

AMERICAN UNIVERSITY OF BEIRUT

ALGORITHMIC AND COMPUTATIONAL MODELS FOR
PROBING CONGENITAL HEART DISEASES THROUGH
TLL1 GENE AND GATA TRANSCRIPTION FACTOR

by
ATLAL MOHAMMAD EL-ASSAAD

A dissertation
submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
to the Department of Electrical and Computer Engineering
of the Faculty of Engineering and Architecture
at the American University of Beirut

Beirut, Lebanon
April 2016

AMERICAN UNIVERSITY OF BEIRUT

ALGORITHMIC AND COMPUTATIONAL MODELS FOR
PROBING CONGENITAL HEART DISEASES THROUGH
TLL1 GENE AND GATA TRANSCRIPTION FACTOR

by
ATLAL MOHAMMAD EL-ASSAAD

Approved by:

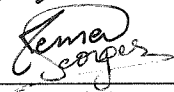
Dr. M. Adnan Al-Alaoui, Professor
Electrical and Computer Engineering


Chair

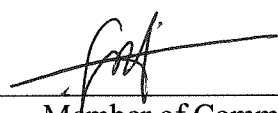
Dr. Zaher Dawy, Professor
Electrical and Computer Engineering


Advisor

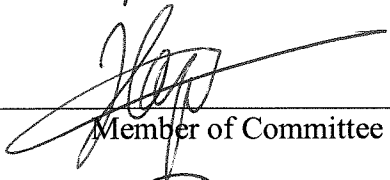
Dr. Georges Nemer, Professor
Biochemistry and Molecular Genetics (AUB)


Co-Advisor

Dr. Fadi Karamah, Associate Professor
Electrical and Computer Engineering


Member of Committee

Dr. Hazem Hajj, Associate Professor
Electrical and Computer Engineering


Member of Committee

Dr. Firas Kobeissy, Assistant Professor
Biochemistry and Molecular Genetics (AUB)


Member of Committee

Dr. Nashat Mansour, Professor
Computer Science, Lebanese American University


Member of Committee

Prof. Tayssir Hamieh, Professor
Faculty of Science, Lebanese University


Member of Committee

Date of dissertation defense: April 8th, 2016

AMERICAN UNIVERSITY OF BEIRUT

THESIS, DISSERTATION, PROJECT RELEASE FORM

Student Name:

Last

First

Middle

Master's Thesis
Dissertation

Master's Project

Doctoral

I authorize the American University of Beirut to: (a) reproduce hard or electronic copies of my thesis, dissertation, or project; (b) include such copies in the archives and digital repositories of the University; and (c) make freely available such copies to third parties for research or educational purposes.

I authorize the American University of Beirut, to: (a) reproduce hard or electronic copies of it; (b) include such copies in the archives and digital repositories of the University; and (c) make freely available such copies to third parties for research or educational purposes

after :

One ---- year from the date of submission of my thesis, dissertation, or project.

Two ---- years from the date of submission of my thesis, dissertation, or project.

Three ---- years from the date of submission of my thesis, dissertation, or project.

Signature

Date

ACKNOWLEDGMENTS

My first gratitude and sincere appreciation is addressed to my advisor, Prof. Zaher Dawy, for his continuous support and advice throughout this work. I appreciate all the time, effort, and patience he put into this.

My second appreciation is addressed to my insightful co-advisor, Prof. Georges Nemer. I am grateful for his continuous support. I also extend my appreciation to the dissertation committee: Prof. Adnan Al-Alaoui, Prof. Hazem Hajj, Prof. Fadi Karamah, Prof. Firas Kobeissy, Prof. Nashaat Mansour, and Prof. Tayssir Hamieh. Thank you for your constructive comments and advice that greatly enhanced the quality of this work.

I also express my warmest gratitude to my sisters and brother who helped me in many ways. Thank you for your continuous support and patience all through. I am so blessed to have such sisters and brother by my side.

Last but not least, I express my deepest gratitude to my parents, whose life-long love, guidance, and encouragement were the key factors of my success throughout my life.

This work was supported by a grant from the National Council for Scientific Research (CNRS) in Lebanon [Grant title: Computational Modeling of DNA Repair Mechanism (MMR) for Understanding Defects by MMR Leading to Congenital Heart Diseases (CHD)] and by a Biomedical Research Program Project within AUBMC [Grant title: The Identification of the “Compensatory Septal Proteome” (CSP) and Tll1 Substrates Involved in Cardiac Septation: Potential for Cardiac Defect Biomarkers.

AN ABSTRACT OF THE DISSERTATION OF

Atlal Mohammad El-Assaad for Doctor of Philosophy
Major: Machine Intelligence and Software
Engineering, Biomedical Engineering

Title: Algorithmic and Computational Models for Probing Congenital Heart Diseases through TLL1 Gene and GATA Transcription Factor

Degradomics - the proteomics analysis of proteases - is reforming the understanding of proteases function. By revealing their substrate repertoire, also called the substrate degradome, the crucial biological roles of proteases is becoming discoverable. Thus, an interesting utility of degradomics is the outcome of protein biomarkers whose role can be symbolic like calpain and caspase proteases, injurious like Matrix Metalloproteinases (MMP-2 and MMP-9), or constructive like the Tll1 gene, depending on the corresponding biological process. In this thesis, we elaborate on the role of the Tll1 protease in Congenital Heart Disease (CHD) and the role of calpain and caspase proteases in brain injury and neuronal cell death types.

It has been a challenge to identify the protease cleaved fragments with high precision and efficiency. Recently, advanced proteomics techniques have shown a remarkable progress in identifying them experimentally. We present in this thesis a detection method that identifies them accurately and efficiently, with validation against experiments from the literature. The method aims at predicting the consensus sequence occurrences and their variants in a large set of experimentally detected protein sequences, based on state-of-the-art sequence matching and alignment algorithms. After detection, the method generates all the potential cleaved fragments. This space and time efficient algorithm is flexible to handle the different orientations that the protein and consensus sequences can take before cleavage by the protease. Subsequently, this knowledge will feed into the development of a novel web tool for researchers to detect the diverse types of biomarkers online, and that will guide in the diagnosis and treatment of related diseases.

Protein-DNA interactions are of fundamental importance in molecular biology, playing roles in functions as diverse as DNA transcription, DNA structure formation, and DNA repair. Protein-DNA association is also important in medicine and understanding protein-DNA binding can assist in identifying disease root causes, contributing to drug

development. In this perspective, we focus on the transcription process by the GATA Transcription Factor (TF) GATA4, which has possible implications in CHD. GATA TF binds to DNA promoter region represented by ‘G, A, T, A’ nucleotides sequence, and initiates transcription of target genes. When proper regulation fails due to some mutations on the GATA TF protein sequence, or due to some mutations on the DNA promoter sequence (weak promoter), deregulation of the target genes might lead to various disorders. In this thesis, we aim to understand the electrostatic mechanism behind GATA TF and DNA promoter interactions, in order to predict protein-DNA binding in the presence of mutations, while elaborating on non-covalent binding. To generate a family of mutants for the GATA:DNA complex, we replaced every charged amino acid, one at a time, with a neutral amino acid like Alanine (Ala). We then applied Poisson-Boltzmann electrostatic calculations feeding into free energy calculations, for each mutation. These calculations delineate the contribution to binding from each Ala-replaced amino acid in the GATA:DNA interaction complex. After analyzing the obtained data in view of a two-step model, we are able to identify potential key amino acids in binding. Finally, we applied the model to GATA3:DNA (crystal structure with PDB-ID: 3DFV) binding complex and validated it against experimental results from the literature.

CONTENTS

ACKNOWLEDGEMENTS.....	v
ABSTRACT.....	vi
LIST OF ILLUSTRATIONS.....	xii
LIST OF TABLES.....	xiv

Chapter

1. INTRODUCTION.....	1
1.1. Dissertation Objectives	1
1.1.1. First Dissertation Objective	2
1.1.2. Second Dissertation Objective	4
1.2. Dissertation Organization	6
2. BACKGROUND	7
2.1. Congenital Heart Disease (CHD).....	7
2.1.1. Tll1 Gene	8
2.1.2. GATA Family	8
2.1.3. GATA4 Transcription Factor	9
2.2. Neuronal Cell Death Types	10
2.2.1. Apoptosis and Necrosis	10
2.2.2. Calpain and Caspase Proteases	11
2.3. Degradomics	14

2.4. Sequence Matching Algorithms	17
2.4.1. Smith-Waterman (SW) Algorithm	18
2.4.2. Other Algorithms	19
2.5. Force Field in Protein-DNA Interactions	22
2.5.1. Intramolecular Bonded Interactions	23
2.5.1.1. Bond Stretching	23
2.5.1.2. Angle Bending	24
2.5.1.3. Torsion/Twisting	25
2.5.2. Inter- and Intramolecular Non-bonded Interactions	25
2.5.2.1. Electrostatic Interactions	26
2.5.2.2. van der Waals (vdW) Forces	28
2.5.2.3. Solvation Effect	29
3. DEGRADOMICS METHODS AND RESULTS	31
3.1. Cleaved Fragments Prediction Algorithm with Application to Tll1....	31
3.1.1. Matches, Mismatches, and Pruning of Gaps	35
3.1.2. Excluding Paths with Three-or-More-Mismatches.....	37
3.1.3. Sub-Matches within Two Mismatches and Overlaps.....	38
3.1.4. INDEL after Consensus Occurrences.....	40
3.1.5. Handling of 4-Way Protein and Consensus Orientations...	41
3.1.5.1. Protein and Consensus Initial Orientations (NN).....	42
3.1.5.2. Protein Initial Orientation and Consensus Orientation Reversed (NC).....	42
3.1.5.3. Protein Orientation Reversed and Consensus Orientation Reversed (CN).....	43
3.1.5.4. Protein and Consensus Orientations Reversed (CC).....	44
3.1.6. Application to Tll1 Metalloprotease.....	45
3.1.6.1. Problem Definition.....	45
3.1.6.2. Data: Mouse Genome.....	46
3.1.6.3. Computational Results.....	46
3.1.6.4. Experimental Validation.....	49
3.1.6.5. Summary.....	49

3.2. Cleaved Fragments Prediction Algorithm with Applications.....	52
3.2.1. CFPA-CalpCasp Algorithm Functionality.....	54
3.2.2. α II-spectrin Application to Calpain and Caspase Proteases	55
3.2.2.1. Problem Definition.....	55
3.2.2.2. Data: α II-spectrin.....	57
3.2.2.3. Computational Results.....	58
3.2.2.4. Experimental Validation.....	62
3.2.3. β II-spectrin Application to Calpain and Caspase Proteases	62
3.2.3.1. Problem Definition.....	62
3.2.3.2. Data: β II-spectrin.....	63
3.2.3.3. Computational Results.....	64
3.2.3.4. Experimental Validation.....	69
3.2.4. Summary.....	70
4. PROTEIN-DNA METHODS AND RESULTS.....	72
4.1. Protein-DNA Models	74
4.2. Binding Free Energy Calculations	74
4.3. Application to Charged-Mutants GATA3.....	76
4.3.1. Problem Definition	76
4.3.2. Data: GATA3 Crystal Structure	76
4.3.3. GATA3 Binding Energy Calculations	78
4.3.4. GATA3 Intermolecular Contacts	79
4.3.5. GATA3 Mutational Analysis	80
4.3.6. GATA3 Experimental Validation	81
4.3.7. Summary	81
4.4. Application to All-Mutants GATA3.....	82
4.4.1. Problem Definition	82
4.4.2. GATA3 Binding Energy Calculations.....	83
5. CONCLUSION AND FUTURE DIRECTIONS	86
REFERENCES.....	90

Appendix.....	106
1. Instances of T111 Fragments Generated by CFPA.....	106
2. All Combinations Generated by CFPA-CalpCasp.....	110
3. β II-spectrin Cleavage by Calpain-2.....	116

ILLUSTRATIONS

Figure	Page
2.1. Ventricular septal defect (VSD) type of CHD.....	7
2.2. Atrial septal defect (ASD) type of CHD.....	8
2.3. Smith-Waterman algorithm.....	18
3.1. Simulated data for matches, mismatches, and pruning of Gaps.....	35
3.2. Dynamic table comprising scoring and alignments.....	36
3.3. Indexes along the consensus and the protein.....	36
3.4. Simulated data for excluding 3-or-more mismatches.....	37
3.5. Two paths, one mismatch, and three mismatches.....	38
3.6. Simulated data for sub-matches within 2 mismatches and overlaps.....	39
3.7. Two accepted paths, one Mismatch, and two mismatches.....	39
3.8. Simulated data for INDEL after consensus occurrences.....	40
3.9. Alignment with an INDEL at the last base.....	41
3.10. Simulated data for protein-consensus NN orientations.....	42
3.11. Simulated data for protein-consensus NC orientations.....	43
3.12. Simulated data for protein-consensus CN orientations.....	44
3.13. Simulated data for protein-consensus CC orientations.....	44
3.14. Histogram of consensus occurrences with all types of matches.....	48
3.15. Protein sequences handled through NN-CC and NC-CN.....	48
3.16. Control neurons (α II-spectrin) undergoing cell death pathways.....	56
3.17. Cleavage sites of α II-spectrin (PDB ID: 2FOT crystal structure).....	57

3.18.	Cleavage sites of α II-spectrin ((PDB ID: 3FB2 crystal structure).....	57
3.19.	Control neurons (β II-spectrin) undergoing cell death pathways.....	63
3.20.	β II-spectrin encoded gene.....	64
3.21.	Cleavage sites of β II-spectrin by caspase-3.....	66
3.22.	Cleavage sites of β II-spectrin by calpain-2.....	69
4.1.	Crystal structure of GATA3:DNA.....	77
4.2.	GATA3 electrostatic free energy differences (charged AA).....	79
4.3.	Molecular graphics of the GATA3:DNA complex.....	80
4.4.	Electrostatic free energy differences of Ala all-mutants.....	84
4.5.	Electrostatic free energy differences of Arg all-mutants.....	85

TABLES

Table	Page
2.1. Calpain-2 and caspase-3 cleavage properties.....	12
3.1. Output of simulated data with pruning of deletes and Inserts.....	36
3.2. Generated fragments based on consensus occurrences in Table 3.1.....	37
3.3. Output of simulated data with pruning of at least three mismatches.....	38
3.4. Generated fragments based on consensus occurrences in Table 3.3.....	38
3.5. Output of simulated data with overlaps of consensus occurrences.....	39
3.6. Generated fragments based on consensus occurrences in Table 3.5.....	40
3.7. Output of simulated data with an INDEL after a consensus.....	41
3.8. Generated fragments based on consensus occurrences in Table 3.7.....	41
3.9. Output of simulated data for protein-consensus NN orientations.....	42
3.10. Output of simulated data for protein-consensus NC orientations.....	43
3.11. Output of simulated data for protein-consensus CN orientations.....	44
3.12. Output of simulated data for protein-consensus CC orientations.....	45
3.13. Mouse proteome output in NN and CC orientations.....	49
3.14. Mouse proteome output in NC and CN orientations.....	51
3.15. CFPA-CalpCasp generated data on 2FOT by calpain.....	59
3.16. CFPA-CalpCasp generated data on 2FOT by caspase.....	60
3.17. CFPA- CalpCasp generated data on 3FB2 by caspase.....	61
3.18. CFPA-CalpCasp generated data on M96803 by caspase-3.....	68
3.19. Few records of CFPA-CalpCasp data on M96803 by calpain.....	68

4.1. Types of protein-DNA interactions.....	76
---	----

CHAPTER 1

INTRODUCTION

Congenital heart defects (CHD) are the most frequent form of major birth defects in newborns, affecting close to 1% of newborn babies (8 per 1,000) [1]. While it is known that the risk of congenital heart defects is higher when there is a close relative with one [2], signs and symptoms are related to the type and severity of the heart defect. Symptoms frequently appear early in life, but it's possible for some CHDs to go undetected throughout life [3]. Some children have no signs while others may exhibit shortness of breath, cyanosis, syncope [4], heart murmur, under-developing of limbs and muscles, poor feeding or growth, or respiratory infections. Atrial septal defects (ASD) and ventricular septal defects (VSD) are the most common types of defects and constitute our areas of focus [5,6] for experimental validation.

1.1. Dissertation Objectives

In this thesis, we are tackling CHD from the roles of two different genes. The first one is the *Tll1* gene and we are tackling it from the degradomics perspective, where cleaved fragments play crucial roles as disease biomarkers. The second one is the GATA4 gene and we are tackling it from the protein-DNA interactions perspective, where mutated amino acids can give insights into the disease.

1.1.1 First Dissertation Objective

One practical application of proteomics is the identification of biomarkers. A biomarker is defined to be a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, pharmacologic responses to a therapeutic intervention, or other diseases [7-9]. Protease activation has been highly associated with protein degradation, leading to the appearance of biomarkers or signature Breakdown Products (BDPs). Thus, the term *degradomics* [10] has been introduced to evaluate those potential BDPs of different protein substrates. Proteases, depending on their cleavage mode, truncate proteins at specific amino acid sequences, and the fragmented proteins, or BDPs, represent molecular signatures that are specific to each protease. Instances include *Tlll* BDPs as biomarkers of Congenital Heart Disease (CHD), calpain and caspase BDPs as biomarkers of Apoptosis and Necrosis [11-13], and Matrix Metalloproteases MMP-2 to MMP-9 BDPs as biomarkers of cancer [14-16].

Tlll gene has major role in heart septal development (i.e., membrane that divides the left and right partition of the heart). The presence of its wild type is active to truncate specific extracellular substrate proteins representing BDPs. Those BDPs can leak into the blood, and upon identification, indicate the absence of heart malfunction. Currently, there are no existing methods to detect septal defects in patients except through imaging, and that is only performed after clinical manifestations of the disease. Thus, the purpose in this thesis is to establish a new and fast methodology for detecting computationally signs of such malformations. Then, the detected results are validated

experimentally from the literature. Nonetheless, the detected results can be validated clinically based on a single blood test.

To achieve this goal, it is crucial to understand the mechanisms involved in septal formation - a process governed by multiple proteins. It involves an investigative, discovery-based study to identify the substrates of the *Tlll* metalloprotease (i.e., cleaving catalytic enzyme) and their BDPs in human blood. The exact targets (or substrates) of the *Tlll* gene in the heart, and specifically in the septum, are detected experimentally in advance. The goal becomes to develop an efficient and accurate algorithm that detects the consensus sequence and its variants in this large set of experimentally detected protein sequences, or *Tlll* substrates.

In addition to CHD, the above suggested algorithm can be applied to neuroscience for distinguishing between the different neuronal cell death types. While morphological changes and biochemical techniques are used to be the most practical discriminators of cell death types [17], they are restricted to experimental studies [18]. A major consequence of these cell death events is the activation of cellular proteases (calpain and caspase), leading to specific protein fragmentation and generation of BDPs. Hence, finding the distinction between different cell death types became contingent on finding the biochemical markers that result from the degradation of substrates by proteases, like calpain and caspase. An instance substrate is α II-spectrin which is degraded by calpain and caspase enzymes [19,20]; the degraded fragments are biomarkers in diseases like Traumatic Brain Injury (TBI), stroke, and aneurysmal subarachnoid hemorrhage [21,22-25]. Specific cleavage sites of this protein are reported in the literature [26,27-29].

Consequently, we are interested in this work in the computational identification of fragment biomarkers generated upon cleavage by different proteases. Resulting PDBs are validated experimentally. Yet, computational predictive models of those BDPs can feed into clinical applications to monitor various normal and abnormal behaviors [30]. Such research will advance fundamental understanding of medicinal important biological systems, contributing to the diagnosis and treatment of related diseases. Furthermore, the degradomics method can feed into the development of a novel web tool that will help scientists detect diverse types of biomarkers dynamically; the tool can also be extended to detect biomarkers related to diseases and biological processes other than CHD and neuronal cell death types.

1.1.2 Second Dissertation Objective

Protein-DNA interactions play a crucial role in several biological processes. In this thesis, we are mainly focusing on the GATA gene and the DNA transcription biological process. GATA gene encodes a zinc finger transcription factor, and mutations in GATA4 in particular are associated with cardiac septal defects [31]. An important target of the GATA TF is the DNA promoter sequence and its variants. When proper regulation fails due to some mutations on the GATA TF, or due to some mutations on the GATA promoter (weak promoter), the target genes are under-expressed/over-expressed, compromising proper regulation, and contributing to CHD. Our methodology seeks to develop an efficient and reliable model for predicting protein-DNA association and binding in the presence of mutations, while elaborating on non-covalent binding. The model is then applied to study, test, and validate against experimentation, the

binding interactions in both the adjacent and opposite GATA3:DNA complexes (PDB Id: 3DFV and 3DFX), due to the unavailability of the GATA4:DNA complex. The computation delineates the driving forces of recognition and binding through Poisson-Boltzmann electrostatic binding free energy calculations. The calculations include all possible mutations of charged amino acids on GATA TF. This study assists in gaining insight into the dynamic and physicochemical characteristics of the GATA:DNA complex, in addition to providing insight on the relation between binding and protein function.

Two major steps comprise the electrostatic association of protein and DNA molecules, recognition and binding [32]. Recognition is characterized by nonspecific long-range electrostatic interactions, whereas binding is characterized by specific favorable local short-range electrostatic interactions, such as hydrogen bonds, salt bridges, medium-range coulombic interactions, hydrophobic interactions, and van der Waals interactions. In recognition, an accelerated weak encounter complex is formed. In binding, the protein and DNA are locked into their final bound conformation, after local side change rearrangements, and exclusion of solvent atoms from their binding interface.

The impact of charged amino acids is key to binding [33] and has been demonstrated in several diseases, such as the eye disease known as Age-related Macular Degeneration (AMD) [34], the kidney disease known as atypical Hemolytic Uremic Syndrome (aHUS), the Dense Deposit Disease (DDD) known as membranoproliferative glomerulonephritis [35], and the autoimmune disease [36]. In addition, earlier studies showed the electrostatic type of interactions in complexes like C3d-CR2 [37-40] and

C3d-EfbC/Ehp [39-41] association, and in interactions with viral proteins VCP/SPICE [42,43] and Kaposica [44].

Accordingly, we study in this thesis the role of charged amino acids in non-specific recognition and in specific binding in the interactions between GATA TF and DNA; we study the role of charge in the regulation function of GATA TF to DNA and the implications to diseases when proper regulation fails.

The protein-DNA method is validated experimentally using data from the literature and is applied to other protein-DNA complexes; the goal is to study the malfunction caused by mutations on their amino acid sequence or on their nucleic acid sequence, as related to other diseases. Not only will this research assist in gaining insight into the physicochemical characteristics of protein-DNA complexes, but it will also provide insight on the relation between binding and protein function. Furthermore, this research might assist in designing new targeted molecules contributing to the discovery of new medications.

1.2. Dissertation Organization

The rest of the dissertation is organized as follows: Chapter 2 gives background on different topics and surveys previous work in the area of sequence matching and alignment. Chapter 3 presents the developed computational degradomics methods. It also covers different applications of the method from neuroscience, including data, results, and analysis. Chapter 4 presents the computational protein-DNA interaction methods and the application to GATA3:DNA complex, including data, results, and analysis. Finally, Chapter 5 presents some conclusions and future work.

CHAPTER 2

BACKGROUND

In this chapter, we first cover the types of diseases in detail in Section 2.1 and Section 2.2. Section 2.3 gives some background about the importance of degradomics in biomedical research and as related to our first thesis objective. Section 2.4 elaborates first on Smith-Waterman algorithm and then gives some background on other sequence matching and alignment algorithms. Finally, Section 2.5 gives some background about the importance of electrostatics and force field used in analyzing protein-DNA interactions, as related to our second thesis objective.

2.1 Congenital Heart Disease (CHD)

Septum defect is a major malfunction (Fig. 2.1 and Fig. 2.2) and early detection is vital for treatment. While current detection techniques are limited to imaging, such techniques subject the patients to radiations, in addition of being costly. Hence, the need to detect such heart defect via alternative methods becomes vital.

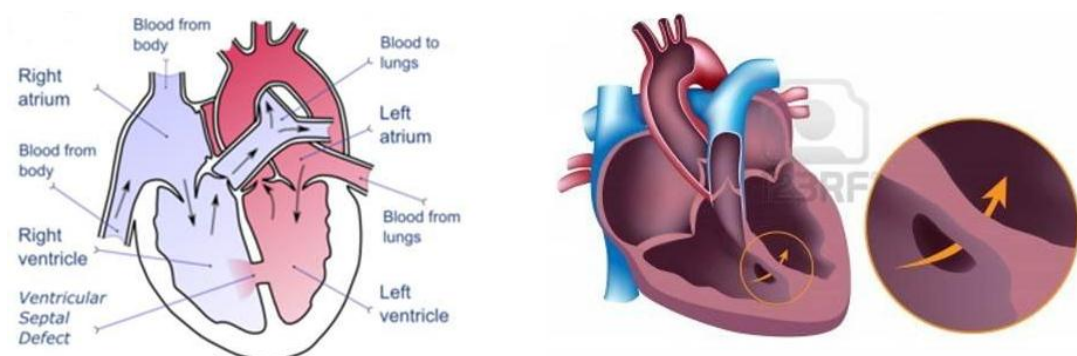


Fig. 2.1 - Ventricular septal defect (VSD) type of CHD

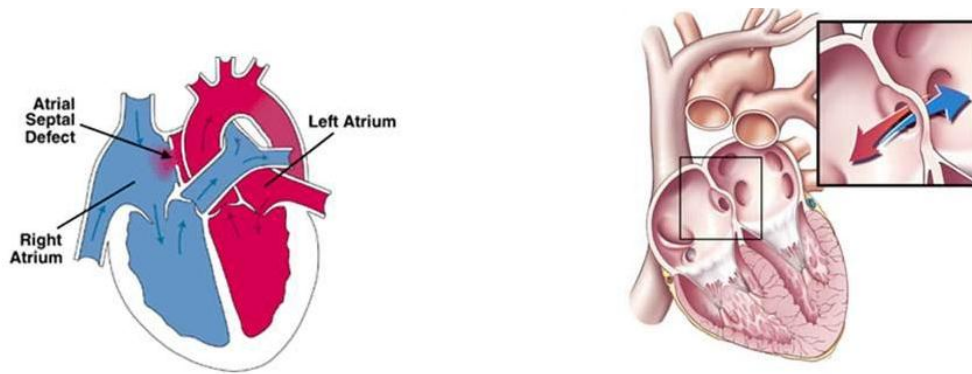


Fig. 2.2 - Atrial septal defect (ASD) type of CHD

2.1.1 *Tll1* Gene

Recent studies have added genes encoding different classes of proteins like growth factors (i.e., a substances stimulating cell growth) and their receptors (i.e., take role in the communication between cell inside and outside) to the list of potential players in septal formation. The latest addition is the *Tll1* (or Tolloid-like 1) gene that encodes a metalloprotease (i.e., a protein that is a catalytic enzyme) [45]. Metalloproteases are documented to be active in tissue remodeling during embryogenesis (early development) as well as in various diseases [46]. Based on the primary and secondary structure, *Tll1* was classified in the family of Astacin, a group of metalloproteases that include tolloids [47]. Typically, tolloids are characterized by C-terminal CUB and EGF-like domains, which are thought to be important for substrate recognition, binding, and cleavage [48].

2.1.2 *GATA* Family

GATA family (GATA1-6) comprises a set of transcription factors (i.e. responsible of turning a gene on/off) containing zinc fingers in their DNA binding

domain [49]. GATA1 is involved in cell growth and cancer; mutations in this gene have been associated with X-linked dyserythropoietic anemia and thrombocytopenia [50]. GATA2 is expressed in hematopoietic progenitor; mutations in this gene have been associated with MonoMAC Syndrome and leukemia [51]. GATA3 is involved in the regulation of luminal epithelial cell differentiation [52]; mutations in this gene have been associated with breast cancer [53]. GATA4 plays a critical role in heart septal deficiency and will be elaborated on in the next section. GATA5 is found to be involved in the activation of the intestinal lactase-phlorizin hydrolase promoter. Finally, GATA6 mutations have been associated with pancreatic agenesis and congenital heart defects [54].

2.1.3 GATA4 Transcription Factor

Despite the major role *Tll1* gene plays in heart septal deficiency, it is not the only gene. Gross phenotypical disorders such as Noonan syndrome (i.e., narrowing in pulmonary valve) also include incomplete atrial and/or ventricular septum [55,56]. Different genes encoding mainly cardiac-enriched transcription factors (i.e. responsible of turning a gene on/off) have been linked to septal defects, such as the Tbx5 protein. Other genes include members of the GATA family of zinc finger transcription factors. Members of this family recognize the GATA motif which is present in the promoters of many genes. GATA4, in particular, is a critical transcription factor for proper mammalian cardiac development and works in combination with other essential cardiac transcription factors, such as Nkx2-5 and Tbx5. It is expressed in both embryo and adult cardiomyocytes, where it functions as a transcriptional regulator for many cardiac genes, and also regulates hypertrophic growth of the heart [57]. Mutations in GATA4

gene have been associated with cardiac septal defects as well as reproductive defects [31,58]. Other mutational defects include a variety of cardiac problems, such as Congenital Heart Disease (CHD), abnormal ventral folding, and defects in the cardiac septum separating the atria and ventricles.

2.2. Neuronal Cell Death Types

2.2.1. Apoptosis and Necrosis

Making a distinction between the major cell death types (Autophagic, Oncotic/Necrotic, and Apoptotic) in different neurological diseases, such as traumatic brain injury (TBI), stroke, and Alzheimer's disease [59,21,60,61], will help better understand the injury mechanisms of each specific pathoneurological condition. In addition, this distinction will guide better diagnosis and help develop future targeted therapy [26,62,63].

Apoptosis, or Programmed Cell Death (PCD), which is characterized by a well-defined series of morphological and biochemical traits [64], is a physiological cell death process used by the organism during development; it includes the activation of a specific family of proteases known as caspases [65]. Besides, certain physical and chemical conditions, such as hypoxia or toxins, might trigger spontaneous apoptotic events. Necrosis, on the other hand, is characterized by cell and organelle swelling leading to nuclear degradation and, eventually, to disruption of the cell membrane and cell lysis [66,67]; it usually occurs when cells are injured by extreme physical stress or chemical challenges, to the point where they cannot be repaired. Further, it is associated with an increase in the activation of members of the proteolytic calpain family [26,68].

Yet, atypical conditions might cause both proteases, calpain and caspase, to be activated simultaneously, leading to a fusion of apoptosis and necrosis as described in traumatic brain injury [69-73].

A number of experimental techniques allow to test for specific proteins spatiotemporal alteration dynamics during a particular biological process, including diseases. Among those techniques are the antibodies-based ones and include: Western blot, enzyme linked immunosorbent assay (ELISA), and immunocytochemistry. Other more sensitive techniques include mass spectrometry/proteomics.

The two cell death pathways apoptosis and necrosis are experimentally demonstrated by examining the computational α II-spectrin cleavage Breakdown Products (BDPs) against calpain-mediated fragments (SBDP145, SBDP150) and caspase3-mediated fragments (SBDP120, SBDP150i), showing all specific cleavage sites [26,74]. β II-spectrin [75] is another substrate that calpain and caspase cleave and its cleaved BDPs are examined within apoptosis and necrosis.

2.2.2. Calpain and Caspase Proteases

Proteases Cleaving Modes: Calpain and caspase proteases cleave substrates in different manners upon separate or combined activation. Table 2.1 represents the consensus pattern necessary for each protease to cleave, in addition to the cleavage site represented by the symbol (*).

Table 2.1 - Calpain-2 and caspase-3 cleavage properties [76]

	Calpain	Caspase
Protease Class	Cysteine Protease	Cysteine Protease
Preferred Cleavage Site (*)	Asp ^x xAsp [*] x	(Leu, Val, Ile) ^x *x
Common Substrates	α II-spectrin 280kDa β II-spectrin 260kDa	α II-spectrin 280kDa β II-spectrin 260kDa
Fragments Produced by αII-spectrin	SBDP150 kDa SBDP145 kDa	SBDP150i kDa SBDP120 kDa
Fragments Produced by βII-spectrin	SBDP110 kDa SBDP85 kDa	SBDP108 kDa SBDP80 kDa
Cell Death Involvement	Most forms of necrosis Some forms of apoptosis	Most forms of apoptosis

Calpain Cleavage (Necrosis and Apoptosis): Calpain-mediated proteolysis and its association to necrotic neuronal death gained major research focus, as manifested in ischemic and excitotoxic neuronal injury [77]. Nonetheless, calpain-mediated α II-spectrin breakdown to a 150 kDa and 145 kDa doublet is not only present in necrotic neuronal death, but also in most forms of neuronal apoptosis, as manifested in thymocytes, in staurosporine-treated neuroblastoma SH-SY5Y cells, in NGF-deprived rat PC12 cells, and in low-K1-treated rat cerebellar granule neurons [78-85].

Calpain cleaves preferentially at Val, Leu, or Ile residues in the second position (P2 position) of the consensus sequence found within the target protein sequence. The amino acids in the first position (P1 position) of the consensus sequence are rather diverse (for example, Tyr, Gly, Arg).

Caspase Cleavage (Apoptosis): A large number of proteases have roles as mediators of apoptosis in a wide range of cell types [81], but caspase-3 is of particular interest as it appears to be a common downstream apoptosis effector; it is solely responsible for the

production of SBDP120 upon α II-spectrin cleavage. Caspase-3, like calpain, is a cytosolic cysteine protease, but does not require Ca^{2+} for activity.

Caspase cleaves preferentially at Asp in the fourth position (P4 position) of the consensus sequence found within the target protein sequence. The amino acid in the first position (P1 position) of the consensus sequence is Asp. The amino acids in the second and third positions (P2 and P3 positions) of the consensus sequence can be diverse.

Calpain and Caspase Cleavages Combined: It is common for caspase to become activated after calpain's activation. Such order allows them to cleave the same substrate at different sites. The SBDP145 (calpain2-specific) and SBDP120 (caspase3-specific) α II-spectrin breakdowns can be easily distinguished by SDS-PAGE, whereas SBDP150 and SBDP150i fragments, generated by calpain-2 and caspase-3 respectively, are within a nine-residue span [85]. Instances of neuronal injury models manifesting the activation of both proteases, include NMDA and kainate and glucose-oxygen-deprivation challenged cerebrocortical neurons [82], in addition to an in vivo impact model of TBI in all affected areas of the brain [83,84].

Nevertheless, it might also be possible, in a special and sporadic case, for calpain and caspase to become activated simultaneously. In such case, two possibilities can happen. Either one protease inhibits the cleavage of the other, or one protease cleaves within the substrate cleaved by the other.

2.3. Degradomics

Degradomic behaviors based on the proteolytic activities and the vital analysis of the produced peptidomes have been conducted in several studies and for different biological processes. The authors in [85], show that homologous granzymes do not necessarily contribute to similarities in their substrate repertoires, in their specificities, and in their efficiency of cleaving, but to differences in the resulting biochemical cascades. In [86], targeted proteomics - a new experimental technique – has been applied to reveal more specificity and more sensitivity in determining degradomics substrates. This revealed some proteins that are up-regulated due to some mutations, turning them into potential degradation targets. The cleavage site specificities of the cell membrane Type II Transmembrane Serine Proteases (TTSPs) are identified in [87] using tandem mass spectrometry. They are sequenced afterwards to reveal a strong cleavage specificity for Arginine in the first position (P1 position) and Lysine in the nineteenth position (P19 position) on the prime side of unfolded peptides. The major application of quantitative proteomics in the regulation of Programmed Cell Deaths (PCD) is elucidated in [88] through the modulation of proteins by protein cleavage. PCD includes apoptosis, autophagy, and necrosis. Those types of cell deaths are genetically determined and form complex processes in multi-cellular organisms; they have been associated with problems in regulation in a number of diseases.

Degradomics studies have been noticeable in the literature of heart malfunction. One application focus is on the genetical aspects of CHD, which are the most frequent forms of major birth defects in newborns. A typical gene, in this context, is the *Tll1* (or Tolloid-like 1). *Tll1* encodes a metalloprotease (i.e., a protein that is a catalytic enzyme)

and plays an important role in heart septal deficiency [89]. The embryonic lethal phenotype of *Tlll* double homozygous null mice displays cardiac deformities [15]. Upon activation, *Tlll* truncates specific extracellular substrate proteins in the heart, and specifically the septum, representing BDPs. The presence of those BDPs may signify putative markers of the absence of CHD malfunction.

Degradomics studies have shown their applications in neuroscience. The molecular basis of calpain-catalyzed proteolysis between α II-spectrin and β II-spectrin is discerned in [90] in the presence of different catalysts. Calpain proteolysis is linked to apoptotic and non-apoptotic neuronal cell death following excessive glutamate exposure. It preferentially cleaves α II-spectrin *in vitro* in repeat 11 between residues Y1176 and G1177. Bound CaM induces a second α II-spectrin cleavage at G1230*S1231. β II-spectrin, on the other hand, is cleaved at four sites. One cleavage occurs in the absence of CaM at high enzyme-to-substrate ratios near the β II-spectrin COOH-terminus. Other β II-spectrin cleavages, that CaM promotes, are at Q1440*S1441, S1447*Q1448, and L1482*A1483. Yet, when stimulated by calcium influx (via maitotoxin), α II- and then β II-spectrin are rapidly and sequentially cleaved, coinciding with the onset of non-apoptotic cell death.

Degradomics studies have been implied in other diseases as well, such as cancer. It is shown in [91] how the degradomics-peptidomics profiling of blood plasma can be highly sensitive to changes not evidenced by conventional bottom-up proteomics, and provides unique signatures of diagnostic utility. The peptidome-degradome profiles of pooled blood plasma, sampled from Breast Cancer Patients (BCP) and control Healthy Persons (HP), were characterized to reveal the ratios of the

peptidome peptide relative abundances. The ratios vary as much as 4000 fold between BCP and HP and the experimental results show differential degradation of substrates in the BCP sample functional domain. The important roles of Matrix Metalloproteinases (MMPs) is uncovered in crucial alterations of cellular signaling in cancer [76]. MMPs, similar to other proteases (caspases, calpains and cathepsins) truncate proteins at specific amino acid sequences. Fragmented proteins represent molecular signatures designated as BDPs, specific to each protease, and distinguished by their protease-generated molecular weight. MMPs production by cancer cells lead to tumour-cell invasion and, subsequently, to metastasis. The selective tumour-cell resistance to apoptosis, driven by selective MMPs production by cancer cells, and the absence of MMPs roles in the presence of MMPs inhibitors, are additional instances. Similarly, [92] presents a review of the intense role of MMPs in cancer disease besides the local degradation of Extracellular Matrix (ECM) components; ECM forms the physical barrier for cell migration. The review elaborates on the cellular and molecular mechanisms of MMPs that influence tumour cell growth, invasion, and metastasis; all this leading to cancer progression. Novel MS-driven proteomics techniques are used in [93] for the identification of certain kallikrein-related peptidase (KLK) as biomarkers for cancer disease. That step is achieved by elucidating KLK substrates using degradomics studies and then making a correlation between proteolysis of biological substrates and tissue-related consequences. A review of the feasibility of MMPs as possible prognostic markers and as potential therapeutic targets in cancer disease is presented in [94]. For instance, elevated levels of distinct MMPs are detected in tumour tissue and in serum of patients with advanced cancerous stage. On the other hand,

studies are critically considering MMP inhibitors for therapeutic intervention of tumour growth and invasion. Through the characterization of all proteases, inhibitors, and protease substrates by proteomic and degradomic techniques, [95] demonstrates that the substrate repertoire of MMPs is not restricted to extracellular proteins, but is expanded to include intracellular ones too; this led MMPs to be modulators of multiple new signaling pathways. In addition, [10] discovers the action of MMPs on intracellular proteins and the exploration of their substrate repertoire using multidimensional degradomics technology; such technology is developed by the integration of broadly available biotechniques. For instance, some of MMP-9 identified candidates are novel substrates, like actin and tubulin, whereas many others are autoantigens, like annexin I and nucleolin. The latter ones are described in multiple autoimmune conditions and in cancer; this fact led to consider MMP-9 with novel regulatory properties.

The release of BDPs can also affect other biochemical processes in cell behavior, such as cell proliferation, adhesion, migration, growth factor bioavailability, chemotaxis, differentiation, angiogenesis, host defense, and signaling; all these processes are mainly due to proteolytically modifying the signaling environment of the cell [96]. In addition, BDPs play an important role in tissue remodeling associated with various physiological or pathological processes, such as morphogenesis, tissue repair, cirrhosis, arthritis, and metastasis.

2.4. Sequence Matching Algorithms

In exploring the most efficient and accurate algorithm that can locate consensus occurrences for the degradomics analysis, we first elaborate on Smith-Waterman (SW)

Algorithm, which we modified for the degradomics approach (see Chapter 3, Section 3.1). Then, we present other sequence matching algorithms.

2.4.1. Smith-Waterman (SW) Algorithm

Smith–Waterman algorithm [97] performs local sequence alignment. It finds regions in an input protein sequence that match a consensus sequence. It is based on a dynamic programming solution that builds a scoring table from the recurrence relation, shown in Fig. 2.3:

Smith-Waterman Algorithm	
$H(i, 0) = 0, 0 \leq i \leq m$	
$H(0, j) = 0, 0 \leq j \leq n$	
$H(i, j) = \max$	$\begin{cases} 0 & \\ H(i-1, j-1) + \text{Similarity_Score}[C_i, S_j] & \text{Match/Mismatch} \\ H(i-1, j) - d & \text{Deletion} \\ H(i, j-1) - d & \text{Insertion} \end{cases}$
if $C_i = S_j$ then $\text{Similarity_Score} = \text{Score}(\text{Mismatch}) = 2$	
if $C_i \neq S_j$ then $\text{Similarity_Score} = \text{Score}(\text{Mismatch}) = -1$	
$d = -1 = \text{Score}(C_i, -) = \text{Score}(-, S_j)$ for a gap in the case of insertion or deletion where	
$1 \leq i \leq m$ and $1 \leq j \leq n$	
$C = \text{Consensus Sequence}$ and $S = \text{Input Sequence}$	
$m = \text{length}(C)$ and $n = \text{length}(S)$	

Fig. 2.3 - Smith-Waterman algorithm

For all N *input* protein sequences, SW algorithm can find all occurrences of a *consensus* sequence exact matches, if executed N times. When a consensus sequence is found in an input sequence, proteases cut the input protein sequence at the *cleavage site* of the consensus occurrence, resulting in fragment sequences or PDBs. The location of the cut, called *cleavage site*, is predefined in the consensus sequence. While finding exact matches of the consensus sequence among all input proteins sequences is the

primary target, variants of exact matches to one or two mismatches, within any position of the consensus, are investigated too. Variants of a consensus sequence can also result in cleaving the input protein sequence into fragments when the variant sequence is found in the input sequence.

Since multiple occurrences of a specific consensus can be found in one protein sequence, all combinations of potential cuts, resulting in different output fragments are considered. For instance, if two consensus occurrences are matched in an input sequence, then the number of combinations to consider, resulting in different output fragments, will be: 1) As if occurrence 1 is only found and breaks the input sequence at the cleavage site, or 2) as if occurrence 2 is only found and breaks the input sequence at a different cleavage site, or 3) as if occurrence 1 and occurrence 2 are both found and break the input sequence at two different cleavage sites. Two simultaneous occurrences result in many short fragments than a few long ones resulting from only one occurrence.

2.4.2. Other Algorithms

The requirement for sequence matching has been recognized in several papers. Similar work includes heuristic methods, hashing methods, suffix tree methods, and sequence alignment methods. The following present some algorithms that have been developed to search for an exact or similar match of a specific subsequence (length m) in a larger sequence (length n). Knuth, D. [98] developed an algorithm that finds exact matches of a subsequence of size m in a sequence of size n in $O(m+n)$. Despite the efficient time complexity of this method, other fast algorithms of comparable complexity are present and can find variants in addition of exact matches. Lipman, D.J.

and Pearson, W.R. [99] devised the algorithm Fast Protein (FASTP) which finds similarities between an amino acid sequence and sequences in the database. It is rapid, sensitive, and similarities are detected based on alignments. However, this algorithm starts with an anchoring scheme - a heuristic that identifies identical regions using a replaceability matrix. Altschul, S.F. [100] developed BLAST and its variations; they are considered as improvements in time over previous methods like FASTP, but with a comparable sensitivity. Such methods focus on identifying those sequences that share high similarity with the query sequence by seeking segment pairs that contain a word pair with a given score. Similarly to FASTP, they use certain seeds for basic anchoring and so, they are probabilistic in nature. Ning, Z. [101] designed sequence search and alignment based on the hashing method Sequence Search and Alignment by Hashing Algorithm (SSAHA). This method performs fast searches for a query sequence in databases in the giga range; it can be three to four times faster than Fast-All method (FASTA) or BLAST. This method can add overhead though; it is based on preprocessing the sequences in the database by breaking them into consecutive k-tuples of k contiguous bases, and then uses a hash-table to store the position of each occurrence of k-tuple. Ma, B. [102] devised a suboptimal homology (i.e. similarity based on common origin or descendent) search algorithm that finds homologies between large sequences. Compared with BLAST, it can perform faster with a modest memory usage and higher sensitivity. Still though, the resultant accuracy is based on a wide seeding model, which makes this algorithm based on heuristics, compromising this accuracy to some extent. Kurtz, S. [103] articulated suffix tree-based methods for similarity sequence search. The corresponding algorithms use low search-time

complexity, but suffer from two main drawbacks; the first one is their limitation to precise matches and the second one is their intrinsic large space requirement. In addition, such methods can be tedious in updating data due to the implementation of linked lists. Lecroq, T. [104] is an algorithm based on hashing methods; it uses the Q-gram hashing, which is an efficient indexing technique against a sequence database. It is considered the fastest, especially on small size alphabet, but it is limited to exact matches. While both of suffix tree and hashing-based methods are used to improve the computational time, the hashing-based method is still preferred due to its ease in updating the data. Needleman, S. and Wunsch, C. in [105], devised one of the first applications of dynamic programming and applied it to biological sequence comparison. Sometimes, it is referred to as the optimal matching algorithm; the disadvantage is that it is based on global alignment, and that makes it more suitable for sequences of comparable sizes.

Most of the above techniques FASTA, FASTP, BLAST, and SSHAH are efficient on existing genomic sequences and databases, but they do not scale up well on large datasets like the human genome. Srikantha, A. in [106] addressed the efficiency issue on a large dataset with an algorithm that finds exact sequence match. The reference sequence can range from several million (individual chromosomes) to several billion bases (whole genome). It is based on down sampling of the large sequence and on polyphase decomposition of the specific sequence. It also uses hash tables and Q-grams to refine the search region, leading to reduction in space and time complexity. The limitation of this method is handling exact matches only. Other algorithms came out to handle similar large datasets like the method of Li, H. and Durbin, R. [107],

which handles long reads (up to 1Mb) efficiently in the comparison against a large sequence database. Such alignment algorithm is based on dynamic programming and demonstrates more accuracy than any of the heuristic methods, but might suffer from memory requirements.

2.5. Force Field in Protein-DNA Interactions

The most accurate method to model biological systems of interest, and the interactions occurring in them, is Quantum Mechanics (QM). QM methods are based on the Schrödinger equation and consider the electrons in the system explicitly. However, complex biological systems, consisting of several thousands of atoms and a large amount of water molecules in the case of proper solvation of biological molecules, are too large to be considered by QM approaches. Moreover, the general difficulty of solving the equation makes QM unpractical and very computationally expensive.

On the other hand, force field methods, also called Molecular Mechanics (MM), are simple models based on the Born Oppenheimer (BO) approximation. BO states that the electrons adapt immediately to any changes in the nuclear positions, especially with the proton being roughly 1800 times as heavy as the electron. MM methods ignore the electrons and write the energy as a sum of potential energies based on bonded degrees of freedom and non-bonded interactions as represented by $V(\mathbf{r}^N)$ in Eqn.1 [108]. $V(\mathbf{r}^N)$ is the potential energy of a system as a function of the positions (\mathbf{r}) of N particles (usually atoms). Bonded terms in $V(\mathbf{r}^N)$ comprise stretching of bonds (1st term), bending of angles (2nd term), and rotation of single bonds or torsions (3rd term), whereas non-bonded terms in $V(\mathbf{r}^N)$ include non-bonded interactions between atom pairs, specifically van der Waals (4th term) and electrostatic interactions (5th term). A

typical MM force field can be accurate and comparable to QM at a much lower computational cost for some properties. Subsequent sections elaborate on its component terms in more detail (see Sections 2.5.1.1, 2.5.1.2, 2.5.1.3, 2.5.2.1, and 2.5.2.2).

$$V(r^N) = \sum_{bonds} \frac{k_i}{2} (l_i - l_{i,0})^2 + \sum_{angles} \frac{k_i}{2} (\theta_i - \theta_{i,0})^2 + \sum_{torsions} \frac{V_n}{2} (1 + \cos(n\omega - \gamma))^2 + \sum_{i=1}^N \sum_{j=i+1}^N \left(4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right) \quad (1)$$

2.5.1. Intramolecular Bonded Interactions

2.5.1.1. Bond Stretching

Morse potential, as represented by Eqn.2 and Eqn.3 [108], is an accurate potential energy curve for bonds. It is computationally expensive since it requires three parameters per bond.

$$v(l) = D_e [1 - \exp(-a[l - l_0])]^2 \quad (2)$$

$$a = \omega \sqrt{\mu/2D_e} \quad (3)$$

where,

D_e : is the depth of the potential energy minimum.

μ : is the reduced mass.

ω : is the frequency of the bond vibration.

l_0 : is the reference length of the bond.

l : is the length of the bond.

A simplified Hooke's law (or harmonic potential) is often used, as in Eqn.4 [108], to replace Morse potential since bond lengths rarely change significantly from the

reference length. Cubic, quartic, and higher order terms are sometimes included in the Hooke's law potential to increase the accuracy compared to Morse potential. Higher order equations, though, can be more complicated and computationally expensive, as additional parameters are introduced.

$$v(l) = \frac{k}{2}(l - l_0)^2 \quad (4)$$

where,

k: is the force constant.

l: is the bond length.

l_0 : is the reference bond length.

2.5.1.2. Angle Bending

Hooke's law potential is also used to represent the potential for angle bending, as in Eqn.5 [108], and higher order terms can be included to increase the accuracy of the force field, but again, at the cost of increased computational expense.

$$v(\theta) = \frac{k}{2}(\theta - \theta_0)^2 \quad (5)$$

where,

k: is the force constant and is typically 1/10 of the force constants used to model bond stretching, since it takes less energy to bend an angle than to change a bond length.

θ : is the bond angle.

θ_0 : is the reference bond angle.

2.5.1.3. Torsion/Twisting

The potential energy formula used to represent torsions or rotational barriers is represented in Eqn.6 [108]. Organic molecules have a torsional component for each quartet A-B-C-D, where the torsion angle is the twisting of the bond B-C. The energy required to rotate a bond is significantly smaller than deviations in bond length or bond angle, and non-bonded interactions are usually sufficient to provide this energy barrier. For this reason, it is not strictly necessary to include explicit torsional terms.

$$v(\omega) = \sum_{n=0}^N \frac{V_n}{2} [1 + \cos(n\omega - \gamma)] \quad (6)$$

where,

V_n : is referred to as the barrier height.

ω : is the torsion angle.

n : is the multiplicity and defines the number of minimum points through all 360°.

γ : is the phase factor and defines the torsion angle when it passes through the energy minima.

2.5.2. *Inter- and Intramolecular Non-bonded Interactions*

Electrostatic and van der Waals contributions are mainly what comprise non-bonded interactions and what drive the corresponding potential energy formulas. With the partial atomic point charges in electrostatics, the charge distribution in molecules is represented in many ways [108]. van der Waals interactions, on the other hand, are electrostatic in nature; their attractive forces are explained with QM as fluctuations in the electron cloud around the atom, which in turn induce dipoles in nearby atoms. The

effect of the resulting induced-dipole/induced-dipole interactions increases with the number of electrons. Further details on each type of interactions are elucidated in subsequent sections.

2.5.2.1. Electrostatic Interactions

These interactions arise from the attraction and repulsion of electrically charged particles. Electrostatic interactions are further divided into different types: Salt bridges, dipole-dipole, ion-dipole, dipole-induced dipole, and hydrogen bonds. In addition to providing overall stability to molecular complexes, electrostatic interactions also play an important role in providing specificity.

Salt bridges are formed when two ionizable amino acids of opposite charges are at a short-range distance in space, typically around 3 Å [108]. They are of significant importance for protein stability as well as for protein-protein complex formation. They connect different parts of the protein by connecting separated amino acids within the protein sequence. Oppositely, they can destabilize the protein by interacting with the solvent ions when exposed to solvent. For this reason, completely buried salt bridges contribute more to the stability; they contribute several kcal/mol to the free energy of folding.

Dipoles interactions occur due to differences in electronegativity between elements, resulting in uneven charge distributions in molecules. Dipole-dipole interactions are classified as electrostatic interactions between the partial charges of the atoms on dipoles of both molecules. Dipole-charge interactions are classified as electrostatic interactions between the partial charges of the atoms on one dipole and the

charge on the other molecule. Dipole/induced-dipole interactions are classified as electrostatic interactions between the partial charges of the atoms on one dipole and the electrostatic influence of this dipole inducing a dipole in otherwise neutral atoms, and that is by polarizing the covalent bond between them.

Hydrogen bonds make another type of electrostatic interactions, special type of dipole-dipole. A hydrogen bond is typically described as an attractive force (a lone pair of electrons) between a hydrogen atom from one molecule and an adjacent N, O or F atom of a second molecule or part of the same molecule; the hydrogen atom is covalently bound to another N, O or F atom. This type of bond is usually depicted as: $X-H\cdots Y$ where the dots denote the bond and (X, Y) pair denotes any of N, O, or F atoms. Typical bond lengths measured from H to Y are about 2 Å. The International Union of Pure and Applied Chemistry (IUPAC) recently updated their definition of hydrogen bonds [109] to give a more general and universal view of the phenomenon.

Hydrogen bonds are mainly electrostatic, but also include dispersion effects and contributions that are covalent in nature [109]. They play a crucial role in stabilizing structures and, accordingly, are considered perhaps the most important interactions in biological systems. Among their applications are the peculiar properties of water, the pivotal role in stabilizing secondary structure elements of proteins, DNA base-pairing, and interfaces of protein-protein/protein-DNA complexes. They are more favorable in protein-ligand systems than in protein-solvent ones [110]. In addition to their role in stabilization, hydrogen bonds play an important role in determining specificity [110,111].

Electrostatic contributions are calculated mathematically using Coulomb's law as in Eqn.7 [108]. This equation forms the basis for understanding interactions between ions. The corresponding potential energy is illustrated as a sum of interactions between pairs of point charges between two molecules, or between different parts of the same molecule.

$$v(q) = \sum_{i=1}^{N_A} \sum_{j=1}^{N_B} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \quad (7)$$

where,

q_i, q_j : are point charges.

ϵ_0 : is the vacuum permittivity.

r_{ij} : is the distance between the point charges.

N_A : is the number of point charges in the first molecule.

N_B : is the number of point charges in the second molecule.

2.5.2.2. van der Waals (vdW) Forces

vdW forces are composed of attractive and repulsive forces, as in Eqn.8. The attractive component is first explained through QM by Fritz London in 1930, and is sometimes called the London force [112], or dispersive force. It is the result of instantaneous dipole interactions caused by fluctuations of the electron clouds. The repulsive component is explained by Pauli's exclusion principle, and states that no two electrons in a given system can occupy the same area of space. When the electrons come close to each other and overlap, the electron density is reduced, resulting in less shielding between the positively charged nuclei, leading to electrostatic repulsion [108].

Not only do vdW interactions have a significant role in complexes interaction, but also in complexes stability. Since vdW interactions occur between any pair of atoms, they can increase in strength as the number of atoms increase, contributing to the overall stability of complexes. When the distance between the interacting particles increases, these interactions decay rapidly, but again the large number of interacting particles can still result in a significant contribution to the overall stability in complexes.

$$v(r) = 4\varepsilon \left[\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^6 \right] \quad (8)$$

where,

r: is the distance between two atoms from center to center.

ε : is the depth of the well.

σ : is the collision diameter.

2.5.2.3. Solvation Effect

The polar effect of solvation is the effect of ions from the solvent on complex formation. Due to the large number of solvent molecules, methods embedding Explicit Solvation effect can be very computationally expensive. Thus, implicit solvation or continuum solvation is a method of representing the solvent as a continuous medium instead of representing it as individual explicit solvent molecules. With the continuum method, the potential of mean force is applied to approximate the averaged behavior of many highly dynamic solvent molecules while considering the interiors of biological membranes (or proteins) as media with specific solvation or dielectric properties.

Although several implicit solvent models exist and various modifications and combinations of the different methods are possible, they are approximate methods with

certain limitations related to parameterization and treatment of ionization effects. We are considering recent continuum electrostatics models based on Poisson-Boltzmann equation (PB). These models describe the electrostatic environment of a solute in a solvent containing ions; they only include the enthalpic (bonded and non-bonded interactions) component of free energy, and are often applied to estimate the free energy of solute-solvent interactions in structural and chemical processes, such as the association of biological macromolecules with ligands. They proved efficient and accurate when applied to protein-ligand complexes; the computed/predicted binding free energy values are similar to the experimental ones.

The effect of solvent on complex formation might not be direct as in the case of polar effect. Non-polar effect involves the hydrophobic effect of the solvent on the complex when the solute is placed in solvation. Hydrophobic interactions are not attractive forces between particles, but rather a resulting effect from the solvent upon dissolving hydrophobic substances in it. Non-polar molecules or non-polar parts of the molecules will be surrounded by solvent (water) molecules that form a network of hydrogen bonds between themselves. This causes the hydrophobic units to aggregate together, maximizing the hydrophobic contacts, and resulting in a reduced combined surface area, when compared to a whole surface area formed from the separate units of the complex. This reduction in the exposed surface area minimizes the disruption of water, making the interaction between solutes favorable by way of increasing entropy. Hence, hydrophobic interactions facilitate complex formation. Proteins are known to have hydrophobic cores and protein-protein binding interfaces are found to be more hydrophobic than the solvated surface area [113].

CHAPTER 3

DEGRADOMICS METHODS AND RESULTS

The main target of the first objective of this thesis is to identify a number of Breakdown Products (BDPs), utilizing data extracted from experimental methods through the detection of substrate proteins. The goal is to computationally search through the set of detected proteins for potential breakdown sites, subject to fragmentation. Experience with tools developed based on some of the previously mentioned algorithms has pointed to their particular power in the context of preliminary results. However, due to the additional accuracy and efficiency requirements that are needed for our degradomics application, considering: a) large dataset (human genome size of ~30k protein sequences), b) short consensus sequences (motifs of ~8 amino acids), and c) flexibility of integrating consensus variants, this work presents a dynamic programming solution based on modifications to Smith-Waterman (SW) algorithm.

3.1. Cleaved Fragments Prediction Algorithm (CFPA)

The solution is based on local alignment and runs in time and space complexity of $O(mn)$ per one protein sequence of size n and one consensus sequence of size m . In addition, the algorithm is optimized to handle efficient memory utilization. For scalability, the mouse genome is used for substrate analysis. The computational search is performed using a set of pre-determined consensus sequences for *TllI*, which includes *SYAA↓DTAG* (\downarrow represents the proteolysis site corresponding to a specific cleavage mode) with its variants up to two mismatches. The different studies concluded that the

consensus site *SYAA↓DTAG* is required for *TllI* optimal activity although other derived sites could also be cleaved with a much reduced efficiency [15].

Next, we present the overall developed algorithm Cleaved Fragments Prediction Algorithm (CFPA) for all consensus occurrences and fragments identification. It embeds a modification version of Smith-Waterman algorithm (SW) [97]. SW is based on dynamic programming that builds a scoring table, as shown in Fig. 2.3, and performs local sequence alignment. It finds regions in the input protein sequence that match a consensus sequence, including variants of the consensus sequence up to two mismatches.

Moreover, the algorithm omits all occurrences with INserts/DELetes or INDELs (as shown in the mathematical formulation of SW algorithm). Finally, the algorithm includes a fragment generation module that generates all fragments based on all consensus occurrences, their different orientations, and all their combinations (since different occurrences might occur separately or simultaneously).

Cleaved Fragments Prediction Algorithm (CFPA)

Algorithm: Read Protein Sequences and a Consensus sequence and Output Occurrences & Fragments

Input: String Consensus Sequence Initial C-N, String Consensus Sequence Reversed C-C, String Input Protein Sequence Initial S-N, String Input Protein Sequence Reversed S-C, int Cleavage Site Initial I, int Cleavage Site Reversed I_Rev, N Input Protein Sequences, int Mismatch Penalty Cost

Output: Occurrences and Fragments with Start and End Positions per Protein Seq. based on 4-WAY

BEGIN

While Not End Of Input Protein Sequences {Process the 4-WAY of each Protein sequence}

 Process_Seq_Cons(S-N, C-N, I); Flush RAM;

 Process_Seq_Cons(S-N, C-C, I_Rev); Flush RAM;

 Process_Seq_Cons(S-C, C-N, I); Flush RAM;

 Process_Seq_Cons(S-C, C-C, I_Rev); Flush RAM;

End While

Process_Seq_Cons (string In_Seq, string Cons_Seq, int Clvg_Site)

 Run SW Algorithm on Cons_seq and In_seq, including generation of index tables

For All Paths tracked

While Not End Of Path

If an INDEL is encountered {INDEL stands for an INsert/DElete gap in the Scoring Table}

 Set Path \leftarrow Skip-Path {Reject Path, Modification to SW Alg.}

End While

If the Path is a Skip-Path

 Continue

Else Reset to Start of Path

 {Control Mismatches #}

While Not End Of Path

If a Mismatch is encountered

 Path_Mismatch_Count \leftarrow Path_Mismatch_Count + 1

End While

If Path_Mismatch_Count \leq 2 {Accept Path, Modification to SW Alg}

 Add Consensus Occurrence node to Occurrences_Linked_List

 Update node with Occurrence string, Match Type, Start & End Positions in Sequence

 Cut \leftarrow Path_Index - Clvg_Site + 1 {Compute Consensus Occurrence Cut }

 Add Cut to vector Cuts {Add to Vector of all possible Cuts in the Protein Sequence}

 Occurrence_Count \leftarrow Occurrence_Count + 1

End For

 {Generate fragments based on Occurrence_Count; different cuts generate different combinations}

For Possible_Occurrence_Count \leftarrow 1 to Occurrence_Count

 {k=Possible_Occurrence_Count}

 Generate_Fragments(Possible_Occurrence_Count, In_Seq)

End For

Generate_Fragments(int Possible_Occurrence_Count, string In_Seq)

 {There is a vector for each combination generated, & and it is based on Cuts vector elements.}

If Possible_Occurrence_Count \leftarrow 0 {Exit Condition of Recursive Function}

 Sort Combination vector containing one combination of cuts

 Generate fragments for a specific combination of Cuts in the Combination vector

 Add each fragment to a Fragments_Linked_List

 Output all fragments per specific combination of Cuts

Else

 {Generate a new combination based on the modified Possible_Occurrence_Count value}

For i \leftarrow 0 to (Cuts.Size - Possible_Occurrence_Count)

 Push into Combination vector the Value of Cuts[i]

 Generate_Fragments(Possible_Occurrence_Count - 1, In_Seq)

 Pop last element from Combination vector

End For

End Elseif

END

The above algorithm executes in space complexity $O(mn)$ and time complexity $O(N[mn + q(q^k)(k \log k + n)])$, where N is the total number of input protein sequences, n is the size of each protein sequence, m is the size of the consensus sequence, q is the number of consensus occurrences, and k is an integer value from 1 to q . It takes $O(N)$ to process all input protein sequences. For each protein sequence and each consensus sequence, it takes $O(mn)$ in space and time to find all occurrences of a specific consensus within a single protein sequence. It takes $O(q * T(k))$ to generate all fragments corresponding to q , where $T(k)$ is a recursive function that entails generating all possible combinations of occurrences and sorting them in $O(k \log k)$. Since $q \ll n$ and $k \ll n$, the overall algorithm time complexity reduces to $\sim O(Nmn)$ for processing N protein sequences; the space complexity remains $O(mn)$, the size of the dynamic table.

Since multiple occurrences of a specific consensus can be found in one protein sequence, all combinations of potential cuts, resulting in different output fragments are considered. For instance, if two consensus occurrences are matched in an input sequence, then the number of combinations to consider for different output fragments is as follows: 1) as if occurrence 1 is only found and breaks the input sequence at the cleavage site, or 2) as if occurrence 2 is only found and breaks the input sequence at a different cleavage site, or 3) as if occurrence 1 and occurrence 2 are both found and break the input sequence at the two different cleavage sites simultaneously, resulting in three fragments compared to two fragments with occurrence 1 alone or occurrence 2 alone.

Small-size dataset is created to be used for preliminary testing of the exceptional pitfalls of the method. The different tests are illustrated in subsequent

sections, Sections 3.1.1-3.1.5. Each of these subsections illustrates a specific test case where the input sequence and the consensus sequence are presented. As shown below, the consensus sequence shows the *cleavage site*, illustrated with a green arrow. It is selected at random in these scenarios for testing purposes, but can be in any position within the consensus sequence.

3.1.1. Matches, Mismatches, and Pruning of Gaps

This case tests for matches, mismatches, and how any occurrence including gaps - due to deletes or inserts - is omitted from the output. It uses the simulated data shown in Fig. 3.1. Since CFPA is based on local alignment, three resulting alignments are shown in Fig. 3.2. The alignments are outputted in Table 3.1 and the corresponding fragments are outputted in Table 3.2. The red traces in the Scoring and Alignment Table show two consensus occurrences with no mismatches and one consensus occurrence with one mismatch. The Indices tables, on the other hand, show the steps to follow the trace in the Scoring and Alignment Table along the i and j directions, as shown in Fig. 3.3. Blue color in Fig. 3.2 represents two alignments that have gaps or INDELS, and so are not shown in the output Table 3.1. Orange color in the Indices Table (Fig. 3.3) represents the common steps shared by one red alignment and two blue ones.

Consensus Sequence with Cleavage Site
 >A↑GC

Input Sequence and Consensus Occurrences
 1 2 3 4 5 6 7 8 9 10 11
 >A↑GCTA↑GCT↑GC

Fig. 3.1 - Simulated data for matches, mismatches, and pruning of gaps

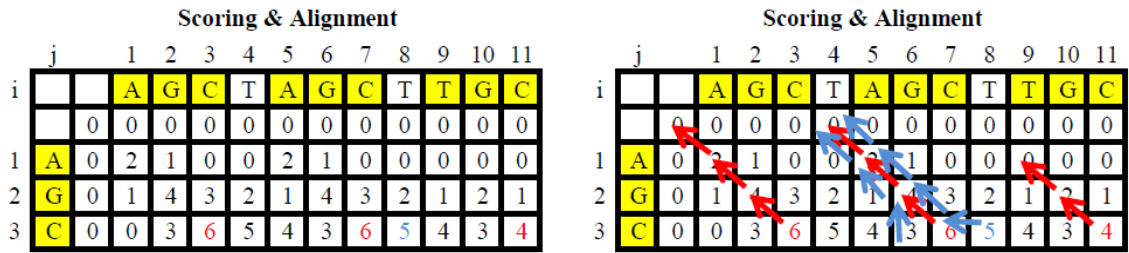


Fig. 3.2 – Dynamic table comprising scoring and alignments. Representation of the dynamic table comprising scoring and alignments. Blue paths represent alignments with gaps (insertion or deletion). Red paths represent alignments with matches or mismatches.

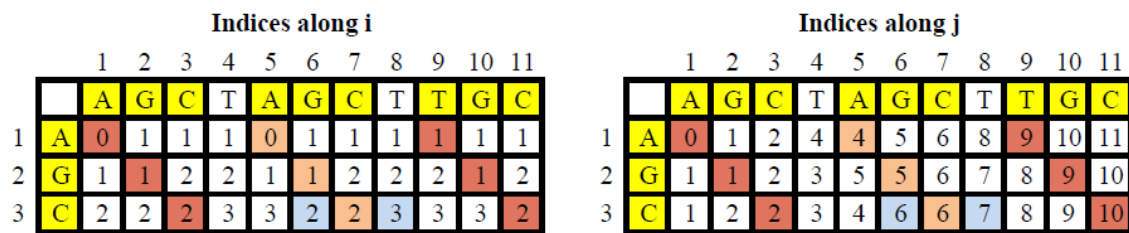


Fig. 3.3 – Indices along the consensus and the protein. Representation of the index i (next step location) along the consensus sequence and the index j (next step location) along the protein sequence to track each path alignment. Red boxes represent steps of an acceptable path (alignment), blue boxes represent gaps (insertion or deletion), and orange boxes represent overlaps among paths (i.e., here covering blue and red paths).

Table 3.1 - Output of simulated data with pruning of deletes and inserts

Count	Occurrence	Start	End	Type
1	AGC	1	3	Match
2	AGC	5	7	Match
3	TGC	9	11	Mismatch

Table 3.2 - Generated fragments based on consensus occurrences in Table 3.1

Combination	Fragment	Start	End
Occurrence 1	A	1	1
	GCTAGCTTGC	2	11
Occurrence 2	AGCTA	1	5
	GCTTGC	6	11
Occurrence 3	AGCTAGCTT	1	9
	GC	10	11
Occurrences 1&2	A	1	1
	GCTA	2	5
	GCTTGC	6	11
Occurrences 1&3	A	1	1
	GCTAGCTT	2	9
	GC	10	11
Occurrences 2&3	AGCTA	1	5
	GCTT	6	9
	GC	10	11
Occurrences 1&2&3	A	1	1
	GCTA	2	5
	GCTT	6	9
	GC	10	11

3.1.2. Excluding Paths with Three-or-More-Mismatches

This case tests for omission or pruning of alignments that correspond to three mismatches. It uses the simulated data shown in Fig. 3.4. Two resulting alignments are shown in Fig. 3.5. The red traces in Fig. 3.5 show one occurrence with one mismatch and another with three mismatches. As shown in Table 3.3, only the occurrence with one mismatch is outputted by the algorithm. Its corresponding fragments, resulting from cleavage, are shown in Table 3.4.

Consensus Sequence with Cleavage Site
 >GGATT↑CA

Input Sequence and Consensus Occurrences
 1 2 3 4 5 6 7 8 9 10 11
 > GAACGCATTCA
 > GAACG↑CATTCA
 > GAACGCATT↑CA

Fig. 3.4 - Simulated data for excluding 3-or-more mismatches

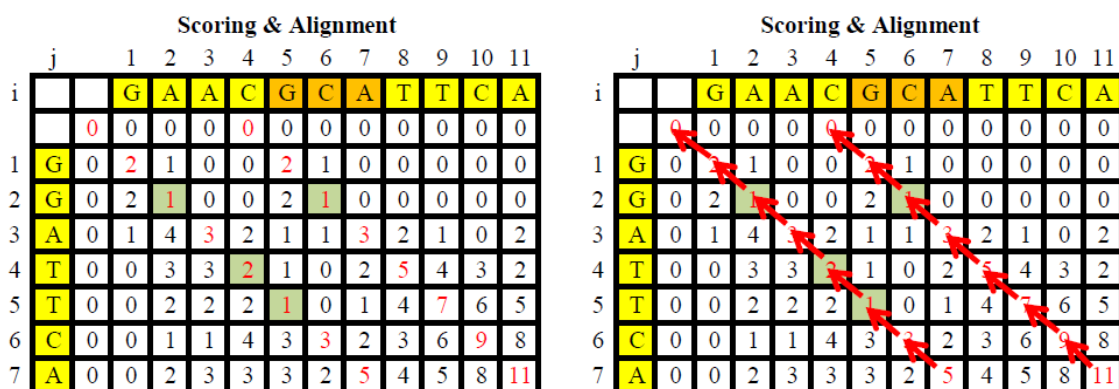


Fig. 3.5 – Two paths, one mismatch, and three mismatches. Representation of two paths: One path with one mismatch (indicated by one green box) and another path with three mismatches (indicated by three green boxes).

Table 3.3 - Output of simulated data with pruning of at least three mismatches

Count	Occurrence	Start	End	Type
1	GCATTCA	5	11	Mismatch

Table 3.4 - Generated fragments based on consensus occurrences in Table 3.3

Combination	Fragment	Start	End
Occurrence 1	GAACGCATT	1	9
	CA	10	11

3.1.3. Sub-Matches within Two Mismatches and Overlaps

This case tests for partial-matches up to two mismatches. It uses the simulated data shown in Fig. 3.6. Looking at one of the two occurrences in Fig. 3.6, BLAST outputs ‘ATTCA’ as the best matching occurrence it can find. As shown in Fig. 3.7, the algorithm adjusts the match to extend the occurrence with two mismatches (shown as green boxes). In addition, this test shows the case of overlaps. Overlaps among the different occurrences of one consensus sequence are probable to appear within the same protein sequence. As shown in output Table 3.5, the purple color is presented to

highlight the overlapping regions detected by the algorithm. Table 3.6 shows the fragments generated after cleavage of the consensus occurrences.

Consensus Sequence with Cleavage Site

>GGATT↑CA

Input Sequence and Consensus Occurrences

1 2 3 4 5 6 7 8 9 10 11

> GAATTCATTCA

> GAATT↑CATTCA

(One occurrence with one mismatch is shown in Table 3.5. Generated fragments are shown in Table 3.6)

> GAAC↑TCATTCA

(A second consensus occurrence with two mismatches is shown in Table 3.5. Generated fragments are shown in Table 3.6)

Fig. 3.6 - Simulated data for sub-matches within 2 mismatches and overlaps

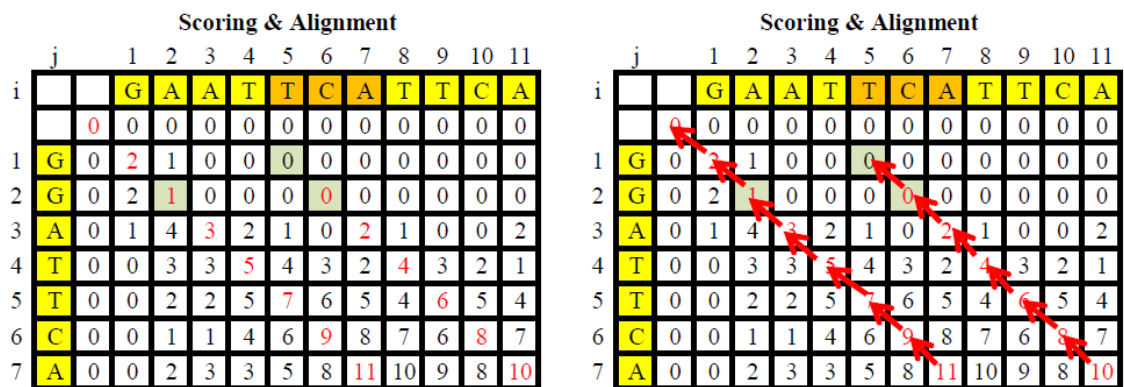


Fig. 3.7 – Two accepted paths, one mismatch, and two mismatches. Representation of two paths accepted by the algorithm: One path with one mismatch (indicated by one green box) and another path with two mismatches (indicated by two green boxes). The two-mismatches-path reveals two mismatches at the end of the path; those two mismatches are missed in BLAST. In addition, the two paths shown depict an instance of two overlapping alignments.

Table 3.5 - Output of simulated data with overlaps of consensus occurrences

Count	Occurrence	Start	End	Type
1	GAATTC	1	7	Mismatch
2	TCATTC	5	11	Mismatch

Table 3.6 - Generated fragments based on consensus occurrences in Table 3.5

Combination	Fragment	Start	End
Occurrence 1	GAATT	1	5
	CATTCA	6	11
Occurrence 2	GAATTCATT	1	9
	CA	10	11
Occurrences 1&2	GAATT	1	5
	CATT	6	9
	CA	10	11

3.1.4. INserts/DELetes (INDEL) after Consensus Occurrences

This case tests for deletes or inserts in the path after a consensus occurrence is found and uses the simulated data shown in Fig. 3.8. While the algorithm excludes any path with deletes or inserts, it makes an exception if they are found after a consensus occurrence. Fig. 3.9 shows the consensus occurrence path with an INDEL in the first row. Table 3.7 shows the consensus occurrence and Table 3.8 shows the corresponding fragments after cleavage.

Consensus Sequence with Cleavage Site

>GATD↑AAYS

Input Sequence and Consensus Occurrences

1 2 3 45678 910111213141516

> FVGLLATDAGYSELFM

> FVGLLATD↑AGYSELFM

(One consensus occurrence with two mismatches is shown in Table 3.7. Generated fragments are shown in Table 3.8)

Fig. 3.8 - Simulated data for INDEL after consensus occurrences

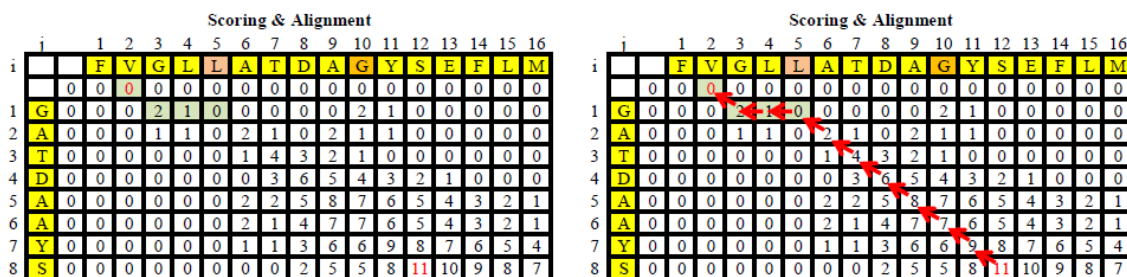


Fig. 3.9 – Alignment with an INDEL at the last base. Representation of a special alignment showing a gap/INDEL at the last base of the consensus occurrence. This path is not excluded by the algorithm and adds the last base as a mismatch to the consensus occurrence.

Table 3.7 - Output of simulated data with an INDEL after a consensus

Count	Occurrence	Start	End	Type
1	LATDAGYS	5	12	Mismatch

Table 3.8 - Generated fragments based on consensus occurrences in Table 3.7

Combination	Fragment	Start	End
Occurrence 1	FVGLLATD	1	8
	AGYSEFLM	9	16

3.1.5. Handling of Four-Way Protein and Consensus Orientations

The protein sequence and consensus sequence can take different orientations, driving many more different occurrences of the consensus, and consequently different fragmentations of the protein sequence [114]. The following four subsections depict all four possibilities of both the protein and consensus sequences orientations, with the corresponding occurrences and fragments. Section 3.1.5.1 illustrates the case when the protein and consensus sequences are in their initial orientations. Section 3.1.5.2 illustrates the change when the consensus sequence reverses its orientation, but the protein sequence keeps its initial orientation. Section 3.1.5.3 illustrates the change when the protein sequence reverses its orientation, but the consensus sequence keeps its initial

orientation. Finally, Section 3.1.5.4 illustrates the change when both of the protein and consensus sequences reverse their orientations compared to their initial orientations.

3.1.5.1. Protein and Consensus Initial Orientations (NN)

The following presents a sample simulated data of a protein sequence and a consensus sequence in their initial orientations N-terminal/N-terminal (NN), as shown in Fig. 3.10. With this arrangement, the algorithm shows one possible combination of the consensus occurrence within the protein sequence, including its type and position, as shown in the output Table 3.9. In addition, the corresponding generated fragments are shown with their positions.

```

Consensus Sequence Initial Orientation with Cleavage Site
>GGATT↑CA

Protein Sequence Initial Orientation
 1 2 3 4 5 6 7 8 9 10 11 12 13
>ACTTACATTCATT

One Occurrence with Two Mismatches
 1 2 3 4 5 6 7 8 9 10 11 12 13
>ACTTACATT↑CATT
  
```

Fig. 3.10 - Simulated data for protein-consensus NN orientations

Table 3.9 - Output of simulated data for protein-consensus NN orientations

Count	Occurrence	Start	End	Type	Combination	Fragment	Start	End
1	ACATTCA	5	11	Mismatch	Occurrence 1	ACTTACATT CATT	1 10	9 13

3.1.5.2. Protein Initial and Consensus Reversed (NC)

The following presents a sample simulated data of a protein sequence and a consensus sequence in their initial and reversed orientations N-terminal/C-terminal

(NC) respectively, as shown in Fig. 3.11. With this arrangement, the algorithm shows one possible combination of the consensus occurrence within the protein sequence, including its type and position, as shown in the output Table 3.10. In addition, the corresponding generated fragments are shown with their positions.

```

Consensus Sequence Orientation Reversed with Cleavage Site Changed
>AC↑TTAGG

Protein Sequence Initial Orientation
 1 2 3 4 5 6 7 8 9 10 11 12 13
>ACTTACATTCATT
One Consensus Occurrence with Two Mismatches
 1 2 3 4 5 6 7 8 9 10 11 12 13
>AC↑TTACATTCATT

```

Fig. 3.11 - Simulated data for protein-consensus NC orientations

Table 3.10 - Output of simulated data for protein-consensus orientations

Count	Occurrence	Start	End	Type	Combination	Fragment	Start	End
1	ACTTACA	1	7	Mismatch	Occurrence 1	AC TTACATTCATT	1 3	2 13

3.1.5.3. Protein Reversed and Consensus Initial (CN)

The following presents a sample simulated data of a protein sequence and a consensus sequence in their reversed and initial orientations C-terminal/N-terminal (CN) respectively, as shown in Fig. 3.12. With this arrangement, the algorithm shows one possible combination of the consensus occurrence within the protein sequence, including its type and position, as shown in the output Table 3.11. In addition, the corresponding generated fragments are shown with their positions.

Consensus Sequence Initial Orientation with Cleavage Site
 >GGATT↑CA

Protein Sequence Orientation Reversed
 1 2 3 4 5 6 7 8 9 10 11 12 13
 >TTACTTACATTCA

One Occurrence with Two Mismatches
 1 2 3 4 5 6 7 8 9 10 11 12 13
 >TTACTTACATT↑CA

Fig. 3.12 - Simulated Data for Protein-Consensus CN Orientations

Table 3.11 - Output of simulated data for protein-consensus CN orientations

Count	Occurrence	Start	End	Type	Combination	Fragment	Start	End
1	ACATTCA	7	13	Mismatch	Occurrence 1	TTACTTACATT CA	1 12	11 13

3.1.5.4. Protein and Consensus Orientations Reversed (CC)

The following presents a sample simulated data of a protein sequence and a consensus sequence in their reversed orientations C-terminal/C-terminal (CC), as shown in Fig. 3.13. With this arrangement, the algorithm shows one possible combination of the consensus occurrence within the protein sequence, including its type and position, as shown in the output Table 3.12. In addition, the corresponding generated fragments are shown with their positions.

Consensus Sequence Orientation Reversed with Cleavage Site Changed
 >AC↑TTAGG

Protein Sequence Orientation Reversed
 1 2 3 4 5 6 7 8 9 10 11 12 13
 >TTACTTACATTCA

One Occurrence with Two Mismatches
 1 2 3 4 5 6 7 8 9 10 11 12 13
 >TTAC↑TTACATTCA

Fig. 3.13 - Simulated data for protein-consensus CC orientations

Table 3.12 - Output of simulated data for protein-consensus CC orientations

Count	Occurrence	Start	End	Type	Combination	Fragment	Start	End
1	ACTTACA	5	11	Mismatch	Occurrence 1	TTAC	1	4
						TTACATTCA	5	13

3.1.6. Application to Tll1 Metalloprotease

3.1.6.1. Problem Definition

The goal is to search for a consensus sequence in a huge set of input protein sequences. A consensus sequence is a specific subsequence that allows an enzyme to cut an input protein sequence at a specific location, resulting into fragment subsequences or BDPs. The location of the cut, or *cleavage site*, is predefined in the consensus sequence. When a match is found between a consensus sequence and its occurrence in an input sequence, the cut of the input sequence is performed at the *cleavage site* in the occurrence sequence, at the same position of the *cleavage site* located in the consensus sequence.

While finding exact matches of the consensus sequence in all input proteins sequences is the primary target, variants of exact matches are studied too. Variants can be within one to two mismatches and within any position of the consensus; they might also result in cleaving the input sequence into fragments when such variants occurrences are found in the input sequence. Since either the protein sequence or the consensus sequence might be recognized from the C-terminal end or the N-terminal end, the suggested solution is required to handle the four different orientations (called 4-WAY). Moreover, it is recommended for the solution to handle possible overlapping occurrences of the consensus. Finally, the solution must also be accurate, efficient, and validated against experimental findings.

3.1.6.2. Data: Mouse Genome

The need for an efficient algorithm is highly marked with real data. For testing a large dataset, the whole Mouse Genome [115] is used as input. It comprises ~35k protein sequences after preprocessing the dataset, excluding redundant and erroneous data. The largest protein sequence size in this dataset is 5,379 chars/amino acids and the shortest protein sequence size is 7 chars/amino acids. The *Tlll* subsequence **SYAA↓DTAG** is considered as the consensus test sequence, with the down-arrow showing the *cleavage site*. The consensus **SYAA↓DTAG** [15], including its variants up to two mismatches, is searched for by the *Tlll* Metalloproteinase within all input protein sequences.

The output is expected to show all hits of the consensus in all input protein sequences, including exact matches and variants up to two mismatches. In addition, the output is expected to show the generated fragments upon cleavage. Nevertheless, the algorithm considers outputting the occurrences that can happen in any orientation, as illustrated in Section 3.1.5.

3.1.6.3. Computational Results

Results statistics of CFPFA run on the complete mouse genome are shown in Fig. 3.14 and Fig. 3.15. Fig. 3.14 shows histograms of the number of consensus occurrences. The output data includes all consensus occurrences, considering whether each occurrence type is an exact match (*hit*), with one mismatch, or with two mismatches. Furthermore, the algorithm outputs the start and end position of each occurrence within each protein sequence, and outputs the generated fragments based on

each possible combination of occurrences, with each fragment start and end position; due to the large sized output, this data is not presented here.

Moreover, the algorithm outputs the occurrences that can appear in any one of the four possible orientations (NN, CN, NC, and NN) of each of the consensus and protein sequences. Fig. 3.15 shows an output spectrum of all protein sequences with the number of mismatches, and according to the four possible orientations. Accordingly, the figure shows how an occurrence that shows in NN/CC orientation does not show in NC/CN orientation and vice versa, among all protein sequences.

Table 3.13 and Table 3.14 show all the consensus occurrences or hits (with the mismatched amino acids shown in red) and the description of the corresponding generated fragments (also known as biomarkers or identified peptides). Table 3.13 shows the consensus occurrences for NN and CC orientations. The row in blue, which corresponds to a hit with one mismatch and labeled Collagen Alpha-I chain (VII) chain (Col7a1), is verified experimentally.

Table 3.14, on the other hand, shows the consensus occurrences for NC and CN orientations. Since it is less likely to have cleavage occurring in the presence of two mismatches, the experimental validation is limited to the consensus occurrence with one mismatch (Col7a1).

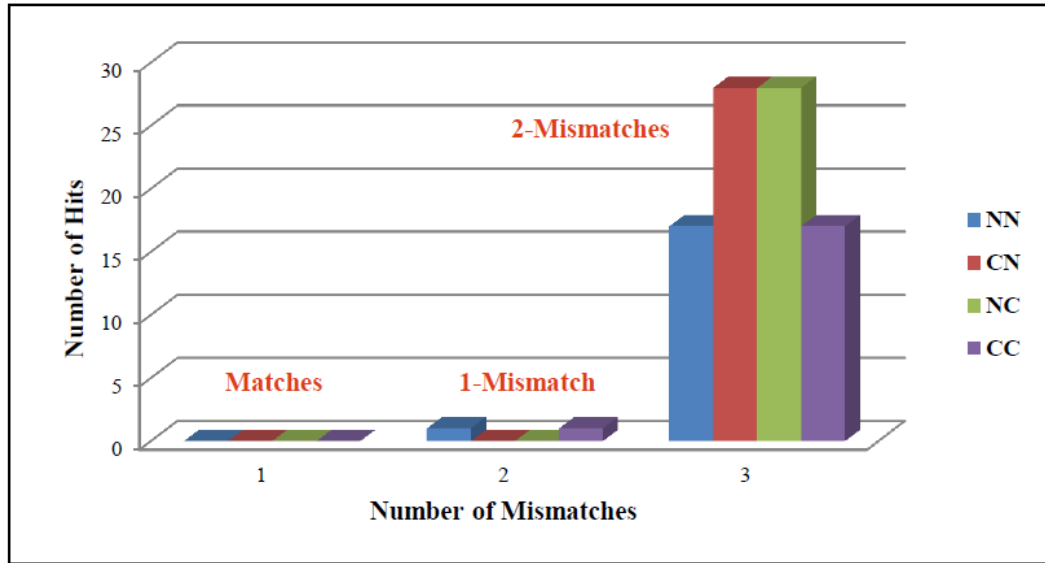


Fig 3.14 – Histogram of consensus occurrences with all types of matches. Histogram depicts all consensus occurrences with complete matches, one mismatch, or two mismatches. Blue colored bar depicts the statistics when both of the protein sequence and the consensus sequence orientations are the same (initial). Purple bar depicts the case when both of the consensus sequence and the protein sequence orientations are reversed. Red bar depicts the case when only the consensus sequence orientation is reversed, and green bar depicts the case when only the protein sequence orientation is reversed. As can be seen, the NN and CC orientations represent the same statistics, and so is the case for CN and NC orientations.

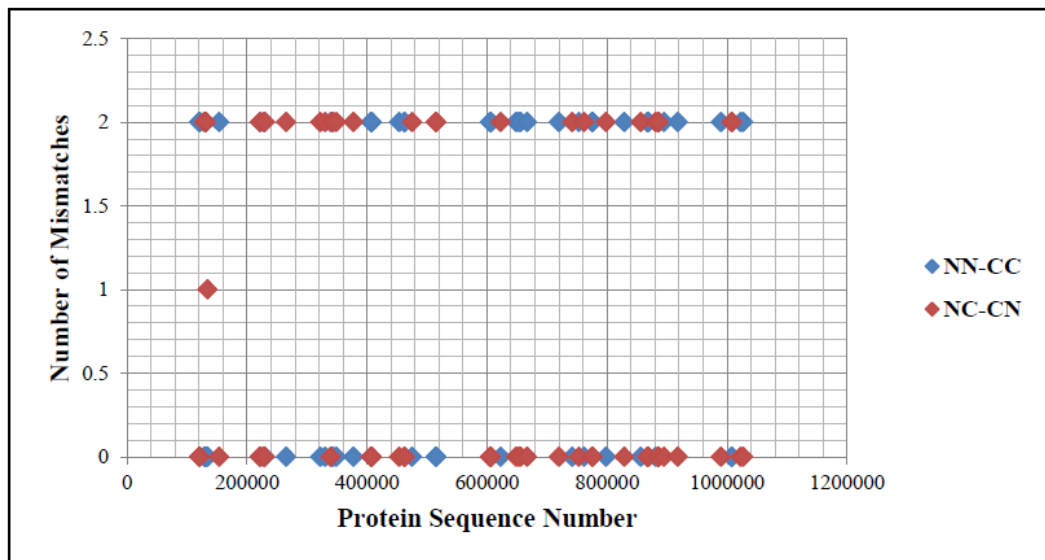


Fig. 3.15 – Protein sequences handled through NN-CC and NC-CN. Representation of some of the protein sequences handled through NN-CC and NC-CN. Figure illustrates the fact that protein sequences handled by NN-CC, are not handled by NC-CN, and vice-versa throughout the mouse genome data (i.e., 1 or 2 mismatches found through NN-CC correspond to zero mismatches found through NC-CN and vice-versa).

3.1.6.4. Experimental Validation

Elevated expression levels of collagen are diagnosed experimentally in healthy individuals upon studying Congenital Heart Disease (CHD) and other aortic disorders [116,117]. Such data is used for validating our computational result of the generated biomarker col7a1 (Collagen alpha-1(VII) chain) upon cleavage of substrate by the *Tlll* gene, indicating the absence of CHD, and shown in Table 3.13.

Table 3.13 - Mouse proteome output in NN and CC orientations

4-WAY Hits for DNA-Sequence Orientation Reversed OR Protein-Sequence Orientation Reversed			
Gene Symbol	Description	NN	CC
Atp6ap1	V-type proton ATPase subunit S1	SYASDCAG	GACDSAYS
Es22	Liver carboxylesterase 22	SLAAFTAG	GATFAALS
Aga	N(4)-(beta-N-acetylglucosaminy)-L-asparaginase	AYADDTAG	GATDDAYA
Catsper2	Cation channel sperm-associated protein 2	STAADTAF	FATDAATS
Col7a1	Collagen alpha-1(VII) chain	GYAADTAG	GATDAAYG
Olfml2a	Olfactomedin-like protein 2A	SKAQDTAG	GATDQAKS
Recql	Isoform Beta of ATP-dependent DNA helicase Q1	SHAADTAA	AATDAAHS
Nlr1	NLR family member X1	SYAARTMG	GMTRAAYS
Kank2	Isoform 1 of KN motif and ankyrin repeat domain-containing protein 2	SQAADGAG	GAGDAAQS
Fbxo38	F-box only protein 38	STAASTAG	GATSAATS
Scaf1	Isoform 1 of Splicing factor, arginine/serine-rich 19	SGAADTAT	TATDAAGS
Mesdc2	LDLR chaperone MESD	AYAADTPG	GPTDAA YA
Mbd6	methyl-CpG-binding domain protein 6	SSAADRAG	GARDAASS
Ewsr1	RNA-binding protein EWS	SYAAQTAY	YATQAAYS
Kank2	Isoform 2 of KN motif and ankyrin repeat domain-containing protein 2	SQAADGAG	GAGDAAQS
Scaf1	Isoform 2 of Splicing factor, arginine/serine-rich 19	SGAADTAT	TATDAAGS
Mbd6	Methyl-CpG-binding domain protein 6	SSAADRAG	GARDAASS
Aga	N(4)-(beta-N-acetylglucosaminy)-L-asparaginase isoform 2 precursor	AYADDTAG	GATDDAYA

3.1.6.5. Summary

The proposed CFPA algorithm is simple, robust, and efficient. It is based on Smith-Waterman algorithm with a few modifications to consider variants (one to two mismatches) in addition to exact matches between the consensus sequence and all input protein sequences. The algorithm prunes all alignments with deletions and insertions. After all consensus occurrences (hits) are found, further modules are added to generate

all fragments as a result of cleaving each input sequence at the cleavage site; this specific site is predefined within the consensus sequence, and mapped accordingly with every consensus occurrence. Due to different protein conformations, the cleavage can happen in all different combinations of occurrences. Hence, fragments resulting from all possible combinations are generated.

Heuristic algorithms produce faster alignments, but are at the cost of reduced sensitivity. Increasing seed size decreases sensitivity and, inversely, decreasing seed size increases sensitivity. CFPA is based on dynamic programming; this guarantees high sensitivity as it is not based on heuristics.

While pairwise alignment algorithms are classified as optimal or heuristic, they are further classified as local or global. Smith-Waterman algorithm is based on local alignment and, accordingly, is selected within CFPA to support consensus searches of a few bases within input protein sequences of thousands amino acids.

In an effort of assessing the utility of the mouse genome proteolysis and its characteristic *TIII* breakdowns as potential markers of CHD, the developed algorithm CFPA is applied to the whole mouse genome (~30k protein sequences with each protein sequence up to ~5k amino acids) to locate consensus sequence occurrences and generate and identify corresponding fragments from such bulky records.

For the current application and data size, CFPA proved its high level performance. The results of the mouse genome showed that CFPA can detect efficiently and with high sensitivity (see hits in Tables 3.13 and 3.14) the regions in the input protein sequences that are similar (within 1 to 2 mismatches) to the query sequence (consensus sequence).

Table 3.14 - Mouse proteome output in NC and CN orientations

4-WAY Hits for Both of DNA-Seq. and Protein-Seq. Orientations Normal OR Reversed			
Gene Symbol	Description	CN	NC
A2bp1	Isoform 1 of RNA binding protein fox-1 homolog 1	PATAAAYS	SYAAATAP
C8b	Isoform 1 of Complement component C8 beta chain	TATDFAYS	SYAFDTAT
Brd8	Isoform 1 of Bromodomain-containing protein 8	KATDAAYQ	QYAADTAK
D11Bwg0517e	Isoform 1 of RNA binding protein fox-1 homolog 3	AATAAAYS	SYAAATAA
Parp12	Poly [ADP-ribose] polymerase 12	FARDAAYS	SYAADRAF
Parp11	Isoform 1 of Poly [ADP-ribose] polymerase 11	FARDAAYS	SYAADRAF
Zyg11b	Isoform 1 of Protein zyg-11 homolog B	LATDAGYS	SYGADTAL
Astn2	astrotactin-2 isoform a	GATAAAAS	SAAAATAG
C8b	Isoform 2 of Complement component C8 beta chain	TATDFAYS	SYAFDTAT
D11Bwg0517e	Isoform 3 of RNA binding protein fox-1 homolog 3	AATAAAYS	SYAAATAA
Ctnbnp2nl	CTTNBP2 N-terminal-like protein	GPTTAAYS	SYAATTPG
Astn2	Isoform 1 of Astrotactin-2	GATAAAAS	SAAAATAG
A2bp1	Isoform 7 of RNA binding protein fox-1 homolog 1	PATAAAYS	SYAAATAP
A2bp1	Isoform 3 of RNA binding protein fox-1 homolog 1	PATAAAYS	SYAAATAP
A2bp1	Isoform 6 of RNA binding protein fox-1 homolog 1	PATAAAYS	SYAAATAP
D11Bwg0517e	RNA binding protein fox-1 homolog 3 isoform 3	AATAAAYS	SYAAATAA
Brd8	Isoform 2 of Bromodomain-containing protein 8	KATDAAYQ	QYAADTAK
Dnajb14	Isoform 3 of DnaJ homolog subfamily B member 142	GATDAFKS	SKFADTAG
Lgr4	Leucine-rich repeat-containing G-protein coupled receptor 4	GATDAANA	ANAADTAG
Parp11	Isoform 2 of Poly [ADP-ribose] polymerase 11	FARDAAYS	SYAADRAF
Zyg11b	Isoform 3 of Protein zyg-11 homolog B	LATDAGYS	SYGADTAL
Zyg11b	Isoform 2 of Protein zyg-11 homolog B	LATDAGYS	SYGADTAL
Astn2	Isoform 2 of Astrotactin-2	GATAAAAS	SAAAATAG
A2bp1	Isoform 5 of RNA binding protein fox-1 homolog 1	PATAAAYS	SYAAATAP
A2bp1	Isoform 4 of RNA binding protein fox-1 homolog 1	PATAAAYS	SYAAATAP
A2bp1	Isoform 2 of RNA binding protein fox-1 homolog 1	PATAAAYS	SYAAATAP
D11Bwg0517e	Isoform 2 of RNA binding protein fox-1 homolog 3	AATAAAYS	SYAAATAA
Tmem132c	transmembrane protein 132C precursor	GATDIAVS	SVAIDTAG

Table 3.13 shows the consensus occurrences (*hits*) within 1 mismatch (Col7a1) and within 2 mismatches for the two orientations NN and CC. Conversely, Table 3.14 shows the consensus occurrences (*hits*) within 2 mismatches for the two orientations NC and CN. Instances of the corresponding generated fragments are shown in Appendix 1.

Future work includes the development of a web-based front end for online users, with a database back end for storing all useful protein, consensus, and fragments sequences. This potential CFPA application can be a valuable, prognostic, and diagnostic tool for molecular biologists and general users, with more functionality to be added on.

3.2. Cleaved Fragments Prediction Algorithm with Applications

CFPA-CalpCasp algorithm is a modification of CFPA algorithm [118] for specific application to calpain and caspase cleavage modes. It is based on dynamic programming and performs Local Sequence Alignment [97] after building a scoring table and removing all occurrences with INserts or DELetes (INDELs). It finds all local regions in each input protein sequence that match a consensus sequence. Instead of searching for a single consensus with all its possible matches and variants as in CFPA algorithm, CFPA-CalpCasp looks for exact matches of every consensus; the consensus variants are built within the consensus patterns specific to calpain and caspase, representing a new cleavage mode. The cleaving style is based on fixing certain amino acids and varying others, making the *cleavage site* right after consensus *hit*. For all N *input* protein sequences, CFPA-CalpCasp algorithm finds all occurrences of exact matches of every *consensus* sequence. To process N input protein sequences, the algorithm executes N times.

In the next section, Section 3.2.1, we explain the CFPA-CalpCasp algorithm in detail. In addition of finding exact matches, the algorithm includes a fragment generation module that finds all fragments based on all occurrences and their combinations. Then, Section 3.2.2 presents an application of the algorithm cleaving α II-spectrin by calpain and caspase proteases. Afterwards, Section 3.2.3 presents another application of the algorithm cleaving β II-spectrin by calpain and caspase proteases.

Cleaved Fragments Prediction Algorithm (CFPA-CalpCasp)

Algorithm: Read Protein Sequences and Consensus sequences and Output Occurrences & Fragments based on the specific cleavage mode of calpain-2 and caspase-3.

Input: N Protein Sequences, N' Consensus Sequences

Output: Occurrences and Fragments with Start and End Positions

BEGIN

While Not End Of Input Protein Sequences

While Not End Of Consensus Sequences

 Run SW Algorithm on Cons_Seq and In_Seq

For All Paths tracked

While Not End Of Path

If an INDEL is encountered

 {Reject Path, Modification to SW Alg.}

 Set Path \leftarrow Skip-Path **fi**

End While

If the Path is a Skip-Path

 Continue

Else Reset to Start of Path **fi**

While Not End Of Path-O(m)

If a Mismatch is encountered

 Path_Mism_Cnt \leftarrow Path_Mism_Cnt + 1 **fi**

End While

 {Accept Path, Modification to SW Alg}

If (Path_Mism_Cnt == 0)

 Add Cons. Occ. node to Occ_Linked_List

 Update node w/ Occ. Details

 {Compute Consensus Occurrence Cut}

 Cut \leftarrow Path_Index - 1

 {Add to vector of all Cuts on Protein Sequence}

 Add Cut to vector Cuts

 Occ_Cnt \leftarrow Occ_Cnt + 1 **fi**

End For

 {Generate Fragments based on Occ_Cnt}

For Poss_Occ_Cnt \leftarrow 1 to Occ_Cnt

Gen_Frags(Poss_Occ_Cnt, In_Seq)

End For

End While {All Consensus Sequences}

End While {All Input Protein Sequences}

Gen_Frags(In: Poss_Occ_Cnt, In_Seq)

 {There is a vector for each combination generated}

If Poss_Occ_Cnt \leftarrow 0 {Exit Cond. of Recur. Funct.}

 Sort Comb. vector w/ 1 Combination of Cuts

 Gen. Frags. for a specific Comb. of Cuts

 Add each Fragment to Frags_Linked_List

 Output all Frags. per specific Comb. of Cuts

Else

 {Gen. new Comb. based on Poss_Occ_Cnt value}

For i \leftarrow 0 to (Cuts.Size - Poss_Occ_Cnt)

 Push into Comb. vector the Value of Cuts[i]

Gen_Frags(Poss_Occ_Cnt - 1, In_Seq)

 Pop last element from Comb. vector

End For

End Elseif

END

The above algorithm executes in space complexity $O(mn)$ and time complexity $O(NN'[mn + q(q^k)(k \log k + n)])$, where N is the total number of input protein sequences, N' is the total number of consensus sequences, n is the size of each protein sequence, m is the size of each consensus sequence, q is the number of consensus occurrences, and k is an integer value from 1 to q . It takes $O(N)$ to process all input protein sequences and $O(N')$ to process each consensus sequence within each protein sequence. For each protein sequence and each consensus sequence, it takes $O(mn)$ in space and time to find all occurrences of a specific consensus within a protein sequence. It takes $O(q * T(k))$ to generate all fragments corresponding to q , where $T(k)$ is a recursive function that entails generating all possible combinations of occurrences and sorting them in $O(k \log k)$. Since $q \ll n$ and $k \ll n$, the algorithm time complexity reduces to $\sim O(NN'mn)$ and the space complexity to $O(mn)$, the size of the dynamic table.

3.2.1. CFPA-CalpCasp Algorithm Functionality

The developed method CFPA-CalpCasp [13] includes identifying all consensus occurrences of a predefined consensus sequence with its predefined cleavage site, in addition of identifying all the generated fragments upon cleavage.

CFPA-CalpCasp looks for an exact match of every consensus, representing the specific cleavage mode of calpain-2 and caspase-3; the cleaving style is based on fixing certain amino acids and varying others, making the *cleavage site* right after consensus *hit*.

Since multiple occurrences of a specific consensus can be found in one protein sequence, all combinations of potential cuts, resulting in different output fragments are considered. Each combination is a possible cleavage incidence by nature. Accordingly, the fragment generation module, within the developed algorithm, generates different fragments based on the different combinations of consensus occurrences. For instance, if two consensus occurrences are matched in an input sequence, then the number of combinations to consider for different output fragments will be: 1) as if occurrence 1 is only found and breaks the input sequence at the cleavage site, or 2) as if occurrence 2 is only found and breaks the input sequence at a different cleavage site, or 3) as if occurrence 1 and occurrence 2 are both found and break the input sequence simultaneously at the two different cleavage sites, resulting in many short fragments than few long fragments due to one occurrence alone. For space limitations, we are presenting the fragments generated from all combinations on one sequence of the mouse genome in Appendix 2.

3.2.2. α II-spectrin Application to Calpain and Caspase Proteases

3.2.2.1. Problem Definition

Calpains and caspases truncate proteins at specific amino acid sequences and the generated fragmented proteins represent molecular signatures that are designated as BDPs, specific to each protease, and distinguished by their protease-generated molecular weight.

The term *degradomics* [119] has been introduced to evaluate the potential BDPs of the different protease substrates. α II-spectrin, with 280 kDa molecular weight,

is a known brain injury marker that generates a 145 kDa BDP upon calpain activation and is associated with necrosis. It generates 120 kDa upon caspase activation and is associated with apoptosis (Fig. 3.16) [120,121]. A crucial and interesting practice of degradomics is how to use specific protein biomarkers to identify specific biological processes, assisting in diagnosis, and subsequently in specific treatments.

In this section, the objective is to utilize a computational approach that can complement and guide the experimental studies; the challenge is to develop an algorithm that can predict the breakdowns of α II-spectrin by calpain and caspase proteases in an accurate and efficient manner (in terms of space and time complexity), tailored to calpain and caspase specific cleavage modes. Moreover, the proposed algorithm should be validated against available experimental studies from the literature to demonstrate its effectiveness.

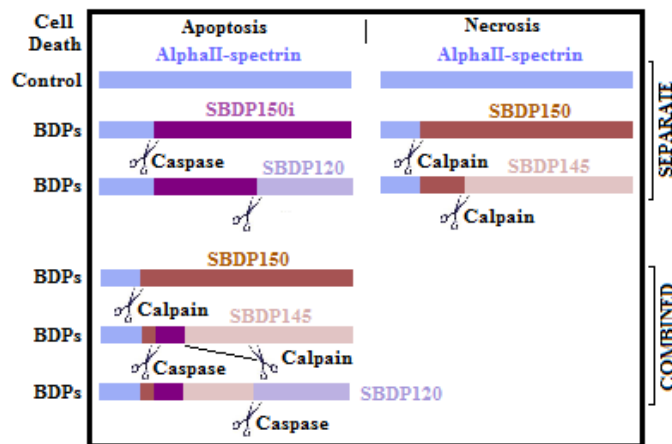


Fig. 3.16 - Control neurons (α II-spectrin) undergoing cell death pathways. Schematic of control neurons (α II-spectrin) undergoing different cell death pathways: Either necrosis with calpain-specific fragments SBDP150 and SBDP145, or apoptosis with caspase-specific fragments SBDP150i and SBDP120, showing approximate sizes of the various breakdowns. Furthermore, the figure shows cleavages by order. Calpain first cleaves α II-spectrin to create SBDP150. Then, caspase and calpain (if both activated) cleave SBDP150 to create SBDP150i and SBDP145 respectively. Finally, caspase cleaves SBDP145 to generate SBDP120 (apoptosis-specific) [74].

We examine first α II-spectrin cleavages and predict computationally the BDPs/cleavage products with each cysteine protease, calpain (necrosis and apoptosis), caspase (apoptosis), and calpain and caspase combined. The computation is based on CFPA-CalpCasp algorithm – a modification to the Cleaved Fragments Prediction Algorithm (CFPA) (Section 3.2) – which is used to accommodate the calpain/caspase particular cleavage mode.

3.2.2.2. Data: α II-spectrin

The demonstration of an accurate algorithm is highly marked with real data. The used input data consists of α II-spectrin substrates, as shown in Fig. 3.17 and Fig. 3.18.

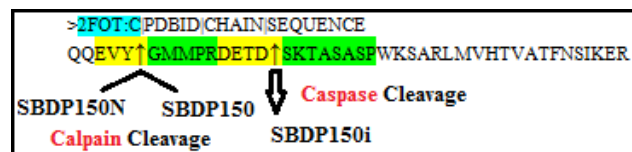


Fig. 3.17 - Cleavage sites of α II-spectrin (PDB ID: 2FOT crystal structure). 2FOT is the structure of the calmodulin α II-spectrin complex, repeat 11. The partial amino acid sequence pictured here highlights the cleavage sites by calpain-2 and caspase-3 [74] and demonstrates the approximate sizes (molecular weight) of the various spectrin breakdown products (SBDPs). Chain C is spectrin and shows one *hit* for calpain-2 and one *hit* for caspase-3 [26].

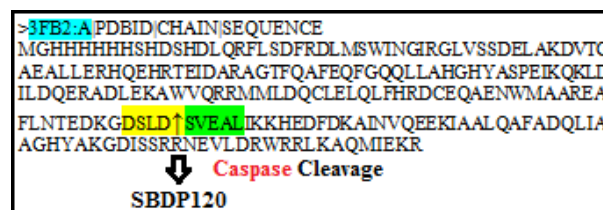


Fig. 3.18 - Cleavage sites of α II-spectrin (PDB ID: 3FB2 crystal structure). 3FB2 is the structure of the human brain α -spectrin, repeats 15 & 16. The partial amino acid sequence pictured here highlights the cleavage sites by caspase-3 (as shown in [74]) and demonstrates the approximate sizes (molecular weight) of the various spectrin breakdown products (SBDPs). Chain A is spectrin and shows one *hit* for caspase-3 (shown in [26] as repeat 13).

The consensus sequences for calpain protease form the patterns LX↑X, VX↑X, and IX↑X, where X can represent any amino acid from the 20 available ones, and the symbol ↑ represents the cleavage site, corresponding to 60 patterns in total. On the other hand, the consensus sequences for caspase protease form the patterns DXXD↑X, corresponding to 400 patterns in total.

3.2.2.3. Computational Results

The results details of CFPA-CalpCasp algorithm are presented in the following subdivisions, showing degradation of α II-spectrin by calpain alone, caspase alone, and calpain and caspase combined. The corresponding output is illustrated beneath the input data. Due to the different conformations a protein can take, the cleaving can occur in any combination. Therefore, the output includes all consensus occurrences with all possible combinations. In addition, the output shows the start and end position of each occurrence exact match within the protein sequence; it also shows the start and end position of each generated fragment, within every possible combination.

Table 3.15 depicts the output generated by CFPA-CalpCasp algorithm based on the α II-spectrin input protein sequence 2FOT-ChainC and calpain protease cleavage mode, as elucidated in Fig. 3.17. It shows all the consensus occurrences (hits) that enable calpain protease to cleave the input protein sequence, and subsequently, show all the corresponding generated fragments, including their positions within the input protein sequence.

Table 3.15 - CFPA-CalpCasp generated data on 2FOT by calpain-2

	AA Sequence	Start Position	End Position
Consensus	VA	33	34
Fragment	QQ...VA	1	34
	TF...ER	35	42
Consensus	VH	30	31
Fragment	QQ...VH	1	31
	TV...ER	32	42
Consensus	VY	4	5
Fragment	QQEVY	1	5
	GM...ER	6	42
Consensus	LM	28	29
Fragment	QQ...LM	1	29
	VH...ER	30	42
Consensus	IK	39	40
Fragment	QQ...IK	1	40
	ER	41	42

Among the consensus occurrences, is the particular sequence motif ‘VY’ which starts at position P4 and ends at position P5 in α II-spectrin input protein sequence, Chain C. Since only one occurrence of it is found, then only one possible combination exists. For motif ‘VY’, the generated sequence fragments are ‘QQEVY’, which extends through positions P1 to P5, and ‘GMMPRDETDSKTASASPWKSARLMVHTVATFNSIKER’, which extends through positions P6 to P42.

The significance of the ‘VY’ occurrence is its validation against manual expected sequence fragments shown in [74], but more importantly its validation against experimentally generated fragments shown on repeat 11 in [26,74]. The other predicted occurrences ‘VA’, ‘VH’, ‘LM’, and ‘IK’ did not show in experimental results and this could be associated with the fast rate of the cleaving transitions, especially at adjacent

cleaving sites, rendering those cleavages simultaneous, and consequently undetected by experimental techniques.

Table 3.16 depicts the output generated by CFPA-CalpCasp based on protein sequence 2FOT-ChainC and caspase protease cleavage mode, as elucidated in Fig. 3.17. It shows all the consensus occurrences (*hits*) that enable caspase protease to cleave the input protein sequence, and subsequently, show all the generated fragments, including their positions in the input protein sequence.

Table 3.16 - CFPA-CalpCasp generated data on 2FOT by caspase-3

	AA Sequence	Start Position	End Position
Consensus	DETD	11	14
Fragment	QQ...TD	1	14
	SK...ER	15	42

The consensus occurrences, in the above results, show only one motif sequence ‘DETD’; it starts at position P11 and ends at position P14. The resulting fragments (or breakdowns) are ‘QQEVYGMMPRETD’, which extends through positions P1 to P14, and ‘SKTASASPWKSARLMVHTVATFNSIKER’, which extends through positions P15 to P42. This occurrence is validated against the manual expected sequence fragments, as shown in [60]. More importantly, this occurrence is validated against the experimental results, as shown on repeat 11 in [26,74].

Table 3.17 depicts the output generated by CFPA-CalpCasp algorithm based on protein sequence 3FB2-ChainA and caspase protease cleavage mode, as elucidated in Fig. 3.18. In a similar way to previous cleavages, the table shows all consensus occurrences (*hits*) with their start and end positions within the protein sequence.

Additionally, the table shows all the generated fragments with their start and end positions.

Table 3.17 - CFPA-CalpCasp generated data on 3FB2 by caspase-3

	AA Sequence	Start Position	End Position
Consensus	DILD	95	98
Fragment	MG...LD	1	98
	QE...KR	99	218
Consensus	DKGD	146	149
Fragment	MG...GD	1	149
	SL...KR	150	218
Consensus	DFRD	21	24
Fragment	MG...RD	1	24
	LM...KR	25	218
Consensus	DSHD	11	14
Fragment	MG...HD	1	14
	LQ...KR	15	218
Consensus	DSL D	149	152
Fragment	MG...LD	1	152
	SV...KR	153	218

The above results show another caspase cleavage in α II-spectrin input protein sequence on repeat 13, as shown in [26,74] and witnessed in PDB ID: 3FB2 (referred to as repeats 15 & 16). Among the different consensus occurrences ‘DILD’, ‘DKGD’, ‘DFRD’, ‘DSHD’, and ‘DSL D’ is the particular sequence motif ‘DSL D’ which starts at position P149 and ends at position P152 within α II-spectrin input protein sequence, ChainA. Similarly to previous occurrences, a sole occurrence corresponds to a sole combination. The resulting fragments are given by: ‘MGHHHHHSHDSDLQRFLSDFRDLMSWINGIRGLVSSDELAKDVTGAEALL ERHQEHRTEIDARAGTFQAFEQFGQQLLAHGHYASPEIKQKLDILDQERADLEK AWWQRRMMLDQCLELQLFHRDCEQAENWMAAREAF LNTE DKGDSL D’,

which extends through positions P1 to P152, and ‘SVEALIKKHEDFDKAINVQEEKIAALQAFADQLIAAGHYAKGDISSRRNEVLDR WRRLKAQMIEKR’, which extends through positions P153 to P218. This occurrence is also of high significance as it is validated against the manual expected sequence fragments, as shown in [74]; it is also validated against the experimental results, as shown in [26,74].

Computational results reflecting the potential simultaneous activation of both proteases (calpain and caspase) is also generated and provide similar insights to the two presented scenarios.

3.2.2.4. Experimental Validation

An in silico digestion of rat brain lysates with in vitro calpain-2 and caspase-3, is probed with α II-spectrin antibody and demonstrated in [74]. Calpain-2 generated α II-spectrin fragments are SBDP150 and SBDP145 and caspase-3 generated fragments are SBDP150i and SBDP120. These experimental results, including cleavage sites and breakdown products, are used for validating our proposed computational approach.

3.2.3. *β II-spectrin Application to Calpain and Caspase Proteases*

3.2.3.1. Problem Definition

The problem is to locate all consensus occurrences of a consensus subsequence in a set of protein sequences. The consensus sequence allows a protease enzyme to cleave a protein substrate at the cleavage site, as shown in Fig. 3.19. The cleavage results into fragment subsequences (or BDPs), signifying disease biomarkers.

The output is expected to show all occurrences (hits) of the consensus sequences among all input protein sequences, in addition to the corresponding cleaved fragments. The hits correspond to exact matches of the consensus models which have variants within them (shown in Table 2.1 as a combination of fixed and variable amino acids).

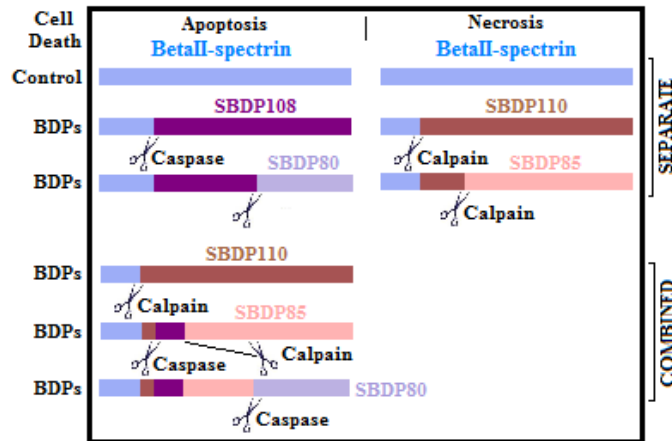


Fig. 3.19 - Control neurons (β II-spectrin) undergoing cell death pathways. Schematic shows necrosis with calpain-2 specific fragments SBDP110 and SBDP85 and apoptosis with caspase-3 specific fragments SBDP108 and SBDP80. Furthermore, the figure shows approximate sizes of the various breakdowns and the cleavages by order. Calpain-2 first cleaves β II-spectrin to create SBDP110. Then, caspase-3 and calpain-2 (if both activated) cleave SBDP110 to create SBDP108 and SBDP85 respectively. Finally, caspase-3 cleaves SBDP85 to generate SBDP80 (apoptosis-specific) [122,123].

3.2.3.2. Data: β II-spectrin

The algorithm needs to be validated with real data to verify its accuracy and effectiveness. The substrate β II-spectrin is used for input data, as shown in Fig. 3.20.

>BetaII-spectrin from GeneBank M96803

```
MTTTVATDYDNIEIQQQYSDVNNRWDVDDWDNENSSARLFERSRIKALADEREAVQKKTFTKWVNSHLARVSCRITDLYTDLRD
GRMLIKLLEVLGGERLPKPTKGRMRIHCLENVDKALQFLKEQRVHLENMGSHDIVDGNHRLTLGLIWTIILRFQIQDISVETED
NKEKKSADALLLWCQMKTAGYPNVNIHNFTTSWRDGMFANLIHKHRPDLIDFDKLLKSNAHYNLQNAFNLAEQHLGLTKLLD
PEDISVDHPDEKSIITYVVVYHYFSKMKALAVEGKRIGKVLDNAIETEKMIEKYESLASDLEWIEQTIILNRRKFANSLVG
VQQQLQAFNTYRTVEKPKFTEKGNLEVLFTIQSKMRANNQKVYMPREGKLI SDINKAWERLEKAEHERELALRNELIRQEKL
EQLARRFDRKAAMRETWLSNQRLVSQDNFGFDLPAVEAATKKHEA IETDIAAYEERVQAVVAVARELEAENYHDIKRITARKD
NVIRLWEYLLLELRARRQRLEMNLGLQKIFQEMLYIMDWMDEMVLVLSQDYGKHLGVEDLLQKHTLVEADIGIQAERVGVN
ASAQKFATDGEQYKPCDPQVIRDRVAHMEFCYQELCQLAAERRARLEESRRLWKFFWEMAE EEGWIREKEKILSSDDYGKDLTS
VMRLLSKHFEDEMSGRSGHFEQAIKEGEDMIAEEHFGSEKIRERI IYIREQWANLEQLSAIRKKRLEEASLLHQFQADADDI
DAWMLDILKIVSSSDVGHDEYSTQSLVKKHKDVAEEIANYRPTLDTLHEQASALPQEHAE SPDVRGRLSGIEERYKEVAELTRL
RKQALQDTLALYKMFSEADACELWIDEKEQWLNMQIPEKLEDELEVIQHRFESLEPEMNNQASRVAVVNIARQLMHSQHPSEK
EIKAQQDKLNTRWSQFRELVDKRDALLSALS IQNYHLECNETKSWIREKTKVIESTQDLGNDLAGVMALQRKLTGMERDLVAI
EAKLSLDLQKEAEKLESEHPDQAQAIL SRLAEISDVWEEMKTTLNREASLGEASKLQQFLRDLDDFQSWLSRTQTATAIASEDMPN
TLTEAEKLLTQHENIKNEIDNYEEDYQKMRDMGEMVTGGQTD AQYMF LRQRLQALDTGWNELHKMWNENRQNLSSQSHAYQQFLR
DTKQAEAFNNQEVVLAHTEMP TLEGAEAAIKKQEDFMTMDANEKINAVVETGRRLVSDGNINSDRIQEKVDSIDDRHRKN
RETASELLMRLKDNRDQLKFLQDCQELSLWINEKMLTAQDMSYDEARNLHVKLKHQAFMAELASNKEWLDKIEKGMQLISEK
PETEAVVKEKLTGLHKMWEVLESTQTKAQRLFDANKAELFTQSCADLDKWLHGLESQIQSDDYGKHLTSVNI LLKQQLLENQ
MEVRKKEIEELQSQAQALSQEGKSTDEVD SKRLTVQTKFMELLEPLNERKHNLLASKEIHQFNRDVEDEILWVGEMRPLATSTD
HGHNLTQTVQLLIKKNQTLQKEIQGHQPRIDDI FERSQNI VTDSSSLSAEAI RQLADLKLWGLLIEETEKRHRRLLEEAHRAQQ
YYFDAEAEAWMSEQELYMMSEKAKDEQSAVSM LKKHQILEQAVEDYAETVHQLSKTSRALVADSHPESE RISMRSKVDKLY
AGLKD LAEERRGKLDERHRLFQLNREVDDLEQWIAEREVVAGSHELGDYEHVTMLQERFREFARDTGNIGQERVDTVNHLADE
LINSGHSDAATIAEWKDG LNEAWADLLELIDTRTQILAAS YELHKFYHDAKEIFGRIQDKHKKLPEELGRDQNTVETLQRMHTT
FEHDIQALGTQVRQLQEDAARLQAAYAGDKADDIQKRENEVLEAWKSLLDACESRRVRLVDTGDKFRFFSMVRDLMLW MEDVIR
QIEAQEKPRDVSSVELLMNNHQGIKAEIDARNDSTTCIELGKSL LARKHYASEEIKEKLLQLTEKRKEMIDKWEDRWEWLRLI
LEVHQFSRDASVAEAWLLGQEPYLSREIGQSVDEVEKLIK RHEAFEKSAATWDERFSALERLTLELLEVRRRQEEEEERKRRP
PSPEPSTKVSEEAESQQQWDT SKGEQVSQNGLP AEQGS PRMAETVDTSEM VNGATEQRTSSKES S PIPSP TSDRKAKTALPAQS
AATL PARTQETPSAQMEGFLNRKHEWEAHNKASSRSWHNVYCVINNQEMGFYKDAKTAASGIPYHSEVPVSLKEAVCEVALDY
KKKKHVFKLRLNDGNEYLFQAKDDEEMNTW IQAISSAIS SDKHEVSASTQSTPASSRAQTLPTSVVTTITSESSPGKREKDKEDK
KEKRFSLFGKKK
```

Fig. 3.20 - β II-spectrin encoded gene [75]. The figure shows the FASTA amino acid sequence of β II-spectrin from the GeneBank. This protein sequence is used as input to the algorithm CFP A-CalpCasp. The algorithm first finds consensus occurrences in this protein sequence, and subsequently, cleaves the input sequence at the cleavage sites, generating fragments or BDPs.

3.2.3.3. Computational Results

The pattern $DXXD\uparrow X$ corresponds to the consensus sequence for caspase-3 protease, where X represents any amino acid from the 20 available ones, symbol \uparrow represents the cleavage site, and D maps to Asp amino acid, corresponding to 400 expected patterns in total. The partial amino acid subsequences, pictured in Fig. 3.21, highlight the cleavage sites by caspase-3, showing two hits in red that are validated experimentally [122,123]. Table 3.18 illustrates the output generated by CFP A-

CalpCasp [13] based on β II-spectrin input protein sequence and caspase-3 protease cleavage mode, as depicted in Fig. 3.21. It shows all the consensus occurrences (hits) that enable caspase-3 protease to cleave the input protein sequence at the specific cleavage site, including their start and end positions within the protein sequence. In addition, it shows all the generated fragments with their start and end positions.

Among the consensus occurrences (hits), are the particular subsequences 'DEVD' and 'DSID', which are validated against the experimentally generated fragments from [122,123]. Motif 'DSID' starts at position P1251 and ends at position P1254 within β II-spectrin input protein sequence. The corresponding generated sequence fragments are: 'MTTT...DSID', which extends through positions P1 to P1254, and 'DRHR.... GKKK', which extends through positions P1255 to P2364. On the other hand, motif 'DEVD' starts at position P1454 and ends at position P1457 within β II-spectrin input protein sequence. The corresponding generated sequence fragments are: 'MTTT...DEVD', which extends through positions P1 to P1457, and 'SKRL...GKKK', which extends through positions P1458 to P2364.

>BetaII-spectrin from GeneBank M96803

MTTIVATDYDNIIEIQQQYSDVNNRWVDVDFWENENSSARLFERSRIKALADEREAVQKKTFTKWVNSHLARVSCRITDLYTDLRD
GRMLIKLLEVLVSGERLPKPTKGRMRIHCLENVDKALQFLKEQRVHLENMGSHDIVDGNHRLTLGLIWTIILRFQIQDISVETED
NKEKSAKDALLLWCQMKTAGYPVNIHNFTTSWRDGMFALIHKHRPDLIDFDKLKKSNAHYNLQAFNLAEQHLGLTKLLD
PEIISVDHPDEKSIITYVVVYYHYFSKMKALAVEGKRIGKVLDNAIETEKMIKEYESLASDLEWIEQTIIILNNRKFANSLVG
VQQQLQAFNTYRTVEKPKPFTEKGNLEVLFTIQSKMRANNQKVYMPREGKLISDINKAWERLEKAEHERELALRNELIRQEKL
EQLARRFDRKAAMRETWLSNQRLVSDNFGFDLPAVEAATKKHEAIEETDIAAYEERVQAVVAVARELEAENYHDIKRITARKD
NVIRLWEYLLLELRARRQRLEMNLGLQKIFQEMLYIMDWMDEMVKLVLSQDYGKHLGVEDLLQKHTLVEADIGIQAEVRVGN
ASAQKFATDGEQYKPCDPQVIRDRVAHMEFCYQELCQLAERRARLEESRRLWKFVEMAE EEGWI REKEKILSSDDY GKD L TS
VMRLLSKHRAFEDEMSSGRSGHFEQAIKEGEDMIAEEHFGSEKIRERIIYIREQWANLEQLSAIRKKRLEEASLLHQFQAADADDI
DAWMLDILKIVSSSDVGHDEYSTQSLVKKHKDVAEEIANRYPTLDTLHEQASALPQEHASPDVVRGRLSGIEERYKEVAELTRL
RKQALQD TLALYKMFSEADACELWIDEKEQWLNMQIPEKLEDELEVIQHRFESLEPEMNNQASRVAVVNIARQLMHSHPSEK
EIKAQQDKLNTRWSQFRELVDKRDALLSALSIQNYHLECNETKSWIREKTKVIESTQDLGNDLAGVMALQRKLTGMERDLVAI
EAKLSLDQKEAEKLESEHPDQAAILSLRAEISDVWEEMKTTLNREASLGEASKLQQFLRDLDFFQSWLSRTQTAIASEDMPN
TLTEAEKLLTQHENIKNEIDNYEEDYQKMRDMGEMVTQQTDAQYMLRQRLQALDTGWNELHKMWENRQNLSSQSHAYQQFLR
DTKQAEAFLNQEQYVLAHTEMPPTLEGAEAAIKKQEDFMTMDANEKINAVVETGRRLVSDGNINSRIQEKVDSIIDRHRKN
RETASELLMRLKDNRD LQKFLQDCQELSLWINEKMLTAQDMSYDEARNLHSHKWLKHQAFMAELASNKEWLDKIEKEGMQLISEK
PETEAVVKEKLTGLHKMWEVLESTTQTKAQRLFDANKAELFTQSCADLDKWLHGLESQIQSDDYGKHLTSVNIILLKKQOMLENQ
MEVRKKEIEELQSQAQALSQEGKSTDEVIDSKRLTVQTKFEMELLEPLNERKHNLLASKEIHQFNFDVELEIILWVGERMPLATSTD
HGHNLTQTVQLLIKKNQTLQKEIQGHQPRIDDI FERSQNI VTDSSLSAEAIRQRLADLQKLGWLLIETEKRHRRLLEAHRAQQ
YYFDAEAEAWMSEQELYMMSEKAKDEQSAVSMLKHKQILEQAVEDYAEVTVHQLSKTSRALVADSHPESEIRSMRQSKVDKLY
AGLKDLAEERRGKLDERHRLFQLNREVDDLEQWIAEREVVAGSHELGDYEHVTMLQERFREFARDTGNIGQERVDTVNHLLADE
LINSGHSDAATAEAWKDLNEAWADLLELIDTRTQILAASYELHKFYHDAKEIFGRIQDKHKKLPEELGRDQNTVETLQRMHTT
FEHDIQALGTQVRQLQEDAARLQAAYAGDKATDIQKRENEVLEAWKSLLDACESRRVRLVDTGFKFRFFSMVRDLMLWMEDVIR
QIEAQEKPRDVSVELLMNNHQIKAEIDARNDSTTCIELGKSL LARKHYASEEIKEKLLQLTEKRKEMIDKWEDRWELRLI
LEVHQFSRDASVAEAWLLGQEPYLSREIGQSVDEVEKLIKREHAEFESAAATWDERFSALERLTTLELLEVRQQEERKRRP
PSPEPSTKVSEEAESQQQWDTSKGEQVSNGLPAEQGSPRMAETVDTSEMVGATEQRTSSKESPISPTSDRKAKTALPAQS
AATLPARTQETPSAQMEGFLNRKHEWEAHNKKASSRSWHNVYCVINNQEMGFYKDAKTAASGIPYHSEVPVSLKEAVCEVALDY
KKKKHVFKRLRNDGNEYLFQAKDDEEMNTWIAISSAISSDKHEVSASTQSTPASSRAQTLPTS SVTITSESSPGKREKDKEDK
KEKRFSLFGKKK

▼ Caspase Cleavage
Caspase Cleavage ▼

Fig. 3.21 - Cleavage sites of β II-spectrin by caspase-3 [75]. The figure shows all the consensus subsequences identified by the algorithm, surrounded in boxes, and obeying the amino acid sequence DXXD (X can be any amino acid). In particular, the red boxes represent the consensus occurrences validated experimentally. In addition, the figure shows the cleavage site by caspase-3.

The patterns LX \uparrow X, VX \uparrow X, and IX \uparrow X correspond to the consensus sequences for calpain-2 protease, where X represents any amino acid from the 20 available ones, \uparrow symbol represents the cleavage site, and (L,V,I) maps to (Leu, Val, Ile) amino acids triplet, corresponding to 60 expected patterns total. The partial amino acid subsequences, pictured in Fig. 3.22, highlight the cleavage sites by calpain-2 showing one hit in red that is validated experimentally [122,124]. Table 3.19 illustrates a few records of the whole output (see Appendix 3) generated by CFP-CalpCasp [13] based

on β II-spectrin input protein sequence and calpain-2 protease cleavage mode, as depicted in Fig. 3.22. The table shows all the consensus occurrences (hits) that enable calpain-2 protease to cleave the input protein sequence with the specific cleavage site, including their start and end positions within the protein sequence. In addition, it shows all the generated fragments with their start and end positions.

Among the consensus occurrences (hits), is the particular subsequence 'ETVD', which is validated against the experimentally generated fragments from [122,124]. Motif 'ETVD' starts at position P2143 and ends at position P2146 within β II-spectrin input protein sequence. The corresponding generated sequence fragments are: 'MTTT...ETVD', which extends through positions P1 to P2146, and 'TSEM... GKKK', which extends through positions P2147 to P2364. The other predicted occurrences like 'VA' and 'VH' (as shown in Fig. 3.22) did not show in experimental results and this could be associated with the fast rate of the cleaving transitions, especially at adjacent cleaving sites making both cleavages simultaneous, and consequently undetected by experimental techniques.

Computational results reflecting the potential simultaneous activation of both proteases (calpain-2 and caspase-3), is also generated and provide similar insights to the two presented scenarios. However, when validated experimentally, two possibilities can happen in such case; either one protease inhibiting the cleavage of the other protease, or one protease cleaving within the substrate cleaved by the other protease.

Table 3.18 - CFPA-CalpCasp generated data on M96803 by caspase-3

		AA Sequence	Start	End			AA Sequence	Start	End
1	Consensus	DVDD	26	29	10	Consensus	DDID	754	757
	Fragment	MTTT.... DVDD	1	29		Fragment	MTTT.... DDID	1	757
	Fragment	WDNE.... GKKK	30	2364		Fragment	AWML.... GKKK	758	2364
2	Consensus	DDWD	28	31	11	Consensus	DLDD	1070	1073
	Fragment	MTTT.... DDWD	1	31		Fragment	MTTT.... DLDD	1	1073
	Fragment	NENS.... GKKK	32	2364		Fragment	FQSW.... GKKK	1074	2364
3	Consensus	DLRD	81	84	12	Consensus	DSID	1251	1254
	Fragment	MTTT.... DLRD	1	84		Fragment	MTTT.... DSID	1	1254
	Fragment	GRML.... GKKK	85	2364		Fragment	DRHR.... GKKK	1255	2364
4	Consensus	DIVD	137	140	13	Consensus	DNRD	1273	1276
	Fragment	MTTT.... DIVD	1	140		Fragment	MTTT.... DNRD	1	1276
	Fragment	GNHR.... GKKK	141	2364		Fragment	LQKF.... GKKK	1277	2364
5	Consensus	DLID	218	221	14	Consensus	DEVD	1454	1457
	Fragment	MTTT.... DLID	1	221		Fragment	MTTT...DEVD	1	1457
	Fragment	FDKL.... GKKK	222	2364		Fragment	SKRL...GKKK	1458	2364
6	Consensus	DPED	252	255	15	Consensus	DVED	1493	1496
	Fragment	MTTT.... DPED	1	255		Fragment	MTTT.... DVED	1	1496
	Fragment	ISVD.... GKKK	256	2364		Fragment	EILW.... GKKK	1497	2364
7	Consensus	DHPD	259	262	16	Consensus	DKAD	1877	1880
	Fragment	MTTT.... DHPD	1	262		Fragment	MTTT.... DKAD	1	1880
	Fragment	EKSL.... GKKK	263	2364		Fragment	DIQK.... GKKK	1881	2364
8	Consensus	DWMD	542	545	17	Consensus	DTGD	1909	1912
	Fragment	MTTT.... DWMD	1	545		Fragment	MTTT.... DTGD	1	1912
	Fragment	EMKV.... GKKK	546	2364		Fragment	KFRF.... GKKK	1913	2364
9	Consensus	DADD	752						
	Fragment	MTTT.... DADD	1						
	Fragment	IDAW.... GKKK	756						

Table 3.19 - A few records of CFPA-CalpCasp generated data on M96803 by calpain (see Appendix 3 for all output records).

		AA Sequence	Start	End			AA Sequence	Start	End
43	Consensus	PDVR	820	821	98	Consensus	ETVD	2145	2146
	Fragment	MTTT.... PDVR	1	821		Fragment	MTTT...ETVD	1	2146
	Fragment	GRLS.... GKKK	822	2364		Fragment	TSEM... GKKK	2147	2364
44	Consensus	KEVA	834	835	99	Consensus	EMVN	2151	2152
	Fragment	MTTT.... KEVA	1	835		Fragment	MTTT.... EMVN	1	2152
	Fragment	ELTR.... GKKK	836	2364		Fragment	GATE.... GKKK	2153	2364
45	Consensus	LEVI	886	887	100	Consensus	HNVY	2225	2226
	Fragment	MTTT.... LEVI	1	887		Fragment	MTTT.... HNVY	1	2226
	Fragment	QHRF.... GKKK	888	2364		Fragment	CVIN.... GKKK	2227	2364

>BetaII-spectrin from GeneBank M96803

MTTTVAFDYDNIETQQQYSDVNRNRVDWDNENSSARLFERSRIKALADEREAVCKKFTFKWVNSHLARVSCRITDLYTDLRD
 GRMLIKLLEVI SGERLPKPTKGRMRIHCLENVDKALQFLKEQRVHLENMGSHDIVDGNHRLTLGLIWTIILRFQIQDISVETED
 NKEKKSAKDALLLWCQMKTAGYPNVNIHNFTTSWRDGMFANALIHKHPDLIDFDKLLKSNAHYLNQAFNLAEQHLGLTKLLD
 PEDISVHPDEKSIITYVVTYYHYFSKMKALAVEGKRIGKVIDNAIETEKMIKEYESLASDLEWIEQTI IILNRRKFANSLVGV
 VQQQLQAFNTYRTVEKPPKFTTEKGNLEVTLFTIQSKMRANNQKVYMPREGKLISDINKAWERLEKAEHERELALRNELIRQEKL
 EQLARRFDRKAAMRETWLSNQRLVSDQDNFGFDLPVAEATKKHEA IETDIAAYEERVQAVMAVAARELEAENYHDIKRITARKD
 NVIRLWEYLLELLRARRRQRLMNLGLQKIFQEMLYIMDWMDEMVI VLSQDYGKHLGVEDLLQKHTLVEADIGIQAERVVGVN
 ASAQKFATDGEQYKPCDPQVIRDRVAHMEFCYQELCQLAERRARLEESRRLWKFVEMAE EEGWIREKEKILSSDDYGKDLTS
 VMRLLSKSHRAFEDEMSGRSGHFEQAIKEGEDMIAEEHFGSEKIRERIIYIREQWANLEQLSAIRKKRLEEASLLHQFQADADDI
 DAWMLDILKIVSSDVCHDEYSTQSLVKKHKDVAEETIANYRPTLDTLHEQASALPQEHAEVSPVGRLSGIEERYKEVAELTRL
 RKQALQDTLALYKMFSEADACELWIDEKEQWLNMMQIPEKLEDLVIQHRFESLEPEMNNQASRVAVVNIQIARQLMHSQHPSEK
 EIKAQDKLNTRWSQFRELVRKDKDALLSALSIQNYHLECNETKSWIREKTKVTESTQDLGNDLAGVMALQRKLTGMERDLVAI
 EAKLSDLQKEAEKLESEHPDQAQAILSR LAEISDVWEEMKTTLKNREASLGEASKLQQLFRDLDDFQSWLSRTQTAIASEDMPN
 TLTEAEKLLTQHENIKNEIDNYEEDYQKMRDMGEMVTQGGTDAQYFRLRQLQALDTGWNELHKMWENRQNLSSQSHAYQQFLR
 DTKQAEAFNNQEQYVIAHTEMPTTLEGAEAAIKKQEDFMTTMDANEKINA VVEVETGRRLVSDGNINSRDIQEKVSIIDDRHKN
 RETASELLMRLKDNRLQKFLQDCQELSLWINEKMLTAQDMSYDEARNLHSHKWLKHQAFMAELASNKEWLDKIEKGMQLISEK
 PETEAVVKEKLTGLHKMWEVLESTTQTKAQLRFDANKAELFTQSCADLDKWLHGLSQIQSDDYGKHLTSVNI LLKQOQMLENQ
 MEVVRKKEIEELQSQQAALSQEGKSTDEVT SKRLTVCTKFMELLEPLNERKHNLLASKEIHQFNRDVEDEILVWGERMPLATSTD
 HGHNLTQVQLLIKKNQTLQKEIQGHQPRIDDI FERSQNI VTDSSSLSAEAI RQRLADLKLWGLLIEETEKRHRREEAHRQQ
 YFDAAEAEAWMSEQELYYMSEKAKDEQSAVSM LKKHQILEQAVE DYAEVHQLSKTSRALVADSHPESERISMRQSKVDKLY
 AGLKDLAEERRGKLDERHRLFQLNREVDLEQWIAEREVVA GSHELQDYEHVIMLQERFREFARDTGNIGQERVITVNIHLADE
 LINSGHSDAATAEWKDLNEAWADLLELIDTRTQILAASYELHKFYHDAKEIFGRIQDKHKKLPEELGRDQNTVETLQRMHTT
 FEHDIQALGTQVQLQEDAARLQAAYAGDKADDIQKRENEVLEAWKSLLDACESRVRVLTGDKFRFFSMVFDLMLWMEVIR
 QIEAQEKPRDVSSVILLMNNHQGIKAEIDARNDSTFTTCIELGKSL LARKHYASEEIKEKLLQLTEKRKEMIDKWEDRWEWLRLI
 LEVHQFRSDASVAEAWLLGQEPYLSREIGQSVI EVELIKIRHEAFEKSAATWDERFSALERLTLELLEVRQOQEEERKRRP
 PSPEPSTKVS EEAESQQQWDTSKGEQVSONGLPAEQGSPRMAETVDTSEM VNGATEQRTSSKESP I P SPTSDRKAKTALPAQS
 AATLPARTQETPSAQMEGFLNRKHEWEAHNKKASSRSWHNVYCVINNQEMGFYKDAKTAASGIPYHSEVSVLSKEAVCEVALDY
 KKKKHVFKLRLNDGNEYLFQAKDDEEMNTWIQAISSAISSDKHEVSASTQSTPASSRAQTLPTS VVITITSESSPGKREKDEKED
 KEKRFSLFGKKK

▼ Calpain Cleavage

Fig. 3.22 – Cleavage sites of β II-spectrin by calpain-2. The figure shows all the consensus subsequences identified by the algorithm, surrounded in boxes, and obeying the amino acid sequence VX (X can be any amino acid). In particular, the red box represents the consensus occurrence validated experimentally. In addition, the figure shows the cleavage site by calpain-2.

3.2.3.4. Experimental Validation

For additional assessment of the algorithm effectiveness, the utility β II-spectrin proteolysis, and its characteristic breakdowns as potential markers of the different cell death types (necrosis and apoptosis), the algorithm CFPA-CalpCasp is applied to β II-spectrin substrate; the matched consensus occurrences are identified and the corresponding fragments are generated.

Calpain-2 and caspase-3 are probed with β II-spectrin and demonstrated in [122,123,124]. Calpain-2 generated β II-spectrin fragments are SBDP110 and SBDP85

and caspase-3 generated fragments are SBDP108 and SBDP80. These experimental results, including cleavage sites and breakdown products, are used for validating our proposed computational approach.

3.2.4. Summary

The concept underlying CFPA-CalpCasp algorithm [13] is simple, robust, and efficient. It is a modified version of Smith-Waterman algorithm and based on CFPA algorithm with a few modifications in favor of calpain and caspase special cleavage modes. It searches for local subsequences similarities of the consensus subsequences in a set of protein sequences. Consequently, alignments with deletions and insertions are pruned. For every acceptable alignment, cleavage of the corresponding protein sequence occurs at the cleavage site - predefined within the consensus sequence - and results in cleaved fragments, identified by the algorithm. The consensus occurrence variants are built within the consensus pattern, such as in subsequence DXXD where D is fixed for Asp, but X can be any amino acid. Due to the different protein conformations, the cleavage can happen in all different combinations of occurrences, and so fragments resulting from all possible combinations are generated.

In an effort of assessing the utility of α II-spectrin proteolysis and its characteristic breakdowns as potential markers of different cell death types (necrosis and apoptosis), CFPA-CalpCasp algorithm is applied to α II-spectrin crystal structure (PDB ID: 2FOT) from the Protein Data Bank (PDB). It is also applied to β II-spectrin [75] and the matching consensus sequences occurrences and the corresponding fragments are generated.

For the current application of neuronal cell deaths and the peculiarity of calpain-2 and caspase-3 cleaving modes, CFPA-CalpCasp results proved their accuracy through validation with experimental data; they also proved their efficiency in performance through detection of consensus occurrences and generation of corresponding cleaved fragments in time complexity $O(N'Nm)$, where N' is the number of consensus sequences, N is the number of protein sequences, m is the length of a consensus sequence, and n is the length of a protein sequence.

Computational prediction of biomarkers is becoming a priority to biologists since it conserves time and cost that would have otherwise been spent on experiments needed to probe for biomarkers. Not only can the generated data and results of this research guide future experiments, but they will also be shared with the scientific community through the development of a web-based front end for online users, with a database backend for storing all input protein, consensus, and fragments sequences. Furthermore, the above degradomics strategies can be applied to other biological disciplines contributing to understand, diagnose, and treat related diseases.

CHAPTER 4

PROTEIN-DNA METHODS AND RESULTS

Poisson-Boltzmann electrostatic study is applied for the analysis of many protein-protein complexes using Adaptive Poisson-Boltzmann Solver (APBS) [125] and within the Analysis of Electrostatic Similarities Of Proteins (AESOP) framework [39,40,126] (see Chapter 2, Section 2.5.2.1). Due to the similar type of force field between protein-protein and protein-DNA interactions, we are applying the same methodology for analyzing protein-DNA complexes. We are studying, in this regard, the role of GATA transcription factor (TF) in binding to DNA. In particular, we are studying the critical mutations on GATA TF amino acid sequence, leading to malfunction in transcription of target genes.

Due to the unavailability of the structural information of GATA4:DNA, we are applying our method to GATA3:DNA structure, and will deploy it to GATA4:DNA crystal structure when it is available. Not only will this study assist in gaining insight into the physicochemical characteristics of the GATA:DNA complex, but it will also provide insight of relation between binding and protein function. Furthermore, it will assist in designing new targeted molecules which might contribute to work related to the discovery of new medications.

AESOP framework comprises APBS and includes:

- Preparation of an alanine scan.
- Calculation of Poisson–Boltzmann electrostatic potentials.
- Calculation of electrostatic free energies of binding.

- Data visualization.

We elaborate on some of the above steps as follows:

- (i) An R script is implemented to generate the alanine scan mutants. It uses the original crystal structure from the Protein Data Bank (PDB) and replaces every ionizable amino acid that is expected to be charged at physiological pH (Asp, Glu, Arg, Lys, and His), one at a time, with Ala.
- (ii) APBS [125], which is based on the linearized Poisson–Boltzmann equation, is used to calculate the electrostatic potentials, as described in previous studies [38,39,40,126]. Within APBS calculations, atomic radii and charges are calculated using the PDB2PQR [127] program, according to AMBER force field parameters [124,128].
- (iii) Electrostatic free energies of binding are calculated based on electrostatic potentials, according to a thermodynamic cycle, as described in [39,40,126] in the form of the following equations:

$$\Delta\Delta G_{solvation}^{association} = \Delta G_{solvation}^{GATA:DNA} - \Delta G_{solvation}^{GATA} - \Delta G_{solvation}^{DNA} \quad (9)$$

$$\Delta\Delta G_{solvation}^{association} = \Delta G_{solvation}^{association} - \Delta G_{reference}^{association} \quad (10)$$

$$\Delta G_{solvation}^X = \Delta G_{solution}^X - \Delta G_{reference}^X \quad (11)$$

$$\Delta G_Y^{association} = \Delta G_Y^{GATA:DNA} - \Delta G_Y^{GATA} - \Delta G_Y^{DNA} \quad (12)$$

where Eqn.9 presents the binding free energy component of the complex in solvent, Eqn.10 presents the binding free energy component of the complex after eliminating artifacts, Eqn.11 presents the energy effect of the solvent after subtracting artifacts, and Eqn.12 presents the energy effect of the complex after subtracting the individual

modules. It is worth noting that Eqn.11 and Eqn.12 feed into Eqn.9 and Eqn.10. Also, Eqn.9 presents the final form of the complex GATA:DNA binding free energy calculation.

4.1. Protein-DNA Models

We study different protein-DNA models, comprising different Molecular Mechanics (MM) models and different Solvation models. MM model entails bonded and non-bonded interactions. Bonded interactions drive energy associated with bond stretching, angle bending, and torsion, comprising mainly *conformational* changes. Non-bonded interactions drive energy associated with electrostatic and vdW interactions. Continuum solvation models, on the other hand, entail efficient methods to compute the effect of solvent on the complex formation.

For a proper selection of a model among the ones that exist, several factors are considered. Among those are:

- Parameterization of the MM force field.
- The specific application or biological process of interest (binding, folding, etc.).
- Complex type (protein-protein, protein-DNA, protein-ligand, protein-RNA, etc.).
- Approximations in continuum methods.

4.2. Binding Free Energy Calculations

Binding free energy calculations are computed on a protein-DNA model comprising a parameterized force field and a solvation model. Most calculation methods estimate the free energy according to *Gibbs Free Energy* discovered by J. Willard

Gibbs, as in Eqn.13 [108]. The symbol ΔG is used to define the *Free Energy* of a system. It is the amount of heat energy released in some biological process to do some work minus the change in entropy ΔS . Entropy S is a thermodynamic state function representing the dispersal of energy and matter, and so the greater the disorder, the higher the entropy, and vice versa.

$$\Delta G = \Delta H - T\Delta S \quad (13)$$

where,

T is the floating point value of the temperature for calculation.

ΔH is represented by Eqn.14 [108].

$$\Delta H = E_{MM} + E_{solvation} - TS_{MM} \quad (14)$$

where,

E_{MM} : is the average energy obtained from a typical MM Force-Field with contributions from bond stretching, angle bending, torsions, electrostatic and van der Waals terms (see Chapter 2, Sections 2.5.1.1, 2.5.1.2, 2.5.1.3, 2.5.2.1, and 2.5.2.2).

$E_{Solvation}$: is the solvation free energy and consists of polar and non-polar contributions (see Chapter 2, Section 2.5.2.3).

S_{MM} : is the solute entropy.

After separately computing the free energy of each molecule in the complex (DNA molecule and GATA TF molecule), and then computing the free energy of the complex (GATA:DNA), binding free energy is estimated as in Eqn.15 [108].

$$\Delta\Delta G_{Binding} = \Delta G_{Complex} - \Delta G_{GATA} - \Delta G_{DNA} \quad (15)$$

Our system represents a molecular complex (protein-DNA) in solvation (GATA3:DNA). For validation of the binding free energy calculations, comparisons are evaluated against the binding affinity values from experimental methods.

4.3. Application to Charged-Mutants GATA3

4.3.1. Problem Definition

Binding free energy calculations of many protein-protein interactions are implemented using the integrated Analysis of Electrostatic Similarities Of Proteins (AESOP) framework [39,40,126]. The types of protein-protein interactions include bonded interactions (bond, angle, torsion) and non-bonded interactions (short-range and long-range electrostatic, vdW, and hydrogen bonds). Due to the similar type of interactions between protein-protein and protein-DNA complexes, as illustrated in Table 4.1 [129], we are applying AESOP framework to the electrostatic study of protein-DNA interactions.

Table 4.1 - Types of protein-DNA interactions

Intra-molecular			
Bonded	Bond	Angle	Torsion
Non-bonded	vdW	H-bond	Ionic
Inter-molecular			
Non-bonded Specific ^a	H-bond		
Non-bonded Non-specific ^b	vdW	H-bond	Ionic

^aSpecific refers to interactions between Amino Acid (AA) and DNA bases.

^bNon-specific refers to interactions between AA and DNA backbone.

4.3.2. Data: GATA3 Crystal Structure

Due to the unavailability of the crystal structure GATA4:DNA in the public database, we are studying the binding free energy of the two forms of the crystal structure GATA3:DNA (as shown in Fig. 4.1), for proof of concept. The two complexes depict two different conformations of the GATA3:DNA complex; they are found in the Protein Data Bank (PDB) under the PDB Ids: 3DFV (GATA factors adjacent on DNA) and 3DFX (GATA factors on opposite sides of DNA) [129].

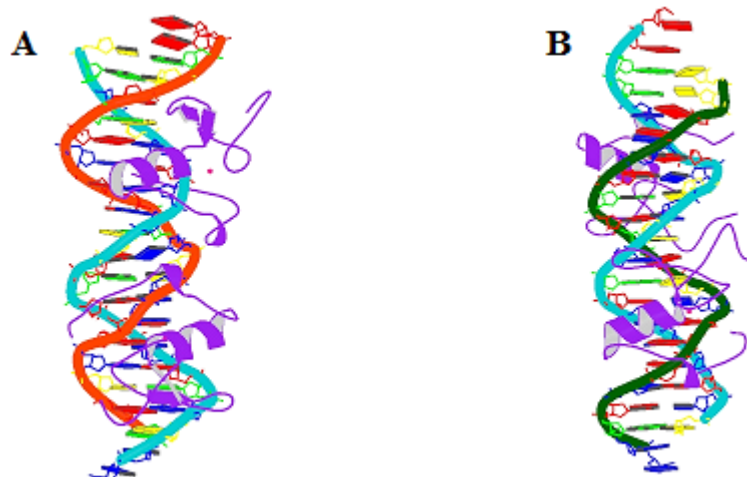


Fig. 4.1. Crystal structure of GATA3:DNA complex. A) N-finger and C-finger of GATA3 bind to DNA in an adjacent way; complex has PDB-ID: 3DFV. B) N-finger and C-finger of GATA3 bind to DNA in an opposite way; complex has PDB-ID: 3DFX [129].

For electrostatic calculations, we used the crystal structure 3DFV.pdb. In this complex, the GATA factors comprise the coordinates of amino acids Arg311 through Arg366 for each of Chain-D and Chain-C. The DNA module in this complex, on the other hand, comprises the coordinates of nucleic acids Thy1 through Cyt20 for each of the two chains Chain-Y and Chain-Z.

The parameters of our calculations are specifically set as follows: The probe radii, which define the dielectric, are set to 1.4 \AA ; the ion accessibility surfaces are set to 2.0 \AA ; the dielectric coefficient for the protein interior is set to 2; and the dielectric coefficient for the solvent is set to 78.54.

The grid used in the APBS calculations contains $129 \times 161 \times 161$ grid points, with coarse grid lengths of $82 \text{ \AA} \times 97 \text{ \AA} \times 104 \text{ \AA}$, and fine grid lengths of $68 \text{ \AA} \times 77 \text{ \AA} \times 81 \text{ \AA}$. Thus, a grid resolution of $\leq 1 \text{ \AA}$ is achieved.

4.3.3. GATA3 Binding Energy Calculations

In an effort to understand the effect of every amino acid on Congenital Heart Disease (CHD), we induce mutations in all possible charged amino acids (AA) of the GATA transcription factor (TF). Mutations are induced one AA at a time because it is very unlikely to have more than one mutation at one time in any normal biological process. Afterwards, the binding free energy is recomputed for each mutation.

Since each mutation perturbation can alter the overall binding ability in a complex, we generated a family of mutants from the crystallographic structure GATA3:DNA [129] at the atomic detail; this is accomplished through an alanine scan in which each charged amino acid is replaced by Ala. We then performed Poisson–Boltzmann electrostatic calculations, feeding into electrostatic free energy calculations on each mutation, in order to reveal the contribution of each charged amino acid to binding. Comparison of site-mutations calculations with parent/wild protein gives an indication of key amino acids in binding.

The mutations dataset consists of one protein family for GATA3 (parent/wild and 26 mutants on both of Chain-C and Chain-D of the 3DFV structure). The structures of the GATA TF protein mutants family are superimposed using the backbone C α atoms and centered on the same grid used for the parent structure (GATA3:DNA). Fig. 4.2 presents the electrostatic free energy calculations of the complex GATA3:DNA, with GATA3 mutants at 150mM ionic strength. The calculated solvation free energy difference for each mutant, computed from Eqn.9, is compared against the parent/wild protein solvation free energy. This comparison serves as a physicochemical classifier of binding ability. Eqn.9 is based on the thermodynamic cycle described in [39,40,126]

and computes the free energy difference $\Delta\Delta G$ of the two different (solvation and vacuum) free energy differences ΔG . Accordingly, an increase in solvation binding free energy $\Delta\Delta G$ is considered favorable, whereas a decrease is considered unfavorable.

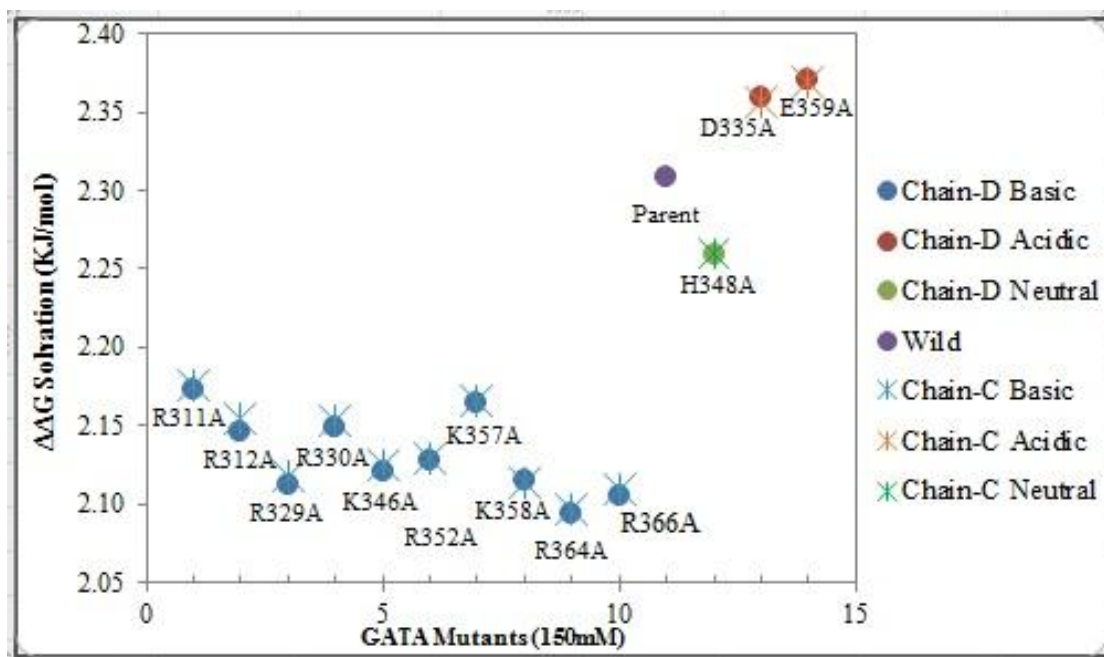


Fig. 4.2 - GATA3 electrostatic free energy differences (charged AA). Plot presents the solvated binding free energy calculations of GATA3 mutants in both of Chain-D and Chain-C. Blue and red colors correspond to basic and acidic mutants respectively. Mutants predicted to enhance binding are shown above the parent (wild) and mutants predicted to reduce binding are shown below it. The x-axis (index) corresponds to the order of the mutant (mutants are numbered and ordered sequentially).

4.3.4. GATA3 Intermolecular Contacts

We identify all intermolecular contacts between GATA TF and DNA. Contacts comprise charged bonds or hydrogen bonds. Crystal structures 3DFV and 3DFX contacts are defined experimentally in [129]. Even though identifying contacts computationally (Fig. 4.3) do not feed directly into the free energy calculations, we still do this step in order to check the correctness of our resulting data against the published

increase binding, and all basic mutants of Chain-D, R311A, R312A, R329A, R330A, K346A, R352A, K357A, K358A, R364A, and R366A, are predicted to decrease binding.

Thus, if the computed solvation binding free energy $\Delta\Delta G$ increases due to a mutation, this implies that the corresponding specific residue does not have a significant impact on binding initially, and so its mutation is not considered a hotspot. The opposite is argued for a residue whose mutation causes a decrease to $\Delta\Delta G$, implying the importance of the corresponding specific residue to original binding. On the other hand, the lower the computed $\Delta\Delta G$ for a specific mutation, the more important the corresponding residue is to binding. This latter case is seen specifically with basic residues R329 and R364 of GATA3 in Fig. 4.2, which shows a reduction in binding free energy upon mutation.

4.3.6. GATA3 Experimental Validation

The computational results for the inhibitors Arg329 (R329) and Arg364 (R364), after being mutated to Alanine (Ala), are validated experimentally in [129] (Section ‘Protein/DNA interactions’). The experimental results elaborate on the loss of binding function of Arg329 and R364 to DNA after being mutated to Alanine, which demonstrate the effectiveness of the utilized approach.

4.3.7. Summary

We presented a detailed method for predicting key residues dominating binding in a biomolecular protein-DNA complex. This method comprises an alanine scan

mutagenesis, electrostatic potential calculations using APBS, and free energy calculations in view of a two-step model. The method is then applied to the structural information from the GATA3:DNA complex [129]. Corresponding results depict which residues are crucial for the complex intermolecular interactions through the analysis of the free energy calculations. Moreover, this study will form the basis for designing future experiments, possibly feeding into biopharmaceutical design studies for enhanced regulation of the GATA target genes.

4.4. Application to All-Mutants GATA3

4.4.1. Problem Definition

In the previous section - Section 4.3 - we studied the effect of charged amino acids only on the binding of the GATA3:DNA complex. While those amino acids play a crucial role in binding, the rest of GATA3 amino acids, DNA bases, and DNA backbone play a comparable role. When mutations occur on the GATA-3 TF or on the DNA sequence, deregulation of the target genes might lead to different disease phenotypes. In order to predict protein-DNA binding in the presence of mutations, we examine the electrostatic mechanism behind the interactions between GATA-3 and DNA, mainly characterized by non-covalent binding and using (AESOP) framework [39,40,126].

Accordingly, we generate a family of *all* mutants of the GATA-3:DNA complex; we replace every amino acid of GATA-3, one at a time, with all other nineteen amino acids. Similarly, we replace every DNA base of the sequence ‘XGATAY’, one at a time, with all other three nucleotides. We then compute Poisson-

Boltzmann electrostatic calculations on each mutation, and subsequently, compute the free energy calculations. Each calculation delineates the contribution to binding of GATA-3:DNA complex from either a mutated amino acid or from a mutated DNA nucleotide. The crystal structure with PDB-ID: 3DFV is applied. Key amino acids and key DNA bases are identified after analyzing the calculations in view of a two-step model. Furthermore, they are validated experimentally and associated with disease phenotypes.

4.4.1. GATA3 Binding Energy Calculations

We are limiting the study of all-mutant effect in Section 4.4 to Ala and Arg amino acids only; we will resume with the study of the rest of amino acids, DNA bases, and DNA backbone in future work.

Since each mutation perturbation can alter the overall binding ability in a complex, we generated an all-mutant family of mutants for both of Ala and Arg from the crystallographic structure GATA3:DNA [129] at the atomic detail; this is accomplished through replacing Ala or Arg with all the other nineteen amino acids. We then performed Poisson–Boltzmann electrostatic calculations, feeding into electrostatic free energy calculations on each mutation, in order to reveal the contribution of each mutant to binding. Comparison of site-mutations calculations with parent/wild protein gives an indication of key amino acids in binding.

Fig. 4.4 presents the electrostatic free energy calculations of the complex GATA3:DNA, with GATA3 Ala mutants at 150mM ionic strength. The calculated solvation free energy difference of each mutant, computed from Eqn.9, is compared

against the parent/wild protein solvation free energy. This comparison serves as a physicochemical classifier of binding ability, where an increase in solvation binding free energy $\Delta\Delta G$ is considered favorable, and a decrease is considered unfavorable.

All acidic amino acid mutants of Alanine (Ala/A) in Chain-D, like A340R (Ala Residue #340 mutated to Arg), A340K, A332R, A332K, A318R, A318K, A313R, and A313K are enhancers (in red) and are predicted to increase binding, whereas all basic mutants of Ala in Chain-D, like A340D, A340E, A332D, A332E, A318D, A318E, A313D, and A313E are inhibitors (in blue), and are predicted to decrease binding. The rest of Ala mutants (in green) lay around the parent/wild region, and therefore, do not have a major impact on binding.

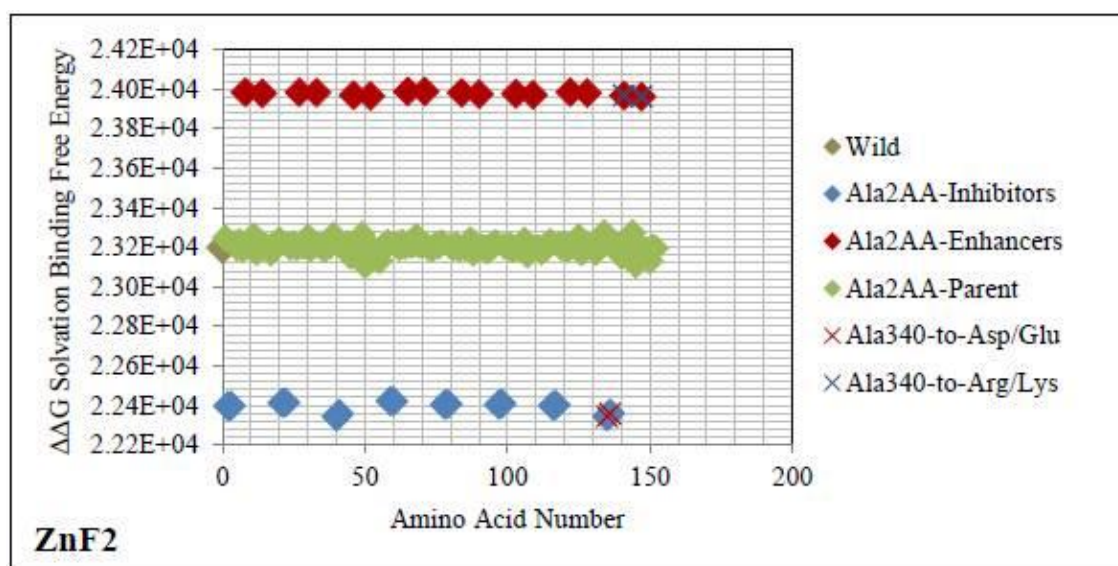


Figure 4.4. Electrostatic free energy differences of Ala all-mutants. Plot presents the solvated binding free energy calculations of GATA-3 Alanine (Ala/A) amino acid mutants in both of Chain-D and Chain-C within GATA-3:DNA complex. Blue and red colors correspond to basic and acidic mutants respectively. Mutants predicted to enhance binding are shown above the parent (wild) and mutants predicted to reduce binding are shown below it. The x-axis (index) corresponds to the order of the mutant (mutants are numbered and ordered sequentially).

In Fig. 4.5, all basic amino acid mutants of Arginine (Arg/R) in Chain-D, like R364D, R364E, R366D, R366E, R329D, R329E, R352D, R352E, R312D, R312E, R330D, and R330E are inhibitors (in blue), and so are predicted to decrease binding. The rest of Arg mutants (in green) lay around the parent/wild region and thus do not have a major impact on binding.

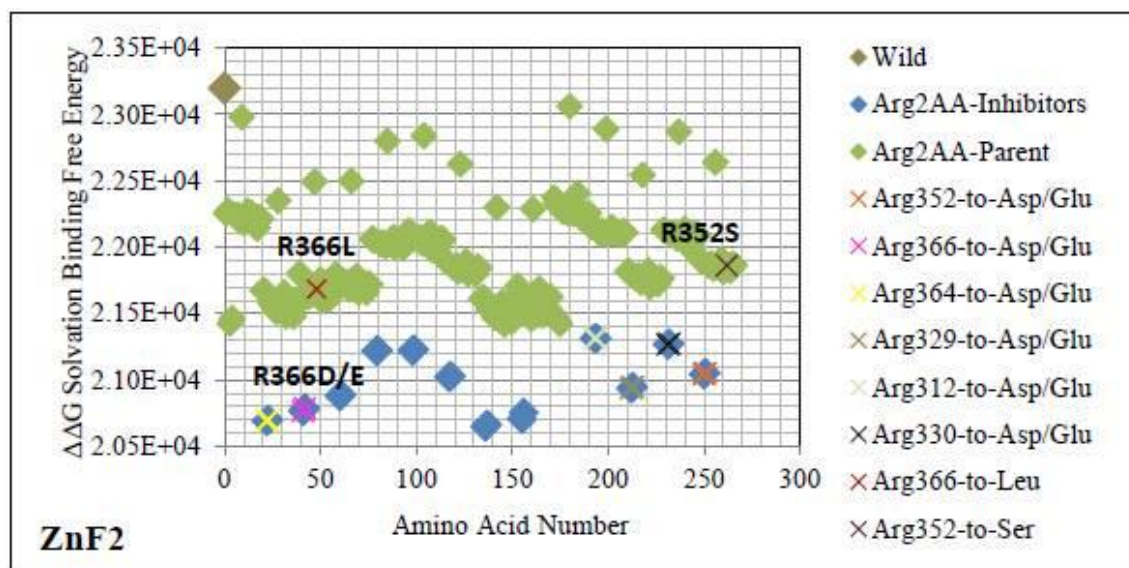


Figure 4.5. Electrostatic free energy differences of Arg all-mutants. Plot presents the solvated binding free energy calculations of GATA-3 Arginine (Arg/R) amino acid mutants in both of Chain-D and Chain-C within GATA-3:DNA complex. Blue colors correspond to basic mutants. Mutants predicted to reduce binding are shown below the parent (wild). The x-axis (index) corresponds to the order of the mutant (mutants are numbered and ordered sequentially).

CHAPTER 5

CONCLUSION AND FUTURE DIRECTIONS

For the first thesis objective, we introduced the degradomics method via Cleaved Fragments Prediction Algorithm (CFPA). The method allows the computational identification of proteases signature Breakdown Products (PDBs). Those PDBs are specific to each protease and represent biomarkers of specific biological processes, pathogenic processes, or diseases. Instances include the Tll1 protease whose BDPs are biomarkers of Congenital Heart Disease (CHD), and calpain and caspase proteases whose BDPs are biomarkers of apoptosis and necrosis.

The concept underlying CFPA is simple, robust, and efficient. It is based on Smith-Waterman algorithm with few modifications to consider variants in addition to exact matches between an input protein sequence and a consensus sequence (variants within one to two mismatches). Every alignment in the constructed dynamic table represents an exact match or a similarity; the developed algorithm prunes all alignments with Inserts or Deletes (INDELS). After all consensus occurrences are found, further modules are added to the algorithm to generate all fragments as a result of cleaving each input protein sequence by proteases at the cleavage site (predefined on the consensus sequence). Every consensus occurrence corresponds to a different cleavage of the input protein sequence. Due to the different possible conformations a protein can take, the actual cleavage becomes one possibility of all the different combinations of occurrences. Accordingly, fragments resulting from all possible combinations are generated.

CFPA is based on Smith-Waterman algorithm supporting local alignment. It searches for consensus occurrences of a few bases within input protein sequences of one to two thousands of bases. For the CHD application and the data size (mouse genome), CFPA proved high efficiency ($O(mn)$) per one protein sequence of size n and one consensus sequence of size m) in detecting regions in each input protein sequence that are similar (within one to two mismatches) to the query sequence (consensus sequence). On the other hand, CFPA is based on dynamic programming and this guarantees high sensitivity since it is not based on heuristics; CFPA did not miss any consensus occurrence and the results are validated experimentally from the literature.

For the second thesis objective, we introduced the GATA:DNA model. The model allows the prediction of key charged amino acids in protein-DNA association, while elaborating on non-covalent binding. When mutations occur on the GATA protein charged amino acid sequence, regulation of target genes fails, leading to CHD.

Due to the unavailability of the GATA4:DNA crystal structure, the model is applied to GATA3:DNA crystal structure. GATA3:DNA mutations have been implied in breast cancer and hypoparathyroidism, sensorineural deafness, and renal disease (HDR) syndrome.

The GATA model is based on the electrostatic Molecular Mechanic (MM) force field and uses Poisson-Boltzmann solvation model for the calculations of binding free energy. The calculations of the complex binding energy are performed for all possible mutations of charged amino acids on the GATA Transcription Factor (TF). The results are then validated experimentally from the literature. Such results will give

insight on any probable deficiency in the transcription biological process, and consequently, on its influence in leading to CHD.

Future directions in *degradomics* include:

The development of a web-based degradomics tool implemented for the developed degradomics method. This tool can be very valuable to molecular biologists and to public online users. It represents an automation of the degradomics method that is efficient in generating consensus similarities and cleaved fragments. Moreover, it can be flexible for adding more functionality and enhancements. For instance, it can be extended to comprise different types of cleavage modes based on different proteases; it can also be extended to input any consensus subsequence and any protein sequence instead of being limited to the ones stored.

Future directions in *protein-DNA interactions* and GATA3 include:

A continuation of the computational study of the effect of mutations of all amino acid types of GATA3 transcription factor (TF) on GATA3:DNA binding. This study will also include the effect of GATA3 dimer in the adjacent GATA3:DNA crystal structure (PDB Id: 3DFV). In addition to single amino acids mutations, the study will cover the effect of mutations of the DNA promoter sequence (DNA sequence G,A,T,A) on the binding of GATA3:DNA complex. DNA promoter sequence mutations include bases mutations and side mutations, where side mutations encompass single mutations in position 'X' or 'Y' of the DNA promoter sequence 'XGATAY'. Assessing the effect of mutations in positions around the DNA promoter main sequence (G,A,T,A), might

give insight on further potential hotspots that can affect proper binding, and subsequently proper transcription of target genes.

Afterwards, we will perform experimental validation of the above computational results against published experimental data, and we will design new experiments if necessary. Furthermore, we will link the studied mutations to disease phenotypes; crucial mutations might be visible in more than one disease.

Future directions in *protein-DNA interactions* and GATA4 include:

The prediction of GATA4:DNA pseudo-crystal structure, followed by experimental validation. Subsequently, we will study the effect of mutations of all amino acid types (charged and non-charged) of GATA4 TF on the binding ability of GATA4:DNA complex. In a similar way to GATA3:DNA complex, we will study the effect of mutations of the DNA promoter sequence (DNA sequence G,A,T,A) on the binding of GATA4:DNA complex. Then, we will perform experimental validation of the computational results and we will associate the mutational findings to known disease phenotypes.

REFERENCES

- [1] Congenital Heart Defects in Children Fact Sheet. *American Heart*. **2010**.
- [2] Schoen, F.; Richard, M. In Vinay Kumar, Abul K. Abbas, Nelson Fausto, et al. The Heart. *Robbins and Cotran Pathologic Basis of Disease (8th ed.)*, Saunders Elsevier, ISBN 978-1-4160-3121-5. **2010**.
- [3] Heart Defects: Birth Defects. *Merck*. **2010**.
- [4] *National Heart, Lung, and Blood Institute*. **2010**.
- [5] Hoffman, J.; Kaplan, S. The incidence of congenital heart disease. *Journal of the American College of Cardiology*. **2002**, 39(12), 1890-1900.
- [6] Congenital Cardiovascular Defects. *American Heart*. **2010**.
- [7] Wu, J.; Baldwin, I. New insights into plant responses to the attack from insect herbivores. *Annu Rev Genet*. **2010**, 44, 1-24.
- [8] Sangha, J.; Chen, Y.; Kaur, J. et al. Proteome Analysis of Rice (*Oryza sativa* L.) Mutants Reveals Differentially Induced Proteins during Brown Planthopper (*Nilaparvata lugens*) Infestation. *International Journal of Molecular Sciences*. **2013**, 14(2), 3921-3945.
- [9] Strimbu, K.; Tavel, J. What are biomarkers? *Curr Opin HIV AIDS*. **2010**, 5(6), 463-466.
- [10] Cauwe, B.; Martens, E.; Proost, P. et al. Multidimensional degradomics identifies systemic autoantigens and intracellular matrix proteins as novel gelatinase B/MMP-9 substrates. *Integr Biol (Camb)*. **2009**, 1(5-6), 404-426.
- [11] Ottens, A.; Kobeissy, F.; Fuller, B. et al. Novel neuroproteomic approaches to studying traumatic brain injury. *Progress in Brain Research*. **2007**, 161, 401-418.

- [12] Kobeissy, F.; Ottens, A.; Zhang, Z. et al. Novel Differential Neuroproteomics Analysis of Traumatic Brain Injury in Rats. *Mol. Cell Proteomics*. 2006, 5(10), 1887-1898.
- [13] El-Assaad, A.; Dawy, Z.; Nemer, G., Kobeissy, F. Cleaved Fragments Prediction Algorithm (CFPA) Application to Calpain and Caspase in Apoptosis and Necrotic Cell Death. *Paper presented at the IEEE International Conference on Electro/Information Technology, Northern Illinois University, DeKalb, IL, USA, 21-23 May. 2015.*
- [14] Patterson, N.; Iyer, R.; de Castro Bras, L. et al. Using proteomics to uncover extracellular matrix interactions during cardiac remodeling. *Proteomics Clin. Appl.* **2013**, 7(7-8), 516-527.
- [15] Berry, R.; Jowitt, T.; Ferrand, J. et al. Role of dimerization and substrate exclusion in the regulation of bone morphogenetic protein-1 and mammalian tolloid. *Proc Natl Acad Sci U. S. A.* **2009**, 106(21), 8561-8566.
- [16] Pellikainen, J.; Ropponen, K.; Kataja, V. et al. Expression of Matrix Metalloproteinase (MMP)-2 and MMP-9 in Breast Cancer with a Special Reference to Activator Protein-2, HER2, and Prognosis. *Clinical Cancer Research.* **2004**, 10, 7621–7628.
- [17] Bonfoco, E. et al. Apoptosis and necrosis: Two distinct events induced, respectively, by mild and intense insults with *N*-methyl-D-aspartate or nitric oxide/superoxide in cortical cell cultures. *Proc. Natl. Acad. Sci. U. S. A.* **1995**, 92, 7162–7166.

- [18] Friedman, L.; Furberg, C.; DeMets, D. *Fundamentals of clinical trials, 4th edn., Springer Science, New York, NY. 2010.*
- [19] Povlishock, J. Traumatically induced axonal injury: pathogenesis and pathobiological implications. *Brain Pathol.* **1992**, 2, 1–12.
- [20] Povlishock, J.; Stone, J. Traumatic axonal injury. *In: Langston JW (ed) Head trauma basic, preclinical, and clinical directions, Wiley, New York. 2001*, 281–301.
- [21] Pike, B.; Flint, J.; Dave, J. et al. Accumulation of Calpain and Caspase-3 Proteolytic Fragments of Brain-Derived α II-spectrin in Cerebral Spinal Fluid After Middle Cerebral Artery Occlusion in Rats. *J.Cereb. Blood Flow & Metab.* **2003**, 24, 98–106.
- [22] Lewis, S.; Velat, G.; Miralia, L. et al. Alpha-II spectrin breakdown products in aneurysmal subarachnoid hemorrhage: a novel biomarker of proteolytic injury. *J. Neurosurg.* **2007**, 107, 792–796.
- [23] Ringger, N.; O'Steen, B.; Brabham, J. et al. A Novel Marker for Traumatic Brain Injury: CSF α II-Spectrin Breakdown Product Levels. *J. Neurotrauma.* **2004**, 21, 1443–1456.
- [24] Pike, B.; Flint, J.; Dutta, S.; Johnson, E.; Wang, K.; Hayes, R. Accumulation of non-erythroid α II-spectrin and calpain-cleaved α II-spectrin breakdown products in cerebrospinal fluid after traumatic brain injury in rats. *J. Neurochem.* **2001**, 78, 1297–1306.

- [25] Cardali, S.; Maugeri, R. Detection of alphaII-spectrin and breakdown products in humans after severe traumatic brain injury. *J. Neurosurg. Sci.* **2006**, 50, 25–31.
- [26] Wang, K. Calpain and caspase: can you tell the difference?" *Trends Neurosci.* **2000**, 23, 20–26.
- [27] Dutta, S.; Chiu, Y.; Probert, A.; Wang, K. Selective Release of Calpain Produced α II-Spectrin (α -fodrin) Breakdown Products by Acute Neuronal Cell Death. *Biol. Chem.* **2002**, 383, 785–791.
- [28] Nath, R.; Davis, M.; Probert, A. et al. Processing of cdk5 Activator p35 to Its Truncated Form (p25) by Calpain in Acutely Injured Neuronal Cells. *Biochem. Biophys. Res. Commun.* **2000**, 274, 16–21.
- [29] Wang, K.; Posmantur, R.; Nath, R. et al. Simultaneous Degradation of α II- and β II-Spectrin by Caspase 3 (CPP32) in Apoptotic Cells. *J. Biol. Chem.* **1998**, 273, 22490–22497.
- [30] Petrov, D., Margreitter, C., Grandits, M. et al. A systematic framework for molecular dynamics simulations of protein post-translational modifications. *PLoS Comput Biol.* **2013**, 9(7): e1003154.
- [31] Entrez Gene: GATA4 GATA binding protein 4.
- [32] McCammon, J.A.; Northrup, H.S.; Allison, A.S. Diffusional dynamics of ligand receptor association. *J. Phys. Chem.* 1986, 90(17), 3901–3905.
- [33] Schmidt, C.Q.; Herbert, A.P.; Hocking, H.G.; Uhrin, D.; Barlow, P.N. Translational mini-review series on complement factor H: structural and

- functional correlations for factor H. *Clinical and Experimental Immunology*. **2008**, 151, 14–24.
- [34] Gehrs, K.M.; Anderson, D.H.; Johnson, L.V.; Hageman, G.S. Age-related macular degeneration—emerging pathogenetic and therapeutic concepts. *Annals of Medicine*. **2006**, 38, 450–471.
- [35] de Cordoba., S.R.; de Jorge, E.G. Translational mini-review series on complement factor H: Genetics and disease associations of human complement factor H. *Clinical and Experimental Immunology*. **2008**, 151, 1–13.
- [36] El-Assaad, A.M.; Kieslich, C.A.; Gorham Jr., R.D.; Morikis, D. Electrostatic exploration of the C3d–FH4 interaction using a computational alanine scan. *Molecular Immunology*. **2011**, 48, 1844–1850.
- [37] Zhang, L.; Mallik, B.; Morikis, D. Immunophysical exploration of C3d–CR2(CCP1–2) interaction using molecular dynamics and electrostatics. *J. Mol. Biol.* **2007**, 369, 567–583.
- [38] Cheung, A.S.; Kieslich, C.A.; Yang, J.; Morikis, D. Solvation effects in calculated electrostatic association free energies for the C3d–CR2 complex and comparison with experimental data. *Biopolymers*. **2010**, 93, 509–519.
- [39] Kieslich, C.A.; Yang, J.; Gunopulos, D.; Morikis, D. Automated computational framework for the analysis of electrostatic similarities of proteins. *Biotechnology Progress*. **2011a**, 27(2), 316–325.
- [40] Kieslich, C.A.; Gorham Jr., R.D.; Morikis, D. Is the rigid-body assumption reasonable? Insights into the effects of dynamics on the electrostatic analysis of barnasebarstar. *Journal of Non-Crystalline Solids*. **2011b**, 357, 707–716.

- [41] Gorham, R.D.; Kieslich, C.A.; Morikis, D. Complement inhibition by *Staphylococcus aureus*: electrostatics of C3d/Efb-C and C3d/Ehp association. *Cellular and Molecular Bioengineering*. **2011a**, 5(1), 32-43.
- [42] Sfyroera, G.; Katragadda, M.; Morikis, D.; Isaacs, S.N.; Lambris, J.D. Electrostatic modeling predicts the activities of orthopoxvirus complement control proteins. *The Journal of Immunology*. **2005**, 174, 2143–2151.
- [43] Zhang, L.; Morikis, D. Immunophysical properties and prediction of activities for vaccinia virus complement control protein and smallpox inhibitor of complement enzymes using molecular dynamics and electrostatics. *Biophysical Journal*. **2006**, 90, 3106–3119.
- [44] Pyram, K.; Kieslich, C.A.; Yadav, V.N.; Morikis, D.; Sahu, A. Influence of electrostatics on the complement regulatory functions of Kaposica, the complement inhibitor of Kaposi's sarcoma-associated herpesvirus. *The Journal of Immunology*. **2010**, 184, 1956–1967.
- [45] Clark, T.; Conway, S.; Scott, I.; Labosky, P.; Winnier, G.; Bundy, J.; Hogan, B.; Greenspan, D. The mammalian Toll-like 1 gene, *Tll1*, is necessary for normal septation and positioning of the heart. *Development*. **1999**, 126, 2631-2642.
- [46] Sterchi, E.; Stocker, W.; Bond, J. Membrane-bound and secreted astacin metalloproteinases. *Mol Aspects Med*. **2008**, 29, 309-328.
- [47] Stanczak, P.; Witecka, J.; Szydło, A.; Gutmajster, E.; Lisik, M.; Augusciak-Duma, A.; Tarnowski, M.; Czekaj, T.; Czekaj, H.; Sieron, A. Mutations in mammalian toll-like 1 gene detected in adult patients with ASD. *Eur J Hum Genet*. **2009**, 17, 344-351.

- [48] Wermter, C.; Howel, M.; Hintz, V.; Bombosch, B.; Aufenvenne, K.; Yiallourous, I., Stocker, W. The protease domain of procollagen C-proteinase (BMP1) lacks substrate selectivity, which is conferred by non-proteolytic domains. *Biol Chem.* **2009**, 388, 513-521.
- [49] Tsai, F.Y.; Keller, G.; Kuo, F.C.; Weiss, M.; Chen, J.; Rosenblatt, M. An early haematopoietic defect in mice lacking the transcription factor GATA-2. *Nature.* 1994, 371(6494), 221–226.
- [50] Entrez Gene: GATA1 GATA binding protein 1 (globin transcription factor 1).
- [51] Dickinson, R.E.; Griffin, H.; Bigley, V.; Reynard, L.N.; Hussain, R.; Haniffa, M. et al. Exome sequencing identifies GATA-2 mutation as the cause of dendritic cell, monocyte, B and NK lymphoid deficiency. *Blood.* **2011**, 118(10), 2656–2658.
- [52] Kouros-Mehr, H.; Slorach, E.M.; Sternlicht, M.D.; Werb, Z. GATA-3 maintains the differentiation of the luminal cell fate in the mammary gland. *Cell.* **2006**, 127(5), 1041–55.
- [53] Koboldt, D.C.; Fulton, R.S.; McLellan, M.D. et al. Comprehensive molecular portraits of human breast tumours. *Nature.* **2012**, 490, 61-70.
- [54] Chao, C.S.; McKnight, K.D.; Cox, K.L.; Chang, A.L.; Kim, S.K.; Feldman, B. J. Novel GATA6 Mutations in Patients with Pancreatic Agenesis and Congenital Heart Malformations. *PloS One.* **2015**, 10(2): 0118449.
- [55] Basson, C.; Cowley, G.; Solomon, S.; Weissman, B.; Poznanski, A.; Traill, T.; Seidman, J.; Seidman, C. The clinical and genetic spectrum of Holt-Oram syndrome. *N Engl J Med.* **1994**, 330, 885-891.

- [56] Benson, D.; Basson, C.; MacRae, C. New understandings in the genetics of congenital heart disease. *Curr Opin Pediatr.* **1996**, 8, 505-511.
- [57] Perrino, C.; Rockman, H. GATA4 and the two sides of gene expression reprogramming. *Circulation Research.* **2006**, 98, 837-845.
- [58] Köhler, B.; Lin, L.; Ferraz-de-Souza, B.; Wieacker, P.; Heidemann, P.; Schröder, V.; Biebermann, H.; Schnabel, D.; Grüters, A.; Achermann, J.C. Five novel mutations in steroidogenic factor 1 (SF1, NR5A1) in 46,XY patients with severe underandrogenization but without adrenal insufficiency. *Hum. Mutat.* **2008**, 29(1), 59–64.
- [59] Raghupathi, R.; Graham, D.; McIntosh, T. Apoptosis after traumatic brain injury. *J. Neurotrauma.* **2000**, 17, 927–938.
- [60] Behl, C. Apoptosis and Alzheimer's disease. *J. Neural Transm.* **2000**, 107, 1325–1344.
- [61] Clark, R.; Chen, J.; Watkins, S. et al. Apoptosis-Suppressor Gene bcl-2 Expression after Traumatic Brain Injury in Rats. *J. Neurosci.* **1997**, 17, 9172–9182.
- [62] Beer, R.; Franz, G.; Srinivasan, A. et al. Temporal Profile and Cell Subtype Distribution of Activated Caspase-3 Following Experimental Traumatic Brain Injury. *J. Neurochem.* **2000**, 75, 1264–1273.
- [63] Vanderklish, P.; Bahr, B. The pathogenic activation of calpain: a marker and mediator of cellular toxicity and disease states. *Int. J. Exp. Path.* **2000**, 81, 323–339.

- [64] Kerr, J.; Wyllie, A.; Currie, A. Apoptosis: a basic biological phenomenon with wide-ranging implications in tissue kinetics. *Br. J. Cancer*. **1972**, 26, 239–257.
- [65] Rathmell, J.; Thompson, C. The central effectors of cell death in the immune system. *Annu. Rev. Immunol.* **1999**, 17, 781–828.
- [66] Majno, G.; Joris, I. Apoptosis, Oncosis, and Necrosis. An Overview of Cell Death. *Am. J. Pathol.* **1995**, 146, 3–15.
- [67] Van Cruchten, S.; Van Den Broeck, W. Morphological and Biochemical Aspects of Apoptosis, Oncosis and Necrosis. *Anat. Histol. Embryol.* **2002**, 31, 214–223.
- [68] Lecoœur, H.; Prevost, M.; Gougeon, M. Oncosis Is Associated with Exposure of Phosphatidylserine Residues on the Outside Layer of the Plasma Membrane: A reconsideration of the Specificity of the Annexin V/Propidium Iodide Assay. *Cytometry*. **2001**, 44, 65–72.
- [69] Farber, E. Programmed cell death: necrosis versus apoptosis. *Mod. Pathol.* **1994**, 7, 605–609.
- [70] Levin, S. Apoptosis, Necrosis, or Oncosis: What Is Your Diagnosis? A Report from the Cell Death Nomenclature Committee of the Society of Toxicologic Pathologists. *Toxicol. Sci.* **1998**, 41, 155–156.
- [71] Hirsch, T.; Marchetti, P.; Susin, S. et al. The apoptosis-necrosis paradox. Apoptogenic proteases activated after mitochondrial permeability transition determine the mode of cell death. *Oncogene*. **1997**, 15, 1573–1581.
- [72] Criddle, D.; Gerasimenko, J.; Baumgartner, H. et al. Calcium signalling and pancreatic cell death: apoptosis or necrosis? *Cell Death & Differ.* **2007**, 14, 1285–1294.

- [73] Muller, G.; Stadelmann, C.; Bastholm, L.; Elling, F.; Lassmann, H.; Johansen, F. Ischemia Leads to Apoptosis--and Necrosis-like Neuron Death in the Ischemic Rat Hippocampus. *Brain Pathol.* **2004**, 14, 415–424.
- [74] Zhang, Z.; Larner, S.; Liu, M.; Zheng, W.; Hayes, R.; Wang, K. Multiple alphaII-spectrin breakdown products distinguish calpain and caspase dominated necrotic and apoptotic cell death pathways. *Apoptosis.* **2009**, 14, 1289–1298.
- [75] www.ncbi.nlm.nih.gov/genbank/.
- [76] Lopez-Otin, C.; Overall, C. Protease degradomics: a new challenge for proteomics. *Nat. Rev. Mol. Cell Biol.* **2002**, 3(7), 509-519.
- [77] Yuen, P-W.; Wang, K. Calpain inhibitors, novel neuroprotectants and potential anticataractic agents. *In: Drugs of the Future, Thomson Reuters.* **1998**, 23, 741–749.
- [78] Nath, R. et al. Non-erythroid α -spectrin breakdown by calpain and interleukin 1β -converting-enzyme-like protease(s) in apoptotic cells: contributory roles of both protease families in neuronal apoptosis. *Biochem. J.* **1996**, 319, 683–690.
- [79] Squier, M. et al. Calpain Activation in Apoptosis. *J. Cell. Physiol.* **1994**, 159, 229–237.
- [80] Nath, R. et al. Effects of ICE-like protease and calpain inhibitors on neuronal apoptosis. *NeuroReport.* **1996**, 8, 249–256.
- [81] Nicholson, D.; Thornberry, N. Caspases: killer proteases. *Trends Biochem. Sci.*, 22, 299–306.

- [82] Nath, R. et al. Evidence for Activation of Caspase-3-like protease in Excitotoxin- and Hypoxia/Hypoglycemia-Injured Neurons. *J. Neurochem.* **1998**, 71, 186–195.
- [83] Pike, B. et al. Regional calpain and caspase-3 proteolysis of α -spectrin after traumatic brain injury. *NeuroReport.* **1998**, 9, 2437–2442.
- [84] Pike, B. et al. Temporal Relationships Between De Novo Protein Synthesis, Calpain and Caspase 3-like Protease Activation, and DNA Fragmentation During Apoptosis in Septo-Hippocampal Cultures. *J. Neurosci.Res.* 1998, 52, 505–520.
- [85] Plasman, K.; Demol, H.; Bird, P. et al. Substrate specificities of the granzyme tryptases A and K. *J Proteome Res.* **2014**, 13(12), 6067-6077.
- [86] Majovsky, P.; Nauman, C.; Lee, C. et al. Targeted proteomics analysis of protein degradation in plant signaling on an LTQ-Orbitrap mass spectrometer. *J. Proteome Res.* **2014**, 13(10), 4246-4258.
- [87] Barre, O.; Dufour, A.; Eckhard, U. et al. Cleavage specificity analysis of six type II transmembrane serine proteases (TTSPs) using PICS with proteome-derived peptide libraries. *PLoS One.* **2014**, 9(9): e105984.
- [88] Wang, L.; Xia, L.; Shen, S. et al. Dissecting cell death with proteomic scalpels. *Proteomics.* **2012**, 12(4-5), 597-606.
- [89] Clark, T.; Conway, S.; Scott, I. et al. The mammalian Tolloid-like 1 gene, *Tll1*, is necessary for normal septation and positioning of the heart. *Development.* **1999**, 126(12), 2631-2642.

- [90] Glantz, S.; Cianci, C.; Iyer, R. et al. Sequential degradation of alphaII and betaII spectrin by calpain in glutamate or maitotoxin-stimulated cells. *Biochemistry*. **2007**, 46(2), 502-513.
- [91] Shen, Y.; Tolic, N.; Liu, T. et al. Blood Peptidome-Degradome Profile of Breast Cancer. *PLoS ONE*. **2010**, 5(10): e13133.
- [92] Itoh, Y.; Nagase, H. Matrix metalloproteinases in cancer. *Essays Biochem*. **2002**, 38, 21-36.
- [93] Fuhrman-Luck, R.; Silva, M.; Dong, Y. et al. Proteomic and other analyses to determine the functional consequences of deregulated kallikrein-related peptidase (KLK) expression in prostate and ovarian cancer. *Proteomics Clin. Appl*. **2014**, 8(5-6), 403-415.
- [94] Vihinen, P.; Kahari, V. Matrix metalloproteinases in cancer: prognostic markers and therapeutic targets. *Int. J. Cancer*. **2002**, 99(2), 157-166.
- [95] Butler, G.; Overall, C. Updated biological roles for matrix metalloproteinases and new "intracellular" substrates revealed by degradomics. *Biochemistry*. **2009**, 48(46), 10830-10845.
- [96] Sterchi, E.; Stocker, W.; Bond, J. Meprins, membrane-bound and secreted astacin metalloproteinases. *Mol. Aspects Med*. **2008**, 29(5), 309-328.
- [97] Smith, T.; Waterman, M. Identification of Common Molecular Subsequences. *J. Mol. Biol*. **1981**, 147(1), 195-197.
- [98] Knuth, D.; Morris, J.; Pratt, V. Fast pattern matching in strings. *SIAM J. Comput.*, 6, 323-350.

- [99] Lipman, D., Pearson, W. Rapid and sensitive protein similarity searches. *Science*. **1985**, 227, 1435–1441.
- [100] Altschul, S.; Gish, W.; Miller, W. et al. Basic local alignment search tool. *J. Mol. Biol.* **1990**, 215, 403–410.
- [101] Ning, Z., Cox, A.; Mullikin, J. SSAHA: a fast search method for large DNA databases. *Genome Res.* **2001**, 11, 1725–1729.
- [102] Tromp, J.; Li, M. PatternHunter: faster and more sensitive homology search. *Bioinformatics*. **2002**, 18, 440–445.
- [103] Kurtz, S.; Phillipy, A.; Delcher, A. et al. Versatile and open software for comparing large genomes. *Genome Biol.* **2004**, 5.
- [104] Lecroq, T. Fast exact string matching algorithms. *Information Processing Letters*, **2007**, 102(6), 229-235.
- [105] Needleman, S.; Wunsch, C. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **1970**, 48, 443-453.
- [106] Srikantha, A.; Bopardikar, A.; Kaipa, K. et al. A fast algorithm for exact sequence search in biological sequences using polyphase decomposition. *Bioinformatics*. **2010**, 26(18): i414-i419.
- [107] Li, H.; Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*. **2010**, 26(5), 589-595.
- [108] Leach, A. Molecular modelling: principles and applications. *Prentice Hall*, 2nd ed., **2001**.

- [109] Arunan, E.; Desiraju, G.; Klein, R.; Sadlej, J.; Scheiner, S.; Alkorta, I.; Clary, D.; Crabtree, R.; Dannenberg, J.; Hobza, P.; Kjaergaard, H.; Legon, A.; Mennucci, B.; Nesbitt, D. Definition of the hydrogen bond. *Pure Appl. Chem.* **2011**, 83(8), 1637.
- [110] Fersht, A. The Hydrogen-Bond in Molecular Recognition. *Trends Biochem. Sci.* **1987**, 12(8), 301.
- [111] Fersht, A. Basis of Biological Specificity. *Trends Biochem. Sci.* **1984**, 9(4), 145.
- [112] London, F. On the Theory and Systematic of Molecular Forces. *Z Phys.* **1930**, 63(3-4), 245.
- [113] Jones, S.; Thornton, J. Principles of protein-protein interactions. *Proc. Natl. Acad. Sci. USA.* **1996**, 93(1), 13.
- [114] Blundell, T.; Srinivasan, N. Symmetry, stability, and dynamics of multidomain and multicomponent protein systems. *Proc. Natl. Acad. Sci. USA.* **1996**, 93, 14243–14248.
- [115] Bult, C.; Eppig, J.; Kadin, J. The Mouse Genome Database (MGD): mouse biology and model systems. *Nucleic Acids Research.* **2008**, 36.
- [116] Dobson, R.; Walker, H.; Walker, N. *Biomarkers in Medicine.* **2014**, 8(7), 965–975 ISSN 1752-0363.
- [117] Yuan, S. Aortic Disorders, Facial Dysmorphism and Mental Retardation: Clinical Features and Genetic Conditions. *Acta Medica Mediterranea.* **2013**, 29, 817.
- [118] El-Assaad, A.; Dawy, Z.; Nemer, G.; Hajj, H.; Kobeissy, F. Efficient and Accurate Algorithm for Cleaved Fragments Prediction (CFPA) in Protein

- Sequences Dataset based on Consensus and its Variants: A Novel Degradomics Prediction Application. *In: Methods in Molecular Biology, Springer. 2015*, Submitted and Accepted.
- [119] Lopez, O.; Overall, C. Protease degradomics: a new challenge for proteomics. *Nat. Rev. Mol. Cell Biol.* **2002**, 3, 509-519.
- [120] Ottens, A.; Kobeissy, F.; Fuller, B.; Liu, M.; Oli, M.; Hayes, R.; Wang, K. Novel neuroproteomic approaches to studying traumatic brain injury. *Prog. Brain Res.* **2007**, 161, 401-418.
- [121] Kobeissy, F.; Ottens, A.; Zhang, Z.; M. Liu, Denslow, N.; Dave, J.; Tortella, F.; Hayes, R.; Wang, K. Novel Differential Neuroproteomics Analysis of Traumatic Brain Injury in Rats. *Mol. Cell Proteomics.* **2006**, 5, 1887-1898.
- [122] Wang, K. Simultaneous Degradation of α II- and β II-Spectrin by Caspase 3 (CPP32) in Apoptotic Cells. *J. Biol. Chem.* **1998**, 273, 22490–22497.
- [123] Kobeissy, F.; Liu, M.; Yang, Z.; Zhang, Z.; Zheng, W.; Glushakova, O.; Mondello, S.; Anagli, J.; Hayes, R.; Wang, K. Degradation of β II-Spectrin Protein by Calpain-2 and Caspase-3 Under Neurotoxic and Traumatic Brain Injury Conditions. *Springer Science.* **2014**.
- [124] Cornell, W.D; Cieplak, P.; Bayly, C.I.; Gould, I.R.; Merz Jr., K.M.; Ferguson, D.M.; Spellmeyer, D.C.; Fox, T.; Caldwell, J.W.; Kollman, P.A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, 117(19), 5179–5197.

- [125] Baker, N.A.; Sept, D.; Joseph, S.; Holst, M.J.; McCammon, J.A. Electrostatics of nanosystems: application to microtubules and the ribosome. *PNAS*. 2001, 98(18), 10037–10041.
- [126] Gorham, R.D.; Kieslich, C.A.; Morikis, D. Electrostatic clustering and free energy calculations provide a foundation for protein design and optimization. *Annals of Biomedical Engineering*. **2011b**, 39(4), 1252–1263.
- [127] Dolinsky, T.J.; Nielsen, J.E.; McCammon, J.A.; Baker, N.A. PDB2PQR: an automated pipeline for the setup of Poisson–Boltzmann electrostatics calculations. *Nucleic Acids Research*, 32, Web Server Issue W665–W667.
- [128] Honig, B.; Nicholls, A. Classical electrostatics in biology and chemistry. *Science*. **1995**, 268(5214), 1144–1149.
- [129] Bates, D.L.; Chen, C.Y.; Kim, G.; Guo, L.; Chen, L. Crystal Structures of Multiple GATA Zinc Fingers Bound to DNA Reveal New Insights into DNA Recognition and Self-Association by GATA. *J. Mol. Biol.* **2008**, 381, 1292–1306.
- [130] Pettersen, E.F.; Goddard, T.D.; Huang, C.C.; Couch, G.S.; Greenblatt, D.M.; Meng, E.C.; Ferrin, T.E. UCSF chimera—a visualization system for exploratory research and analysis. *Journal of Computational Chemistry*. **2004**, 25(13), 1605–1612.

Appendix

1. Instances of Tll1 Fragments Generated by CFPA

Sequence Description	Sequence 1			
Gene_Symbol =A2bp1 Isoform 1 of RNA binding protein fox-1 homolog 1	MNCEREQLRGNQEAAAAPDTMAQPYASAQFAPPQNGIPA EYTAPHPHPAPEYTGQTTVPDHTLNLYPPT QTHSEQSADTSAQTVSGTATQTDDAAPT DGQPQTQPSENTESK SQPKRLHVSNI PFRFRDPDLRQMFGQF GKILDVEIIFNERGSKGFGFVTFENSADADRAREKLHGTVVEGRKIEVNNATARVMTNKKTVNPTYNGW KLNPNVVGAVYSPDFYAGTVLLCQANQEGSSMYSGPSSLVYTSAMPGFYPAPAATAAAAYRGAHLRGRGR TVYNTFRAAAPPPPIAYGGVVYQDGFYGADIYGGYAA YRYAQPTPATAAAAYSDSYGRVYAADPYHHT LAPAPTYGVGAMNAFAPLTD AKTRSHADDVGLV LSSLQASIYRGGYNRFAPY			
Consensus Occurrence (CN)	Start	End		
PATA↑AAYS	322	329		
Fragment 1			Start	End
MNCEREQLRGNQEAAAAPDTMAQPYASAQFAPPQNGIPA EYTAPHPHPAPEYTGQTTVPDHTLNLYPPT QTHSEQSADTSAQTVSGTATQTDDAAPT DGQPQTQPSENTESK SQPKRLHVSNI PFRFRDPDLRQMFGQF GKILDVEIIFNERGSKGFGFVTFENSADADRAREKLHGTVVEGRKIEVNNATARVMTNKKTVNPTYNGW KLNPNVVGAVYSPDFYAGTVLLCQANQEGSSMYSGPSSLVYTSAMPGFYPAPAATAAAAYRGAHLRGRGR TVYNTFRAAAPPPPIAYGGVVYQDGFYGADIYGGYAA YRYAQPTPATA			1	325
Fragment 2			Start	End
AAYSDSYGRVYAADPYHHTLAPAPTYGVGAMNAFAPLTD AKTRSHADDVGLV LSSLQASIYRGGYNRF APY			326	396
Sequence Description	Sequence 1 Reversed			
Gene_Symbol =A2bp1 Isoform 1 of RNA binding protein fox-1 homolog 1	YPAFRNYGGRYISAQLSSLVLGVDDAHSRTKADTLPAFANMAGVGYTPAPAL THHYPDAA YVRGYSDS YAAATAPTPQAYRYAAYGGYIDAGYFGDQYVVGGYAPIPPPPAAARFTNYVTRGRGRLHAGRYAAAA TAAPYPFPGPMASTYVLSPPGYSMSSGEQNAQCLLV TGAYFDPSYVAGVVPNLK WGNTYPNVTKKNTM VRATANNVEIKRGEVVTGHLKERARDADASNEFTVFGFGKSGRENFIIEVDLIKGFQGFMRQLDPRFRF PINSVHLRKPQSKSETNESPQTQPQGDTPAADDTQTATGSVTQASTDASQESHTQTPPYLNLTHDPVTTQ GTYEPAPHPHPATYEAPIGNQPPAFQASAYPQAMTDPAAAAEQNGRLQERECNM			
Consensus Occurrence (NC)	Start	End		
SYAA↑ATAP	68	75		
Fragment 1			Start	End
YPAFRNYGGRYISAQLSSLVLGVDDAHSRTKADTLPAFANMAGVGYTPAPAL THHYPDAA YVRGYSDS YAA			1	71
Fragment 2			Start	End
ATAPTPQAYRYAAYGGYIDAGYFGDQYVVGGYAPIPPPPAAARFTNYVTRGRGRLHAGRYAAAA TAAP YPFPGPMASTYVLSPPGYSMSSGEQNAQCLLV TGAYFDPSYVAGVVPNLK WGNTYPNVTKKNTMVRATA NNVEIKRGEVVTGHLKERARDADASNEFTVFGFGKSGRENFIIEVDLIKGFQGFMRQLDPRFRFPINSVH LRKPQSKSETNESPQTQPQGDTPAADDTQTATGSVTQASTDASQESHTQTPPYLNLTHDPVTTQGTYEPA PHPHPATYEAPIGNQPPAFQASAYPQAMTDPAAAAEQNGRLQERECNM			72	396

Sequence Description	Sequence 2
Gene_Symbol =C8b Isoform 1 of Complement component C8 beta chain	MKIGAQVWRALAKSCLLCATLGLHFPGSRGGKPDFFETKAVNGSLVKS RPPVRSVAEAPAPIDCELSTWSSWTACDPCQKKRYRHTYLLRPSQFYGELCDLSDKEVEDCVTNQPCRSQVRCEGFVCAQTGRVCVNRRLCNGDNDCGDSDEANCRRIYKNCQREMEQYWAIDRLASGINLFTNTFEGPVLDRHYAAGGCSPHYILD TNFRKPYNVESYTPQTKCEYEFITL TEYESYSDFERLVIEKKTHMFNFTSGFKVDGVM DLGKVESNEGKN YVTRTKRFAHTQSKFLHARSVLEV AHYKLSRSLMLHYEFLQRVKSLPLEYSYGEYRDLLRDFGTHFITEAVLGGIYEYTLIMNKDAMEQGDYTL SHVTACAGGSFGIGGMVYKVVYKVGVS AKKCS DIMKEINERNK RSTMVEDLVVLRGGTSEDITALAYKELPTPELMEAWGDAVKYNPAI IKIKAEPLYELV TATDFAYS STV KQNLKKALEEFQSEVSSRCAPCRGNGVPVLKGSRCCEICPGGFQGTACEV TYRKDIPIDGKWSCWSDW SACSGGHKTRHRQCNNPAPHKGGSPSGPASETLCN

Consensus Occurrence (CN)	Start	End
TATD↑FAYS	474	481

Fragment 1	Start	End
MKIGAQVWRALAKSCLLCATLGLHFPGSRGGKPDFFETKAVNGSLVKS RPPVRSVAEAPAPIDCELSTWSSWTACDPCQKKRYRHTYLLRPSQFYGELCDLSDKEVEDCVTNQPCRSQVRCEGFVCAQTGRVCVNRRLCNGDNDCGDSDEANCRRIYKNCQREMEQYWAIDRLASGINLFTNTFEGPVLDRHYAAGGCSPHYILD TNFRKPYNVESYTPQTKCEYEFITL TEYESYSDFERLVIEKKTHMFNFTSGFKVDGVM DLGKVESNEGKN YVTRTKRFAHTQSKFLHARSVLEV AHYKLSRSLMLHYEFLQRVKSLPLEYSYGEYRDLLRDFGTHFITEAVLGGIYEYTLIMNKDAMEQGDYTL SHVTACAGGSFGIGGMVYKVVYKVGVS AKKCS D	1	477

Fragment 2	Start	End
FAYSSTVKQNLKKALEEFQSEVSSRCAPCRGNGVPVLKGSRCCEICPGGFQGTACEV TYRKDIPIDGKWSCWSDWSACSGGHKTRHRQCNNPAPHKGGSPSGPASETLCN	478	589

Sequence Description	Sequence 2 Reversed
Gene_Symbol =C8b Isoform 1 of Complement component C8 beta chain	CNLTESAPGSCPSGGKHPAPNNCQRHR TKHGGSCASWDSWCSWKGDIPIDKRYTVECATGQFGGPCICE CRSGKLVPGNGRCPACRCSVESQFEELAKKLNQKVTS SYAFDTAT VLEYLPEAKIKIAPNYKVADGWAEMLEPTPLEKYALATIDESTGGRV LVVLDDEVMTSRKNRENIEKMIDSCKKASVGVKVVYKVMGGI GFSGGACATVHSLTYDQEMADKNMILTYEYIGGLVAETIFHTGFDRLLDREYEGYSYELPLSKVRQLFEYHMLMRSRSLKYHAVELVSR AHLFQSQTAFRKRTRTVYKNGENSEVKIGLDMVGDV KFGSTFNFMTTK KEIVLREFDSYSEYETLTFEYECKTQPTYSEVNYPKRFNTDLIYHPS CGGAYYRHDLVPGEFTNTFLNIGS ALRDIAWYQEMERQCNKYIRRCNAEDSQDGDNDGNCLLRNVCRTGQACVFGECRVQSRCPQNTVC DEVEKDSLDCLEGYFQSPRLLYTHRYRKKQCPDCATWSSWTSLECDIPAPAEAVSRVPRSKVLSGNVAKTEFFDPKGGRS GPFHLCGLTACLCSKALARWVQAGIKM

Consensus Occurrence (NC)	Start	End
SYAF↑DTAT	109	116

Fragment 1	Start	End
CNLTESAPGSCPSGGKHPAPNNCQRHR TKHGGSCASWDSWCSWKGDIPIDKRYTVECATGQFGGPCICE CRSGKLVPGNGRCPACRCSVESQFEELAKKLNQKVTS SYAF	1	112

Fragment 2	Start	End
DTATVLEYLPEAKIKIAPNYKVADGWAEMLEPTPLEKYALATIDESTGGRV LVVLDDEVMTSRKNRENIE KMIDSCKKASVGVKVVYKVMGGI GFSGGACATVHSLTYDQEMADKNMILTYEYIGGLVAETIFHTG FDRLLDRYEGYSYELPLSKVRQLFEYHMLMRSRSLKYHAVELVSR AHLFQSQTAFRKRTRTVYKNGENS EVKIGLDMVGDV KFGSTFNFMTTKKEIVLREFDSYSEYETLTFEYECKTQPTYSEVNYPKRFNTDLIYHPS CGGAYYRHDLVPGEFTNTFLNIGS ALRDIAWYQEMERQCNKYIRRCNAEDSQDGDNDGNCLLRNVC RTGQACVFGECRVQSRCPQNTVC DEVEKDSLDCLEGYFQSPRLLYTHRYRKKQCPDCATWSSWTSLECD IPAPAEAVSRVPRSKVLSGNVAKTEFFDPKGGRS GPFHLCGLTACLCSKALARWVQAGIKM	113	589

Sequence Description	Sequence 3
Gene_Symbol =Atp6ap1 V-type proton ATPase subunit S1	MMAATVVSRI RTGTGRAPVMWLSLSLVAVAAA VATEQQVPLVLWSSDRNLWAPVADTHEGHITSDMQ LSTYLDPALELGPRNVLLFLQDKLSIEDFTAYGGVFGNKQDSAFSNLENALDLAPSSLVLPVVDWY AISTL TTYLQEKLGASPLHVLDLTKELKLNASLPALLLIRLPYTASSGLMAPREVL TNDEVIGQV LSTLKSE DVPTYAALTA VRPSRVARDITMVAGGLGRQLLQTQVASPAIHPPVSYNDTAPRILFWAQNFVAYKDE WKDLTSLTFGVENLNL TGSFWNDSFAMLSLTYEPLFGATVTFK FILASRFYPVSARYWFAMERLEIHSNG SVAHFNVSQVTGPSIYSFHCEYVSSVSKKGNLLVTNVPSVWQMTLHNFQIQAFNVTGEQF SYASDCAG FSPGIWMGLLTTFLMFLFIFTYGLHMILSLKTMDFDDHKGPTITLTQIV

Consensus Occurrence (NN)	Start	End
SYAS↑DCAG	406	413

Fragment 1	Start	End
MMAATVVSRI RTGTGRAPVMWLSLSLVAVAAA VATEQQVPLVLWSSDRNLWAPVADTHEGHITSDMQ LSTYLDPALELGPRNVLLFLQDKLSIEDFTAYGGVFGNKQDSAFSNLENALDLAPSSLVLPVVDWY AISTL TTYLQEKLGASPLHVLDLTKELKLNASLPALLLIRLPYTASSGLMAPREVL TNDEVIGQV LSTLKSE DV PYTAALTA VRPSRVARDITMVAGGLGRQLLQTQVASPAIHPPVSYNDTAPRILFWAQNFVAYKDEWKD L TSLTFGVENLNL TGSFWNDSFAMLSLTYEPLFGATVTFK FILASRFYPVSARYWFAMERLEIHSNGSVAH FNVSQVTGPSIYSFHCEYVSSVSKKGNLLVTNVPSVWQMTLHNFQIQAFNVTGEQFSYAS	1	409

Fragment 2	Start	End
DCAGFFSPGIWMGLLTTFLMFLFIFTYGLHMILSLKTMDFDDHKGPTITLTQIV	410	463

Sequence Description	Sequence 3 Reversed
Gene_Symbol =Atp6ap1 V-type proton ATPase subunit S1	VIQTLTITPGKHDDFRDMTKLSLIMHLGYTFIFLMFLTLLGMWIGPSFF GACDSAYS FQEGTVNFAQIQF NHLTMQWVSPVNTVLLNGKKS VSSVYECHFSYISP GTVQSVNFHAVSGNSHIELREMAFWYRASVPYFR SALIFKFTVTAGFLPEYTL SLMAFSDNWFSGTLNLNEVGFTLSTLDKWKEDKYAVSFNQAWFLIRPATDN YSVPPHIAPSAVQTQLLQRGLGGAVMTIDRAVRSPRVATLAATYPVDESKLTSLVQGVIEDNGTLVERPA MLGSSATYPLRILLAPLSANLKLEKLTALDVHLPSAGLKEQLYTTLSIA YWDVAPLVLSSPALDLANE LNSFASDQKNGFVGGYATFDEISLKDQLFLLVNRPGLELAPDLYTSLQMDSTIHGEHTDAVPAWLN RDS SWLVLPVQQETA VAAA VAVLSLSLWMVPARGTGTRIRSVVTAAMM

Consensus Occurrence (CC)	Start	End
GACD↑SAYS	51	58

Fragment 1	Start	End
VIQTLTITPGKHDDFRDMTKLSLIMHLGYTFIFLMFLTLLGMWIGPSFFGACD	1	54

Fragment 2	Start	End
SAYSFQEGTVNFAQIQFNHLTMQWVSPVNTVLLNGKKS VSSVYECHFSYISP GTVQSVNFHAVSGNSHIE LREMAFWYRASVPYFRSALIFKFTVTAGFLPEYTL SLMAFSDNWFSGTLNLNEVGFTLSTLDKWKEDKYA VSFNQAWFLIRPATDNYSVPPHIAPSAVQTQLLQRGLGGAVMTIDRAVRSPRVATLAATYPVDESKLTSL VQGVIEDNGTLVERPAMLGSSATYPLRILLAPLSANLKLEKLTALDVHLPSAGLKEQLYTTLSIA YWDVAPLVLSSPALDLANE LNSFASDQKNGFVGGYATFDEISLKDQLFLLVNRPGLELAPDLYTSLQMDSTIHG EHTDAVPAWLN RDS SWLVLPVQQETA VAAA VAVLSLSLWMVPARGTGTRIRSVVTAAMM	55	463

Sequence Description	Sequence 4
Gene_Symbol =Atp6ap1 V-type proton ATPase subunit S2	MCLSALILVSLAAFTAGAGHPSSPPMVDTVQGKVLGKYISLEGFTQPVAVFLGVPFAKPPGLSLRFAPPQPAEPWSSVKNATSYPPMCQDPVTGQIVNDLLTNRKEKIPLQFSEDCLYLNIYTPADLTKSDRLPVMVWIHGGGLVLGGASTYDGLVLSTHENVVVVVIQYRLGIWGFSTGDEHSRGNWGHLDQVAALHWVQDNIAKFGGDPGSVTIFGESAGGESVSVLVSPLAKNLFQRAISESGVALTAGLVKKNTRPLAEKIAVISGCKNTTSAAMVHCLRQKTEEEELLGTTLKLNLFKLDLHGDSRQSHPFVPTVLDGVLLPKMPEEILAEKNFNTVPYIVGINKQEFGWILPTMMNYPSPDVKLDQMTAMSLKSSFLLNLPEDAIAVAIEKYLRDKDYTGRNKDQLLELIGDVVFGVPSVIVSRGHRDAGAPTYMYEFQYSPSFSSEMKPDTVVDHGDEIYSVFGAPILRGGTSEEEINLSKMMMFWANFARNGNPNGQGLPHWPEYDQKEGYLQIGATTQQAQKLKEKEVAFWTELLAKQLPTEHTEL

Consensus Occurrence (NN)	Start	End
SLAA↑FTAG	10	17

Fragment 1	Start	End
MCLSALILVSLAA	1	13

Fragment 2	Start	End
FTAGAGHPSSPPMVDTVQGKVLGKYISLEGFTQPVAVFLGVPFAKPPGLSLRFAPPQPAEPWSSVKNATSYPPMCQDPVTGQIVNDLLTNRKEKIPLQFSEDCLYLNIYTPADLTKSDRLPVMVWIHGGGLVLGGASTYDGLVLSTHENVVVVVIQYRLGIWGFSTGDEHSRGNWGHLDQVAALHWVQDNIAKFGGDPGSVTIFGESAGGESVSVLVSPLAKNLFQRAISESGVALTAGLVKKNTRPLAEKIAVISGCKNTTSAAMVHCLRQKTEEEELLGTTLKLNLFKLDLHGDSRQSHPFVPTVLDGVLLPKMPEEILAEKNFNTVPYIVGINKQEFGWILPTMMNYPSPDVKLDQMTAMSLKSSFLLNLPEDAIAVAIEKYLRDKDYTGRNKDQLLELIGDVVFGVPSVIVSRGHRDAGAPTYMYEFQYSPSFSSEMKPDTVVDHGDEIYSVFGAPILRGGTSEEEINLSKMMMFWANFARNGNPNGQGLPHWPEYDQKEGYLQIGATTQQAQKLKEKEVAFWTELLAKQLPTEHTEL	14	562

Sequence Description	Sequence 4 Reversed
Gene_Symbol =Atp6ap1 V-type proton ATPase subunit S2	LETHETPLQKKALLETWFAVEKEKLLQAQQTAGIQLYGEKQDYEPWHPLGQGNPNGNRAFNAWFKM MMKSLNIEEESTGGRLIPAGFVSYIEDGHDGVVTDPKMESSFSPSYQFEYMYTPAGADRHGRSVIVSPVGFVVDGILELLQDKNRGTYDKDRLYKEIAVAIADEPLNLLFSSKLLSMATMQDLKVDSPPNMMPPLIWGFEQKNIGVIYPVTNFNKEALIEEPMKPLLVDLVTVPFPHSQRSDGHLDLKFLNLKLTGGLLEETKQRLCHVMAASTTNKCGSIVAIKEALPRTNKKVLGATLAVGSESIARQFLNKALPSLVLSVSEGGASEGFITVSGPDGGFKAINQVWHLAAVQDLHGWNRSLEDGTSFFGWIGLRYQIVVVVNEHTSLVLGDYTSAGGLVGGHIWVMVPLRDSKTLDAPTYINLYLCESEFQLPIKEKRNTLLDNVIQGTVPDQFCMPPYSTANKVSSWPEAPQPPAFRLSGLPPKAFPVGLFVAVPQTFGELSIYKGLVKGQVTDVMPSSPHGAGATFAALS SVLILASLCM

Consensus Occurrence (CC)	Start	End
GATF↑AALS	546	553

Fragment 1	Start	End
TAGIQLYGEKQDYEPWHPLGQGNPNGNRAFNAWFKMMMKSLEEESTGGRLIPAGFVSYIEDGHDGVVTDPKMESSFSPSYQFEYMYTPAGADRHGRSVIVSPVGFVVDGILELLQDKNRGTYDKDRLYKEIAVAIADEPLNLLFSSKLLSMATMQDLKVDSPPNMMPPLIWGFEQKNIGVIYPVTNFNKEALIEEPMKPLLVDLVTVPFPHSQRSDGHLDLKFLNLKLTGGLLEETKQRLCHVMAASTTNKCGSIVAIKEALPRTNKKVLGATLAVGSESIARQFLNKALPSLVLSVSEGGASEGFITVSGPDGGFKAINQVWHLAAVQDLHGWNRSLEDGTSFFGWIGLRYQIVVVVNEHTSLVLGDYTSAGGLVGGHIWVMVPLRDSKTLDAPTYINLYLCESEFQLPIKEKRNTLLDNVIQGTVPDQFCMPPYSTANKVSSWPEAPQPPAFRLSGLPPKAFPVGLFVAVPQTFGELSIYKGLVKGQVTDVMPSSPHGAGATF	1	549

Fragment 2	Start	End
AALS SVLILASLCM	550	562

2. All Combinations Generated by CFP A-CalpCasp

Input Protein Sequence Seq.#194
MLQDSITGIVNSFNLFFPSTMSRPTLMPTCVAFCSILFLTLATGCQAFPKVERRETAQEYAEKEQSQKMNTDDQENISFAPK YMLQQMSSEAPMVLSEGPSEIPLIKVFSVNKESHLPAGLLHPTSPGVYSSSEPVVSAEQEPGSSLERMSSEHLSKVMILT VAVSSPASLNPQEGPYNSLSTQPIVAAVTDVTHGSLDYLDNQLFAAKSQEAVSLGNSPSSSINTKEPEIHKADAAMGTTVV PGVDSTGDMEDRERPESEMAADDGQSTTTKYLVTIPNFLTTEPTAGSILGDAKVTVSVSTAGPVSSIFNEEWDTKFESIS RGRPPEPGDNAETQMRTKPPHGYESEGTEESPSSTAVLKVAPGHLGGEPALGTALVTALGDERSPVLTHQISFTPMSLAE DPEVSTMKLFPSAGGFRASTQGDRTQLSSETAFSTSQYESVPQQEAGNVLKDITQERKMATQAMNTTSPVVTQEHMATIE VPRGSGEPEEGMPSLSPVPAEVADAELSRRESLATPASTTVVPLSLKLTSSMEDLMDTITGPSEEFIPVLGSPMAPPAMTVE APTISSALPSEGRTPSISRPNATAASYGLEQLESEEEVEDDEDEEEDDEEEEEDEEEDDEEDKETSLSLYKDFDGDTEPPGFT LPGITSQEPDIRSGSMDLLEVATYQVPETIEWEQNQGLVRSWMEKLDKAGYMSGMLVPVGVGIAGALFILGALYSIKV MNRRRRNGFKRHKRQREFNSMQDRVMLLADSSSEDEF

	Consensus	Start	End
Consensus 1	DEED	616	619
Consensus 2	DEED	619	622
Consensus 3	DEED	629	632
Consensus 4	DEED	632	635
Consensus 5	DEED	635	638

Combination			
Consensus 1			
Fragment 1	Start	End	
MLQDSITGIVNSFNLFFPST.....SYGLEQLESEEEVEDDEDEED	1	619	
Fragment 2	Start	End	
EEDDEEEEEDEEEDK.....EFNSMQDRVMLLADSSSEDEF	620	775	

Combination			
Consensus 2			
Fragment 1	Start	End	
MLQDSITGIVNSFNLFFPST.....LEQLESEEEVEDDEEED	1	622	
Fragment 2	Start	End	
EEEEDEEEDKETS.....EFNSMQDRVMLLADSSSEDEF	623	775	

Combination			
Consensus 3			
Fragment 1	Start	End	
MLQDSITGIVNSFNLFFPST.....DDEDEEEDDEEED	1	632	
Fragment 2	Start	End	
EEDKETSLSLYKDFDGDTEPP.....EFNSMQDRVMLLADSSSEDEF	633	775	

Combination			
Consensus 4			
Fragment 1	Start	End	
MLQDSITGIVNSFNLFFPST.....DEEEDDEEED	1	635	
Fragment 2	Start	End	
EEDKETSLSLYKDFDGDTEPP.....EFNSMQDRVMLLADSSSEDEF	636	775	

Combination			
Consensus 5			
Fragment 1	Start	End	
MLQDSITGIVNSFNLFFPST.....DEEEDDEEED	1	638	
Fragment 2	Start	End	
KETSLSLYKDFDGDTEPPGFT.....EFNSMQDRVMLLADSSSEDEF	639	775	

Combination		
Consensus 1 and Consensus 2		
Fragment 1	Start	End
MLQDSITGIVNSFNLFFPST.....SYGLEQLESEEEVEDDEDEED	1	619
Fragment 2	Start	End
EED	620	622
Fragment 3	Start	End
EEEEEEDEEDEDKEDT.....EFNSMQDRVMLLADSSSEDEF	623	775

Combination		
Consensus 1 and Consensus 3		
Fragment 1	Start	End
MLQDSITGIVNSFNLFFPST.....SYGLEQLESEEEVEDDEDEED	1	619
Fragment 2	Start	End
EEEEEEDEEDED	620	632
Fragment 3	Start	End
EEDEEDKETDSLYKDFDGDTEPP.....EFNSMQDRVMLLADSSSEDEF	633	775

Combination		
Consensus 1 and Consensus 4		
Fragment 1	Start	End
MLQDSITGIVNSFNLFFPST.....SYGLEQLESEEEVEDDEDEED	1	619
Fragment 2	Start	End
EEEEEEDEEDED	620	635
Fragment 3	Start	End
EEDKETDSLYKDFDGDTEPP.....EFNSMQDRVMLLADSSSEDEF	636	775

Combination		
Consensus 1 and Consensus 5		
Fragment 1	Start	End
MLQDSITGIVNSFNLFFPST.....SYGLEQLESEEEVEDDEDEED	1	619
Fragment 2	Start	End
EEEEEEDEEDED	620	638
Fragment 3	Start	End
KETDSLYKDFDGDTEPPGFT.....EFNSMQDRVMLLADSSSEDEF	639	775

Combination		
Consensus 2 and Consensus 3		
Fragment 1	Start	End
MLQDSITGIVNSFNLFFPST.....LEQLESEEEVEDDEDEEDED	1	622
Fragment 2	Start	End
EEEEEEDEEDED	623	632
Fragment 3	Start	End
EEDEEDKETDSLYKDFDGDTEPP.....EFNSMQDRVMLLADSSSEDEF	633	775

Combination		
Consensus 2 and Consensus 4		
Fragment 1	Start	End
MLQDSITGIVNSFNLFFPST.....LEQLESEEEVEDDEDEEDED	1	622
Fragment 2	Start	End
EEEEEEDEEDED	623	635
Fragment 3	Start	End
EEDKETDSLYKDFDGDTEPP.....EFNSMQDRVMLLADSSSEDEF	636	775

Combination		
Consensus 2 and Consensus 5		
Fragment 1	Start	End
MLQDSITGIVNSFNLFFPST.....LEQLESEEEVEDDEDEEDED	1	622
Fragment 2	Start	End
EEEEEEDEEDED	623	638
Fragment 3	Start	End
KETDSLYKDFDGDTEPPGFT.....EFNSMQDRVMLLADSSSEDEF	639	775

Combination		
Consensus 3 and Consensus 4		
Fragment 1	Start	End
MLQDSITGIVNSFNLFFPST.....DDEDEEDEDEEEEEDEED	1	632
Fragment 2	Start	End
EED	633	635
Fragment 3	Start	End
EEDKETDSLYKDFDGDTEPP.....EFNSMQDRVMLLADSSSEDEF	636	775

Combination		
Consensus 3 and Consensus 5		
Fragment 1	Start	End
MLQDSITGIVNSFNLFFPST.....DDEDEEDEDEEEEEDEED	1	632
Fragment 2	Start	End
EEDEED	633	638
Fragment 3	Start	End
KETDSLYKDFDGDTEPPGFT.....EFNSMQDRVMLLADSSSEDEF	639	775

Combination		
Consensus 4 and Consensus 5		
Fragment 1	Start	End
MLQDSITGIVNSFNLFFPST.....DDEDEEDEDEEEEEDEED	1	635
Fragment 2	Start	End
EED	636	638
Fragment 3	Start	End
KETDSLYKDFDGDTEPPGFT.....EFNSMQDRVMLLADSSSEDEF	639	775

Combination		
Consensus 1 and Consensus 2 and Consensus 3		
Fragment 1	Start	End
MLQDSITGIVNSFNLFFPST.....SYGLEQLESEEEVEDDEDEED	1	619
Fragment 2	Start	End
EED	620	622
Fragment 3	Start	End
EEEEDEDEED	623	632
Fragment 4	Start	End
EEDEEDKETDSLYKDFDGDTEPP.....EFNSMQDRVMLLADSSSEDEF	633	775

Combination		
Consensus 1 and Consensus 2 and Consensus 4		
Fragment 1	Start	End
MLQDSITGIVNSFNLFFPST.....SYGLEQLESEEEVEDDEDEED	1	619
Fragment 2	Start	End
EED	620	622
Fragment 3	Start	End
EEEEDEDEED	623	635
Fragment 4	Start	End
EEDKETDSLYKDFDGDTEPP.....EFNSMQDRVMLLADSSSEDEF	636	775

Combination		
Consensus 1 and Consensus 2 and Consensus 5		
Fragment 1	Start	End
MLQDSITGIVNSFNLFFPST.....SYGLEQLESEEEVEDDEDEED	1	619
Fragment 2	Start	End
EED	620	622
Fragment 3	Start	End
EEEEDEDEDEED	623	638
Fragment 4	Start	End
KETDSLYKDFDGDTEPPGFT.....EFNSMQDRVMLLADSSSEDEF	639	775

Combination		
Consensus 1 and Consensus 3 and Consensus 4		
Fragment 1	Start	End
MLQDSITGIVNSFNLFFPST.....SYGLEQLESEEEVEDDEDEED	1	619
Fragment 2	Start	End
EEDEEEEEDEED	620	632
Fragment 3	Start	End
EED	633	635
Fragment 4	Start	End
EEDKETDSLYKDFDGDTEPP.....EFNSMQDRVMLLADSSSEDEF	636	775

Combination		
Consensus 1 and Consensus 3 and Consensus 5		
Fragment 1	Start	End
MLQDSITGIVNSFNLFFPST.....SYGLEQLESEEEVEDDEDEED	1	619
Fragment 2	Start	End
EEDEEEEEDEED	620	632
Fragment 3	Start	End
EEDEED	633	638
Fragment 4	Start	End
KETDSLYKDFDGDTEPPGFT.....EFNSMQDRVMLLADSSSEDEF	639	775

Combination		
Consensus 1 and Consensus 4 and Consensus 5		
Fragment 1	Start	End
MLQDSITGIVNSFNLFFPST.....SYGLEQLESEEEVEDDEDEED	1	619
Fragment 2	Start	End
EEDEEEEEDEEDEED	620	635
Fragment 3	Start	End
EED	636	638
Fragment 4	Start	End
KETDSLYKDFDGDTEPPGFT.....EFNSMQDRVMLLADSSSEDEF	639	775

Combination		
Consensus 2 and Consensus 3 and Consensus 4		
Fragment 1	Start	End
MLQDSITGIVNSFNLFFPST.....LEQLESEEEVEDDEDEEDEED	1	622
Fragment 2	Start	End
EEEEEEDEED	623	632
Fragment 3	Start	End
EED	633	635
Fragment 4	Start	End
EEDKETDSLYKDFDGDTEPP.....EFNSMQDRVMLLADSSSEDEF	636	775

Combination		
Consensus 2 and Consensus 3 and Consensus 5		
Fragment 1	Start	End
MLQDSITGIVNSFNLFFPST.....LEQLESEEEVEDDEDEEDEED	1	622
Fragment 2	Start	End
EEEEEEDEED	623	632
Fragment 3	Start	End
EEDEED	633	638
Fragment 4	Start	End
KETDSLYKDFDGDTEPPGFT.....EFNSMQDRVMLLADSSSEDEF	639	775

Combination		
Consensus 2 and Consensus 4 and Consensus 5		
Fragment 1	Start	End
MLQDSITGIVNSFNLFFPST.....LEQLESEEVEDDEDEDEED	1	622
Fragment 2	Start	End
EEEEEEDEDEED	623	635
Fragment 3	Start	End
EED	636	638
Fragment 4	Start	End
KETDSLYKDFDGDTEPPGFT.....EFNSMQDRVMLLADSSSEDEF	639	775

Combination		
Consensus 3 and Consensus 4 and Consensus 5		
Fragment 1	Start	End
MLQDSITGIVNSFNLFFPST.....DDEDEDEDEDEEEEEDEED	1	632
Fragment 2	Start	End
EED	633	635
Fragment 3	Start	End
EED	636	638
Fragment 4	Start	End
KETDSLYKDFDGDTEPPGFT.....EFNSMQDRVMLLADSSSEDEF	639	775

Combination		
Consensus 1 and Consensus 2 and Consensus 3 and Consensus 4		
Fragment 1	Start	End
MLQDSITGIVNSFNLFFPST.....SYGLEQLESEEVEDDEDEED	1	619
Fragment 2	Start	End
EED	620	622
Fragment 3	Start	End
EEEEEEDEED	623	632
Fragment 4	Start	End
EED	633	635
Fragment 5	Start	End
EEDKETDSLYKDFDGDTEPPGFT.....EFNSMQDRVMLLADSSSEDEF	636	775

Combination		
Consensus 1 and Consensus 2 and Consensus 3 and Consensus 5		
Fragment 1	Start	End
MLQDSITGIVNSFNLFFPST.....SYGLEQLESEEVEDDEDEED	1	619
Fragment 2	Start	End
EED	620	622
Fragment 3	Start	End
EEEEEEDEED	623	632
Fragment 4	Start	End
EEDEED	633	638
Fragment 5	Start	End
KETDSLYKDFDGDTEPPGFT.....EFNSMQDRVMLLADSSSEDEF	639	775

Combination		
Consensus 1 and Consensus 2 and Consensus 4 and Consensus 5		
Fragment 1	Start	End
MLQDSITGIVNSFNLFFPST.....SYGLEQLESEEVEDDEDEED	1	619
Fragment 2	Start	End
EED	620	622
Fragment 3	Start	End
EEEEEEDEDEED	623	635
Fragment 4	Start	End
EED	636	638
Fragment 5	Start	End
KETDSLYKDFDGDTEPPGFT.....EFNSMQDRVMLLADSSSEDEF	639	775

Combination		
Consensus 1 and Consensus 3 and Consensus 4 and Consensus 5		
Fragment 1	Start	End
MLQDSITGIVNSFNLFFPST.....SYGLEQLESEEEVEDEDEED	1	619
Fragment 2	Start	End
EEDEEEEEDEED	620	632
Fragment 3	Start	End
EED	633	635
Fragment 4	Start	End
EED	636	638
Fragment 5	Start	End
KETDSLYKDFDGDTEPPGFT.....EFNSMQDRVMLLADSSSEDEF	639	775

Combination		
Consensus 2 and Consensus 3 and Consensus 4 and Consensus 5		
Fragment 1	Start	End
MLQDSITGIVNSFNLFFPST.....LEQLESEEEVEDEDEDEED	1	622
Fragment 2	Start	End
EEEEEEDEED	623	632
Fragment 3	Start	End
EED	633	635
Fragment 4	Start	End
EED	636	638
Fragment 5	Start	End
KETDSLYKDFDGDTEPPGFT.....EFNSMQDRVMLLADSSSEDEF	639	775

Combination		
Consensus 1 and Consensus 2 and Consensus 3 and Consensus 4 and Consensus 5		
Fragment 1	Start	End
MLQDSITGIVNSFNLFFPST.....SYGLEQLESEEEVEDEDEED	1	619
Fragment 2	Start	End
EEDEEEEEDEED	620	632
Fragment 3	Start	End
EEEEEEDEED	633	642
Fragment 4	Start	End
EED	643	645
Fragment 5	Start	End
EED	646	648
Fragment 6	Start	End
KETDSLYKDFDGDTEPPGFT.....EFNSMQDRVMLLADSSSEDEF	649	775

3. β II-spectrin Cleavage by Calpain-2

		AA Sequence	Start	End			AA Sequence	Start	End
1	Consensus	TTVA	5	6	56	Consensus	NAVV	1228	1229
	Fragment	MTTTVA	1	6		Fragment	MTTT.... NAVV	1	1229
	Fragment	TDYD.... GKKK	7	2364		Fragment	ETGR.... GKKK	1230	2364
2	Consensus	SDVN	21	22	57	Consensus	AVVE	1229	1230
	Fragment	MTTT.... SDVN	1	22		Fragment	MTTT.... AVVE	1	1230
	Fragment	NRWD.... GKKK	23	2364		Fragment	TGRR.... GKKK	1231	2364
3	Consensus	WDVD	27	28	58	Consensus	RLVS	1236	1237
	Fragment	MTTT.... WDVD	1	28		Fragment	MTTT.... RLVS	1	1237
	Fragment	DWDN.... GKKK	29	2364		Fragment	DGNI.... GKKK	1238	2364
4	Consensus	EAVQ	55	56	59	Consensus	EKVD	1250	1251
	Fragment	MTTT.... EAVQ	1	56		Fragment	MTTT.... EKVD	1	1251
	Fragment	KKTF.... GKKK	57	2364		Fragment	SIDD.... GKKK	1252	2364
5	Consensus	KWVN	64	65	60	Consensus	EAVV	1350	1351
	Fragment	MTTT.... KWVN	1	65		Fragment	MTTT.... EAVV	1	1351
	Fragment	SHLA.... GKKK	66	2364		Fragment	KEKL.... GKKK	1352	2364
6	Consensus	ARVS	71	72	61	Consensus	AVKK	1351	1352
	Fragment	MTTT.... ARVS	1	72		Fragment	MTTT.... AVVK	1	1352
	Fragment	CRIT.... GKKK	73	2364		Fragment	EKLT.... GKKK	1353	2364
7	Consensus	LEVL	94	95	62	Consensus	WEVL	1364	1365
	Fragment	MTTT.... LEVL	1	95		Fragment	MTTT.... WEVL	1	1365
	Fragment	SGER.... GKKK	96	2364		Fragment	ESTT.... GKKK	1366	2364
8	Consensus	ENVD	116	117	63	Consensus	TSVN	1415	1416
	Fragment	MTTT.... ENVD	1	117		Fragment	MTTT.... TSVN	1	1416
	Fragment	KALQ.... GKKK	118	2364		Fragment	ILLK.... GKKK	1417	2364
9	Consensus	QRVH	128	129	64	Consensus	MEVR	1431	1432
	Fragment	MTTT.... QRVH	1	129		Fragment	MTTT.... MEVR	1	1432
	Fragment	LENM.... GKKK	130	2364		Fragment	KKEI.... GKKK	1433	2364
10	Consensus	DIVD	139	140	65	Consensus	DEVV	1456	1457
	Fragment	MTTT.... DIVD	1	140		Fragment	MTTT.... DEVD	1	1457
	Fragment	GNHR.... GKKK	141	2364		Fragment	SKRL.... GKKK	1458	2364
11	Consensus	ISVE	164	165	66	Consensus	LTVQ	1463	1464
	Fragment	MTTT.... ISVE	1	165		Fragment	MTTT.... LTVQ	1	1464
	Fragment	TEDN.... GKKK	166	2364		Fragment	TKFM.... GKKK	1465	2364
12	Consensus	PNVN	193	194	67	Consensus	RDVE	1494	1495
	Fragment	MTTT.... PNVN	1	194		Fragment	MTTT.... RDVE	1	1495
	Fragment	IHNF.... GKKK	195	2364		Fragment	DEIL.... GKKK	1496	2364
13	Consensus	ISVD	258	259	68	Consensus	LWVG	1501	1502
	Fragment	MTTT.... ISVD	1	259		Fragment	MTTT.... LWVG	1	1502
	Fragment	HPDE.... GKKK	260	2364		Fragment	ERMP.... GKKK	1503	2364
14	Consensus	TYVV	270	271	69	Consensus	QTVQ	1520	1521
	Fragment	MTTT.... TYVV	1	271		Fragment	MTTT.... QTVQ	1	1521
	Fragment	TYYH.... GKKK	272	2364		Fragment	LLIK.... GKKK	1522	2364
15	Consensus	YVVT	271	272	70	Consensus	NIVT	1552	1553
	Fragment	MTTT.... YVVT	1	272		Fragment	MTTT.... NIVT	1	1553
	Fragment	YYHY.... GKKK	273	2364		Fragment	DSSS.... GKKK	1554	2364
16	Consensus	LAVE	285	286	71	Consensus	SAVS	1628	1629
	Fragment	MTTT.... LAVE	1	286		Fragment	MTTT.... SAVS	1	1629
	Fragment	GKRL.... GKKK	287	2364		Fragment	MLKK.... GKKK	1630	2364
17	Consensus	GKVL	293	294	72	Consensus	QAVE	1641	1642
	Fragment	MTTT.... GKVL	1	294		Fragment	MTTT.... QAVE	1	1642
	Fragment	DNAL.... GKKK	295	2364		Fragment	DYAE.... GKKK	1643	2364
18	Consensus	SLVG	335	336	73	Consensus	ETVH	1648	1649
	Fragment	MTTT.... SLVG	1	336		Fragment	MTTT.... ETVH	1	1649
	Fragment	VQQQ.... GKKK	337	2364		Fragment	QLSK.... GKKK	1650	2364
19	Consensus	VGVS	337	338	74	Consensus	ALVA	1659	1660
	Fragment	MTTT.... VGVS	1	338		Fragment	MTTT.... ALVA	1	1660
	Fragment	QQLQ.... GKKK	339	2364		Fragment	DSHP.... GKKK	1661	2364
20	Consensus	RTVE	350	351	75	Consensus	SKVD	1676	1677
	Fragment	MTTT.... RTVE	1	351		Fragment	MTTT.... SKVD	1	1677
	Fragment	KPPK.... GKKK	352	2364		Fragment	KLYA.... GKKK	1678	2364

21	Consensus	LEVL	364	365	76	Consensus	REVD	1707	1708
	Fragment	MTTT.... LEVL	1	365		Fragment	MTTT.... REVD	1	1708
	Fragment	LFTL.... GKKK	366	2364		Fragment	DLEQ.... GKKK	1709	2364
22	Consensus	QKVY	380	381	77	Consensus	REVV	1719	1720
	Fragment	MTTT.... QKVY	1	381		Fragment	MTTT.... REVV	1	1720
	Fragment	MPRE.... GKKK	382	2364		Fragment	AGSH.... GKKK	1721	2364
23	Consensus	LRVS	445	446	78	Consensus	EVVA	1720	1721
	Fragment	MTTT.... LRVS	1	446		Fragment	MTTT....EVVA	1	1721
	Fragment	QDNF.... GKKK	447	2364		Fragment	GSHE....GKKK	1722	2364
24	Consensus	PAVE	457	458	79	Consensus	EHVT	1733	1734
	Fragment	MTTT.... PAVE	1	458		Fragment	MTTT.... EHVT	1	1734
	Fragment	AATK.... GKKK	459	2364		Fragment	MLQE.... GKKK	1735	2364
25	Consensus	ERVQ	478	479	80	Consensus	ERVD	1755	1756
	Fragment	MTTT.... ERVQ	1	479		Fragment	MTTT.... ERVD	1	1756
	Fragment	AVVA.... GKKK	480	2364		Fragment	TVNH.... GKKK	1757	2364
26	Consensus	QAVV	481	482	81	Consensus	DTVN	1758	1759
	Fragment	MTTT.... QAVV	1	482		Fragment	MTTT.... DTVN	1	1759
	Fragment	AVAR.... GKKK	483	2364		Fragment	HLAD.... GKKK	1760	2364
27	Consensus	AVVA	482	483	82	Consensus	NTVE	1839	1840
	Fragment	MTTT...AVVA	1	483		Fragment	MTTT.... NTVE	1	1840
	Fragment	VARE.... GKKK	484	2364		Fragment	TLQR.... GKKK	1841	2364
28	Consensus	VAVA	484	485	83	Consensus	TQVR	1860	1861
	Fragment	MTTT.... VAVA	1	485		Fragment	MTTT.... TQVR	1	1861
	Fragment	RELE.... GKKK	486	2364		Fragment	QLQE.... GKKK	1862	2364
29	Consensus	DNVI	506	507	84	Consensus	NEVL	1889	1890
	Fragment	MTTT.... DNVI	1	507		Fragment	MTTT.... NEVL	1	1890
	Fragment	RLWE.... GKKK	508	2364		Fragment	EAWK.... GKKK	1891	2364
30	Consensus	MKVL	549	550	85	Consensus	RRVR	1905	1906
	Fragment	MTTT.... MKVL	1	550		Fragment	MTTT.... RRVR	1	1906
	Fragment	VLSQ.... GKKK	551	2364		Fragment	LVDT.... GKKK	1907	2364
31	Consensus	VLVL	551	552	86	Consensus	RLVD	1908	1909
	Fragment	MTTT.... VLVL	1	552		Fragment	MTTT.... RLVD	1	1909
	Fragment	SQDY.... GKKK	553	2364		Fragment	TGDK.... GKKK	1910	2364
32	Consensus	LGVE	563	564	87	Consensus	SMVR	1920	1921
	Fragment	MTTT.... LGVE	1	564		Fragment	MTTT.... SMVR	1	1921
	Fragment	DLLQ.... GKKK	565	2364		Fragment	DLML.... GKKK	1922	2364
33	Consensus	TLVE	573	574	88	Consensus	EDVI	1930	1931
	Fragment	MTTT.... TLVE	1	574		Fragment	MTTT.... EDVI	1	1931
	Fragment	ADIG.... GKKK	575	2364		Fragment	RQIE.... GKKK	1932	2364
34	Consensus	ERVR	584	585	89	Consensus	RDVS	1943	1944
	Fragment	MTTT.... ERVR	1	585		Fragment	MTTT.... RDVS	1	1944
	Fragment	GVNA.... GKKK	586	2364		Fragment	SVEL.... GKKK	1945	2364
35	Consensus	RGVN	587	588	90	Consensus	SSVE	1946	1947
	Fragment	MTTT.... RGVN	1	588		Fragment	MTTT.... SSVE	1	1947
	Fragment	ASAQ.... GKKK	589	2364		Fragment	LLMN.... GKKK	1948	2364
36	Consensus	PQVI	608	609	91	Consensus	LEVH	2019	2020
	Fragment	MTTT.... PQVI	1	609		Fragment	MTTT.... LEVH	1	2020
	Fragment	RDRV.... GKKK	610	2364		Fragment	QFSR.... GKKK	2021	2364
37	Consensus	DRVA	613	614	92	Consensus	ASVA	2028	2029
	Fragment	MTTT.... DRVA	1	614		Fragment	MTTT....ASVA	1	2029
	Fragment	HMEF.... GKKK	615	2364		Fragment	EAWL....GKKK	2030	2364
38	Consensus	TSVM	673	674	93	Consensus	QSVD	2049	2050
	Fragment	MTTT.... TSVM	1	674		Fragment	MTTT.... QSVD	1	2050
	Fragment	RLLS.... GKKK	675	2364		Fragment	EVEK.... GKKK	2051	2364
39	Consensus	KIVS	767	768	94	Consensus	DEVE	2052	2053
	Fragment	MTTT.... KIVS	1	768		Fragment	MTTT.... DEVE	1	2053
	Fragment	SSDV.... GKKK	769	2364		Fragment	KLIK.... GKKK	2054	2364
40	Consensus	SDVG	772	773	95	Consensus	LEVR	2087	2088
	Fragment	MTTT.... SDVG	1	773		Fragment	MTTT.... LEVR	1	2088
	Fragment	HDEY.... GKKK	774	2364		Fragment	RQQE.... GKKK	2089	2364
41	Consensus	SLVK	783	784	96	Consensus	TKVS	2109	2110
	Fragment	MTTT.... SLVK	1	784		Fragment	MTTT....TKVS	1	2110
	Fragment	KHKD.... GKKK	785	2364		Fragment	EEAE.... GKKK	2111	2364
42	Consensus	KDVA	789	790	97	Consensus	EQVS	2127	2128
	Fragment	MTTT.... KDVA	1	790		Fragment	MTTT.... EQVS	1	2128
	Fragment	EEIA.... GKKK	791	2364		Fragment	QNGI.... GKKK	2129	2364

43	Consensus	PDVR	820	821	98	Consensus	ETVD	2145	2146
	Fragment	MTTT.... PDVR	1	821		Fragment	MTTT...ETVD	1	2146
	Fragment	GRLS.... GKKK	822	2364		Fragment	TSEM.... GKKK	2147	2364
44	Consensus	KEVA	834	835	99	Consensus	EMVN	2151	2152
	Fragment	MTTT.... KEVA	1	835		Fragment	MTTT.... EMVN	1	2152
	Fragment	ELTR.... GKKK	836	2364		Fragment	GATE.... GKKK	2153	2364
45	Consensus	LEVI	886	887	100	Consensus	HNVY	2225	2226
	Fragment	MTTT.... LEVI	1	887		Fragment	MTTT.... HNVY	1	2226
	Fragment	QHRF.... GKKK	888	2364		Fragment	CVIN.... GKKK	2227	2364
46	Consensus	SRVA	905	906	101	Consensus	YCVI	2228	2229
	Fragment	MTTT....SRVA	1	906		Fragment	MTTT.... YCVI	1	2229
	Fragment	VVNQ.... GKKK	907	2364		Fragment	NNQE.... GKKK	2230	2364
47	Consensus	VAVV	907	908	102	Consensus	SEVP	2253	2254
	Fragment	MTTT.... VAVV	1	908		Fragment	MTTT.... SEVP	1	2254
	Fragment	NQIA.... GKKK	909	2364		Fragment	VSLK.... GKKK	2255	2364
48	Consensus	AVVN	908	909	103	Consensus	VPVS	2255	2256
	Fragment	MTTT.... AVVN	1	909		Fragment	MTTT.... VPVS	1	2256
	Fragment	QIAR.... GKKK	910	2364		Fragment	LKEA.... GKKK	2257	2364
49	Consensus	ELVD	944	945	104	Consensus	EAVC	2261	2262
	Fragment	MTTT.... ELVD	1	945		Fragment	MTTT.... EAVC	1	2262
	Fragment	RKKD.... GKKK	946	2364		Fragment	EVAL.... GKKK	2263	2364
50	Consensus	TKVI	977	978	105	Consensus	CEVA	2264	2265
	Fragment	MTTT.... TKVI	1	978		Fragment	MTTT....CEVA	1	2265
	Fragment	ESTQ.... GKKK	979	2364		Fragment	LDYK....GKKK	2266	2364
51	Consensus	AGVM	991	992	106	Consensus	KHVF	2274	2275
	Fragment	MTTT.... AGVM	1	992		Fragment	MTTT.... KHVF	1	2275
	Fragment	ALQR.... GKKK	993	2364		Fragment	KLRL.... GKKK	2276	2364
52	Consensus	DLVA	1006	1007	107	Consensus	HEVS	2313	2314
	Fragment	MTTT....DLVA	1	1007		Fragment	MTTT.... HEVS	1	2314
	Fragment	IEAK....GKKK	1008	2364		Fragment	ASTQ.... GKKK	2315	2364
53	Consensus	SDVW	1043	1044	108	Consensus	TSVV	2333	2334
	Fragment	MTTT.... SDVW	1	1044		Fragment	MTTT.... TSVV	1	2334
	Fragment	EEMK.... GKKK	1045	2364		Fragment	TITS.... GKKK	2335	2364
54	Consensus	EMVT	1128	1129	109	Consensus	SVVT	2334	2335
	Fragment	MTTT.... EMVT	1	1129		Fragment	MTTT.... SVVT	1	2335
	Fragment	QGQT.... GKKK	1130	2364		Fragment	ITSE.... GKKK	2336	2364
55	Consensus	EYVL	1191	1192					
	Fragment	MTTT.... EYVL	1	1192					
	Fragment	AHTE.... GKKK	1193	2364					