

University of Missouri, St. Louis
IRL @ UMSL

Theses

Graduate Works

10-5-2017

2-D Electrophoresis Modeling of Multienzyme Cutting of Polypeptides

Howard Mayes
hgmb76@mail.umsl.edu

Follow this and additional works at: <https://irl.umsl.edu/thesis>

 Part of the [Biochemistry Commons](#), and the [Chemistry Commons](#)

Recommended Citation

Mayes, Howard, "2-D Electrophoresis Modeling of Multienzyme Cutting of Polypeptides" (2017). *Theses*. 312.
<https://irl.umsl.edu/thesis/312>

This Thesis is brought to you for free and open access by the Graduate Works at IRL @ UMSL. It has been accepted for inclusion in Theses by an authorized administrator of IRL @ UMSL. For more information, please contact marvinh@umsl.edu.

2-D Electrophoresis Modeling of Multienzyme Cutting of Polypeptides

Howard Mayes

Student Education/Degrees

B.S. Chemistry, University of Missouri – St. Louis, 2015

M.S. Computer Science, University of Missouri – Rolla, 1986

B.S. Electrical Engineering, Southern Illinois University – Edwardsville, 1980

B.A. Psychology, Valparaiso University, 1972

A Thesis Submitted to The Graduate School at the University of Missouri-St. Louis
In partial fulfillment of the requirements for the degree
Master of Science in Chemistry

December
2017

Advisory Committee

Chung Wong, Ph.D.
Chairperson

Cynthia Dupureur, Ph.D.

Michael Nichols, Ph.D.

Table of Contents

Table of Figures.....	iii
Table of Tables	iii
Abstract.....	iv
Acknowledgement	iv
Chapter 1 -Introduction and Review.....	1
1.1 Introduction	1
1.2 Purpose	3
1.3 Review of Simulation and Modeling Programs.....	4
1.3.1 Simulation with GPMAW	4
1.3.2 Simulation with 2-D Electrophoresis for Multiple Proteins	7
1.3.3 Determination of Peptide Sequence Cutting for Multiple Enzymes.....	8
1.3.4 Determination of Isoelectric Point Using ExpASy	10
1.3.5 Determination of Isoelectric Point Using IPC.....	13
Chapter 2 – Developing and Testing of Program	14
2.1 Multi-enzyme 2-Dimensional Electrophoresis Program Description	14
2.1.1 Program Display and Operation.....	14
2.1.2 Program Logic	18
2.1.3 Match to Enzyme Logic	21
2.1.4 Enzyme Function in Program	23
2.1.5 Isoelectric Point Determination	27
2.2 Validation	29
Chapter 3 – Conclusion and Future Direction.....	32
References	34

Table of Figures

Figure 1: Electrophoresis with Standards and α -Synuclein Protein.....	1
Figure 2: 2-D Electrophoresis.....	2
Figure 3: GPMAW lite Input Screen	6
Figure 4: GPMAW lite Output Screen	6
Figure 5: 2-D Electrophoresis Program by Fisher et. al.....	7
Figure 6: Protein Information Window for Fisher Program.....	8
Figure 7: Tandem Mass Spectrometer Simulation from Fisher et. al.	8
Figure 8: PeptideCutter Input Webpage.....	9
Figure 9: PeptideCutter Output Webpage.....	10
Figure 10: ProtParam Tool Input Webpage	11
Figure 11: ProtParam Webpage Output	12
Figure 12: Protein Isoelectric Point Input Webpage.....	13
Figure 13: Protein Isoelectric Point Calculator Output.....	13
Figure 14: MEEP Program Display.....	14
Figure 15: MEEP Output Text File	15
Figure 16: Input Control Section of MEEP Display	17
Figure 17: Output Section of MEEP Display	18
Figure 18: Program Flowchart	19
Figure 19: Match to Enzyme Flowchart	22

Table of Tables

Table 1: Enzymes and Cutting Rules	25
Table 2: Trypsin Exceptions.....	27
Table 3: Test Cases for IEP and Molecular Weight	30

Abstract

2-Dimensional Electrophoresis is one of the tools in the identification of proteins by molecular weight and pH. The display of molecular weight allows the researcher to quickly identify whether a specific protein or peptide string is in the sample. The pH measurement allows even better resolution between different species in the sample. The MultiEnzyme ElectroPhoresis (MEEP) program tries to model that by providing a graph that displays separated protein strings by both molecular weight and pH. The ability to cleave the protein with 43 different enzyme variations allows the researcher to analyze appropriate enzymes to isolate a protein subsequence before the actual experiment or to compare the experimental data with the simulated electrophoresis. This thesis reviews protein cutting simulations that have been done in the past or are currently available. It then describes the MEEP program: how it appears to the user, how the user makes it operate, and how it is structured. The thesis provides validation information for the calculation of molecular weight and isoelectric point. The program will hopefully provide a useful addition for the researcher's work.

Acknowledgement

I want to thank Prof. Chung F. Wong for his help with this research project. He made this program possible by giving me small goals that were achievable to finally reach the goal of this program and this thesis. He took me from knowing nothing of C#, to learning Visual Studio, to learning about programming graphs, to writing a simple one-dimensional electrophoresis program with a single enzyme and a single amino acid cut point. From there we expanded the program to two-dimensions and then to multiple enzymes. He has asked for flowcharts and explanations on various topics at times that I would not have done otherwise. The smaller goals kept this project within reach.

Chapter 1 -Introduction and Review

1.1 Introduction

Electrophoresis is the movement of charged particles through a gel solution in an electric field.¹ One common type of electrophoresis is SDS-PAGE (sodium dodecyl sulfate polyacrylamide gel electrophoresis) where the gel is polyacrylamide and sodium dodecyl sulfate (SDS) is used to denature the proteins. When a voltage is applied to the gel the proteins separate based on molecular mass. Often you will see an electrophoresis plot in a research paper as that provides essential information as to whether a protein is in a solution. By displaying the relative molecular weight of a test sample versus a reference protein, the researchers can gain confidence that they have isolated a given protein in their experiment. Figure 1² shows electrophoresis used as part of a verification that acetylated α -Synuclein protein had been isolated in genetically modified *Escherichia coli* strains and from human erythrocytes.

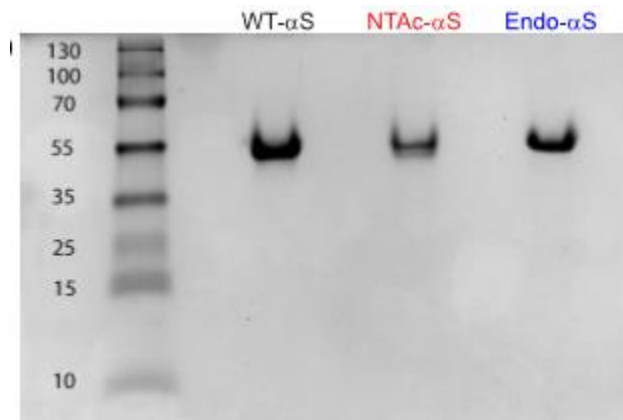


Figure 1: Electrophoresis with Standards and α -Synuclein Protein

2-Dimensional (2-D) Electrophoresis¹ provides the benefits of separating particles based not only on charge or molecular weight but also on a second criterion such as isoelectric point. In the

first step of a two-step process, a voltage is applied in a gel that has a pH gradient, the protein fragments will move to a point where the combined charge of their negative and positive charges is zero at that pH. That point is called the isoelectric point (IEP) and the technique is called isoelectric focusing (IEF). The second step is a voltage applied at right angle to the first to get the separation of the protein fragments by molecular weight.

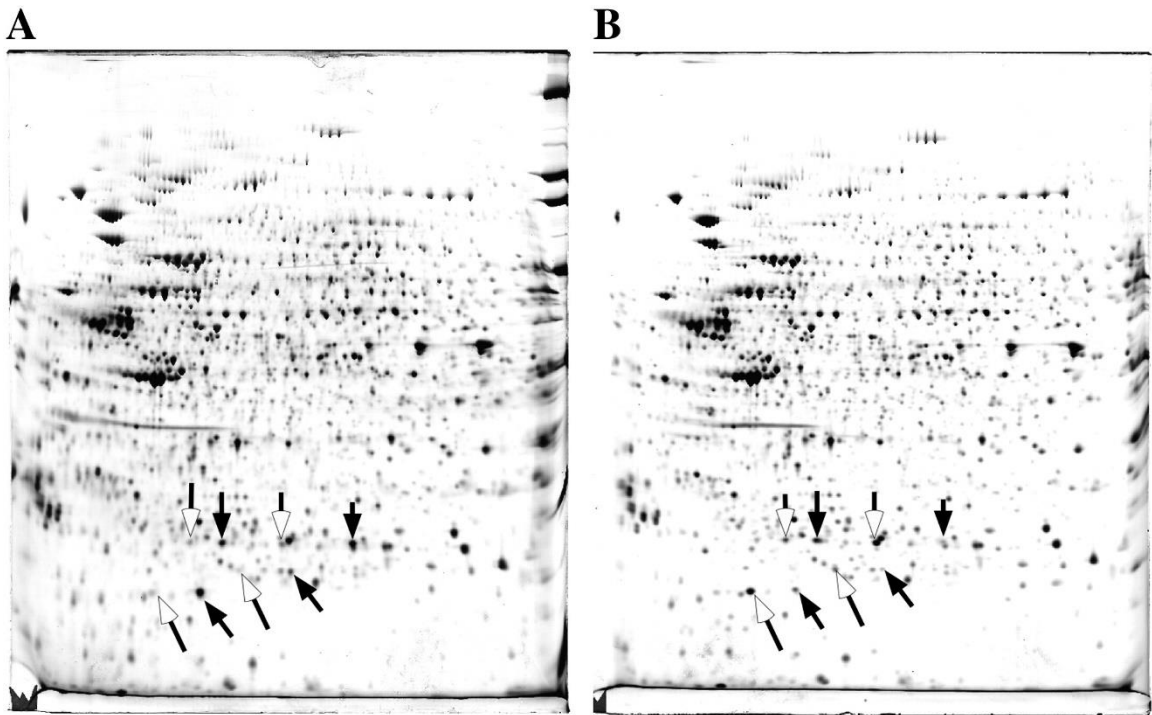


Figure 2: 2-D Electrophoresis

Figure 2³ shows 2-D electrophoresis gels where two samples have been compared and protein differences are noted by the arrows. By having two dimensions the separation of the proteins is easier to analyze although the sheer number of points makes for difficult comparisons.

This paper examines various models or simulations that have become available for the researcher who wants molecular weight or IEP data on a protein or set of proteins. Not only the whole protein but protein segments, when digested by enzymes, have been provided by these

programs and calculators. Most of the time the information is provided in a textual format. The cleavage by one protease out of a small set of proteases helps the researcher isolate the protein. In several cases an algorithm searches the input for matches between an amino acid in the input sequence and the enzyme requirement at the cut point.

This program attempts to provide a different “view” of things by giving the researcher a graphical display of a protein when cleaved or cut by not only one enzyme but by a series of enzymes. Each enzyme cutting process uses up to six amino acids around the cut site. The display provides a one-dimensional view of the molecular weights of the cut protein. To better isolate the protein subsequences, a two-dimensional view is also provided with molecular weight varied on one axis and isoelectric point varied on the other axis. When the researcher wants a more quantitative view of the cut sequences, the program provides a text file that records the molecular weight, the isoelectric point (IEP) for the subsequence, and the string of amino acids (AA) in the subsequence itself.

Since there are at least two ways to calculate the isoelectric point, the program provides two methods: one method scans the sequence and finds the pH point where the charge sum is zero, the other method uses the Henderson-Hasselbalch equation to step through values of pH that would let the equation approach a point where the positive and negative charges of the sequence balance out to zero.

1.2 Purpose

The purpose of this research is to provide a method to visualize the separation of peptide sub-sequences by both molecular weight and by isoelectric point (IEP) as with two-dimensional electrophoresis. An input peptide sequence is cleaved by multiple selectable enzymes to provide a unique display of that protein or peptide sequence when one or more enzymes are used.

One motivation for this work is to help researchers by providing a tool to find out if a protein is present in the 2-D electrophoresis gel. Another motivation is to provide a tool to help students visualize what to expect when a protein of interest is cleaved by a set of enzymes. Simulations are available that provide a textual display of amino acid sequences that have been cut by an enzyme and display the cuts as a list in the order of the sequence of the cuts. This simulation provides a visual display of the cuts in the order of the molecular weights of the subsequences, providing a graph that is similar to a 2-D electrophoresis gel. It also utilizes six positions around the cleavage point, instead of one position, which in turn provides a more realistic simulation. Although a simulation is not the same as running a 2-D electrophoresis gel, it can provide useful information to better identify what to expect when a 2-D electrophoresis gel is run. If a subsequence is present in the simulation and not in the actual electrophoresis, then it may be necessary to analyze the results for differences.

1.3 Review of Simulation and Modeling Programs

1.3.1 Simulation with GPMAW

In 2001, Perl, Steen, and Pandey developed a program⁴ called General Protein/Mass Analysis for Windows (GPMAW) that analyzed proteins and peptides. A company called Lighthouse data was formed and has continued to develop the program. The current version of the program is intended for mass spectrometric analysis of proteins and peptides. A smaller version of the program called GPMAW lite is provided by the company, Alphalyse. The program provides the molecular weight, extinction coefficient, isoelectric point, and hydrophobicity index, and numbers and percentages of amino acids. It also will cleave the protein with one of 6 proteases and display information on molecular weight, isoelectric point, and HPLC retention time.

Figure 3 shows the input screen for the web based program.⁵ The input can be entered as an accession number from NCBI or UniProt or it can be entered as a text string. The user can select one of the six proteases for cleavage. In the text box, the amino acid string “GAVLIMPWFSTNQYCKHRDE”, a sample string with each amino acid, is entered and the protease trypsin, which cleaves the amino acid string at Lysine (K) or Arginine (R), is selected. Once the Calculate button is pressed the program calculates the parameters and displays the output.

Figure 4 shows the output screen which has two display areas. The top area displays the input string in groups of ten. The check boxes allow the user to identify characteristics for the sequence such as those amino acids that are acidic or basic. The bottom box displays the subsequences from the cuts along with the molecular weight, HPLC retention time and the isoelectric point.

Figure 3: GPMaw lite Input Screen

#	from-to	MW	HPLC time	pI	Sequence
1	1-16	1,857.90	26.94	8.99	GAVLIMPWFSTNQYCK
2	17-18	312.18	2.57	10.55	HR
3	19-20	263.09	1.69	3.30	DE

Figure 4: GPMaw lite Output Screen

1.3.2 Simulation with 2-D Electrophoresis for Multiple Proteins

Fisher, Sekera, Payne, and Craig⁶ described the construction of a program in 2012 that simulates 2-D electrophoresis and tandem mass spectrometry for proteins. The program is meant to be used in biochemistry, proteomics, and bioinformatic education. The 2-D electrophoresis program simulates the display of the proteins in a mixture of proteins, locating the proteins by molecular weight and by isoelectric point, as shown in Figure 5. In the simulation, the user can watch the mixture of proteins move across the top of the screen as the simulated isoelectric focusing moves them to different positions based on pH. Once that part of the simulation has completed, the simulation of the SDS PAGE moves the protein dots down the gel based on molecular weight.

The user can then click on a single dot to bring up a second window, Figure 6, which will allow the user to do a Blast search, an NCBI search, a UniProt search, or to run the simulated mass spectrum. In this case the dot selected is the iron-sulfur protein of hydrogenase 3.

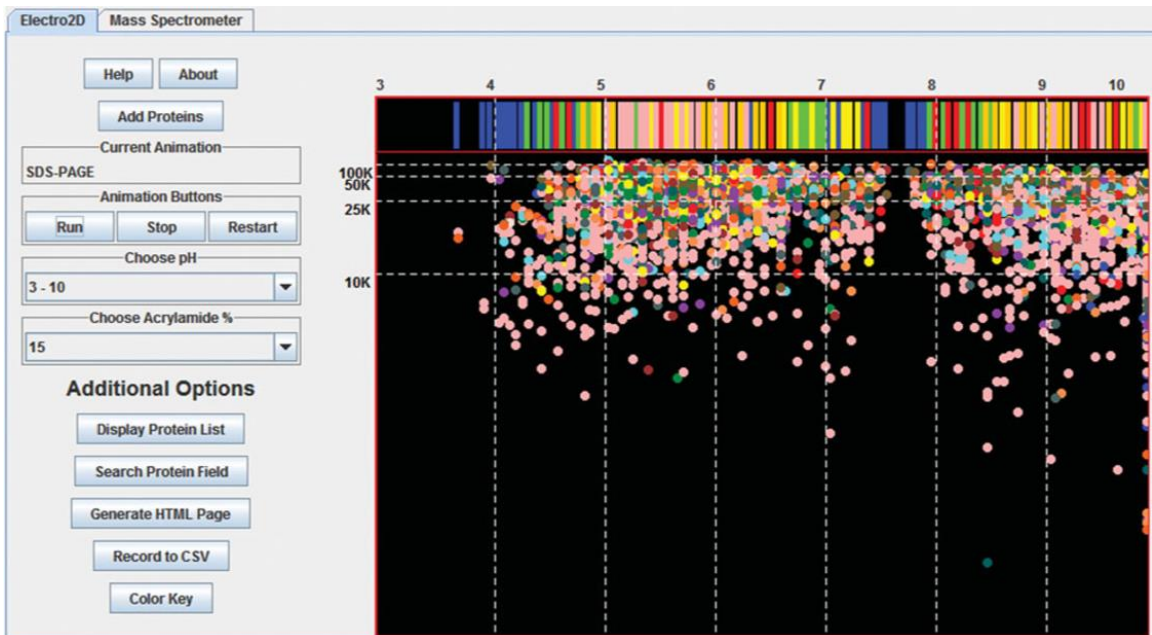


Figure 5: 2-D Electrophoresis Program by Fisher et. al.

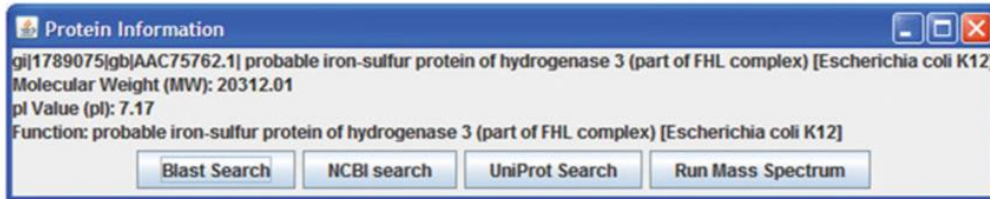


Figure 6: Protein Information Window for Fisher Program

By clicking on the Run Mass Spectrum button, the user then brings up the second program, Figure 7, that simulates a tandem mass spectrometer. The user can then select from four different proteases to observe the effect of those proteases on the mass spectrum.

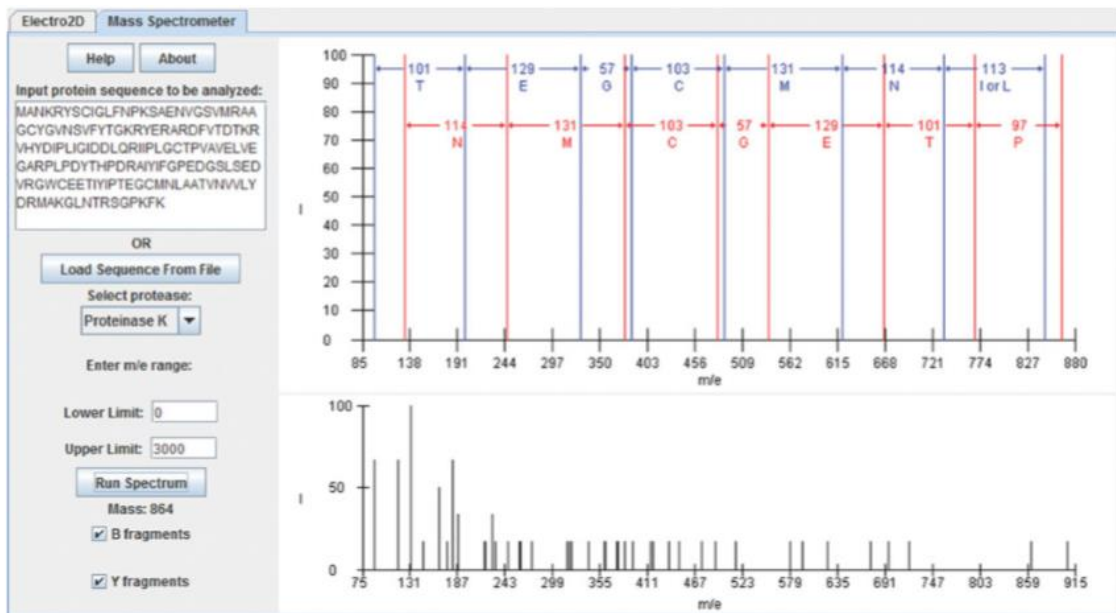


Figure 7: Tandem Mass Spectrometer Simulation from Fisher et. al.

1.3.3 Determination of Peptide Sequence Cutting for Multiple Enzymes

ExpASY is a Swiss Institute for Bioinformatics (SIB) website. The ExpASY title is an abbreviation for Expert Protein Analysis System.⁷ It provides databases and software in the areas of genomics, phylogeny, systems biology, evolution, population genetics, and transcriptomics.⁸ Of particular interest for this research is their work in proteomics and within that is the area of function analysis. They developed a program, PeptideCutter,⁹ which analyzes peptide sequences

cleaved with proteases or chemicals. The input to the program, Figure 8, is a peptide sequence or a FASTA protein file identified with a protein identifier or an accession number. In this case, the sequence “GAVLIMPWFSTNQYCKHRDE” was chosen to include each amino acid. The other main input is the set of proteases and chemicals used to cut the sequence into subsequences, in this example, trypsin.

The output, Figure 9, shows the sequence “GAVLIMPWFSTNQYCKHRDE” cut by trypsin at the Arginine (R) location. For each subsequence, the length of the subsequence is given along with the molecular weight.

PeptideCutter

PeptideCutter [references / documentation] predicts potential cleavage sites cleaved by proteases or chemicals in a given protein sequence. PeptideCutter returns the query sequence positions.

Enter a UniProtKB (Swiss-Prot or TrEMBL) protein identifier, ID (e.g. ALBU_HUMAN), or accession number, AC (e.g. P04406), or an amino acid sequence (e.g. 'SERVELAT'):

GAVLIMPWFSTNQYCKHRDE

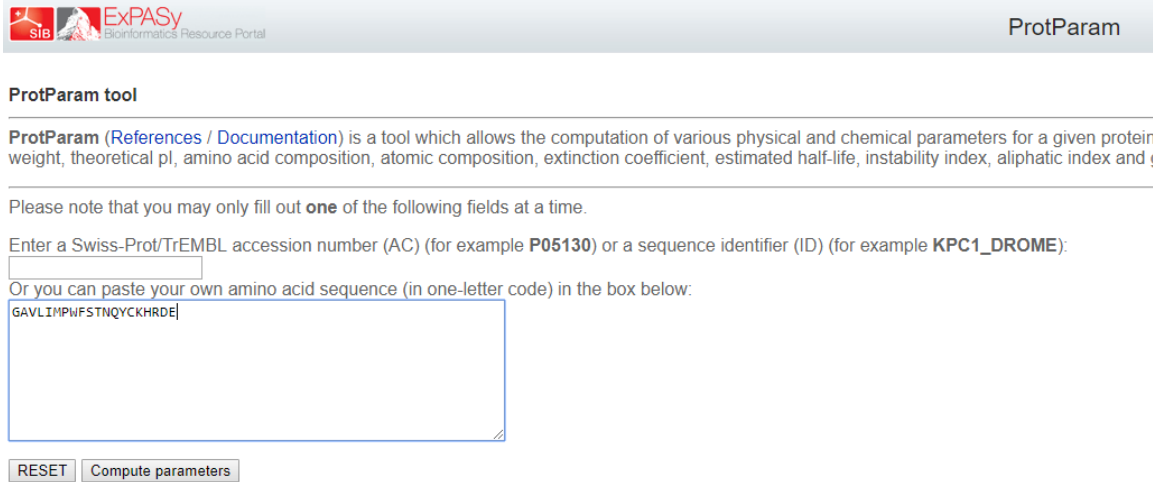
the cleavage of the protein. the fields.

Please, select

- all available enzymes and chemicals
- only the following selection of **enzymes and chemicals**

<input type="checkbox"/> Arg-C proteinase	<input type="checkbox"/> Asp-N endopeptidase	<input type="checkbox"/> Asp-N endopeptidase + N-terminal Glu
<input type="checkbox"/> BNPS-Skatole	<input type="checkbox"/> Caspase1	<input type="checkbox"/> Caspase2
<input type="checkbox"/> Caspase3	<input type="checkbox"/> Caspase4	<input type="checkbox"/> Caspase5
<input type="checkbox"/> Caspase6	<input type="checkbox"/> Caspase7	<input type="checkbox"/> Caspase8
<input type="checkbox"/> Caspase9	<input type="checkbox"/> Caspase10	
<input type="checkbox"/> Chymotrypsin-high specificity (C-term to [FYW], not before P)	<input type="checkbox"/> Chymotrypsin-low specificity (C-term to [FYWML], not before P)	
<input type="checkbox"/> Clostripain (Clostridiopeptidase B)	<input type="checkbox"/> CNBr	<input type="checkbox"/> Enterokinase
<input type="checkbox"/> Factor Xa	<input type="checkbox"/> Formic acid	<input type="checkbox"/> Glutamyl endopeptidase
<input type="checkbox"/> GranzymeB	<input type="checkbox"/> Hydroxylamine	<input type="checkbox"/> Iodosobenzoic acid
<input type="checkbox"/> LysC	<input type="checkbox"/> LysN	<input type="checkbox"/> NTCB (2-nitro-5-thiocyanobenzoic acid)
<input type="checkbox"/> Neutrophil elastase		
<input type="checkbox"/> Pepsin (pH1.3)	<input type="checkbox"/> Pepsin (pH>2)	<input type="checkbox"/> Proline-endopeptidase
<input type="checkbox"/> Proteinase K	<input type="checkbox"/> Staphylococcal peptidase I	<input type="checkbox"/> Tobacco etch virus protease
<input type="checkbox"/> Thermolysin	<input type="checkbox"/> Thrombin	<input checked="" type="checkbox"/> Trypsin

Figure 8: PeptideCutter Input Webpage



ProtParam tool

ProtParam ([References](#) / [Documentation](#)) is a tool which allows the computation of various physical and chemical parameters for a given protein weight, theoretical pI, amino acid composition, atomic composition, extinction coefficient, estimated half-life, instability index, aliphatic index and c

Please note that you may only fill out **one** of the following fields at a time.

Enter a Swiss-Prot/TrEMBL accession number (AC) (for example **P05130**) or a sequence identifier (ID) (for example **KPC1_DROME**):

Or you can paste your own amino acid sequence (in one-letter code) in the box below:

```
GAVLIMPWFSTNQYCKHRDE
```

Figure 10: ProtParam Tool Input Webpage

The output generated from ProtParam program, Figure 11, provides a textual output of the isoelectric point which is of interest for this research.



ProtParam

User-provided sequence:

10 20
 GAVLIMPWFS TNQYCKHRDE

[References and documentation](#) are available.

Number of amino acids: 20

Molecular weight: 2395.74

Theoretical pI: 6.74

Amino acid composition: CSV format

Ala (A)	1	5.0%
Arg (R)	1	5.0%
Asn (N)	1	5.0%
Asp (D)	1	5.0%
Cys (C)	1	5.0%
Gln (Q)	1	5.0%
Glu (E)	1	5.0%
Gly (G)	1	5.0%
His (H)	1	5.0%
Ile (I)	1	5.0%
Leu (L)	1	5.0%
Lys (K)	1	5.0%
Met (M)	1	5.0%
Phe (F)	1	5.0%
Pro (P)	1	5.0%
Ser (S)	1	5.0%
Thr (T)	1	5.0%
Trp (W)	1	5.0%
Tyr (Y)	1	5.0%
Val (V)	1	5.0%
Pyl (O)	0	0.0%
Sec (U)	0	0.0%
(B)	0	0.0%
(Z)	0	0.0%
(X)	0	0.0%

Total number of negatively charged residues (Asp + Glu): 2

Total number of positively charged residues (Arg + Lys): 2

Figure 11: ProtParam Webpage Output

1.3.5 Determination of Isoelectric Point Using IPC

The Isoelectric Point Calculation (IPC) program,¹¹ is described in a journal article by Kozlowski.¹² The webpage that displays the program input form, Figure 12, contains an entry window where a peptide sequence can be entered. The result of the calculation is shown in the display, Figure 13, of the isoelectric point (IEP) relative to the molecular weight, average IEP from different sources, and a listing of those source values.

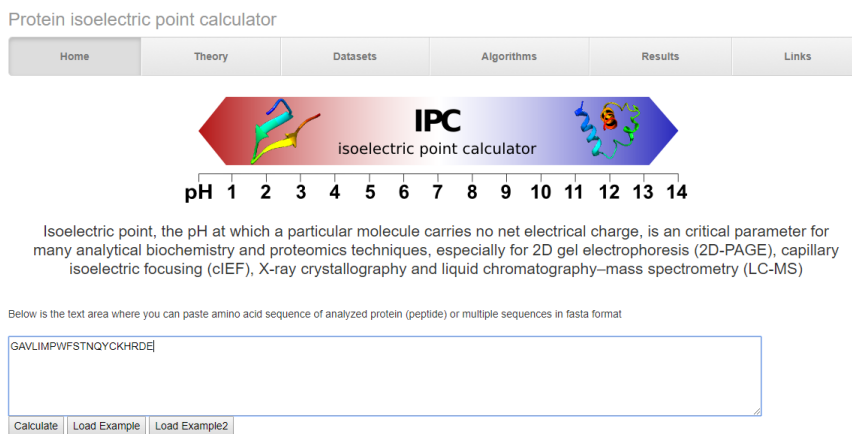


Figure 12: Protein Isoelectric Point Input Webpage

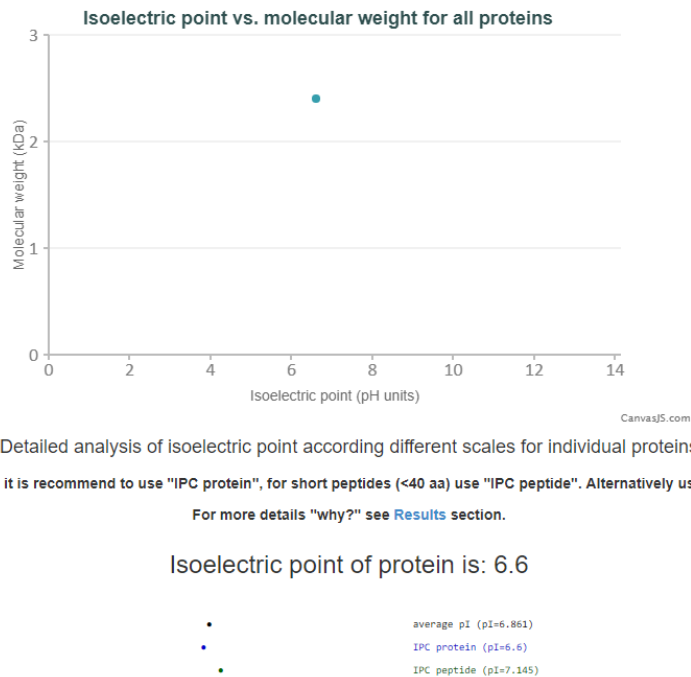


Figure 13: Protein Isoelectric Point Calculator Output

Chapter 2 – Developing and Testing of Program

2.1 Multi-enzyme 2-Dimensional Electrophoresis Program Description

The MultiEnzyme ElectroPhoresis Program (MEEP) is the program developed to provide a two-dimensional representation of the electrophoresis of a protein or peptide sequence. It is a stand-alone program that runs on Windows 10 computers.

2.1.1 Program Display and Operation

Figure 14 shows the display that the user sees when they start the MEEP program. Figure 15 shows the text output that is available to the user when they press the “Specify and Save Optional Output File”. It contains the list of subsequences of the input amino acid sequence sorted by molecular weight. Each row consists of the computed value of the molecular weight, the computed value of the isoelectric point, and the subsequence.

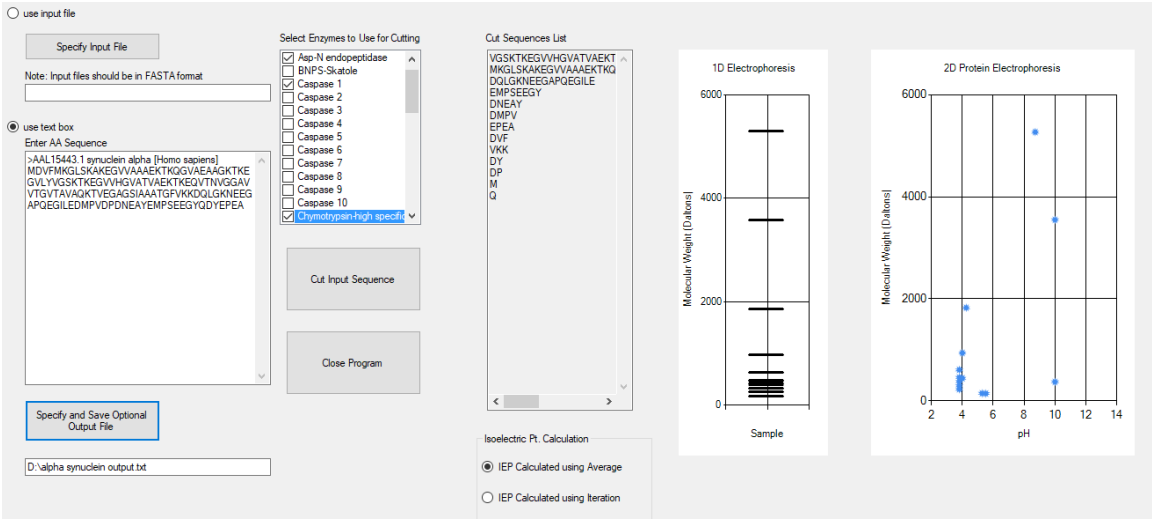


Figure 14: MEEP Program Display

70620 alpha synuclein output.txt - Notepad

File	Edit	Format	View	Help
MolWt	IEP	Subsequence		
146.15	5.525	Q		
149.22	5.275	M		
230.23	3.8	DP		
296.28	3.8	DY		
373.49	10	VKK		
379.41	3.8	DVF		
444.44	4	EPEA		
460.56	3.8	DMPV		
610.65	3.8	DNEAY		
940.97	4	EMPSEEGY		
1826.99	4.25	DQLGKNEEGAPQEGILE		
3549.08	10	MKGLSKAKEGVAAAETKQGVAAEAGKTKEGVLY		
5268.91	8.72	VGSKTKEGVWHGVATVAEKTKEQVTNVGGAVVTGVTAVAQKTVEGAGSIAAATGF		

Figure 15: MEEP Output Text File

Figure 16 shows a closeup of the input side of the program display. There are a pair of radio buttons that are used to select between the specification of an input text file and a sequence that the user enters in the text input area. When the user clicks the button “Specify Input File”, a window pops up that allows the user to locate and select the file. The selected file is displayed in the text box below the button. The input file must be in FASTA format or just contain the sequence to be evaluated. It has limited ability to isolate text from row numbers along the margin as it removes numbers, special characters, spaces, and end of line characters from the input text when the text is cleaned of unwanted characters at the beginning of processing. The text can be in either upper case or lower case or a combination of the two. The other radio button, “use text box”, allows for the input of any amino acid sequence from one amino acid to a whole protein.

The “Specify and Save Optional Output File” button allows the user to specify an output file that will contain the molecular weight, isoelectric point, and subsequences in a sorted list as shown in Figure 15. The name of the selected file is displayed in the text box below the button.

Because the program allows the user to change the input files or input sequences or the enzymes used for each “Cut Input Sequence” button press, it is also possible to generate and save multiple output files.

Under the label “Select Enzymes to Use for Cutting”, a checked list box that lists all the enzymes, provided in the ExPASy website, that are selectable by the user is displayed. Each time the user selects or deselects one of the 43 enzyme variations of the 24 enzymes and chemicals, the selected list is updated for the program.

After the user has provided the input text or text file and selected at least one enzyme to cut the sequence, the “Cut input Sequence” button can be pressed. If the user does not select one of the input forms, the button is not available. If the user does not select an enzyme to cut the sequence before pressing the “Cut Input Sequence” button, a message will pop up telling the user that the first enzyme in the list has been selected.

When the user is done with the program the “Close Program” button can be pressed to exit from the program. It is also possible to select the X in the upper right of the program to exit.

Figure 17 shows the output side of the program. There are three outputs that are generated when the user selects the “Cut Input Sequence” button. The “Cut Sequence List” that provides the sorted list of the cut subsequences from largest to smallest, the graph that displays the idealized representation of what the one-dimensional electrophoresis would look like, and the graph of the idealized two-dimensional electrophoresis. The scale of the molecular weight is automatically adjusted on each graph, depending on the molecular weight of the largest cut subsequence. The pH scale is set to a range of 2 to 14.

There is one pair of input radio buttons on the output control section of the display: the “IEP Calculated using Average” button and the “IEP Calculated using Iteration”. The “IEP Calculated using Average” option is the default option when the program is run. It will calculate the

isoelectric point by scanning the sequence of sorted pKa values and finding the pKa value where the charge sum of the positive charge residues and negative charge residues, along with the charges of the amine end and the carboxyl end sum to zero. It then averages the value of the pKa of that residue and the next pKa to get the value of the IEP.

The “IEP Calculation using Iteration” calculates the isoelectric point by summing the fractional ionization values for both the positive and negative residues at increasing pH values until the difference becomes close to zero. A more complete explanation of the calculations is given in the section on isoelectric point calculations.

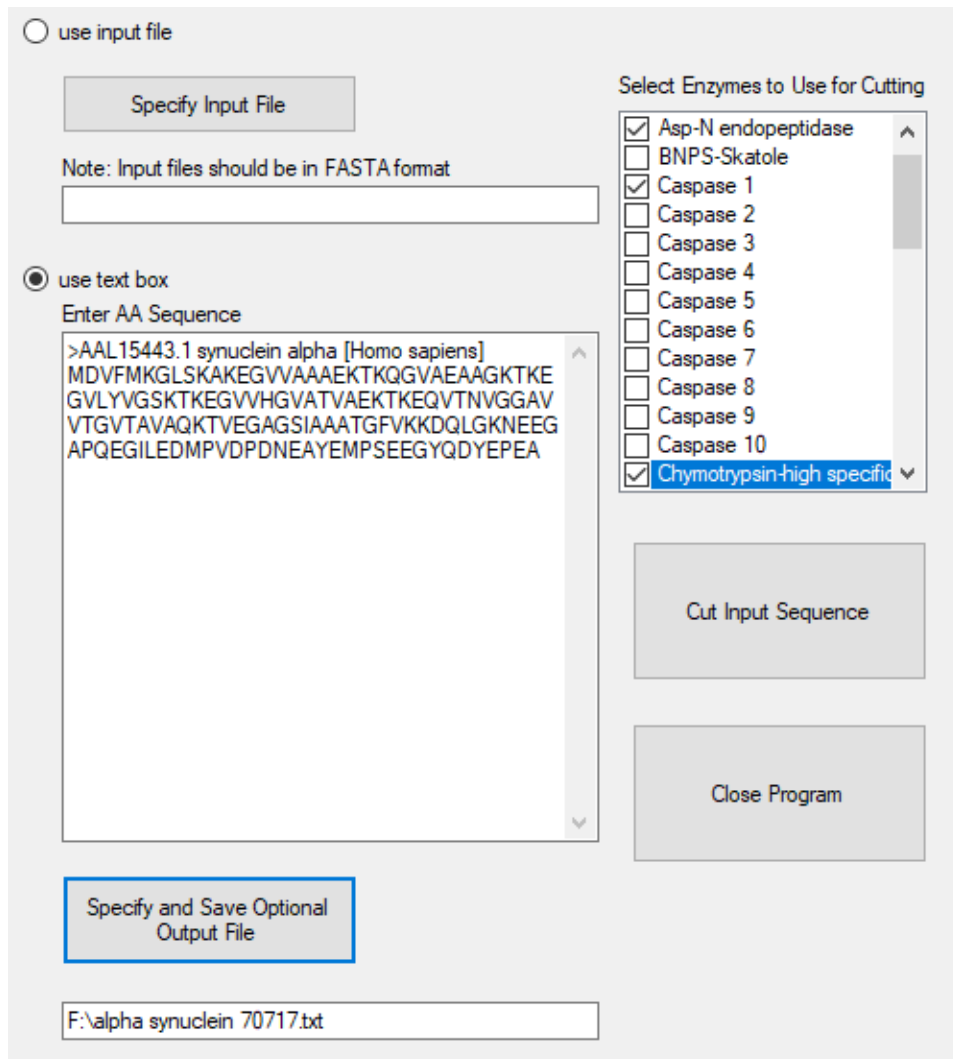


Figure 16: Input Control Section of MEEP Display

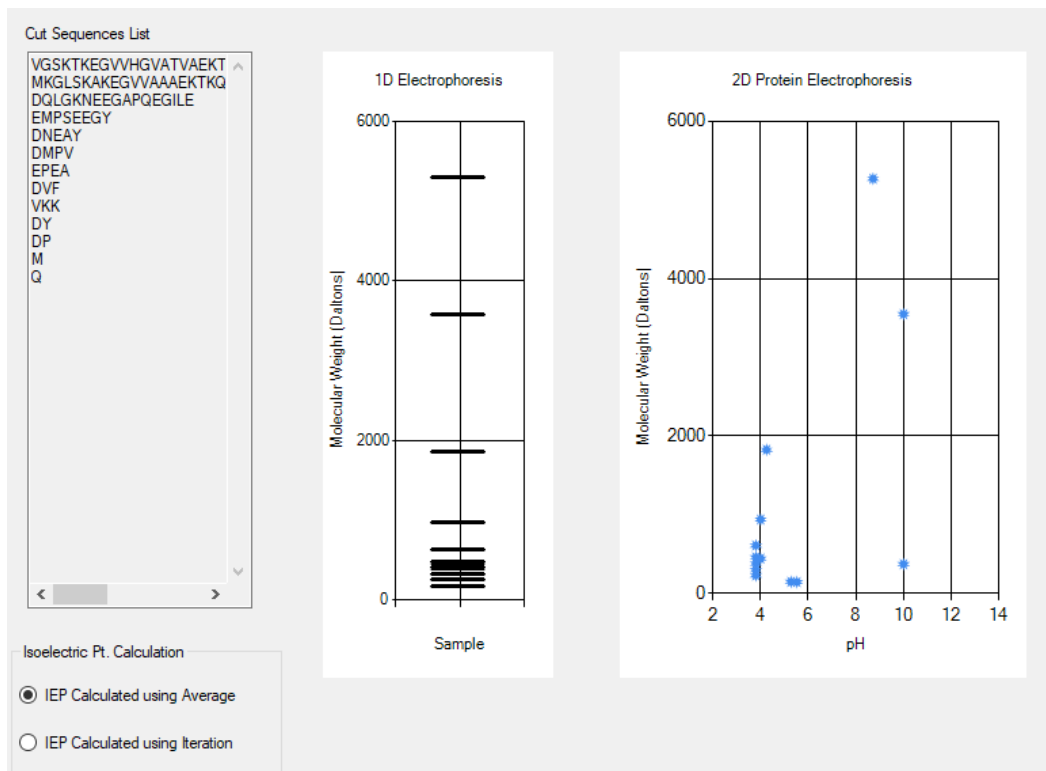


Figure 17: Output Section of MEEP Display

2.1.2 Program Logic

The structure of the MEEP program was based on the programming language, C#, which was developed by Microsoft Corporation. It uses an integrated development environment program (IDE) called Visual Studio that allows the programmer to write the C# code within a structure that has standardized methods to simplify coding and debugging. The MEEP program uses a graphical user interface (GUI) to provide the display that the user sees as well as the method of input. One of the features that Visual Studio provides is a package of programming tools that gives the programmer the ability to “drag and drop” the components of the displayed form as seen in Figure 14. Part of the code is used to interface those components with the other parts of the program. The two graphs used to display the electrophoresis simulation were placed on the form as generic graphs and then modified by specifying components that could be manipulated to get the program to display the appropriate titles, scales, and output symbols.

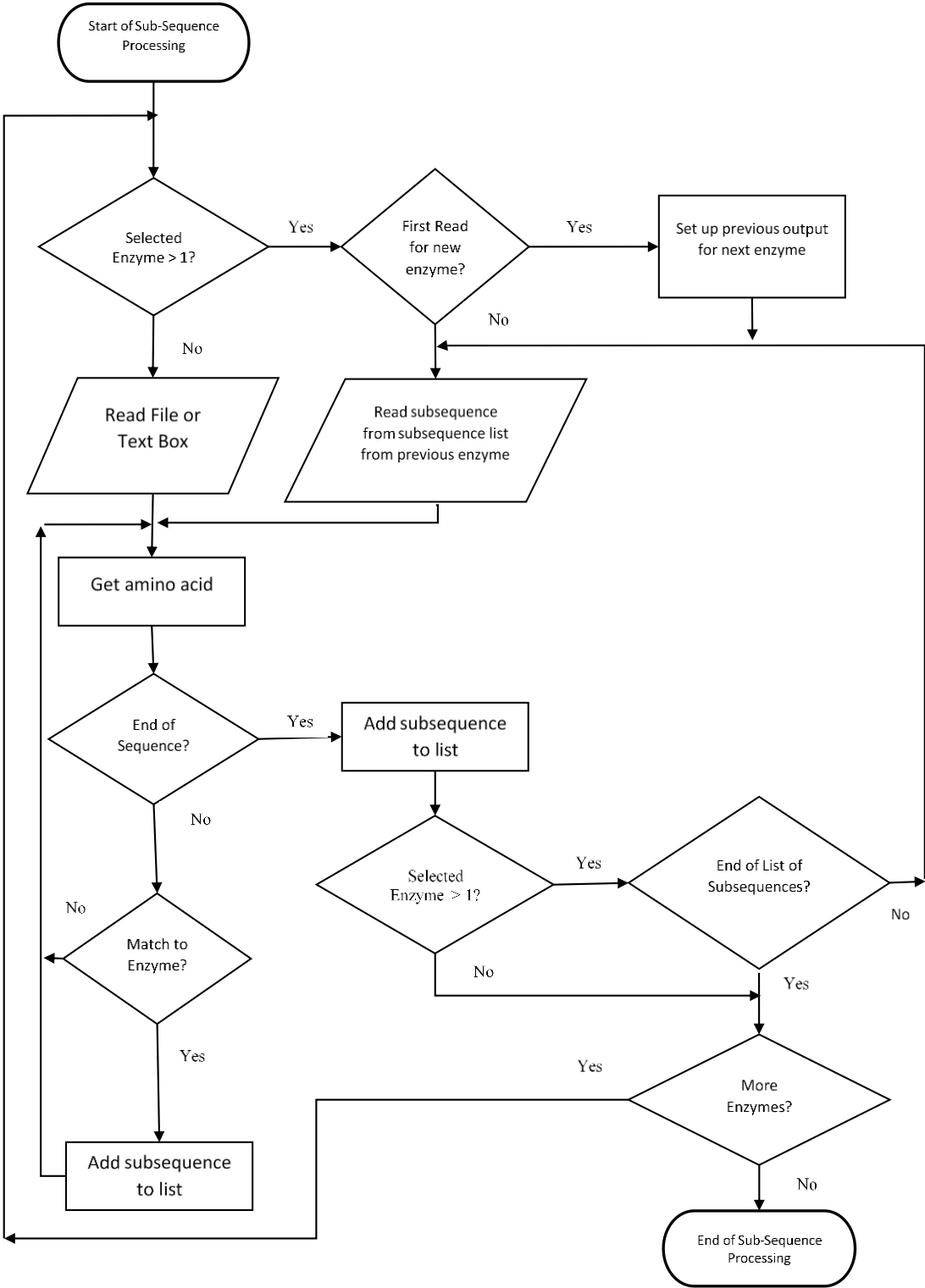


Figure 18: Program Flowchart

An overview flowchart of the functional logic that provided the interpretation of the input and calculation of the output is shown in Figure 18. When the “Cut Input Sequence” button is pressed, the program must establish whether there is more than one enzyme that is cutting the input string. The reason is that when multiple enzymes are used it is necessary for the output of one set of cuttings of the sequence to be used as inputs for the next enzyme to cut. If it is the first time through the logic it is necessary to read the text file that has the peptide string or protein. Regardless of whether the input comes from a file or a text box, the input data must be cleaned, such as removing introductory lines, as in the case of FASTA files, or end of line characters, or spaces, or converting lower case letters to upper case letters. Once the input has been cleaned, the single letter abbreviation amino acid string is taken one amino acid at a time and evaluated. After the check for the end of the sequence, the process of matching the amino acid with the enzyme is performed. If the match is successful, the amino acid is added to the subsequence and the looping continues. If the check of all the amino acids in the sequence has been completed, then any remaining portion of the amino acid sequence that is not part of one of the cut subsequences is added as another subsequence.

After the amino acid sequence has been checked against an enzyme, there is a check to see if the selected enzyme is not the first enzyme, as the later enzymes must evaluate the subsequences generated by the first enzyme and each of those subsequences must be treated as a separate sequence by the later enzymes. If the input list of subsequences generated by the previous enzyme has not all been read, then the program needs to read in the next subsequence for processing.

When the list of subsequences is completed or if the first enzyme check is completed then a check is made for additional enzymes. If there are more enzymes then the program loops back

to pick up those enzymes, otherwise, the processing is complete. If there are more enzymes, then the loop checks to see if it is the second or subsequent enzyme and the program checks to see if it is the first time with that enzyme. If it is the first time, then the output subsequences from processing of the previous enzyme must be converted to the set of input sequences to be evaluated. After those adjustments, each new subsequence is read in to be processed as before.

2.1.3 Match to Enzyme Logic

The general logic of matching the amino acid sequence to the enzyme is shown in Figure 19. When an amino acid is compared with an enzyme, it is checked with the enzyme's required amino acids at one of the six locations. If it matches, then the program checks to see if all positions have been checked. When that occurs, the program must do additional processing for trypsin because trypsin has exceptions that will block the cutting if the amino acid sequence has a certain combination of amino acids around the cut point. If the enzyme is trypsin and the sequence matches the exclusion, then the program ends the matching process for that amino acid. Otherwise, the amino acid has satisfied all the criteria and the molecular weight and isoelectric point are calculated.

For each amino acid that does not match, it is necessary to determine which position in the enzyme was compared. If the position compared was not the first position for that enzyme, then the program must reset the counters for the amino acid sequence and the enzyme, to allow for a possible shift in the correct sequence starting point. The other decision that needs to be made in the matching routine is whether all positions for the enzymes have been matched. If not, then the processing ends, and the next amino acid is retrieved.

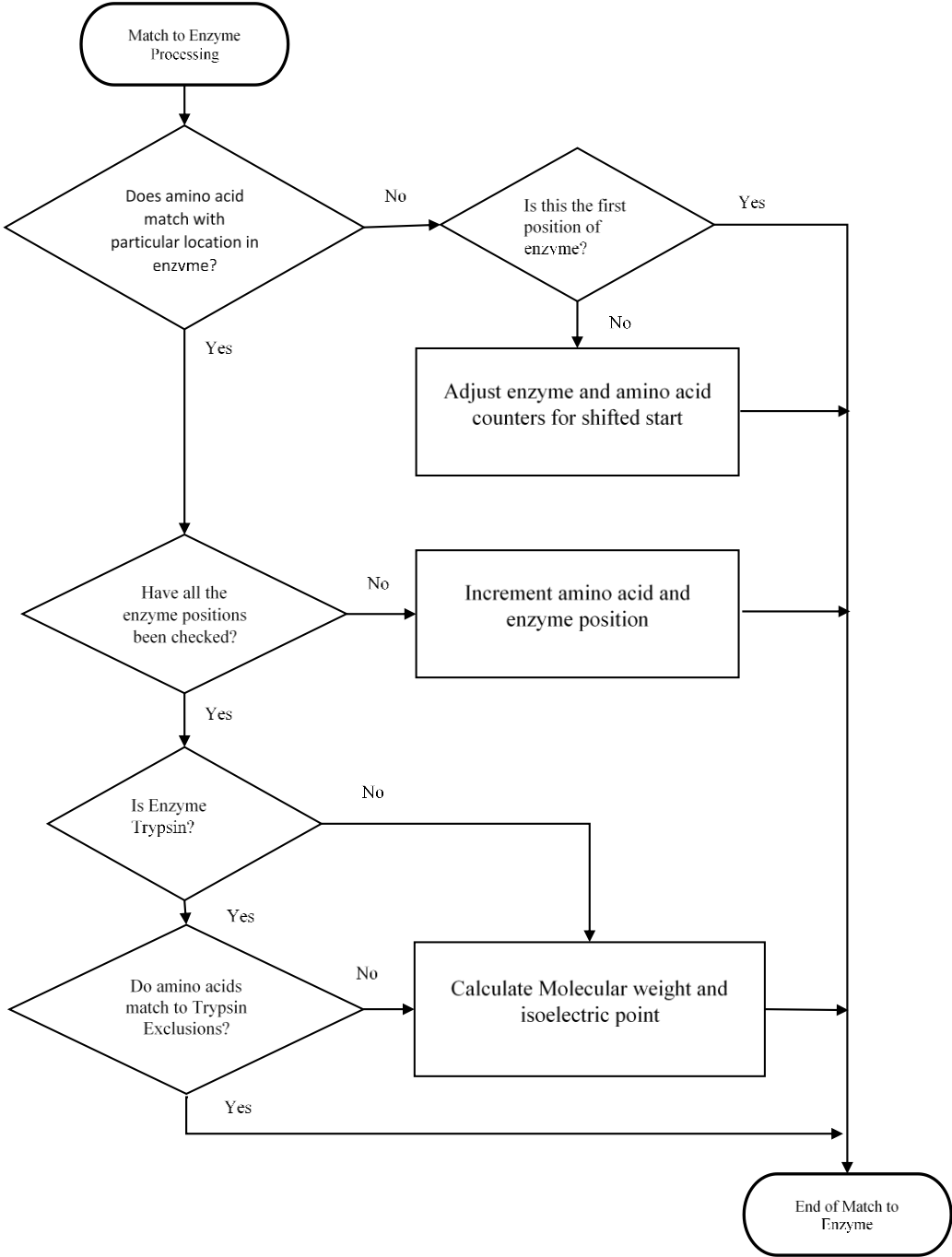


Figure 19: Match to Enzyme Flowchart

2.1.4 Enzyme Function in Program

To improve the quality of the simulation, the enzyme cutting logic involved the use of 43 enzyme variations with the potential evaluation of six amino acid residues around the cut position. The logic used in this program was based on the enzyme cutting protocol in the ExPASy website.¹³ They in turn based their nomenclature on descriptions by Schechter and Berger^{14,15}. The rules for the enzyme function were taken from the work of Keil.¹⁶

Table 1 shows the enzymes used in this program along with the six positions around the cleavage point of the amino acid sequence. The position, four amino acids before the cut point, referenced from the amino acid end of the sequence, is labeled P4. Each successively closer position to the cut is decremented in count, with P1 being the closest to the cut. On the carboxyl side of the cut the amino acid, locations increase from P1' to P2'. At each position, the amino acids are listed as one-character abbreviations that will satisfy the requirements of the enzyme for a cut. At the positions where there is a "Not" written, each of those amino acids cannot be in that position for the cut to work. Arg-C proteinase is one of the simpler enzyme cut sequences, requiring only an arginine (R) at the amino acid position right before the cut. Thrombin (P) is probably the most complex cutting enzyme and requires the amino acid sequence to contain any of the amino acids A, F, G, I, L, T, V, or M at the locations, four and three positions before the cut. At the position two amino acids before the cut, a proline (P) is required and at the position immediately before the cut, an arginine (R) is required. At the two positions after the cut, P1' and P2', the amino acid sequence cannot have the amino acids aspartate (D) or glutamate (E).

The name for the enzyme in the table is the name given by the ExPASy site. The information in parentheses following the name represents my way of distinguishing different unique conditions for the same enzyme. As an example, there are four entries for Pepsin. Two of the

conditions require a pH of 1.3 and two require a pH greater than 2. The label “Pepsin (pH 1.3) (not R)” is used when the pH is at 1.3 and the amino acid is not an arginine (R) at position P1. For the Pepsin with (pH 1.3) (F, L), the amino acid sequence must be at a pH of 1.3 and have either a phenylalanine (F) or a leucine (L) at position P1.

It is necessary to assign names with parenthetical modifications to variations in particular enzymes that are shown in the ExPASy website; specifically, Chymotrypsin, Pepsin, and Trypsin. Each of these enzymes have multiple entries for the cutting rules that reflect modifications that cannot be addressed as one enzyme. For example, Chymotrypsin - high specificity is divided into two groupings: the first entry has phenylalanine (F) or tyrosine (Y) in position P1 and not proline (not P) in position P1'. The second entry has tryptophan (W) in position P1 and not methionine (not M) or not proline (not P) in position P1'. Since the first entry has rules in two positions that do not match up to the rules in the second entry, it is necessary to add these as separate rules for the cutting process. If someone wants to include both sets of rules for Chymotrypsin high specificity, they would need to check both Chymotrypsin - high specificity enzyme selections. For Chymotrypsin - low specificity there are four entries, for Pepsin – pH 1.3 there are two entries, for Pepsin – pH >2 there are two entries, and for Trypsin there are three entries.

Table 1: Enzymes and Cutting Rules

Enzyme	P4	P3	P2	P1	P1'	P2'
Arg-C proteinase				R		
Asp-N endopeptidase					D	
BNPS-Skatole				W		
Caspase 1	F,W,Y, or L		H,A, or T	D	Not P,E,D, Q,K, or R	
Caspase 2	D	V	A	D	Not P,E,D, Q,K, or R	
Caspase 3	D	M	Q	D	Not P,E,D, Q,K, or R	
Caspase 4	L	E	V	D	Not P,E,D, Q,K, or R	
Caspase 5	L or W	E	H	D		
Caspase 6	V	E	H or I	D	Not P,E,D, Q,K, or R	
Caspase 7	D	E	V	D	Not P,E,D, Q,K, or R	
Caspase 8	I or L	E	T	D	Not P,E,D, Q,K, or R	
Caspase 9	L	E	H	D		
Caspase 10	I	E	A	D		
Chymotrypsin high specificity (F, Y)				F or Y	Not P	
Chymotrypsin high specificity (W)				W	Not M or P	
Chymotrypsin Low specificity (F,L,Y)				F,L, or Y	Not P	
Chymotrypsin Low specificity (W)				W	Not M or P	
Chymotrypsin Low specificity (M)				M	Not P or Y	
Chymotrypsin Low specificity (H)				H	Not D,M, P, or W	
Clostripain				R		
CNBr				M		
Enterokinase	D or E	D or E	D or E	K		
Factor Xa	A,F,G,I,L, T,V, or M	D or E	G	R		

Table 1 Cont.: Enzymes and Cutting Rules

Enzyme	P4	P3	P2	P1	P1'	P2'
Formic acid				D		
Glutamyl endopeptidase				E		
GranzymeB	I	E	P	D		
Hydroxylamine				N	G	
Iodosobenzoic				W		
LysC				K		
Neutrophil elastase				A or V		
NTCB					C	
Pepsin (pH 1.3) (not R)		Not H,K, or R	Not P	Not R	F or L	Not P
Pepsin (pH 1.3) (F, L)		Not H,K, or R	Not P	F or L		Not P
Pepsin (pH > 2) (not R)		Not H,K, or R	Not P	Not R	F,L,W or Y	Not P
Pepsin (pH > 2) (F,L,W,Y)		Not H,K, or R	Not P	F,L,W or Y		Not P
Proline-endopeptidase			Not H,K, or R	P	Not P	
Proteinase K				A,E,F,I,L,T, V,W, or Y		
Staphylococcal Peptidase I			Not E	E		
Thermolysin				Not D or E	A,F,I,M, Or V	
Thrombin (G)			G	R	G	
Thrombin (P)	A,F,G,I,L, T,V, or M	A,F,G,I,L,T, V,W, or A	P	R	Not D or E	Not D or E
Trypsin (not P)				K or R	Not P	
Trypsin (W)			W	K	P	
Trypsin (M)			M	R	P	

The exceptions for the trypsin enzyme are given in Table 2. Trypsin not only requires the amino acid sequence to satisfy the conditions in Table 1 but also to “not” have the set of conditions given in Table 2. For the Exception 1 case the amino acid sequence will not be cut if it has a cysteine (C) or an aspartate (D) in the position P2 and has a lysine (K) at position P1 and

has an aspartate (D) at position P1'. If an amino acid sequence satisfies all the conditions for Trypsin (not P) and satisfies all the conditions for Exception 1, it will not be cut.

Table 2: Trypsin Exceptions

Trypsin Exceptions	P4	P3	P2	P1	P1'	P2'
Exception 1			C or D	K	D	
Exception 2			C	K	H or Y	
Exception 3			C	R	K	
Exception 4			R	R	H or R	

2.1.5 Isoelectric Point Determination

The program provides two methods to determine the isoelectric point. One method, the averaging method, calculates the isoelectric point by balancing the total charge on the peptide string. Since the total charge on the peptide string varies as a function of the pH, the list of pKa's of the peptides is sorted and the charge on the peptide string is evaluated at each pKa until the total charge evaluates to zero. That pKa and the next one in the sequence are averaged to get the IEP.

The other method, the iteration method, calculates the IEP¹⁷ by using a form of the Henderson-Hasselbalch equation iteratively until the squared difference between the fractional value of the negative charged amino acids and the fractional value of the positive charged amino acids approaches zero. In the case of the program the value is 0.0000000001.

The Henderson-Hasselbalch equation is typically seen as

$$pH = pKa + \log \frac{[A^-]}{[HA]}$$

However, it can also be written as

$$pH = pKa + \log \frac{\alpha_{A^-}}{\alpha_{A^+}}$$

where α_{A^-} and α_{A^+} are the fractional coefficients and add up to 1.

$$\alpha_{A^-} + \alpha_{A^+} = 1$$

By taking the anti-log of both sides of the Henderson-Hasselbalch equation, the equation becomes

$$10^{pH-pKa} = \frac{\alpha_{A^-}}{\alpha_{A^+}}$$

By adding 1 to both sides, the equation can be rearranged to get the fractional amount of the positive charged peptide, α_{A^+} .

$$1 + 10^{pH-pKa} = \frac{\alpha_{A^-}}{\alpha_{A^+}} + 1$$

$$1 + 10^{pH-pKa} = \frac{\alpha_{A^-}}{\alpha_{A^+}} + \frac{\alpha_{A^+}}{\alpha_{A^+}}$$

$$1 + 10^{pH-pKa} = \frac{1}{\alpha_{A^+}}$$

$$\alpha_{A^+} = \frac{1}{1 + 10^{pH-pKa}}$$

In a similar fashion, it is possible to determine the fractional amount of the negative charged peptide.

$$\alpha_{A^-} = \frac{1}{1 + 10^{pKa-pH}}$$

Because there are multiple peptides in a polypeptide it is necessary to sum all n peptides for the positive charged peptide string and for the negative charged peptide string. The positive charged peptides are those peptides with a charged amino acid in the residue, arginine (R), lysine (K), or histidine (H), along with the amino group at the start of the amino acid sequence. The negative charged peptides are those peptides with a charged carboxyl residue, aspartate (D) or glutamate (E), or a cysteine (C) or tyrosine (Y) along with the carboxyl end of the amino acid sequence.

$$total \alpha_{A^+} = \sum_1^n \frac{1}{1 + 10^{pH-pK_a}}$$

$$total \alpha_{A^-} = \sum_1^n \frac{1}{1 + 10^{pK_a-pH}}$$

The pH is incremented by one pH unit until the squared difference starts to increase. The pH is then backed off by two units, the incrementing value is reduced by a factor of 10 and the incrementing process begins again. This iterating process continues until the squared difference becomes arbitrarily small, 0.0000000001.

2.2 Validation

As a check on the numbers generated by the program, a comparison was made to the same calculations generated by two websites: ExpPASy and IPC. Both ExpPASy and IPC are referenced earlier as websites that generate values for IEP and molecular weight. The IPC website calculates the isoelectric point and the ExpPASy site calculates both IEP and molecular weight. Because the ExpPASy site is considered one of the primary sources for protein analysis, there was an attempt in the program to generate an IEP as close as possible to the value generated by the ExpPASy site. That required the matching of pKa values used by ExpPASy. To match with their calculations, the program uses the iterative method of IEP calculation and uses Bjellqvist, Basse, Olsen, Celis¹⁸ for the values of the pKa's.

Table 3: Test Cases for IEP and Molecular Weight

Peptide	# of AA	MEEP ave IEP (pH)	MEEP iter IEP (pH)	ExpASy IEP (pH)	IPC IEP (pH)	MEEP Mol Wt (Daltons)	ExpASy Mol Wt (Daltons)
Zika polyprotein	3417	9.00	8.79	8.79	7.64	378498.31	378493.00
Bovine serum albumin	607	5.98	5.82	5.82	5.59	69293.7	69293.41
Egg albumin	386	4.45	5.19	5.19	5.09	42882.24	42881.24
Glyceraldehyde-3-phosphate dehydrogenase	335	9	8.57	8.57	7.74	36054.33	36053.21
Carbonic anhydrase II	260	5.98	6.41	6.41	6.12	29114.53	29113.78
Trypsinogen	246	9	8.40	8.40	7.23	25786.29	25785.24
Trypsin inhibitor	216	4.45	4.95	4.95	4.84	24005.85	24005.29
α -Lactalbumin	123	4.45	4.80	4.80	4.67	14156.51	14156.04
Thyroglobulin	2769	4.45	5.48	5.48	5.33	303224.20	303221.61
β - Amylase	499	4.45	5.17	5.17	5.04	56081.49	56079.70
Alcohol dehydrogenase	375	9.00	8.26	8.26	7.08	39859.2	39858.69
Lactate dehydrogenase	332	9.00	8.44	8.44	7.5	36689.58	36688.72
Ferritin heavy chain	183	4.45	5.31	5.31	5.13	21226.38	21225.64
KHRKH	5	11.00	11.17	11.17	10.7	704.8	704.85
CDEYC	5	3.80	3.67	3.67	3.29	631.66	631.62
GAVLIMPWFSTNQ	13	5.53	5.52	5.52	5.98	1463.64	1463.71

Table 3 lists the peptide sequences taken from the UniProt website¹⁹, the number of amino acids in the sequences, the various calculations of the isoelectric points and the molecular weight calculations. Some of the peptides come from standards for an SDS-PAGE calibration curve²⁰ and single polypeptide strings used in a Sephadex standard where FASTA sequences are available. A polypeptide string representing the Zika virus protein strings is used to display a large sequence. In addition, three other cases are sequences used to get a minimum (minimum required by ExpASy site) of five positive charges, KHRKH, and negative charges, CDEYC, and a group that contains the remaining amino acids, GAVLIMPWFSTNQ. The MEEP average (ave) and MEEP iteration (iter) provide a comparison between the two methods used by the program to calculate the IEP. The positive values cause the average calculation to be less than the iterative

method. The negative values cause the average calculation to be greater than the iterative method.

A comparison between the program's iterative calculation and the value calculated by the ExPASy site shows that calculations and the pKa's used, are the same. The IPC website shows much more variability from the program's calculation; since IPC uses the same iterative method, the differences must be related to the pKa's used.

Because no reference was found at the ExPASy for the molecular weights, the molecular weights were taken from the CRC.²¹ The difference between the ExPASy and CRC molecular weights are within 0.1 Daltons for small sequences but grows to 5.31 Daltons for the Zika polyprotein which has 3417 amino acids.

Chapter 3 – Conclusion and Future Direction

There are several programs that provide information that help the researcher determine what to expect with an electrophoresis experiment. This thesis describes several of those programs and their inputs and outputs.

The program MultiEnzyme Electrophoresis (MEEP) is described which includes its inputs, outputs, and its operation. The overall logic for the operation of the cutting and sorting is given as well as the logic associated with the matching of enzymes. A description is provided of the two methods available in the program to determine the isoelectric point for a subsequence. The equations are provided that go from the Henderson-Hasselbalch equation to the equations used in the program. For validation, several test sequences are compared with the values determined from the ExpASy website and the Isoelectric Point Calculator website.

The program provides another tool to help visualize what to expect when cuts are made in a protein sequence. I see it being used in two ways: 1) a fingerprint of the protein which would allow that protein to be isolated from among several proteins if present in a sample, 2) a predictive program to evaluate which enzymes would be best suited to isolate the expected protein fragments. Prior to running the experiment, an analysis can be made as to whether the enzymes chosen will provide an output that most effectively breaks out the fragments of interest.

So where to go in the future? Because the cutting process can be generalized it might be possible to look at nucleotides instead of peptides, both of which can use the 2-D electrophoresis. Another possibility would be to look at a mixture of proteins, so the extension of the program from one protein sequence to multiple proteins could be helpful. A third direction could be to address the statistical issues of cleavage of protein sequences with multiple enzymes at the same time. Currently the program evaluates the cleavage of the first

enzyme in the list and then evaluates the cleavage of the second enzyme, and then the following enzyme; until all the enzymes are addressed. Because not all the enzymes act all the time and the order in which the enzymes act will vary, a statistical approach would be needed to more accurately model that aspect.

References

1. Voet, D.; Voet, J.; Pratt, C. *Principles of Biochemistry*, 4th ed.; John Wiley and Sons: Singapore, 2013; pp 54, 104.
2. Iyer, A.; Roeters, S. J. S. N.; Hommersom, B.; Heeren, R. M. A.; Woutersen, S.; Claessens, M. A. A. E.; Subramaniam, V. The Impact of N-terminal Acetylation of α -Synuclein on Phospholipid Membrane Binding and Fibril Structure. *Journal of Biological Chemistry* **2016**, *291* (40), 21110-21122.
3. Rabilloud, T.; Lelong, C. Two-dimensional gel electrophoresis in proteomics: A tutorial. *Journal of Proteomics* **2011**, *74* (10), 1829-1841.
4. Perl, S.; Steen, H.; Akhilesh, P. GPMaw - a software tool for analyzing proteins and peptides. *TRENDS in Biochemical Sciences* **2001**, *26* (11), 687-689.
5. Visit to use GPMaw, the free protein calculator: Alphalyse.
<http://www.alphalyse.com/customer-support/gpmaw-lite-bioinformatics-tool/start-gpmaw-lite/> (accessed July 7, 2017).
6. Fisher, A.; Sekera, E.; Payne, J.; Craig, P. Simulation of Two Dimensional Electrophoresis and Tandem Mass Spectrometry for Teaching Proteomics. *Biochemistry and Molecular Biology Education* **2012**, *40* (6), 393-399.
7. about ExPASy. www.expasy.org/about (accessed June 16, 2017).
8. ExPASy features. <https://www.expasy.org/features> (accessed June 16, 2017).
9. PeptideCutter.
http://web.expasy.org/peptide_cutter/?_ga=1.251315653.2121197300.1481946381 (accessed June 16, 2017).
10. ProtParam. <https://web.expasy.org/protparam/> (accessed October 22, 2017).
11. Kozlowski, L. P. Protein isoelectric point calculator Home.
<http://isoelectric.ovh.org/index.html> (accessed June 16, 2017).
12. Kozlowski, L. P. IPC - Isoelectric Point Calculator. *Biology Direct* **2016**.
13. PeptideCutter. http://web.expasy.org/peptide_cutter/peptidecutter_enzymes.html (accessed June 23, 2017).
14. Schechter, I.; Berger, A. On the size of active site in proteases. I. Papain. *Biochemical and Biophysical Research Communication* **1967**, *27* (157).
15. Schechter, I.; Berger, A. On the active site of proteases. 3. Mapping the active site of papain; specific peptide inhibitors of papain. *Biochemical and Biophysical Research Communication* **1968**, *32* (898).
16. Keil, B. *Specificity of proteolysis*; Springer-Verlag: Berlin-Heidelberg-NewYork, 1992; p 335.
17. Kozlowski, L. P. theory. <http://isoelectric.ovh.org/theory.html> (accessed July 20, 2017).
18. Bjellqvist, B.; Basse, B.; Olsen, E.; Celis, J. E. Reference points for comparisons of two-dimensional maps of proteins from different human cell types defined in a pH scale where isoelectric points correlate with polypeptide compositions. *Electrophoresis* **1994**, 529-539.
19. * in UniProtKB. <http://www.uniprot.org/uniprot/> (accessed October 20, 2017).
20. Farrell, S. O.; Taylor, L. E. *Experiments in Biochemistry: A Hands-On Approach*, 2nd ed.; Brooks/Cole Cengage Learning: Belmont, 2006.
21. Haynes, W. M., Ed. *CRC Handbook of Chemistry and Physics*, 94th ed.; CRC Press: Boca Raton, Florida, 2016.

