4	1	4	2	4	4	85	634
80	15	26	5	19	15	6	20	651
81	41	14	5	1	8	14	11	656
82	15	26	16	19	15	6	20	236
83	7	21	46	1	1	10	4	3060
84	1	8	10	4	1	9	3	43
85	5	18	35	40	12	9	39	636
86	6	77	11	1	1	15	72	376
87	1	6	8	6	2	6	4	37
88	4	189	16	19	15	6	20	3039
89	15	4	5	19	15	6	20	1461
90	7	104	6	1	17	6	1	1536
91	7	14	6	1	2	6	15	146
92	15	104	16	19	15	100	20	
93	15	14	29	4	21	30	1	242
94	7	1	11	10	1	6	8	156
95	6	77	11	1	1	15	72	376
96	15	1	5	19	15	6	20	8014

97	8	14	13	14	4	17	4	199
98	18	97	12	4	44	14	77	558
99	7	104	388	1	17	6	1	9057
100	7	1	11	10	1	6	8	156
101	15	14	16	19	15	6	20	926
102	2	6	5	90	61	17	130	1175
103	16	18	13	4	4	6	113	695
104	4	26	16	19	68	6	20	2476
105	6	152	11	1	67	15	72	1296
106	7	22	15	2	10	6	1	180
107	4	26	16	19	15	6	20	271
108	7	22	15	2	10	6	1	108
109	4	1	16	19	15	6	20	320
110	7	1	11	10	1	6	8	156
111	8	17	9	2	1	6	1	191
112	8	154	13	14	4	17	12	1341
113	8	207	13	14	4	17	4	1936
114	2	72	128	4	1	14	1	5872
115	8	14	13	14	4	17	4	199
116	15	14	29	4	21	30	1	242
117	5	1	12	29	16	9	39	2705
118	6	14	11	1	67	100	72	
119	6	77	11	1	1	15	72	376
120	7	67	6	9	2	6	1	384
121	4	1	4	2	4	4	1	81
122	1	8	5	4	5	5	27	13
123	7	1	11	10	1	6	8	156
124	6	14	11	1	67	5	72	2270
125	8	244	13	14	4	17	4	2584
126	2	1	13	4	11	100	16	
127	1	6	8	6	2	6	4	37
128	2	14	5	36	12	15	21	2543
129	4	1	16	19	15	6	20	320
130	4	1	4	2	4	4	1	81
131	6	231	11	1	67	15	72	8207
132	4	26	11	1	67	15	72	2541
133	8	14	13	14	4	4	4	2269
134	7	1	11	10	1	6	198	4464
135	1	20	13	9	12	94	4	4150
136	2	14	5	36	12	17	21	63
137	7	14	13	23	6	25	6	3280
138	4	189	16	19	15	6	20	3039
139	8	17	9	2	1	6	1	191
140	16	18	13	4	4	6	113	695
141	6	14	11	1	67	15	72	1339
142	18	14	2	22	16	9	23	654
143	7	316	11	10	1	6	8	4026
144	10	29	20	8	10	6	1	1176
145	18	97	12	4	44	14	77	558
146	61	14	60	67	5	6	12	1374

147	8	14	13	14	4	17	4	199
148	2	205	5	29	1	16	113	
149	7	22	15	2	10	6	1	180
150	15	26	16	19	15	6	20	236
151	4	1	16	19	15	6	20	320
152	6	8	11	1	67	15	72	2268
153	2	14	5	29	12	16	3	
154	4	1	16	19	15	6	20	320
155	2	14	5	36	12	17	21	63
156	13	8	8	13	5	17	4	304
157	1	8	10	4	1	9	3	43
158	15	26	5	19	15	6	20	651
159	5	1	12	29	16	9	39	2750
160	6	260	11	1	67	5	293	4176
161	7	1	11	10	1	6	36	3148
162	5	20	13	9	12	94	28	3676
163	10	14	5	36	12	17	21	5004
164	7	8	13	8	6	1	6	338
165	7	1	11	10	1	6	8	156
166	7	1	11	10	1	6	8	156
167	1	17	1	4	1	18	58	433
168	1	8	5	4	15	5	27	
169	4	1	16	19	15	6	20	320
170	1	17	1	4	1	18	58	433
171	2	14	13	2	1	6	19	1092
172	10	20	13	1	43	98	1	1390
173	1	20	13	9	12	94	28	1292
174	7	8	13	8	6	1	6	338
175	8	14	13	14	4	17	4	199
176	7	56	13	42	6	10	6	268
177	2	1	8	2	4	6	1	66
178	2	14	5	29	12	16	3	62
179	8	14	13	14	4	17	4	199
180	6	77	11	1	1	15	72	376
181	1	6	8	6	2	6	4	37
182	7	104	6	1	17	6	1	1536
183	6	77	11	1	1	15	72	376
184	8	17	9	2	1	6	1	191
185	4	1	4	2	4	4	1	81
186	7	22	15	2	10	6	1	180
187	10	9	5	4	1	7	19	303
188	7	15	5	1	1	10	7	595
189	2	14	5	29	12	16	3	62
190	1	8	5	4	5	5	27	13
191	7	14	5	1	8	10	11	132
192	7	436	6	1	2	6	15	6214
193	7	1	11	10	1	6	8	156
194	1	8	5	1	1	1	1	485
195	1	6	8	6	2	6	4	37
196	7	22	15	2	10	6	1	180

197	5	27	7	4	10	10	1	460
198	7	14	11	10	1	6	76	1269
199	1	8	5	4	5	5	27	13
200	7	15	5	1	1	10	7	595
201	16	18	13	4	4	6	10	899
202	6	152	11	1	67	15	72	1296
203	12	8	12	8	1	9	14	1797
204	7	22	15	2	10	6	1	180
205	7	22	15	2	10	6	1	180
206	7	22	15	2	10	6	1	180
207	10	470	9	4	12	287	426	6934
208	15	156	16	19	15	6	20	1392
209	4	1	4	2	4	4	1	81
210	4	1	4	2	4	4	1	81
211	15	26	16	19	15	6	20	236
212	4	1	16	19	15	6	20	320
213	10	18	5	4	5	13	10	205

Appendix I. Vaxijen results.

Seq. no.	Start	End	Length	Antigenic Score
109	2040394	2042433	680	0.7561
199	1045743	1048207	822	0.4735
117	2020019	2022106	696	0.7222
80	2043269	2045308	680	0.7611
81	1827867	1830475	870	0.4763
82	2037677	2039653	659	0.7819
83	2070179	2072278	700	0.7449
84	2036943	2038497	518	0.9704
85	1973805	1975892	696	0.7298
86	2107878	2110793	972	0.7811
87	2030927	2033733	936	0.6315
88	2035162	2037108	649	0.7529
19	1973814	1975840	676	0.7691
89	2075948	2077987	680	0.7611
101	2101225	2103264	680	0.7611
118	2136864	2138838	658	1.0105
8	2011333	2013450	706	0.8
90	2000206	2002397	731	0.5852
91	2038229	2040255	676	0.7234
92	2028999	2031037	680	0.6047
93	2093146	2095196	684	0.7891
94	2055093	2057234	714	0.7626
95	2136261	2139175	972	0.633
10	1947220	1949708	830	0.5918
96	2015824	2017862	680	0.8116
97	2027649	2029673	675	0.7653
98	1972800	1975790	997	0.4184
119	2083723	2086596	958	0.7184
99	2034067	2036198	711	0.5912
9	1960342	1962399	686	0.8207
11	1991275	1993820	849	0.6773
120	2036538	2038748	737	0.7625
121	2136815	2138473	553	0.6043
122	2041042	2043781	913	0.9505
1	1976482	1978977	832	0.495
123	2036922	2039591	890	0.4144
124	2092457	2095258	934	0.8188
125	2057134	2059100	656	0.6501
126	2016531	2018068	513	1.0574
102	2051939	2053995	686	0.7434
127	2081963	2083893	644	0.7927
128	2036837	2038449	538	0.7362
129	2007307	2009346	680	0.7561
12	1943919	1946006	696	0.7295
130	2139715	2141856	714	0.7579
200	1979749	1981368	540	0.6669
131	2118922	2121762	947	0.8149

132	2078089	2080868	927	0.8197
133	2118213	2120239	676	0.6432
134	2058215	2060728	838	0.3867
135	1532024	1534188	722	0.6212
103	2094092	2096413	774	0.7424
136	2060979	2063008	677	0.5698
137	2058664	2060714	684	0.7831
138	2051810	2053789	660	0.7873
139	1987726	1989769	681	0.8506
201	2090416	2092497	694	0.8332
13	1933609	1935678	690	0.7262
140	2088014	2090215	734	0.7867
141	2106261	2109065	935	0.7022
142	2091647	2093749	701	0.7482
143	2055483	2057876	798	0.48
144	1953892	1956162	757	0.7383
104	1663236	1665273	679	0.8791
145	1977301	1980231	977	0.427
146	2173286	2175400	705	0.7238
147	2044279	2046443	722	0.6093
202	65284	67225	647	0.4809
148	1992501	1994588	696	0.7298
149	2006475	2008503	676	0.9757
14	1943201	1945686	829	0.8139
150	2059811	2061850	680	0.7611
152	2088769	2091576	936	0.8207
153	1958395	1960611	739	0.738
154	2027460	2029499	680	0.7561
105	273576	275517	647	0.4809
155	2032903	2035211	770	0.5825
156	2014344	2015924	527	0.6072
203	2038813	2040914	701	0.6516
157	2041319	2043395	692	0.9359
158	2031490	2033529	680	0.7611
159	2022339	2024426	696	0.7222
15	1954875	1956868	665	0.661
160	2113285	2116006	907	0.9378
161	2066124	2068265	714	0.7626
162	2047380	2050409	1010	0.409
163	2003932	2006148	739	0.7372
106	1994989	1997232	748	0.7185
164	2031114	2033264	717	0.7446
204	1057353	1059500	716	0.7335
165	2041992	2044636	882	0.476
166	2068760	2070901	714	0.7906
167	2068138	2070993	952	0.3989
168	1612524	1614477	651	0.4813
169	2034009	2036048	680	0.7561
16	1993498	1995552	685	0.8309
170	2193373	2196156	928	0.4161

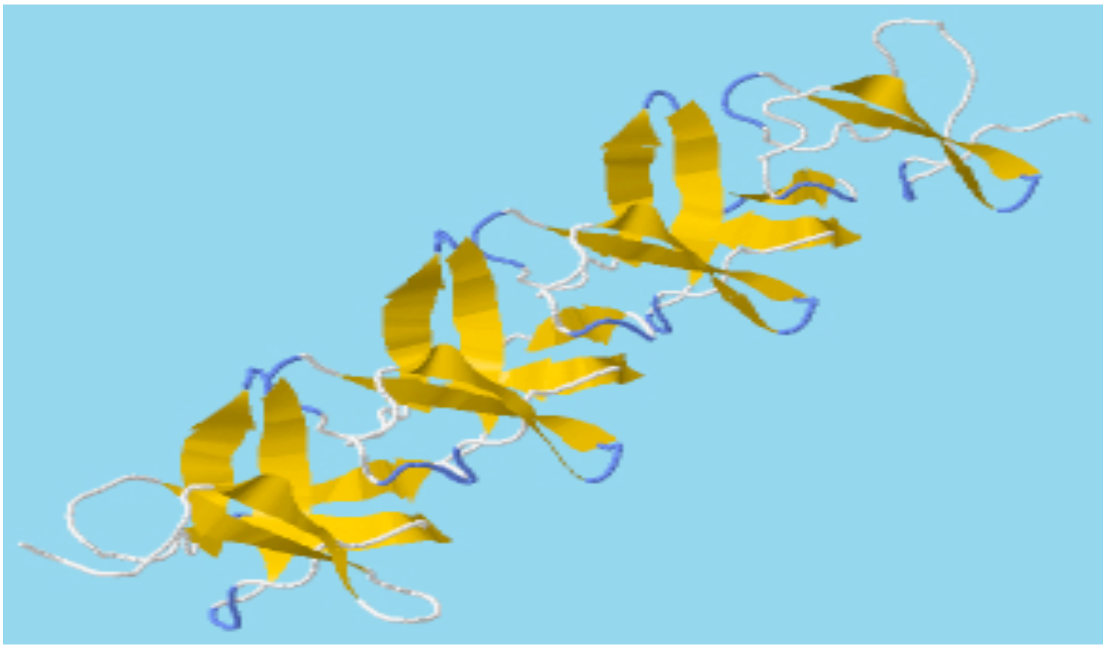
171	2065249	2067379	710	0.9619
172	2090077	2092179	701	0.7474
107	2038692	2040731	680	0.7611
205	1986660	1988675	672	0.9605
173	2030713	2033742	1010	0.4089
174	2116957	2119106	717	0.8045
175	2050671	2052835	722	0.6153
176	2197008	2199110	701	0.7366
177	2069879	2071978	700	0.7433
178	1005047	1007194	716	0.7548
179	2034168	2037344	1059	0.7568
17	1914367	1916105	580	0.9027
180	2091652	2094566	972	0.633
181	2030441	2033173	911	0.7605
191	1152750	1155177	809	0.5068
206	1975833	1977926	698	0.7687
108	819609	821570	654	0.7345
182	2001373	2003565	731	0.7703
183	2092876	2095790	972	0.8287
184	1985740	1987785	682	0.7389
185	2066421	2068562	714	0.7585
186	1986096	1988288	731	0.7916
187	2061360	2063099	580	0.623
188	1987544	1989649	702	0.8138
189	1214300	1216207	636	0.8118
18	370968	373047	693	0.9902
207	2079566	2081489	641	0.6233
190	1198012	1200417	802	0.4774
100	2041006	2042824	606	0.8948
208	2095718	2097610	631	0.7572
110	495690	498359	890	0.4144
209	2093389	2095530	714	0.7626
20	2037096	2039829	911	0.4228
210	1999751	2001892	714	0.7635
211	2061822	2063861	680	0.7611
212	195941	197437	499	0.6164
213	2110015	2112096	694	0.8332
192	1812647	1814609	654	0.4815
21	2002063	2004039	659	0.7819
22	2040504	2042471	656	0.8235
23	2036149	2038188	680	0.7611
24	2096354	2098300	649	0.8295
111	2013418	2015463	682	0.7389
25	2040678	2042717	680	0.7561
26	2026983	2029022	680	0.7561
27	2049806	2051773	656	0.8235
28	2189909	2192707	933	0.7712
29	2047399	2049425	676	0.5872
193	1137342	1139483	714	0.7626
2	1946537	1948588	684	0.741

30	2031138	2033164	676	0.6395
31	2043992	2045886	632	0.8416
32	2134034	2136112	693	0.7899
33	2028632	2030671	680	0.7511
112	2066712	2068679	656	0.8235
34	2034814	2036853	680	0.7561
35	2021945	2023984	680	0.7561
36	2015230	2017269	680	0.7561
37	2078650	2080857	736	0.6817
194	1753749	1755690	647	0.4883
38	1292293	1294758	822	0.4658
39	167795	169354	520	0.4253
3	1962555	1965321	922	0.8929
40	747136	748401	422	0.5134
41	1002497	1004056	520	0.4269
42	1162300	1164357	686	0.7595
113	2048064	2050031	656	0.8235
43	756277	758322	682	0.7389
44	1642022	1643963	647	0.4809
45	1347390	1349357	656	0.8235
195	761979	764574	865	0.9723
46	2158221	2160383	721	0.7502
47	1995044	1997149	702	0.8138
48	2015826	2018600	925	0.406
49	2139758	2141959	734	0.7867
4	1976360	1979130	924	0.8671
50	2046209	2048134	642	0.8071
51	2041048	2043135	696	0.7222
114	2058118	2060181	688	0.7585
52	2037135	2040089	985	0.3814
53	2024808	2026972	722	0.6153
196	840685	842621	646	0.6534
54	2027843	2030152	770	0.7537
55	2095443	2097992	850	0.4338
56	2097885	2100686	934	0.8213
57	1985791	1987982	731	0.5852
58	2124513	2127500	996	0.7765
59	2090589	2092556	656	0.8235
5	1907830	1909758	643	0.6638
60	1989373	1991565	731	0.7703
115	2080623	2082589	656	0.6501
61	1990902	1992878	659	0.7819
197	912598	915624	1009	0.7198
62	2064692	2066719	676	0.8012
63	2091628	2093712	695	0.7613
64	2057132	2059192	687	0.5909
65	2042148	2044340	731	0.7703
66	2037668	2039833	722	0.7683
67	1976352	1978414	688	0.6578

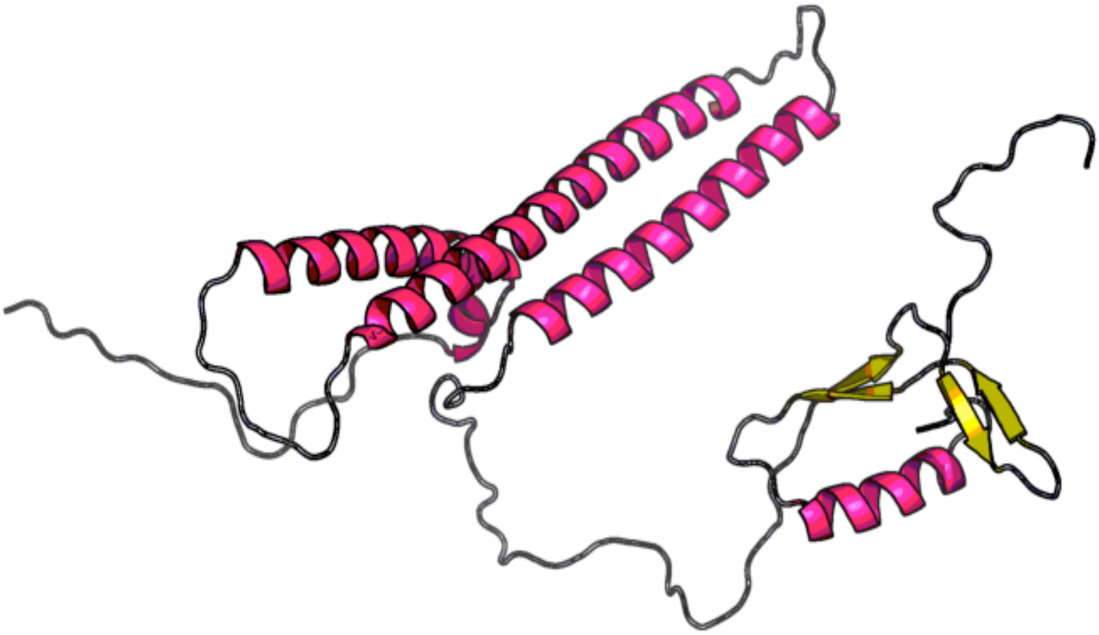
68	2060715	2062796	694	0.8332
69	1979875	1981962	696	0.7298
6	1894817	1898268	1151	0.7892
116	2116874	2118914	680	0.907
198	19507	21562	685	0.976
70	2059441	2061582	714	0.7626
71	2031016	2033055	680	0.7611
72	2068249	2070350	701	0.648
73	2065313	2068107	932	0.8232
74	2111480	2113446	656	0.6501
75	2101458	2103598	714	0.63
77	2017714	2019843	710	0.7242
78	2027782	2030534	918	0.4838
79	2059902	2062043	714	0.7626
7	1968733	1970063	444	0.8635

Appendix J. Selected CbpA structures and their corresponding pneumococcal conserved regions (best domains).

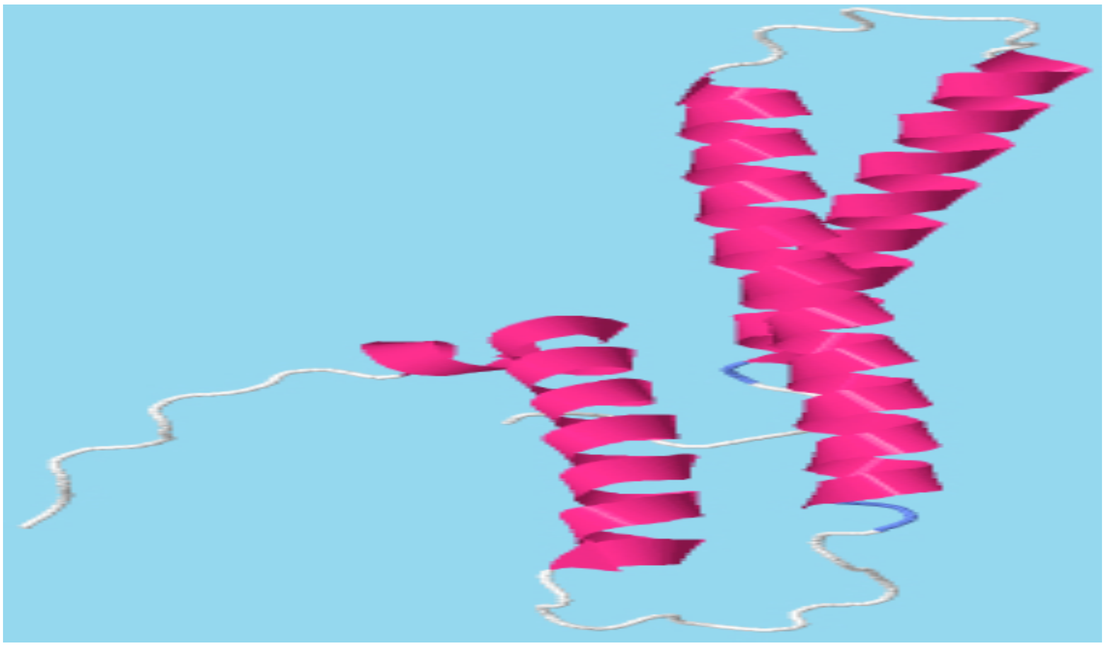
Amino acid sequence 1(CbpA) 3-D structure.



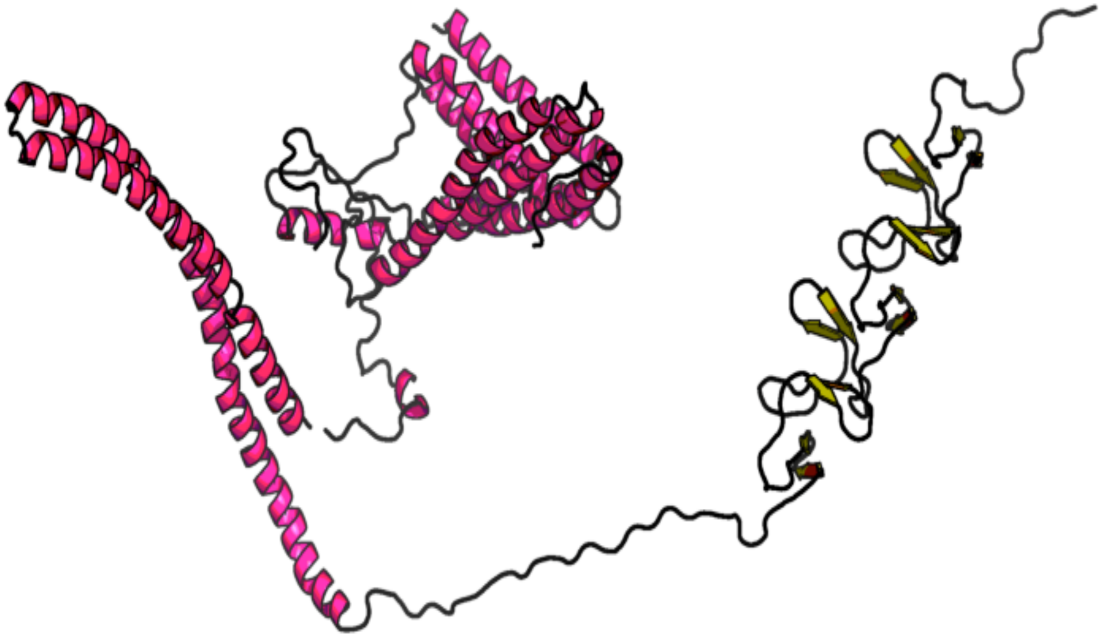
2bib:A 3-D structure.



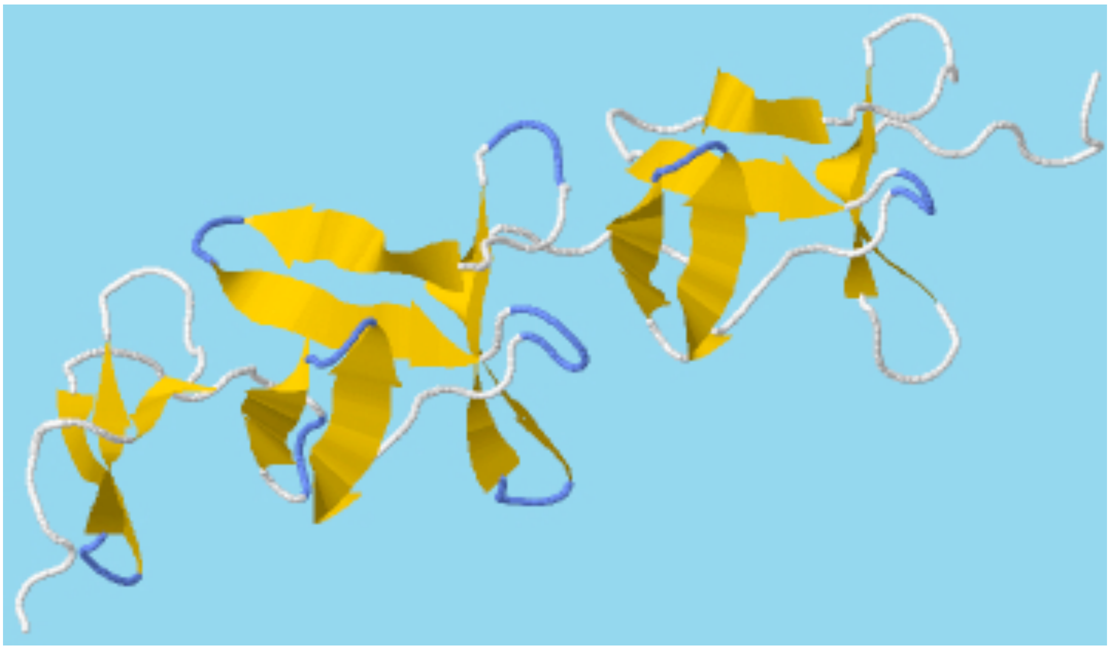
Amino acid sequence 3(CbpA) 3-D structure.



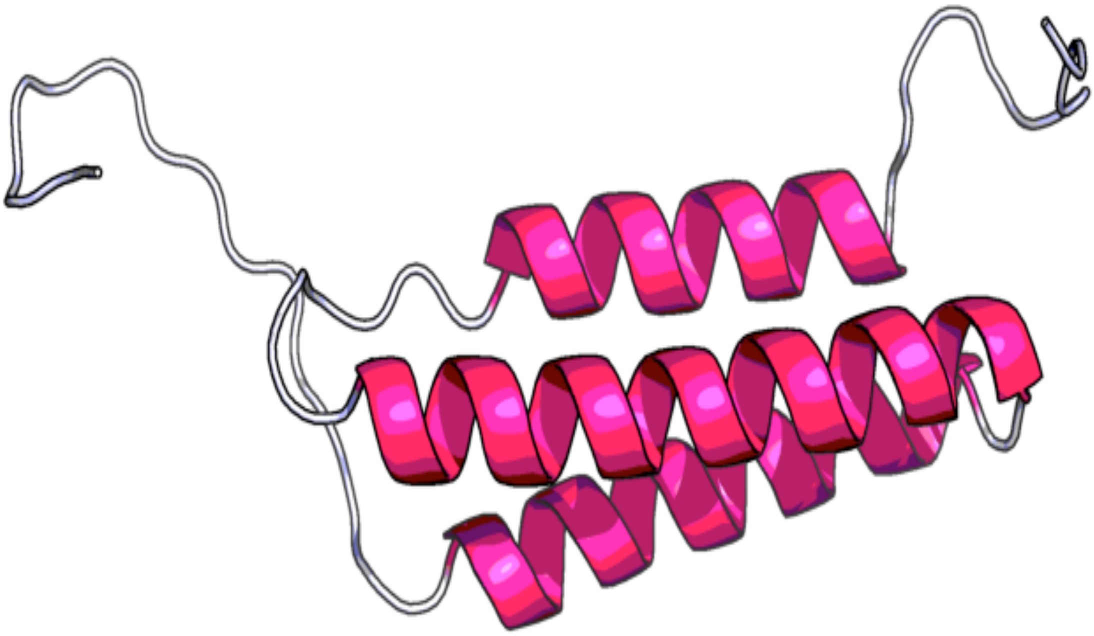
2pms:C 3-D structure.



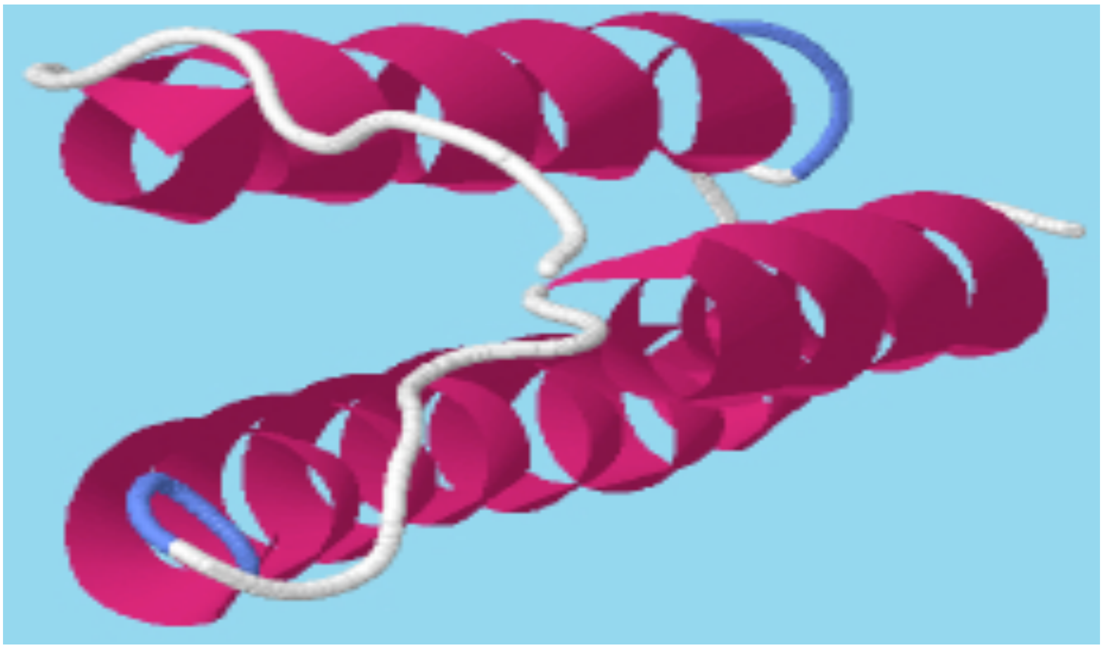
Amino acid sequence 5 (CbpA) 3-D structure.



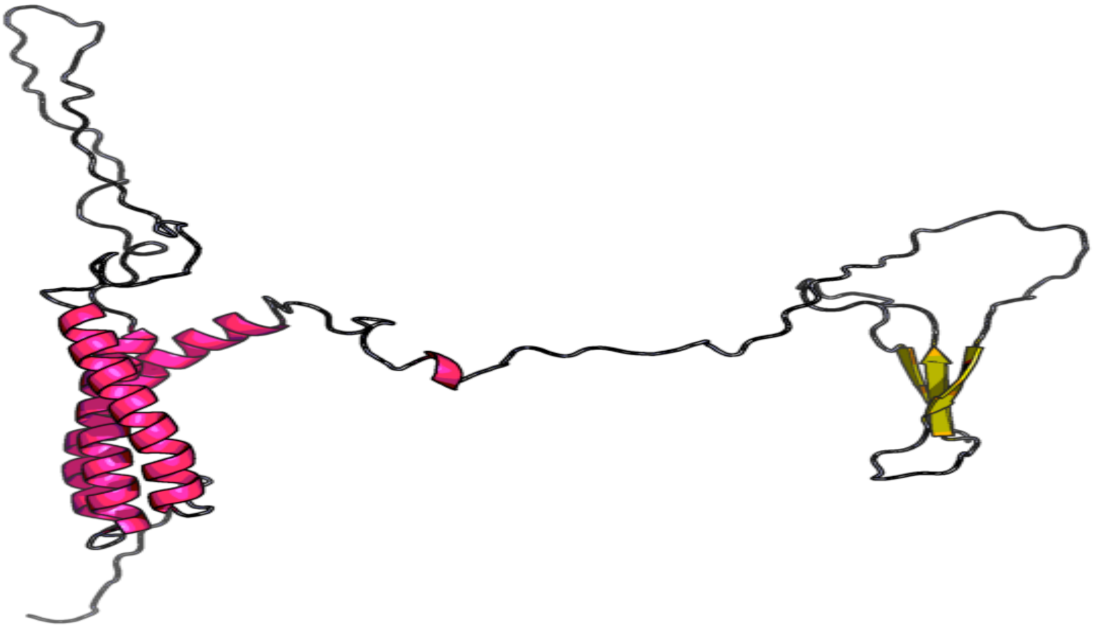
2vyu:A 3-D structure.



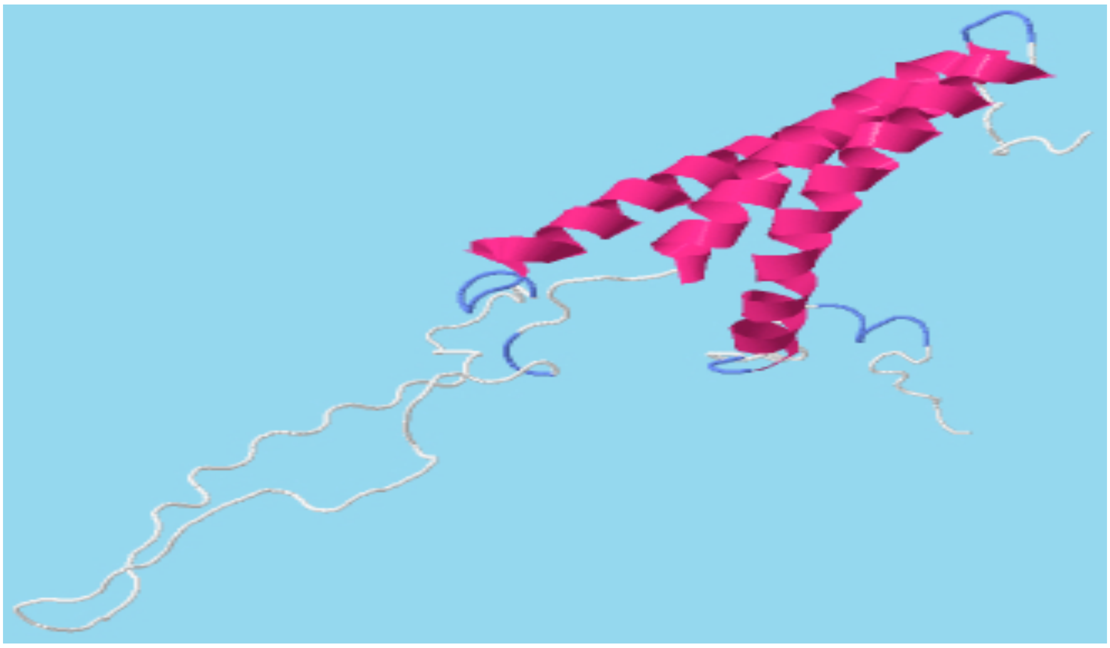
Amino acid sequence 7(CbpA) 3-D structure.



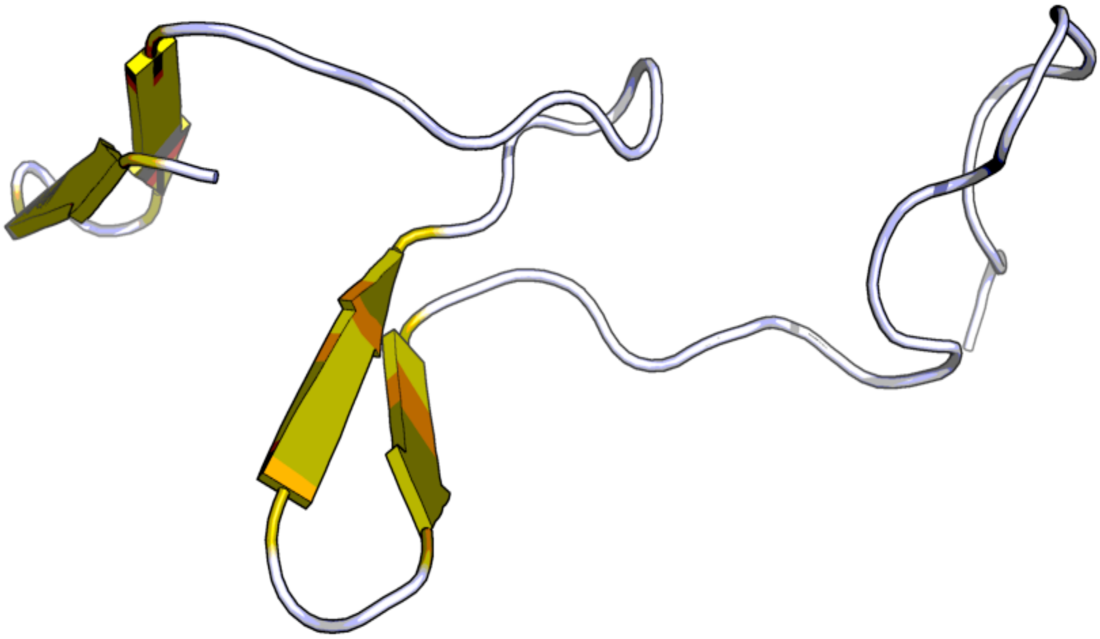
2m6u:A 3-D structure.



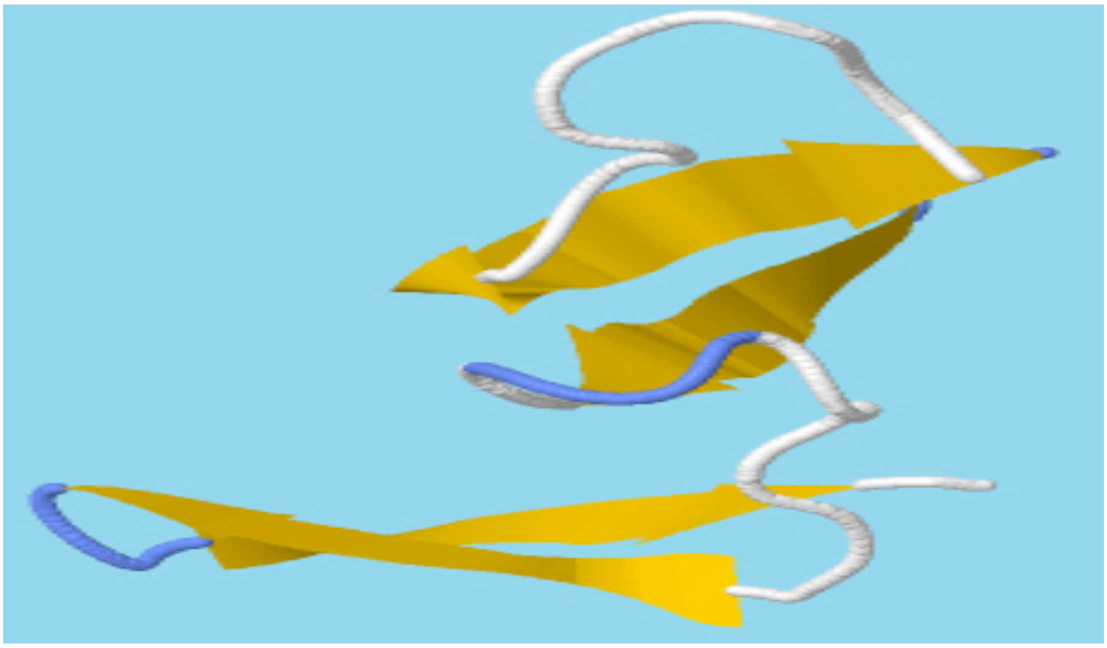
Amino acid sequence 14(cbpA) 3-D structure.



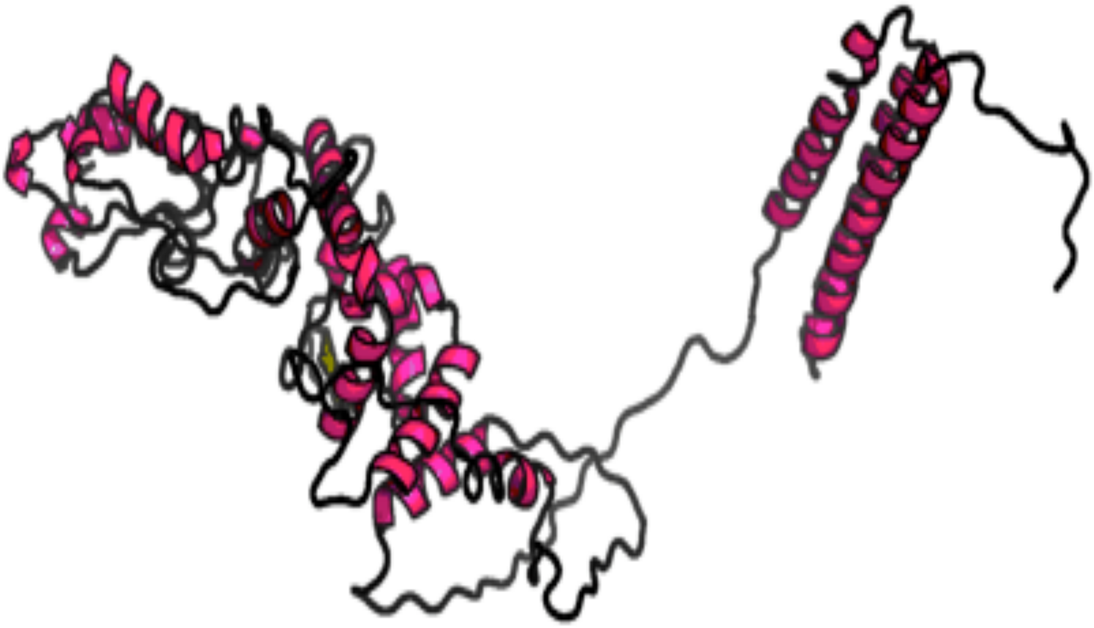
1w9r:A 3-D structure.



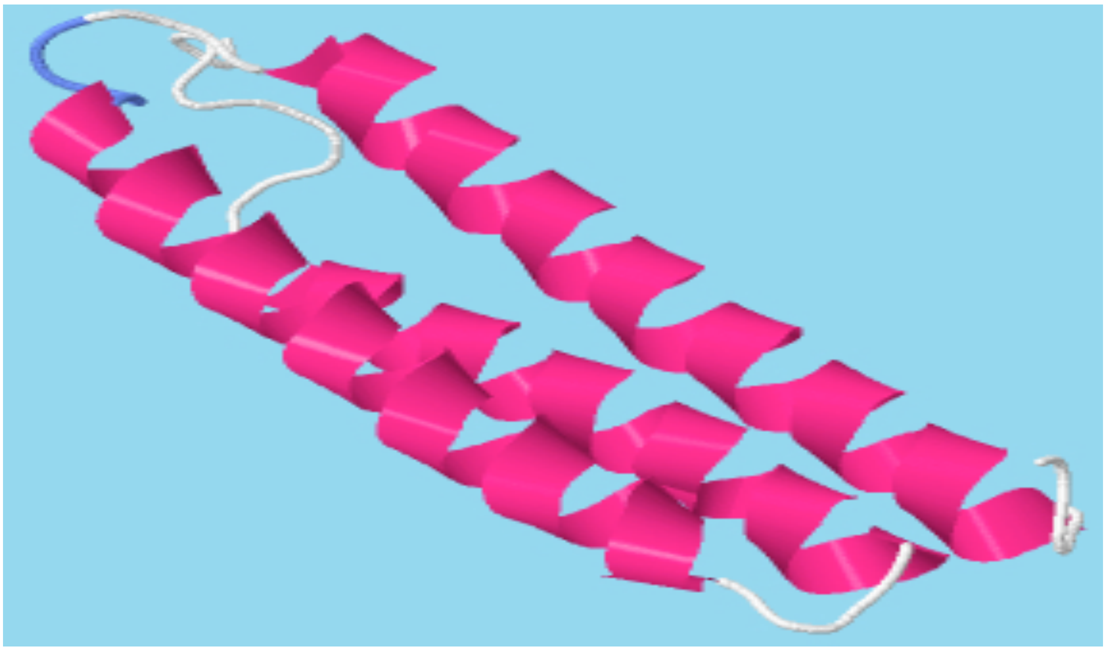
Amino acid sequence 84(CbpA) 3-D structure.



3hia:A 3-D structure.



Amino acid sequence 108 (CbpA) 3-D structure.



4k12:B 3-D structure.

Appendix K. Table of CbpA domains.

Seq.	Domains	Best	Position	P-value	Score	Others
1	4	2bib:A	608-831	1.13E-13	209	1w9r:A , 1jch:A 1w9r:A
2	2	2bib:A	55-683	8.72E-10	208	2h2n:A
3	2	2pms:C	494-628	5.31E-15	50	2vxo:A
4	2	2pms:C	504-638	1.13E-13	40	4np4:A
5	5	2vyuA	447-596	3.09E-12	129	1w9r:A , 2m6u:A 3g7nA , 3k29:A
6	5	2bib:A	459 - 666	4.36E-16	185	1w9r:A,1w9r:A,4np4A 2bibA,4nby:A
7	1	2m6u:A	60-134	2.70E-03	32	
8	4	2bib:A	478-705	1.46E-17	209	1w9r:A , 3kog:A , 1w9r:A
9	1	2bib:A	50-685	2.58E-07	196	
10	4	2bib:A	638 -708	1.15E-09	98	4HpqC , 1qffA , 4np4:A
11	2	3mttA	639-767	1.09E-03	33	4k12:B
12	2	2bib:A	468-694	2.81E-17	200	1w9r:A
13	3	2bib:A	282-679	6.08E-09	143	1w9r:A,2m6u:A
14	2	1w9r:A	469 -646	8.58E-03	52	1wk4:A
15	3	2bib:A	477-650	1.63E-13	102	4hpq:C,1w9r:A
16	1	2bib:A	61-684	5.41E-06	175	
17	2	1w9rA	130-258	8.87E-18	40	2h8n:A
18	1	2m6uA	62 - 153	3.72E-06	18	
19	1	1w9r:A	290 - 418	2.00E-16	44	
20	2	2ixuA	701 -835	1.23E-11	104	2p0w:A
21	4	2bib:A	478-658	1.52E-14	167	1w9r:A,3kog:A,1w9r:A
22	4	2vyu:A	475-655	5.83E-13	150	1w9r:A,1w9r:A,2iho:A
23	2	2bib:A	53-679	3.23E-07	191	2bov:B
24	5	2vyu:A	489-648	6.15E-12	127	1w9r:A,1w9r:A,4k12:B,3g7nA 1uwcA 1dt3A 3o0dA

25	3	2bib:A	475-679	3.03E-16	188	1w9r:A,1w9r:A
26	3	2bib:A	475-679	3.06E-16	188	1w9r:A,1w9r:A
27	4	2vyu:A	475-655	5.83E-13	150	1w9r:A,1w9r:A, 2iho:A
28	2	2bib:A	707-932	3.38E-17	204	2pms:C
29	1	4np4A	392-650	5.03E-08	63	
30	2	1w9r:A	127-255	1.59E-18	42	2bibA 2vyuA
31	2	2vyu:A	463-629	2.64E-12	137	1w9r:A
32	3	2bib:A	466-692	1.33E-16	206	1w9r:A,1w9r:A
33	3	2bib:A	475-679	1.64E-16	188	1w9r:A,1w9r:A
34	3	2bib:A	475-679	3.03E-16	188	1w9r:A,1w9r:A
35	3	2bib:A	475-679	3.03E-16	188	1w9r:A,1w9r:A
36	3	2bib:A	475-679	3.06E-16	188	1w9r:A,1w9r:A
37	4	2vyu:A	473-735	1.18E-14	189	1w9r:A,2imh:A,1w9r:A
38	3	2bib:A	618-821	4.92E-18	193	1w9r:A,1c1g:A
39	1	4kx7:A	1-505	4.64E-03	51	
40	3	1vh4:A	120-270	3.03E-05	63	2vyu:A,2c1l:A
41	2	3fr7:A	48-358	4.38E-03	39	3u44:A
42	4	2bib:A	473-685	8.90E-16	198	1w9r:A,1w9r:A,1i84:S
43	1	2bibA	63-681	8.32E-12	208	
44	3	4biuA	164-450	1.24E-05	237	2wxf:A,3rko:B
45	4	4biu:A	475-655	5.83E-13	150	1w9r:A,1w9r:A,2iho:A
46	1	2bib:A	56-720	6.57E-09	207	
47	3	2bib:A	475-701	4.36E-17	206	1w9r:A,2xnc:A
48	1	2bib:A	722-924	7.63E-18	190	
49	4	2bib:A	507-733	4.26E-18	206	1w9r:A, 2m6u:A,1w9r:A
50	4	2bib:A	475-640	8.15E-13	133	1w9r:A,3vcyA 3sg1A,1w9r:A
51	1	2bib:A	61-695	1.05E-09	209	
52	3	2bib:A	760-984	1.53E-17	209	1w9r:A , 1w9r:A
53	1	4j0xA	572-648	1.14E-02	22	4jxmA

54	1	1w9r:A	165-293	1.75E-11	26	
55	5	2bib:A	667-849	6.18E-16	168	1w9r:A , 2bov:B , 3k29:A,3iox:A
56	3	2bib:A	707-933	1.04E-18	203	1w9r:A , 2pms:C
57	4	4np4:A	487-703	1.02E-06	57	4np4:A,3k29:A,4np4:A
58	3	2bib:A	740-995	6.79E-17	210	2pms:C , 1w9r:A
59	4	2vyu:A	475-655	5.83E-13	150	1w9r:A , 1w9r:A , 2iho:A
60	1	2bib:A	63-730	7.07E-09	210	
61	4	2bibA	52-658	1.52E-14	167	1w9r:A,1w9r:A ,3kog:A
62	2	2bib:A	51-675	1.15E-07	189	2raa:A
63	4	2bib:A	490 - 694	3.23E-16	188	1w9r:A ,1w9r:A , 3kog:A
64	4	2bib:A	341 -570	6.77E-17	208	1w9r:A , 1w9r:A , 1dg3:A
65	1	2bib:A	63 - 730	7.07E-09	210	
66	3	2bib:A	494-721	5.39E-17	206	4h5y:A ,1w9r:A
67	1	1ciiA	1- 617	6.10E-05	67	
68	4	2bib:A	509 - 693	1.24E-14	167	1w9r:A ,1w9r:A ,3lycA 3jx8A
69	3	2bib:A	469 - 695	6.17E-17	206	1w9r:A , 1w9r:A
70	1	2bib:A	59-713	1.27E-07	212	
71	2	2bib:A	53 - 679	3.23E-07	191	2bov:B
72	3	1w9r:A	133 - 261	1.38E-15	38	4h5y:A,4np4:A
73	1	4h5y:A	54-182	2.28E-03	49	
74	4	4np4:A	479-637	6.93E-10	85	1w9r:A , 4hpq:C , 1jad:A
75	1	2bib:A	469 - 671	2.23E-17	95	
77	1	2bib:A	64 - 709	3.39E-09	206	
78	2	4f61I	232-415	3.56E-04	70	3lycA 3petA 3jx8A
79	1	2bib:A	59 - 713	1.27E-07	212	
80	2	2bib:A	53 - 679	3.23E-07	191	2bov:B
81	4	3k29A	359-496	8.15E-04	73	3k29:A,3lycA 3petA,3odtA
82	2	2bib:A	52-658	8.20E-07	168	3kog:A

83	2	2bib:A	187-699	9.43E-11	204	3k29:A
84	1	3hia:A	443 - 482	3.45E-05	31	
85	3	2bib:A	469 -695	6.17E-17	206	1w9r:A , 1w9r:A
86	3	2vyuA	716 - 971	7.14E-16	195	1w9r:A , 2pms:C
87	3	3tnf:B	301 - 610	1.48E-07	100	4j0xA,4ddqA
88	2	2bib:A	444 - 648	3.97E-17	188	1w9r:A
89	2	2bib:A	53 - 679	3.23E-07	191	2bov:B
90	1	4np4:A	487 - 703	1.92E-13	125	
91	2	1w9r:A	124 - 252	5.49E-21	47	2bibA 2vyuA
92	1	4np4:A	442-628	6.27E-13	101	
93	3	4np4:A	493 - 673	3.19E-11	113	1w9r:A ,3k29:A
94	1	2bib:A	59-713	1.27E-07	212	
95	1	3sn6B	456-931	1.15E-05	185	
96	2	4np4:A	474-667	1.16E-12	113	1w9r:A
97	5	2bib:A	462 - 665	1.47E-16	189	1w9r:A , 1jad:A , 2vsa:A , 3k29:A
98	1	2bib:A	771 - 996	1.96E-19	209	
99	4	4np4:A	528 -594	1.14E-06	55	2bib:A ,3k29:A,4np4A
100	3	2bib:A	464-580	4.34E-12	115	1w9r:A, 1c1g:A
101	2	2bib:A	53-697	3.23E-07	191	2bov:B
102	3	2bib:A	545-654	2.00E-08	68	3k29:A , 4np4:A
103	1	2vyu:A	71-773	1.23E-07	194	
104	2	1w9r:A	126-254	1.58E-15	33	1a87:A
105	3	4biu:A	129 - 450	4.10E-05	237	2a65:A , 3rko:B
106	1	2bib:A	1-747	3.23E-07	211	
107	2	2bib:A	53 - 679	3.23E-07	191	2bov:B
108	3	4k12:B	58- 149	1.24E-09	20	1z6t:A ,4mhbA
109	3	2bib:A	475-679	3.03E-16	188	1w9r:A, 1w9r:A

110	4	2bib:A	664-889	4.93E-18	208	1w9r:A , 1w9r:A , 4hpq:C
111	1	2bib:A	63-681	8.32E-12	205	
112	4	2vyu:A	475-655	5.83E-13	112	1w9r:A,1w9r:A,2iho:A
113	4	2vyu:A	475-655	5.83E-13	150	1w9r:A,1w9r:A,2iho:A
114	4	2bib:A	460-687	5.49E-17	210	1w9r:A , 1w9r:A , 1ei3:B
115	4	4np4A	127-255	2.98E-10		4hpq:C , 4np4A , 2bibA 2vyuA , 4f61:I
116	2	1w9r:A	129 - 257	2.18E-18		2rd0A 2wxfA
117	4	2bib:A	470-695	2.56E-17	204	1w9r:A,1dg3:A,1w9r:A
118	1	1w9r:A	357-485	4.64E-17	37	
119	3	2vyuA	688 - 943	4.41E-15	190	2pms:C , 1w9r:A
120	1	2bib:A	57-735	2.66E-07	218	
121	3	2bib:A	327-552	7.62E-17	204	1w9r:A 1jad:A
122	1	2p0wA	713-864	1.70E-03	35	
123	5	2bib:A	664-889	4.93E-18	208	1w9r:A, 1w9r:A,3esi:A, 4hpq:C
124	2	2bib:A	708-933	4.78E-18	202	2pms:C
125	4	4np4A	460-637	2.98E-10	85	1w9r:A, 4hpq:C, 4f61:I
126	1	2m6uA	70-136	2.72E-02	31	
127	1	1w9r:A	77-205	9.11E-22	49	
128	1	1cii:A	61-503	2.27E-05	124	
129	3	2bib:A	475-679	3.03E-16	188	1w9r:A ,1w9r:A
130	3	2bib:A	486-713	1.98E-17	207	4k12:B,1w9r:A
131	2	2bib:A	721-946	7.77E-19	205	2pms:C
132	2	1w9r:A	130-258	7.02E-21	44	3no2A 3vgzA 1l0qA 4o9dA
133	4	1w9r:A	127-255	6.10E-17	42	1cii:A,2bibA 2vyuA,4gnk:E
134	3	2bib:A	612-837	6.10E-17	207	1w9r:A,3k29:A
135	3	3tnfB	59-488	4.98E-08	87	2bry:A,1gxrA 2ymuA 3ei1B 3fm0A 4lg8A
136	1	1u4qA	335-641	1.35E-03	128	

137	1	1w9r:A	128-256	3.01E-21	48	
138	1	2bib:A	52-659	3.96E-06	172	
139	1	1a87A	574-631	1.78E-02	22	
140	4	2bib:A	507-733	4.26E-18	206	2m6u:A 1w9r:A 1w9r:A
141	2	2bib:A	684-907	7.16E-17	204	2pms:C
142	4	2bib:A	475-700	1.21E-17	203	1w9r:A,1w9r:A,2imh:A
143	4	2bib:A	594-797	2.44E-17	196	1w9r:A 2pn5:A 1c1g:A
144	4	2vyu:A	512-756	1.64E-14	188	1w9r:A 1ei3:B 1w9r:A
145	1	2bib:A	773-976	6.76E-18	190	
146	3	2bib:A	477-704	4.07E-17	209	1w9r:A , 1w9r:A
147	1	2ymuA	527-649	1.07E-07	58	
148	3	2bib:A	469-695	6.17E-17	206	1w9r:A, 1w9r:A
149	2	1w9r:A	92 -220	1.96E-12	25	1a87:A
150	2	2bib:A	53 - 679	3.23E-07	191	2bov:B
152	2	2bib:A	710 - 935	6.74E-18	201	2pms:C
153	1	1w9r:A	165 - 293	3.53E-12	27	
154	3	2bib:A	475 -679	3.03E-16	188	1w9r:A , 1w9r:A
155	2	4hpqC	315-734	6.41E-05	205	1w9r:A
156	3	2bib:A	343 - 526	1.22E-14	171	1w9r:A, 4k12:B
157	1	3hvaA	1-136	9.52E-03	52	
158	2	2bib:A	53 - 679	3.23E-07	191	2bov:B
159	4	2bib:A	470-695	2.56E-17	204	1w9r:A,1dg3:A,1w9r:A
160	2	2p01A	108-450	2.19E-17	32	1f8n:A
161	1	2bib:A	59-713	1.27E-07	212	
162	4	2bib:A	786-1009	6.64E-18	210	1w9r:A , 3kog:A ,1cii:A
163	1	1w9r:A	165 - 293	9.58E-12	26	
164	2	2bib:A	490-716	5.61E-17	202	1w9r:A
165	2	4hpq:C	73-512	1.46E-06	131	4c9b:B
166	3	2bib:A	471 - 694	1.66E-17	210	1w9r:A,1c1g:A

167	1	2bib:A	726 - 951	1.05E-18	209	
168	4	4biu:A	164-450	1.24E-05	238	2d4y:A,4ap2:B,1pcxA 2nupB
169	3	2bib:A	475 -679	3.03E-16	188	1w9r:A , 1w9r:A
170	3	2bib:A	702 - 927	8.51E-19	209	1w9r:A , 1c1g:A
171	2	1w9r:A	132-259	6.79E-15	27	4anuA 2y3aA 2rd0A 2wxfA
172	1	2bib:A	57-700	1.86E-09	204	
173	3	2bib:A	786 - 1009	7.04E-18	210	1w9r:A , 3kog:A
174	5	4np4:A	565 - 629	2.02E-06	58	1w9r:A,4np4A 4nbyA 2ixvA,3k29:A,2bibA 2f6eA 2ww5A 2vyuA 4np4A
175	1	4j0xA	572-648	1.14E-02	22	
176	1	2bib:A	19 -700	5.85E-08	211	
177	3	2bib:A	474 - 699	1.05E-16	206	1w9r:A , 1w9r:A
178	1	2bib:A	18 -715	1.33E-08	211	
179	3	2bib:A	791 - 1017	1.31E-17		2pms:C,2mii:A
180	2	4lg9A	669-931	1.60E-11	156	4hpq:C
181	1	2bib:A	436 -908	3.95E-10	204	
182	1	2bib:A	63 -730	7.07E-09	210	
183	1	1w9r:A	130 - 258	4.36E-20	47	
184	1	2bib:A	63-681	8.32E-12	208	
185	3	2bib:A	487 - 713	2.03E-17	207	1w9r:A , 1w9r:A
186	1	2m6u:A	58 - 149	1.61E-10	23	
187	3	2bib:A	354 - 579	8.58E-17	206	1w9r:A , 2m6u:A
188	3	2bib:A	475-701	4.36E-17	206	2xnc:A,1w9r:A
189	4	2vyu:A	471 - 635	4.30E-12	134	1w9r:A , 1w9r:A , 1dg3:A
190	4	2bib:A	618 - 801	8.16E-16	176	1w9r:A , 2ww9:A , 1c1g:A
191	1	1w9r:A	109-237	4.15E-10	21	

192	3	4biu:A	127-454	6.04E-06	234	3mk4:A, 3ptrB 3n9lA 3kv9A 3kv4A
193	1	2bib:A	59-713	2.27E-17	207	
194	3	2bib:A	29-450	1.20E-05	237	1jpr:A, 3vld:A
195	1	1a87:A	709-766	1.60E-02	22	
196	3	4k12:B	58-149	3.81E-09	18	4av3:A , 3zge:A
197	1	2bib:A	785-1008	5.52E-17	207	
198	2	1w9r:A	286-414	1.03E-15	44	4anu:A
199	2	4HpqC	374 -600	1.20E-04	127	2Ymu:A
200	4	2bib:A	314-538	4.18E-16	195	1w9r:A , 1jad:A , 2j3z:A
201	4	2bib:A	509-693	1.24E-14	167	1w9r:A , 2m6u:A, 1w9r:A
202	3	4biu:A	129-450	1.24E-05	237	2a65:A , 3rko:B
203	3	1w9r:A	133 -261	5.95E-16	38	4np4:A , 4o9bA 3k29A ,
204	1	2bib:A	1-750	6.67E-17	208	
205	2	2m6u:A	58-149	8.22E-15	34	2vbe:A
206	2	2m6u:A	58-149	1.24E-11	26	2vbe:A
207	4	2vyu:A	299-468	1.46E-12	140	1w9r:A , 2cxa:A , 4nby:A
208	2	2bib:A	51-630	1.78E-08	196	2imh:A
209	3	2bib:A	486-713	2.27E-17	207	1w9r:A , 4k12:B , 3p52:A
210	1	2bib:A	60 - 713	1.69E-08	207	
211	2	2bib:A	53-679	3.23E-07	191	2bov:B
212	2	2bib:A	315-498	3.15E-14	166	1ww9r:A
213	4	2bib:A	509-693	1.24E-14	167	1w9r:A,1w9r:A,3lycA 3jx8A

Appendix L: Table of 30 best protein domains predicted.

No.	Code	Protein description	Species
1	2bib:A	Modular teichoic acid phosphorylcholine esterase(CBPE)	*S.pneumoniae*
2	2pms:C	Lactoferrin-binding domain of pspA	*S.pneumoniae*
3	2vyuA	Choline binding protein F(CbpF)	*S.pneumoniae*
4	2m6u:A	Choline binding protein A(CbpAN)	*S.pneumoniae*
5	3mttA	Phosphatidylinositol 3-kinase regulatory subunit beta	*Homo sapiens*
6	1w9r:A	Choline binding protein A(Domain r2)	*S.pneumoniae*
7	2ixuA	CPL-1 endolysin	*Streptococcus phage Cp-1*
8	4np4A	Toxin B	*Clostridium difficile*
9	4kx7:A	Glutamyl aminopeptidase	*Homo sapiens*
10	1vh4:A	SufD protein	*E.coli*
11	3fr7:A	Putative ketol-acid reductoisomerase	*Oryza sativa*
12	4biuA	Sensor protein(CPXA)	*E.coli*
13	4j0xA	Ribosomal RNA-processing protein 9	*Saccharomyces cerevisiae S288c*
14	1ciiA	COLICIN IA	*E.coli*
15	4h5y:A	LidA protein, substrate of the Dot/Icm system	*Legionella pneumophila*
16	4f61I	Tubulin alpha chain	*Ovis aries*
17	3k29A	Putative uncharacterized protein	*Chlamydia trachomatis*
18	3hia:A	Choline binding protein	*S.pneumoniae*
19	3tnf:B	Ras-related protein Rab-8A	*Homo sapiens*
20	3sn6B	Guanine nucleotide-binding protein G(s) subunit	*Bos taurus*
21	4k12:B	Choline binding protein A	*S.pneumoniae*

22	2p0wA	Histone acetyltransferase type B catalytic subunit	*Homo sapiens*
23	1u4qA	Spectrin alpha chain	*Gallus gallus*
24	1a87A	COLICIN N	*E. coli k-12*
25	2ymuA	WD-40 REPEAT PROTEIN	*Nostoc punctiforme*
26	4hpqC	Atg31	*Lachancea thermotolerans*
27	3hvaA	Protein FimX	*Aerugenosa pseudomonas*
28	2p01A	Alpha-2-macroglobulin receptor-associated protein	*Homo sapiens*
29	4lg9A	F-box-like/WD repeat-containing protein TBL1XR1	*Homo sapiens*
30	4HpqC	Atg31	*Lachancea thermotolerans*

Appendix M: Published manuscript associated with my research work.