Determining Cleavage Site Sequences to Characterize the Active Site of BACE1 Aspartic Protease

A Thesis
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Nicole Heinks

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

Dr. Joseph L. Johnson

April 2015

Acknowledgements

I would like to acknowledge Dr. Johnson for giving me the opportunity to work in his research lab and for his help with my research. I would like to acknowledge all the students and graduate students who also worked in Dr. Johnson's research lab: Erik Carlson, Chris Bicknese, Amrita Oak, as well as the other undergraduate students in this research group. I would also like to acknowledge the University of Minnesota, Duluth graduate school for accepting my application. The Minnesota Supercomputing institute (MSI) was a source of software for database searching and sample analysis. The Center of Mass Spectrometry at the University of Minnesota, St. Paul Campus performed all mass spectrometry and helped in the initial analysis of samples.

Abstract

Proteases play key roles in physiology and disease development due to their active role in the regulation of other proteins. Understanding and characterizing the active site of relevant proteases provides information vital to the production of inhibitors and elucidates potential native substrates that may be affected by inhibitors. Once the preferred sequence of the active site of a target protease, such as β-secretase (BACE1), has been defined, protease inhibitors can be created to treat or manage diseases such as Alzheimer's disease (AD). BACE1 is an aspartic protease that is overexpressed in the AD brain and is believed to initiate the AD disease pathway. As such, it is a strong candidate for drug design. However, chronic administration of BACE1 inhibitors could result in undesirable side effects due to the impairment of its ability to hydrolyze native substrates; therefore, the amino acid peptide sequence preferentially cleaved by BACE1 needs to be characterized. Not only will this indicate potential substrates that may be affected by BACE1 inhibition, it will also aid in the synthesis of viable inhibitors. Recently, a novel method for determing the cleavage sequence of proteases was reported. Proteomic identification of cleavage sites (PICS) is a method designed to accurately identify cleavage sequence preferences and neighbor interactions in the cleavage site that influence protease cleavage. This method gives additional information with less bias than previous methods used to characterize protease active sites. Multiple controls were used to confirm the validity of the procedure. These controls demonstrated our ability to successfully identify the amino acid sequence preferences for well characterized proteases. We were thus able to confidently use PICS to obtain the sequence of amino

acids preferentially cleaved by BACE. BACE1A has two noticeable characteristics for the amino acid sequence cleaved: aromatic amino acids are preferred in the P1 site and leucine is strongly preferred in P2'. Other preferred amino acids are observed in the sequence, but not to the extent of P1 and P2'. Neighbor interactions were also investigated. Positive cooperativity resulted with leucine or valine in P3 and with phenylalanine or tyrosine in P1. There were strong interactions between valine in P3 and phenylalanine in P1 and between tyrosine in P1 and valine in P2'. Sequence preferences were also investigated for BACE2A, which exhibited both similarities and differences between BACE1A and BACE2A. The next step in this research will be to use knowledge of the preferred cleavage sites to determine physiological substrates of BACE1A. This will reveal more information about the natural function of BACE1A and identify potential side effects of its inhibition.

Table of Contents

List of Tables

List of Figures

Abbreviations

Alzheimer's Disease	AD
Amyloid precursor protein	APP
Amyloid-β	Aβ
Atomic mass unit	amu
Collision Induced Dissociation	CID
Electron Capture Desorption	ECD
Electron Transfer Dissociation	ETD
Electrospray Ionization	ESI
False Discovery Rate	FDR
Fourier transform-ion cyclotron resonance	FT-ICR
Isotope-Coded Affinity Tag	icat
Linear Trap Quadrupole	LTQ
Liquid Chromatography	LC
Mass Spectrometry	MS
Mass to charge ratio	m/z
Matrix-assisted Laser Desorption	MALDI
Number of missed cleavages per Peptide	NMC
Peptide Fragmentation Pathway	PFP
Proteomic Identification of Cleavage Sites	PICS
Proton Affinity	PA
Time of Flight	tof

Trans Proteomic Pipeline	TPP
β-secretase cleaving enzyme 1	BACE1/B1A

Introduction

Protease function and importance in biological systems

Proteases are a complex class of enzymes that make up nearly 2% of the human proteome, including over 500 proteases and their homologs. They vary in size, location of expression, and characteristics such as reaction mechanism and optimal conditions for activity. Proteases cleave other proteins by recognizing a specific sequence of amino acids or a single amino acid.

Proteases are separated into five classes based on their catalytic mechanism: aspartic, metallo, cysteine, serine, and threonine proteases.[1–6] The majority of proteases belong in the metallo, serine, and cysteine protease classes. Threonine and aspartic proteases are more specialized and less numerous.[2] Each class of protease is defined by the active site catalytic residues, e.g., aspartic proteases contain two aspartate groups in the active site. The subsite positions in the active site and the corresponding substrate sequence recognized by the protease are described by their position in relation to the scissile bond (the peptide bond cleaved by the active site of the protease) (Fig. 1). Amino acid residues N-terminal of the scissile bond in the recognition sequence are labeled P1, P2, etc. while the C-terminal positions are indicated as 'prime' (P1', P2', etc.). [3,5] Corresponding pockets that bind the substrate residues are labeled S, e.g., the S2 pocket on the enzyme consists of all of the atoms (and residues) that interact with the P2 residue.

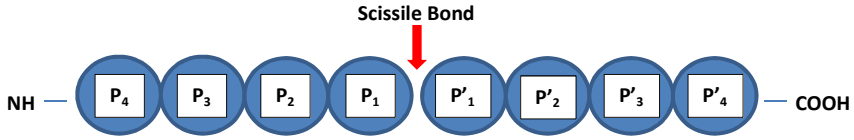


Figure 1: Protease active site nomenclature. The amino acids comprising the substrate sequence cleaved are labeled P or P' depending on which side of the scissile bond they are positioned. The scissile bond, the bond cleaved by the protease, is indicated by the red arrow. The active site positions on the protease are labeled S or S' depending on which side of the scissile bond they are positioned. Both active site and substrate positions are numbered moving away from the scissile bond.[3]

As mentioned before, two active site aspartate residues are responsible for aspartic protease catalytic activity. They employ a general acid-base mechanism to cleave peptide bonds. One of the aspartic residues acts as a general acid, while the other acts as a general base, resulting in the observed peak activity at an acidic pH (4.6). [7,8] The aspartate residue acting as a base accepts a proton from a water molecule. The resulting hydroxyl ion then carries out a nucleophilic attack on the carbonyl carbon adjacent to the scissile bond, while the other aspartate residue donates a proton to the oxygen of the carbonyl carbon to yield a tetrahedral carbon intermediate.[7,8] The two aspartate residues then reverse their function, with one aspartate donating a proton to the nitrogen of the amide bond, and the other aspartate accepting a proton from the newly formed hydroxyl group on the tetrahedral carbon adjacent to the scissile bond (Fig. 2). Cleavage of the peptide bond creates a new carboxyl terminus for one protein/peptide fragment and a new amino terminus for the other.

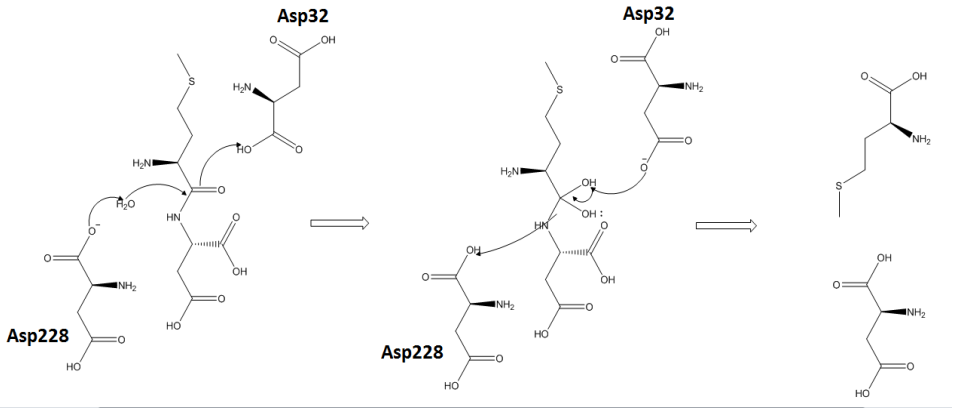


Figure 2: This figure generated using ChemDraw depicts the acid-base mechanism used by BACE1 to cleave substrates. [7] *The two aspartate residues (Asp32 and Asp228) in the active site of BACE1 catalyze the reaction.*

Protease active sites can recognize a single amino acid or class of amino acid, or they can recognize a sequence of amino acids depending on the size and shape of the active site.[5] Trypsin is a serine protease that recognizes and cleaves substrates after arginine or lysine.[1] While trypsin only discriminates based on the amino acid in P1, papain is a cysteine protease that recognizes and cuts within a recognition sequence of 7 amino acids. Papain's active site allows multiple amino acids in each of the 7 positions.[1,5] Active site-substrate interactions are determined by the amino acid sequence of the substrate that is recognized by the protease. Being able to define this sequence enables synthesis of inhibitors and insight into potential substrates of the protease being investigated.[9]

Proteases play a key role in disease development due to their active role in the regulation of other proteins and enzymes. Overexpression, underexpression, or mutations of proteases result in disruptions of biological processes that are implicated in a variety of disorders. In 2000, 14% of human proteases were being pursued as pharmaceutical drug targets, and protease inhibitors have already been synthesized and used to treat

cardiovascular disease, periodontitis, AIDS, thrombosis, cancer, and other disorders. [3,9,10] Angiotensin-converting enzyme inhibitors have been used as a treatment for cardiovascular conditions for the past 20 years.[3,11] Successful inhibition of these proteases underscores the importance of investigating proteases implicated in other diseases such as AD.

Many proteases have multiple substrates; therefore, strong inhibition could interfere with a protease's non-disease function resulting in side effects. Conversely, weak inhibition could have little to no effect in treating the disease. Therefore a more thorough understanding of active site-substrate interactions is important in understanding and developing treatments strong enough to treat different diseases while minimizing significant side effects.

BACE1 protease activity and biological significance

β-secretase cleaving enzyme 1 (BACE1) was identified in 1999 as a 501aa aspartic protease with a type-I transmembrane domain. Its mRNA is predominantly expressed in the brain and pancreas.[12–15] Type-I transmembrane proteases are single pass membrane proteases with a luminal or extracellular N-terminus and a cytoplasmic C-terminus. The extensive exctacellular N-terminal portion of BACE1 contains the active site.[16] The crystal structure of BACE1 has been identified as shown in Fig. 3.

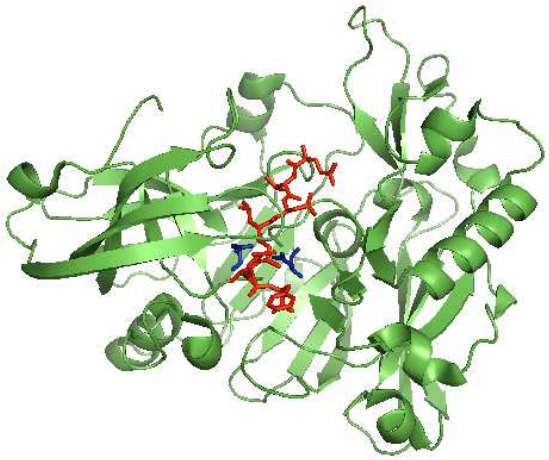


Figure 3: Crystal structure of BACE1A in complex with an inhibitor generated using PyMOL. The inhibitor is shown in red (PDB: 1FKN). The two active site aspartate residues are shown in blue.

BACE1 has five known splicing isoforms resulting from alternative splicing of exons 3 and 4. BACE1A is the full length isoform of 501aa. The other four isoforms yield non frameshift deletions in mRNA splicing, with lengths of 476, 457, 455, and 432 residues known as BACE1B, BACE1C, BACE 455, and BACE1D, respectively.[17–20] The mRNA of all of these isoforms are expressed in vivo.[19] Alternative splicing could negatively affect the activity of the isoforms, but experimental evidence suggests that they all are catalytically active (Johnson, unpublished).

As mentioned, BACE1 is an aspartic protease with optimal activity at low pH values.[21] Aspartic proteases have two aspartyl residues in the active site that use acid-base catalysis to break peptide bonds.[3,4] These aspartates can be seen in Fig. 4. BACE1 is primarily found in intracellular compartments where the acidic environment favors BACE1 cleavage of substrates.[16,22]

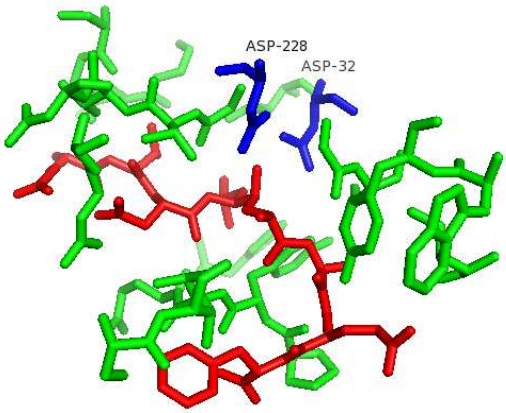


Figure 4: The crystal structure of the BACE1A active site with an inhibitor bound generated using PyMOL. The inhibitor is shown in red. The two active site aspartate residues are shown in blue.

The most widely studied substrate for BACE1 is the amyloid precursor protein (APP) from which the amyloid-β (Aβ) peptides are derived. When BACE1 and another protease, γ-secretase, cleave APP, they form 40 or 42aa peptides known as Aβ-peptides (Fig. 5).[12] These peptides aggregate to form oligomers, which further aggregate to form plaques. Amyloid plaques are one of the hallmark pathologies of Alzheimer's disease (AD). These peptides and their soluble oligomers are toxic to cells and are hypothesized to be directly responsible for some of the symptoms of AD.[23] APP is also a natural substrate for a different protease, α-secretase, which does not lead to the formation of Aβ-peptides (Fig. 5). The processing of APP by BACE1 is minimal in humans without AD, while in brains of AD patients, BACE1 expression is upregulated.[24,25] Though it is involved in the AD pathological pathway, it is likely that there are other native BACE1 substrates in non-disease pathways.

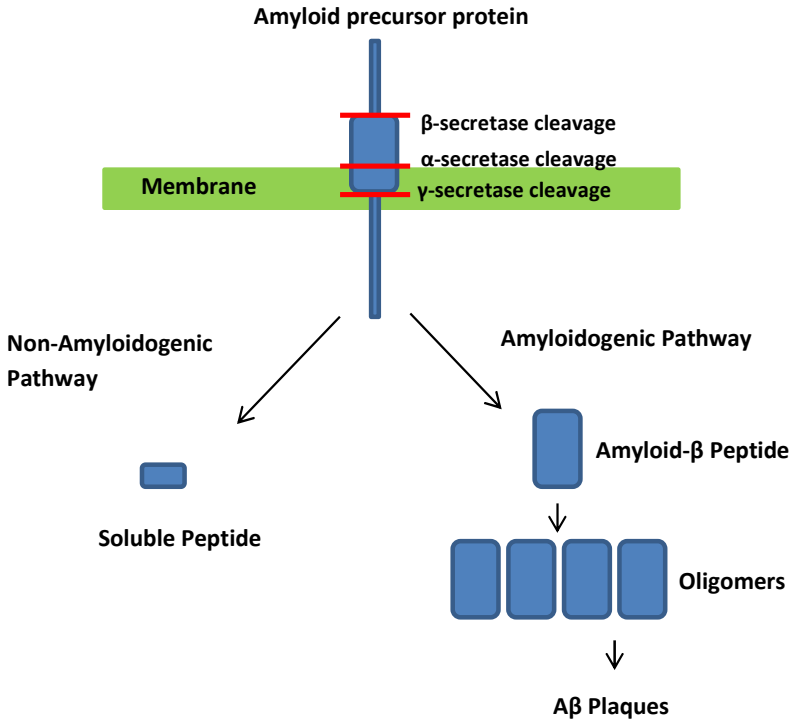


Figure 5: BACE1 in the Alzheimer's disease pathway. Amyloid-β peptides are formed when the amyloid precursor protein is cleaved first by BACE1, then by γ-secretase. These peptides aggregate to plaques, which can be seen in the AD brain. Cleavage by α-secretase and γ-secretase results in non-amyloidogenic processing of APP.

Native BACE1 substrates are currently being studied to determine the potential side effects resulting from BACE1 inhibition. All of the known substrates are transmembrane proteins. BACE1 regulates the activity of neuregulin 1 (NRG1), involved in axonal myelination, the β-subunits of voltage gated sodium ion channels (Na_v1), platelet selectin glycoprotein ligand 1 (PSGL-1), type II-2,6-sialytransferase (ST6Gal-1), and Jagged1 which regulates astrogenesis.[22,26,27] Many of the substrates are involved in stress or injury response pathways in which BACE1 is upregulated.[28]

Known APP mutations in the eight amino acid sequence identified as the β-secretase cut site increase processing by BACE1.[29–31] Three of these mutations, known as the Swedish, Flemish, and London mutations, result in early onset AD.[13,30–32]

Recenctly, a mutation in APP resulting in a change in the amino acid residue corresponding to the BACE1 P2' site was identified that decreased the production of Aβ plaques.[33] Since these mutations occur within the sequence targeted by BACE1, they indicate that BACE1 does show preference for certain amino acid sequences. BACE1 expression is upregulated in the brains of AD patients; therefore, moderate inhibition could, in theory, decrease Aβ peptide production without interfering with the interaction of BACE1 and its other substrates.

BACE2 protease activity and biological significance

BACE2 is also a membrane associated aspartyl protease and is the only known homolog of BACE1. The amino acid sequences for the two proteases share 75% homology.[34] BACE2 is expressed in the heart, kidney, prostrate, brain, and various peripheral tissues, but expression in the brain and peripheral tissues is low.[35,36] BACE2 also has three known splicing isoforms. BACE2A is the full length isoform comprised of 518aas.[37] BACE2B occurs when exon 8 is deleted, and BACE2C occurs when exon 7 is deleted.[35]

BACE2 has the ability to cleave APP, but was not originally implicated in AD due to its lack of expression in neurons. The different expression patterns indicate that BACE2 would likely have different substrates than BACE1, but very few substrates have been identified and are not well understood. It is interesting to note that the BACE2 gene is found in the Down's critical region of chromosome 21, the chromosome that when present in triplicate leads to the formation of Down's syndrome.[35] Although studies have

also shown that BACE2 is influential in insulin receptor tracking, the specific function of BACE2 and its substrates is not well understood. [38]

Methods used to characterize BACE

Proteases are often involved in disease pathways due to their regulatory activity. Characterizing active site−substrate interactions is important if the disease is to be understood and treated. One method used to gain information about a protease is to determine its tissue expression. Expression shows where putative substrates may be located, as proximity is an important consideration in protease-substrate interactions. Expression studies suggest that putative substrates for BACE1 would be within the brain, specifically neurons, and pancreas.[13–15,39]

Another useful method to determine the function of a protease is to knock out the gene in a model organism and identify the phenotypic effects. Though knocking out a gene gives some idea of the function of a protease, it does not identify the substrates or preferred cleavage sequence. BACE1 deficient mice were small and hyperactive compared to wild-type mice, but otherwise they showed a relatively mild phenotype.[40] Once a fundamental understanding of a protease is established, additional in vitro and in vivo methods can be used to identify putative substrates, such as determining the peptide sequence cleaved and relating that to the proteins expressed in the same tissue as the protease.

Several studies have been directed towards identifying the peptide sequence targeted by BACE1. To obtain this data, synthetic octapeptide libraries were created.

This was done to understand the preferred amino acid sequence cleaved by BACE1, hopefully leading to identification of its native substrates as well as to an understanding of potential side effects that might result due to BACE1-targeting therapeutics. In 2001, Turner and Tang did this by creating peptide libraries starting from the peptide sequence derived from the Swedish mutant of APP that results in early onset AD. One amino acid in the sequence was varied while the other seven residues were kept constant.[41,42] This method gave information about the individual amino acids preferred by BACE1 in each part of the sequence; however, it was of necessity biased because it used a set parent sequence. It also was not able to factor in cooperation between or dependence on neighboring amino acids. Though multiple amino acids were identified in each position, the sequence that was hypothesized to be preferentially cleaved by BACE1 was EIDL*MVLD, while the native APP sequence is EVKM*DAEF, and the Swedish mutation sequence, which results in enhanced cleavage by BACE1, is EVNL*DAEF.[41,42]

Quantitative proteomics, specifically stable isotope labeling with amino acids in cell culture (SILAC) has been used to identify potential BACE1 substrates. SILAC involves using at least two different cell cultures. Heavy isotopes are used to grow one culture, and light isotopes are used to grow the other. One cell culture is the control, while BACE1 is overexpressed in the other. Protein differences between the two cultures can be identified using mass spectrometry to discover putative substrates. More than 70 putative substrates were identified.[39] All of the substrates were membrane bound; with three being glycophosphatidylinositol (GPI) linked proteins. This makes sense, because BACE1 is a membrane bound protein, and they would be in close proximity. Several of

the substrates were confirmed in subsequent cell-based studies.[39] One potential problem with this method is that BACE1 is overexpressed, which could lead to the identification of proteins that are only BACE1 substrates under these non-natural conditions.

Proteomic identification of cleavage sites (PICS) is a recently described method designed to accurately identify cleavage sequence preferences and neighbor interactions in the cleavage site that influence protease cleavage (Fig. 6). This method involves creating peptide libraries from organisms with known proteomes, cleaving the peptides in the library with the protease of interest and tagging the newly formed N-termini, isolating and sequencing the peptides corresponding to the original C-terminal (or prime side) sequences by mass spectrometry, and then determing the associated N-terminal peptide sequence through bioinformatcs database searching (Fig. 6). PICS has been used successfully for other proteases and should be useful in determining the preferred cleavage sequence for BACE1.

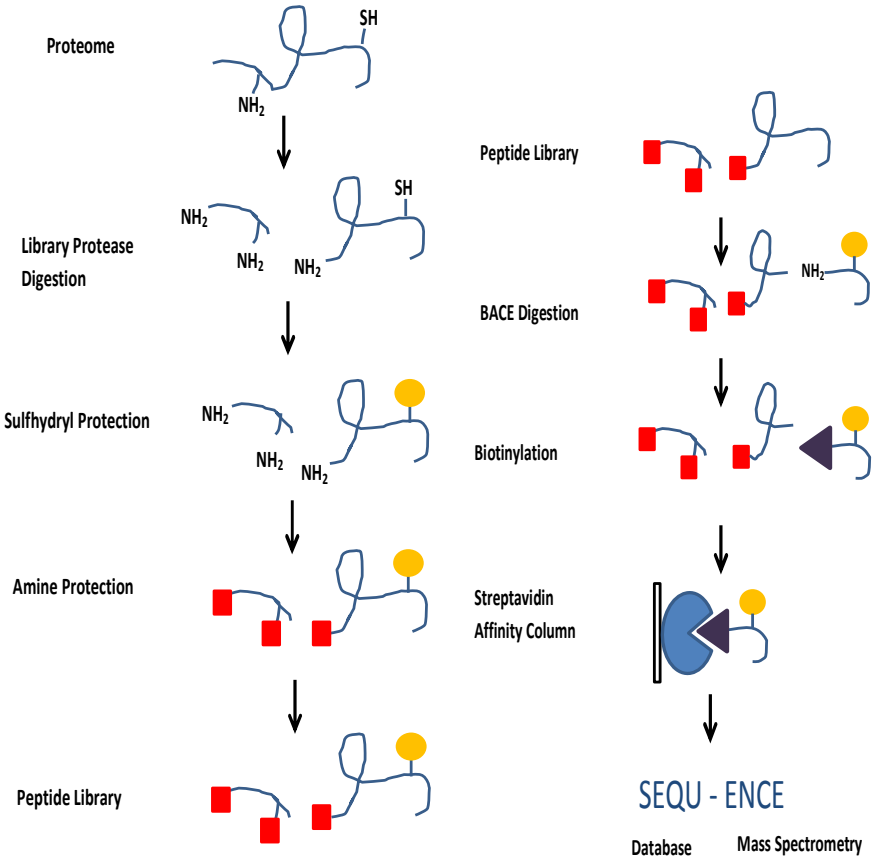


Figure 6: Flowchart for the proteomic identification of cleavage sites method. The proteome of an organism with a sequenced genome is isolated. The proteome is used to make a peptide library, which can be cut by the protease of interest. New N-termini are tagged with biotin, and isolated. The prime peptide sequence is then determined through mass spectrometry, while the non-prime sequence can be found using database searching (figure adapted from ref 59).

Peptide sequencing and Mass spectrometry

Peptide sequencing was first accomplished in the 1950s, when Edman published his method termed 'Edman degradation'. Edman's approach focused on cleaving individual amino acids sequentially from the N-terminus of a peptide, and then identifying the amino acid cleaved using chromatography.[43] This approach was time-consuming, as each cleavage was done individually. Also, it did not work well if the N-terminus was modified or if the sample was not pure. In the 1960's, mass

spectrometry (MS) was first introduced as a new way to sequence peptides. In the 1990s, mass spectrometry essentially replaced Edman degradation, because peptides did not need to be purified before sequencing, modified peptides could be analyzed, peptide fragmentation was much faster, and sensitivity was significantly better.[44–46]

Mass spectrometry has become a powerful tool in peptide sequencing and proteomics, especially since the genome of many organisms have been sequenced. Fig. 7 illustrates the general process of peptide sequencing using mass spectrometry. Most MS instruments that are used for peptide sequencing couple high performance liquid chromatography (HPLC) to the mass spectrometer. This allows peptides to be separated before injection for samples with large numbers of peptides. HPLC elutions can be directly ionized, and the ions can then be guided through the mass spectrometer. The peptides are fragmented within the mass spectrometer, and the mass to charge (m/z) ratios are detected. This information is used to identify the sequence of the peptides. The peptide sequence can then be matched to the organism's proteome, and the original N-terminal or non-prime peptide sequence can be identified.

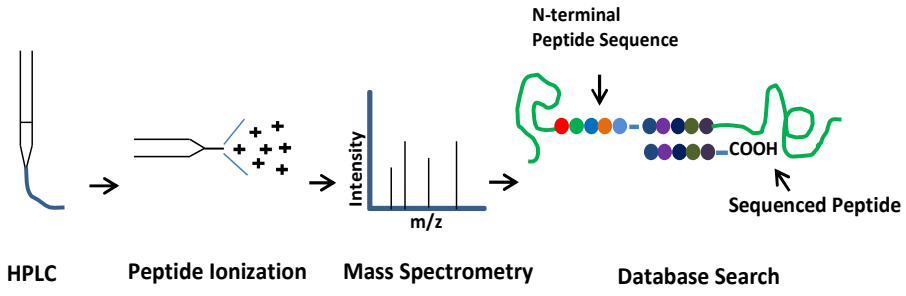


Figure 7: The schematics for peptide sequencing. Peptides are separated by size using high performance liquid chromatography (HPLC), and then ionized. The peptide sequence is determined through mass spectrometry, and then the sequenced peptide is compared to the proteome through bioinformatic database searching to identify the N-terminal sequence.

For peptide sequencing, mass spectrometers need to be tandem MS machines that convert the sample into gas phase ions, select precursor ions based on their mass to charge ratio in an ion trap (m/z), fragment the selected ions in a collision cell, and then detect the m/z ratios of the product ions (Fig. 8). There are multiple mass analyzers that can be used to sequence peptides, such as the linear trap quadrupole (LTQ), time-of-flight (TOF), Fourier-transform ion cyclotron resonance (FT-ICR), and Orbitrap mass analyzers. All of these machines are useful, however using multiple machines or 'hybrid' instruments can increase resolution even more.

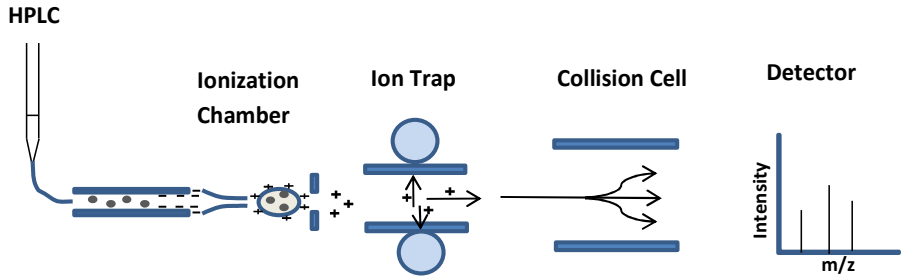


Figure 8: Schematic of a Mass Spectrometer. Peptides are separated by size using HPLC, and then they are ionized within an ionization chamber. The peptide ions enter an ion trap where a single m/z ratio is selected. That peptide ion enters the collision cell, where it is fragmented. The fragment m/z ratios are detected and used to identify the peptide sequence.

In order to enter the mass analyzer, peptides have to be ionized. Electrospray ionization (ESI) is a soft ionization technique that brings peptides into the gas phase without fragmentation. The sample is introduced to the ionization chamber through a tapered needle, which sustains a charge of a few kilovolts compared to the chamber, causing an electric field to be present at the needle's tip. This field creates a charge on the liquid surface, which causes the liquid to be electrostatically dispersed in a spray consisting of multiply charged droplets that are driven towards the mass analyzer. The

conditions within the chamber combine to speed evaporation of the droplets. This reduces the size of the droplets, while increasing their charge density. This creates instability within the droplet, making the droplet explode. This reaction is repeated until ions are desorbed into the gas phase (Fig. 9).[44,45] As tryptic peptides are most commonly used in peptide sequencing, the ions created are generally doubly protonated, though the ions can become multiply charged if they are large.[45]

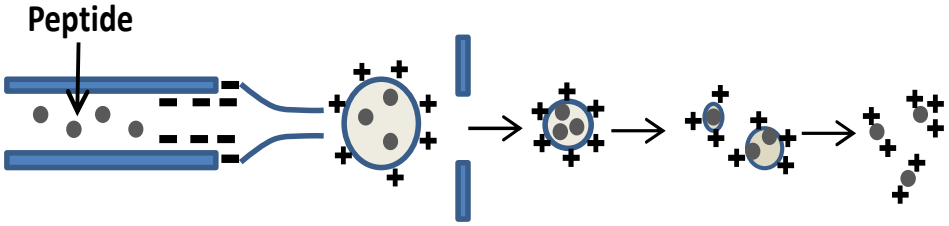


Figure 9: ESI: The peptides in solution are run through a needle with a slightly higher charge than the surroundings. The liquid is thus electrostatically dispersed, which creates multiply charged droplets. The solvent evaporates, which reduces the size of the droplets. This increases the charge density, and sample ions are created due to desorption.[59]

Individual peptide ions of a specific m/z ratio are selected, and then that peptide is fragmented to obtain its sequence.[47] One fragmentation technique is collision-induced dissociation (CID), which can be performed at low energies.[45] In this approach, an inert gas, such as nitrogen or argon, is used to fragment the selected peptide ions by collision. The gas molecules collide with the precursor ions and impart energy, which causes the precursor ion to fragment into product ions.[45] These ions are then scanned in the mass analyzer to produce product ion scans, though scans of fragments less than 7 amino acids long are not analyzed, because they are not very informative.[45] This technique does have limitations; pertinent information is hard to obtain with large, multiply charged peptides, glycosylated and phosphorylated peptides.[44]

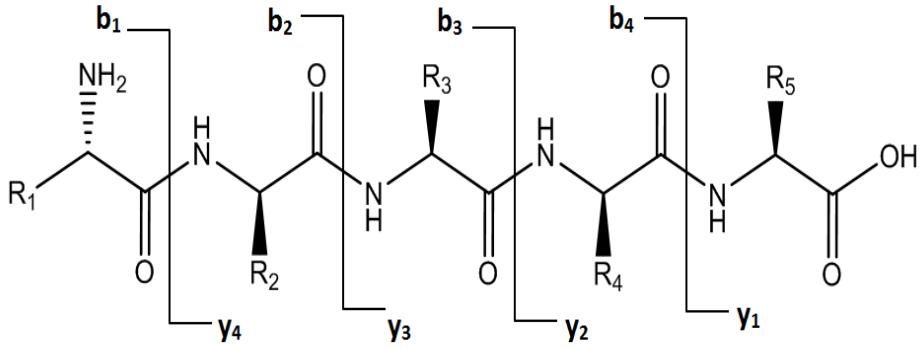


Figure 10: Nomenclature of peptide fragmentation: b and y ions are the most common, as they are produced in CID. These fragments are created when the peptide bond is broken.[48]

B_n and y_n ions (with n = 1, 2, 3, etc.) are the most common ions produced; however, other ion cleavage products can be produced as well (Fig. 10). B_n ions are the N-terminal amide bond cleavage product, and y_n ions are the C-terminal amide bond cleavage product. Fragmentation patterns depend on multiple variables, including the amino acid arrangement, peptide size, and time scale of the instrument used.[49] When a peptide is fragmented into its b and y ions, the structure of the peptide can be determined using bioinformatics and database searching.[44]

There are multiple reaction pathways for the fragmentation of peptides. Charge-remote peptide fragmentation pathways (PFPs) and other ions are formed by the loss of water or ammonia. [49] In some cases, side chain reactions also occur. These reactions can compete with the b-y ion reactions, but for the most part, these reactions do not predominate.[50] The PFP of most interest for general peptide sequencing generates the charge-directed sequence b_n-y_n ions. The formation of these ions can be described by what is known as the 'mobile proton' model, where the proton on precursor ions can migrate between protonation sites, unless there is a basic amino acid, in which case, the

proton stays on the basic amino acid.[50,51] The site of protonation is also the cleavage site in charge-directed pathways, because the protonation helps facilitate the cleavage reaction.[52] All of the PFPs are in competition, which leads to different information appearing in the mass spectra. The most important and informational pathways create b and y ions. These fragments result in sequence information that can be used to determine the primary structure of peptides.

The newest mass analyzer is the Orbitrap. The Orbitrap uses only electrostatic fields to confine and analyze ions.[53] This machine is a very powerful tool in proteomics, specifically peptide sequencing. It is a relatively low cost machine that combines high resolving power, sensitivity, dynamic range, and mass accuracy. When this mass analyzer is coupled to LTQ, it creates a valuable blend of the LTQ and Orbitrap capabilities, mainly the sensitivity and MS/MS capacity of the LTQ, and the high mass accuracy and resolution power of the Orbitrap.[54] The LTQ-Orbitrap is capable of greater than 150,000 mass resolution and mass accuracy measurements of less than 2 ppm when measuring mixtures of multiple peptides.[53,54] In this machine, ionized peptides are directed to the LTQ, which is a linear ion trap. The precursor ions are scanned, and precursor ions with the selected m/z ratio are trapped while all other ions are expelled. The selected ions are fragmented through CID and sent to the C-trap, which injects the ions into the Orbitrap in a pulsed manner.

The Orbitrap is composed of two electrodes, a barrel-shaped outer electrode and a spindle-shaped inner electrode.[53] A DC voltage applied between the two electrodes creates an electrostatic potential distribution, which traps ions. Ions with stable

trajectories can be detected, while ions with unstable trajectories are not detected. Ions with stable trajectories orbit around the inner electrode and move axially along the z-axis. Motion along the z-axis can be described as simple harmonic oscillation. The frequency of axial oscillations is only dependent on the mass and charge of the ions. The axial motion is therefore used in detection, as it is only dependent on the ions themselves. This characteristic is the reason for the Orbitrap's high resolution and mass accuracy.[53] Axial motion of the ions induces an image current in the outer electrodes. This current is acquired as a 'time-domain transient' and Fourier-transform is used to create a frequency spectrum.[53] The resulting spectrum is then converted to a corresponding m/z ratio spectrum.

Once the peptide fragmentation spectra have been acquired, further work is required to determine the peptide sequences. Bioinformatics resources are used to perform these analyses. The bioinformatics software suites use algorithms and statistics to correctly assign sequence identification to mass spectra. In the PICS procedure, only organisms with sequenced genomes are used to create the peptide libraries. Once the C-terminal peptide sequence is sequenced, it can be matched to the original protein through database searching to identify the upstream sequence.

Experimental Procedures

This procedure has been modified from the PICS procedure outlined by Schilling and Overall (Fig. 6).[55] Protease libraries were created by growing cell lines of two sequenced genomes, *E. coli* or *Drosophila melanogaster*. The cells were lysed to release

the proteome, which was then digested with one of three proteases: trypsin, chymotrypsin, or GluC. The new peptides were modified chemically to avoid disulfide bonding and to mask existing amino groups and N-termini, resulting in a peptide library. This library was cleaved by the protease of interest, and new N-termini were biotinylated and then purified by chromatography using a streptavidin column to isolate all the biotinylated peptides. These peptides were sequenced by mass spectrometry and used to search the proteome to determine the entire sequence that was preferentially cleaved by the protease of interest.

Protease Library Creation

Cell growth

Two different cell lines were used for the described studies: *E. coli* DH5α and *Drosophila melanogaster* Schneider 2 cells (S2).

DH5α

E. coli DH5α cells were grown in 5 mL Luria broth (LB) at 37 ^{o}C for 12-14 hours. This was used to inoculate 300 mL of LB broth, which was grown to an optical density of 0.4-0.6 at 37 °C. Once the cells reached the desired density, they were centrifuged at 400xg for 20 min at 4 °C. The cell pellet was washed with 5 mL of 1x PBS and centrifuged using the same settings. The PBS wash and centrifuge step were repeated.

S2

Drosophila S2 cells were cultured and grown in Express 5 Serum Free medium (Gibco) containing 16 mM L-glutamine (SFM). S2 cells were grown at 28 ^{o}C in a 75mm^2 flask. They were then scaled up to 400 mL of SFM in a shaking incubator at 28 ^{o}C and

grown for 3-4 days. Once the cells reached the desired density, they were centrifuged at 400xg for 20 min at 4 °C. The cell pellet was washed with 5 mL of 1x PBS and centrifuged using the same settings. The PBS wash and centrifuge steps were repeated.

Cell lysis

DH5α

The cells were first re-suspended in 20 mL of lysis buffer (10mM HEPES pH 7.5, 10 mM EDTA, 1 mM PMSF, 10 μM E-64). Repeated cycles of ultrasonication (Fisher Scientific Sonic Dismembrator Model 500) were used to lyse the cells on ice. DH5α cells were lysed at 55% sonication 10 times for 10 seconds with 20 second breaks in between. The sample was then centrifuged for 20 min at 20,000xg at 4 °C.

S2

The cells were first re-suspended in 20 mL of lysis buffer (10 mM HEPES pH 7.5, 10 mM EDTA, 1 mM PMSF, 10 μM E-64). Repeated cycles of ultrasonication (Fisher Scientific Sonic Dismembrator Model 500) were used to lyse the cells on ice. S2 cells were lysed at 40% sonication 10 times for 5 seconds with 15 second breaks in between.

Proteome modification and precipitation

DH5α and S2

The resuspended peptide solution was adjusted to a final concentration of 100 mM HEPES and 5 mM DTT. This was incubated at 25 °C for 1 hour. The concentration of iodoacetamide in solution was adjusted to 20 mM. This was incubated

at 25 °C for 1 hour. Again, the concentration of DTT in solution was adjusted to 5 mM (total concentration of 10 mM) and incubated at 25 °C for 15 min. Then, the concentration of trichloroacetic acid (TCA) was adjusted to 15% and the sample was incubated on ice for 1hr.

Protease library creation

DH5α

The sample was centrifuged at 20,000xg for 20 min at 4 °C. The protein pellet was washed twice with 1 mL of chilled methanol and then air-dried for 5 min. The protein pellet was re-suspended with 5 mL of 20 mM NaOH. The sample was sonicated if necessary to completely re-suspend protein. Once the protein was re-solubilized, the concentration of the solution was adjusted to 200 mM HEPES (pH 7.5). The sample was centrifuged at 20,000xg for 10 min at 4 °C. The protein concentration was determined using a micro Bradford assay. One of three proteases (trypsin, chymotrypsin, or GluC) was added at a 1:100 (mass/mass) ratio of protease to library. The digestion was incubated for 16 hrs at 37 °C.

S2

The sample was centrifuged at 20,000xg for 20 min at 4 °C. The protein pellet was washed twice with 1 mL of chilled methanol and then air-dried for 5 min. The protein pellet was re-suspended with 15 mL of 20 mM NaOH. The sample was sonicated on ice for 1 hr at 35% sonication. Once the protein was re-solubilized, the concentration of the solution was adjusted to 200 mM HEPES (pH 7.5). The sample was centrifuged at

20,000xg for 10 min at 4 °C. The protein concentration was determined using a micro Bradford assay. One of three proteases (trypsin, chymotrypsin, or GluC) was added at a 1:100 (mass/mass) ratio of protease to library. The digestion was incubated for 16 hrs at 37 °C.

Acetylation of cysteine residues

DH5α and S2

Protease activity was stopped with 1 mM PMSF and 1M guanidine hydrochloride, and the sample was centrifuged at 20,000xg for 10 min at 4 °C. The concentration of DTT in solution was adjusted to 5 mM, and the sample was incubated for an additional hour at 37 °C. The concentration of iodoacetamide was added to a final concentration of 40 mM, and the sample was incubated at 37 °C for 1.5 hr. The concentration of DTT in solution was adjusted to 15 mM (total concentration of 20 mM) followed by a 10 min incubation. The concentrations of formaldehyde and sodium cyanoborohydride in solution were adjusted to 30 mM incubated for 2 hrs. This step was repeated for a total concentration of 60 mM formaldehyde and sodium cyanoborohydride, and the sample was incubated for 16 hrs.

Size exclusion chromatography of peptide library

The concentration of the solution was adjusted to 100 mM glycine, and the sample was incubated at 25 °C for 0.5 hr. Two G10 columns (GE) arranged in tandem were equilibrated with 10 mM sodium phosphate and 20% acetonitrile (pH 2.7). The system was connected to an Äkta FPLC system monitoring the absorbance at 280 nm (A280) and the conductivity vs. mL loaded plot in real time. The sample was loaded and

washed with the buffer until an increase in A280 was observed. Fractions were collected until an increase in conductivity was observed. The acetonitrile was removed by vacuum evaporation. The samples were acidified to 0.5% trifluoroacetic acid and degassed by applying mild vacuum. The libraries were purified by C18 solid-phase extraction using the manufacturer's protocol with an elution solution of 80% acetonitrile (Thermo Scientific). A volume of 3 mL was used for all wash, elution, and sample loading steps. The peptide concentration was determined using a BCA assay. Acetonitrile was removed from the samples by vacuum evaporation until the solution reached a concentration of 1.5-2 mg/mL. The samples were aliquoted into 200 µg peptide libraries.

Cleavage site sequence determination

Peptide library cleavage and product isolation

The concentration of sodium acetate (pH 4.5) in the peptide libraries was adjusted to 20 mM with 1 M sodium acetate (pH 4.5), and the peptide library concentration was adjusted to 1 mg/mL. BACE1 or BACE2 was added in a ratio of protease to library between 1:50 and 1:1,000 (mass/mass). The digestion was incubated for 16 hr at 37 °C. The protease was deactivated by heating the sample at 80 °C for 20 min. The new N-termini were biotinylated by adding 0.5 mM sulfo-NHS-SS-biotin (Pierce, Inc.), and the sample was incubated for 2 hrs at 25 °C.

Half a milliliter of high capacity streptavidin-Sepharose resin (GE) was equilibrated in a spin column with a 10 µm filter with buffer (50 mM HEPES, 150 mM NaCl, pH 7.5). All spin steps were performed at 200xg for 1 min. The sample was added to the resin and incubated at 22 °C for 30 min, then centrifuged. The eluent was

reapplied, and the centrifuge step repeated. The flow-through was discarded and the column was washed 10 times with 0.5 mL of buffer (50 mM HEPES, 150 mM NaCl, pH 7.5). The sample was incubated with the elution buffer, 0.5 mL (50 mM HEPES, 20 mM DTT, pH 7.5), for 1 hr at 25 °C. The sample was centrifuged, and the eluent was collected. The elution step was repeated, and the two elution fractions were pooled. Another C18 solid-phase extraction was performed as earlier, but with an elution buffer of 80% acetonitrile and 0.1% TFA. The sample was dried down with vacuum evaporation, and the sample was sent to the Center for Mass Spectrometry and Proteomics at the University of Minnesota, St. Paul campus for mass spectrometric analysis with the LTQ Orbitrap Velos.

Mass Spectrometry and Database Analysis

LTQ-Orbitrap settings and database search parameters

The standard settings for Orbitrap peptide analysis were used. CID and single charge mode specifications were used.

The data analysis was performed using software available through MSI, the TINT Proteomics Software pipeline (https://tint.msi.umn.edu/). An integrated identification workflow was used to start SEQUEST searches to analyze the raw mass spectrometric data (Fig 11).

A B

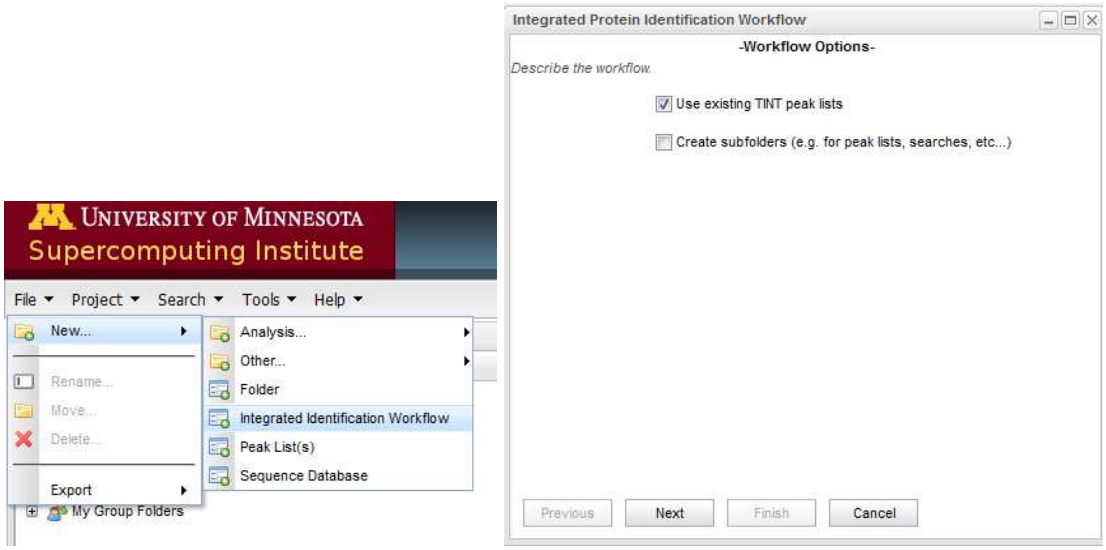


Figure 11: This figure showed the TINT website. A) An integrated identification workflow was started. B) An existing TINT peak list was chosen.

The integrated identification workflow contained multiple windows with options for different parameters (Figs. 11-14). This workflow gave options for the different parameters that were chosen. Fig. 12 illustrates how to select the sample, the database, and start the database search using SEQUEST.

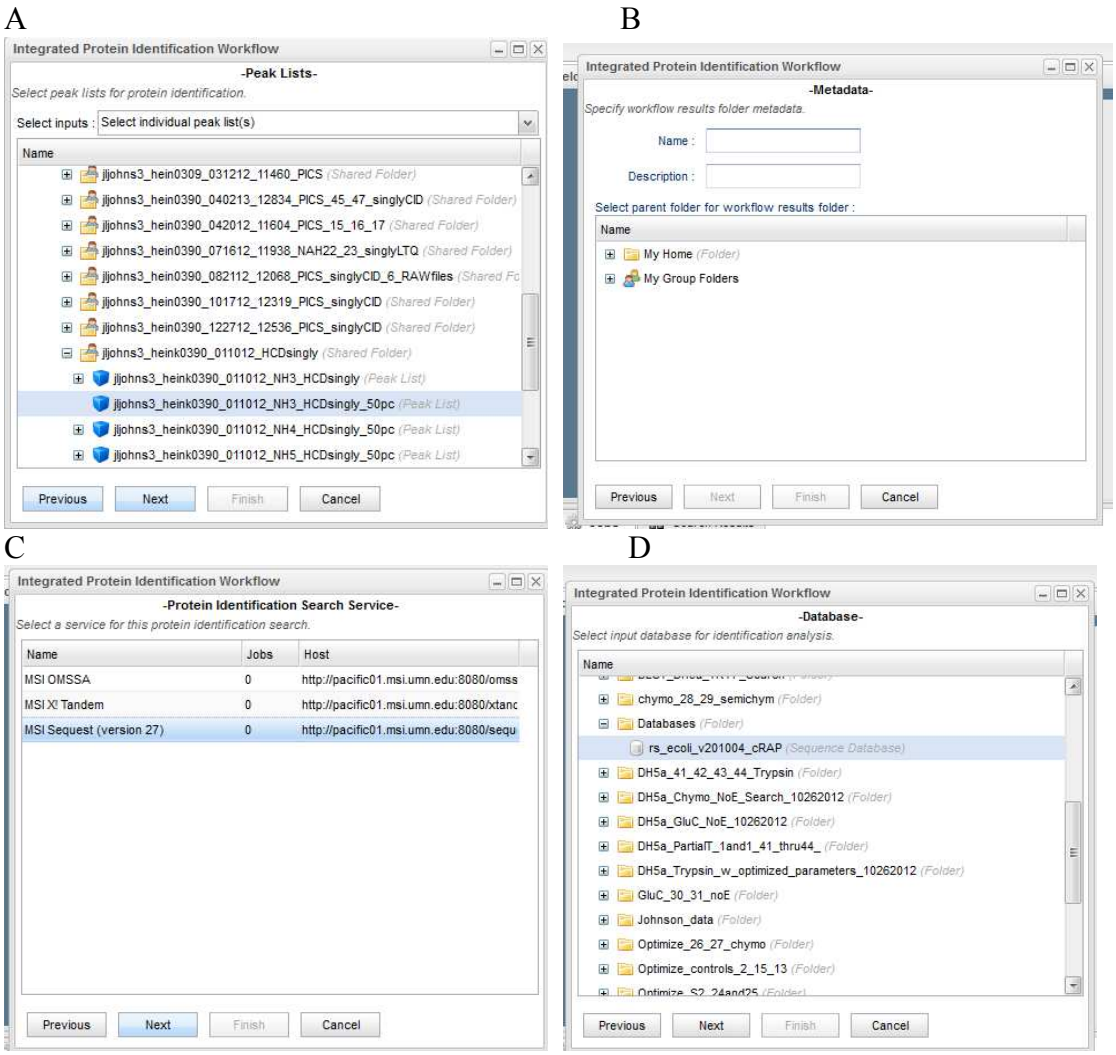


Figure 12: Four of the windows in the integrated identification workflow on TINT were shown. A) The samples of interest were uploaded. B) The output name and parent folder for the samples were entered. C) MSI SEQUEST was the only search program option. D) The database necessary for the sample was entered.

The main search parameters for the integrated identification workflow are shown in Fig. 13. Here a peptide and fragment tolerance of 1 amu, maximum number of cleavage sites as 1, and tryptic search limited to C-termini for all tryptic libraries were chosen (Fig. 13A). There were three fixed amino acid modifications due to the chemical modifications introduced during the procedure: cysteine (57.0215), lysine (28.0313), and

the N-terminal amino acid (87.9983) (Fig. 13 B). There was one variable amino acid modification: methionine (15.994915) (Fig. 13 C). For any library made with chymotrypsin or GluC, a partial no enzyme search was used (Fig. 13 D).

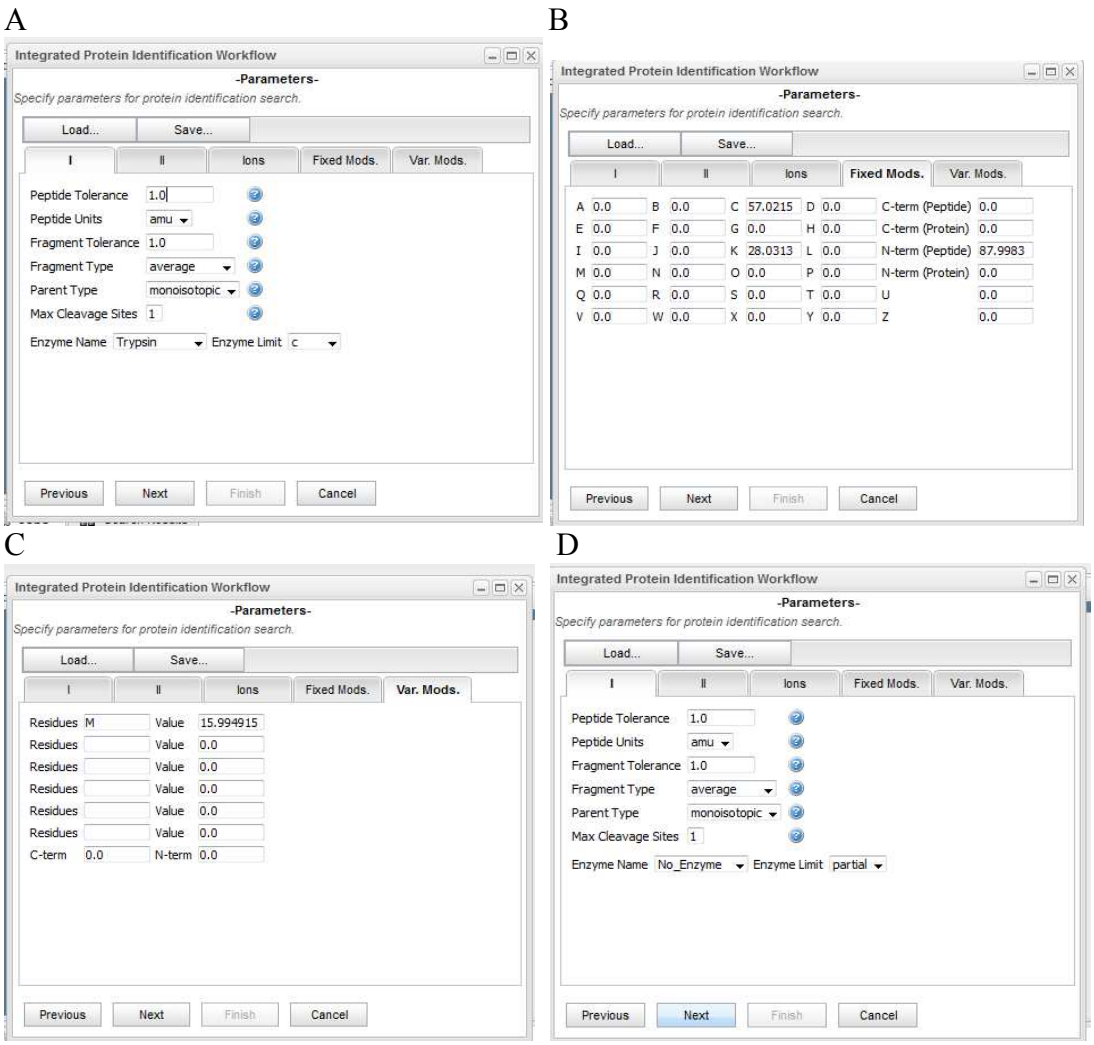
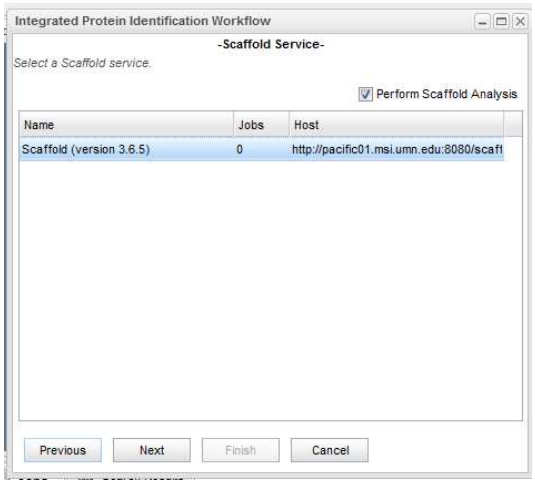


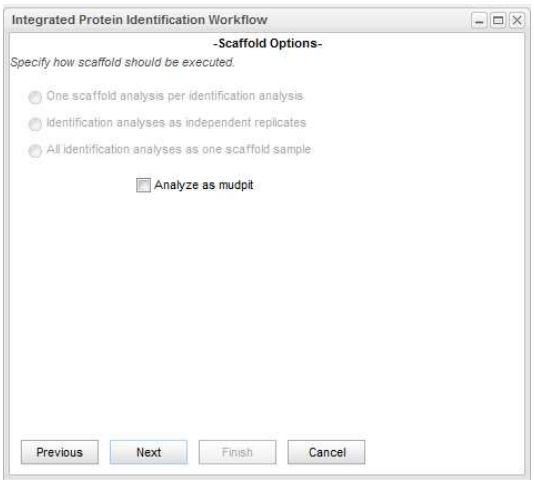
Figure 13: The integrated identification workflow parameter options were displayed here. A) This was the main set of parameters to choose from. B) Fixed modifications for amino acids were entered in this window. C) The variable modifications for amino acids were entered in this window. D) This set of parameters was used for any library that was not created with trypsin.

The final three windows in the integrated identification workflow were shown in Fig. 14. They apply to Scaffold, which was a viewing program. There was only one choice of viewing program (Fig. 14A). In the next window, two different options were used. For individual samples, 'one scaffold analysis per identification analysis' was chosen. For multiple search results, then 'all identification analyses as one scaffold sample' was chosen (Fig. 14B). The parameters in the final window could be changed when the search output file was opened in Scaffold (Fig. 14C).

A



B



C

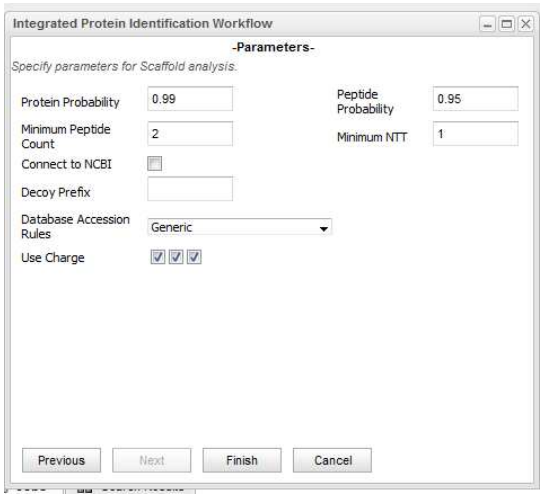


Figure 14: The final windows in the integrated identification workflow were shown. A) There was only one option here: choose Scaffold. B) Either view individual samples or multiple samples added together. C) These options were for the scaffold viewing program and were changed in the scaffold program.

Once the samples were run through the SEQUEST search program, they were opened in Scaffold (Fig. 15). Settings of min. protein of 80%, min. # peptides 2, and min. peptide of 90% were chosen. Then the peptide report was exported as an Excel file.

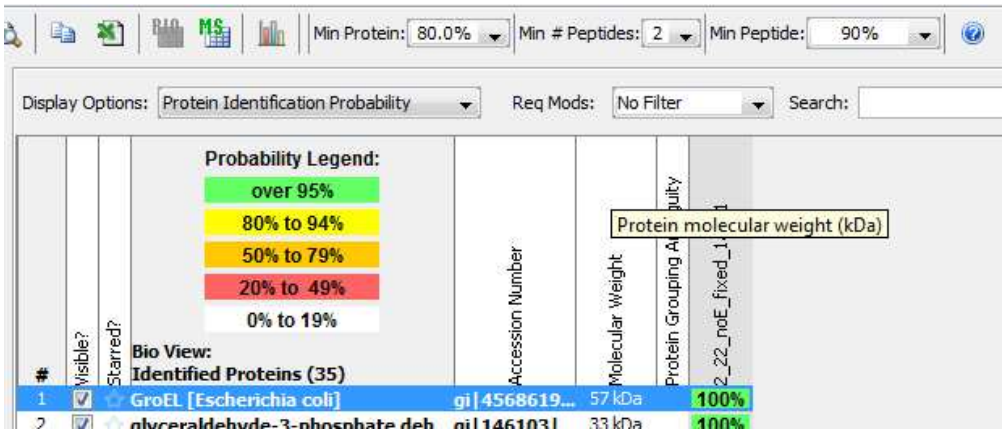


Figure 15: This was the Scaffold program that showed optimal settings for exporting the peptide list.

The peptide list was pasted into the CLIP-PICS website:

http://clipserve.clip.ubc.ca/pics/cgi/PICS.cgi

The analysis tab was opened and the peptide list was uploaded. The sample was named, and the protease and organism used to create the library were chosen. Subsite cooperativity was analyzed for all samples (Fig. 16).

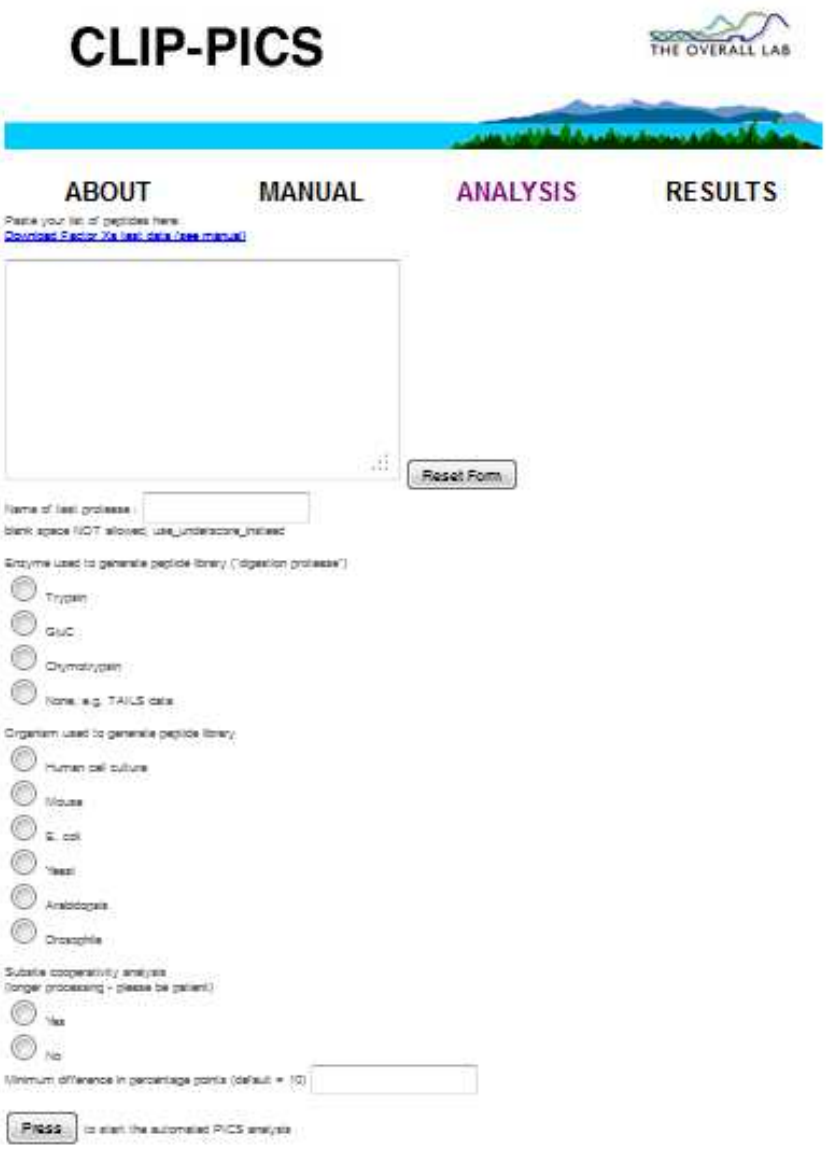


Figure 16: This was the Analysis page on the CLIP-PICS website. To obtain data from the Mass Spectrometric analysis, the peptide list from Scaffold was uploaded, the protease and organism used to generate the peptide library were chosen, and substrate cooperativity was analyzed (http://clipserve.clip.ubc.ca/pics/cgi/PICS.cgi).

Results

Control optimization: Tryptic DH5α library cut with GluC

Two different protease-to-library ratios were used to cut tryptic DH5α libraries with GluC, a very specific protease that only cuts after acidic amino acids. Parameters

were optimized using a sample created at a 1:100 GluC:Library ratio (Table 1-8). This information was used to optimize search parameters so that greater than 95% of the correct peptides were selected.[56] Many search programs are available, each using a different algorithm to identify peptides. The level of error is measured differently for each method. All of the programs matched the raw Orbitrap sequencing data to amino acid sequences in the sequenced proteome used to generate the peptide library. This allowed us to determine the original amino acid sequence cut by the protease of interest. The parameters optimized were: enzyme search specifications, the level of error allowed, and whether the modifications occurring with the PICS procedure were fixed or variable. The enzyme specification could be tryptic, semi-tryptic or non-specific. These affect the results by setting limits on the database search. Setting tryptic enzymatic specificity caused the search program to compare our mass spectrometer sequencing data with a hypothetical tryptic digest of the organism used. This decreased the length of the search; however, the tradeoff is that some peptides could be missed due to the constraints. Our goal was to minimize the search time while maximizing the number of peptides identified.

In the PICS procedure, there were three separate modifications made to the peptide libraries: carboxyamidomethylation of cysteine, dimethylation of lysine, and thioacylation of the N-termini.[55] Thioacylation had to be a fixed modification, as this was what identified the peptides cleaved by the protease of interest. The other two modifications may or may not have been present in all of the peptides. Therefore, fixed and variable modifications were tested for each method. Two different methods were

optimized and compared to determine which one performed better for this data. The two different methods were: SEQUEST search data viewed with Scaffold and SEQUEST search data run through the Trans-Proteomic Pipeline (TPP) with Peptide Prophet. SEQUEST was the first search program to be optimized because it was the most straightforward and fastest method, and it was available free of charge through the University of MN. Scaffold is a viewing program that is used to both view the SEQUEST data and set the percent false discovery rate (FDR) that was used for the fragment data. The second method also used SEQUEST; however, the next step was to put the SEQUEST data through another search program, TPP. TPP is an online data analysis pipeline for processing mass spectrometric data. Peptide Prophet was the most important step in TPP for this data as it validates data from other search programs, such as SEQUEST. The peptides identified using these different methods were uploaded onto the website, CLIP-PICS (http://clipserve.clip.ubc.ca/pics/), specifically designed for the PICS procedure.[57] This website matched the list of peptides with the proteome of the peptide library and the protease used to cut the peptide library. These data were used to re-validate the peptides and list the information in tables.

Two scaffold parameters were optimized using SEQUEST parameters of partial tryptic enzyme specification, fixed modifications on lysine and cysteine, and peptide and fragment tolerance of 1 amu: minimum peptide and minimum protein threshold tolerance, These optimal values gave 80% protein and 90% peptide matches. Of the 241 peptides identified, 95.4% of the peptides had glutamate or aspartate in P1, matching GluC specificity (Table 1). 80% protein match means that 80% of the sequenced peptide

matched a protein from the organism's proteome. 90% peptide match means that 90% of the sequenced peptide identified with a peptide from the program's hypothetical cleavage of the proteome.

Table 1: A tryptic DH5α library was cut with GluC as a control to optimize Scaffold parameters. The parameters that yielded the most peptides without losing specificity were 80% protein and 90% peptide thresholds.

Scaffold optimization for Sample 3			
Probability threshold Protein_Peptide	**80_80**	**80_90**	**80_95**
# of Peptides	269	241	215
% Match	94.8	95.4	95.8
Probability threshold Protein_Peptide	**90_90**	**90_95**	**95_95**
# of Peptides	232	215	118
% Match	95.7	95.8	95.8

Scaffold was also optimized using the same tryptic DH5α library with a 1:100 GluC:Library ratio with double the amount of the control run through the Orbitrap. The results in Table 2 agree with those in Table 1; a threshold of 80% for proteins and 90% for peptides gave the best values for total number of peptides and for the number of peptides. Of the 272 peptides identified 97.8% matched the known GluC specificity. The scaffold parameters for both runs of the sample were the same. An 80% match for proteins and a 90% match for the peptides gave a large percent error; however, later steps revalidated the identified peptides, which allowed us to be confident that the peptides were cleaved by the protease of interest and were relevant.

Table 2: A tryptic DH5α library was cut with GluC as a control to optimize Scaffold parameters. Double the amount of sample was run through the mass spectrometer compared to the sample in Table 2. The parameters that yielded the most peptides without losing specificity were 80% protein and 90% peptide thresholds.

Scaffold optimization for Sample 3_50			
Probability threshold Protein_Peptide	**80_80**	**80_90**	**80_95**
# of Peptides	287	272	243
% Match	97.2	97.8	98.3
Probability threshold Protein_Peptide	**90_90**	**90_95**	**95_95**
# of Peptides	263	243	149
% Match	98.1	98.3	98.6

The various SEQUEST parameters were optimized using the control to determine which parameters resulted in the highest number of peptides matching the specificity of GluC cleavage: glutamate or aspartate in P1. The parameters that were optimized included: enzyme search specifications, peptide and fragment tolerance in atomic mass units (amu), and fixed or variable amino acid modifications. All other SEQUEST parameters were left in their default settings in accordance with the recommendation of Dr. LeAnn Higgins, the research assistant professor responsible for operating the LTQ Orbitrap Velos at the Center for Mass Spectrometry and Proteomics at the University of Minnesota. The results of these optimization searches can be seen in Table 3. The parameters that gave the best results were peptide and fragment tolerance of 1 amu, C-terminal tryptic enzymatic search, and fixed modifications for lysine, cysteine, and the N-termini. These parameters resulted in 243 peptides, 96.3% of which had the GluC preferred glutamate (E) or aspartate (D) in P1.

Table 3: A tryptic DH5α library was cut with GluC as a control. Sequest search parameters were optimized. The most peptides were found using a fragment tolerance of 1amu. The search parameters that gave the highest percent of matching peptides and the most peptides was a C-terminal tryptic digest with a peptide and fragment tolerance of 1 amu..

SEQUEST Data Optimization for Tryptic DH5α Library cut with GluC with Sample 3									
SEQUEST Search Specification	**Partial Tryptic digest with variable modifications**								
Peptide/Fragment Tolerance (amu)	1_1	1_1.5	1_2	1.5_1	1.5_1.5	1.5_2	2_1	2_1.5	2_2
# of Peptides	241	189	178	247	177	180	237	175	170
% Matching Peptides	95	95.3	96.6	94	94.4	95	95	96	96.5
SEQUEST Search Specification	**C-terminal Tryptic digest with variable modifications**								
Peptide/Fragment Tolerance	1_1	1_1.5	1_2	1.5_1	1.5_1.5	1.5_2	2_1	2_1.5	2_2
# of Peptides	243	193	180	244	191	182	244	182	178
% Matching Peptides	94.2	94.8	95	94.2	93.7	94.5	94.3	93.9	94.9
SEQUEST Search Specification	**Non-specific digest with variable modification**								
Peptide/Fragment Tolerance (amu)	1_1	1_1.5	1_2	1.5_1	1.5_1.5	1.5_2	2_1	2_1.5	2_2
# of Peptides	233	172	167	230	173	167	219	155	145
% Matching Peptides	95.3	94.8	94	94.8	94.8	94.6	95.4	95.5	94.5
SEQUEST Search Specification	**Partial Tryptic digest with fixed modifications**								
Peptide/Fragment Tolerance (amu)	1_1	1_1.5	1_2	1.5_1	1.5_1.5	1.5_2	2_1	2_1.5	2_2
# of Peptides	241	182	172	243	182	192	229	174	164
% Matching Peptides	95.4	96.1	97.1	94.7	96.1	96.5	95.6	95.9	95.7
SEQUEST Search Specification	**C-terminal Tryptic digest with fixed modifications**								
Peptide/Fragment Tolerance (amu)	1_1	1_1.5	1_2	1.5_1	1.5_1.5	1.5_2	2_1	2_1.5	2_2
# of Peptides	243	197	189	249	196	179	241	183	164
% Matching Peptides	96.3	95.9	96.2	94.8	95.9	95.5	95.8	95.1	95.7
SEQUEST Search Specification	**Non-specific digest with fixed modification**								
Peptide/Fragment Tolerance (amu)	1_1	1_1.5	1_2	1.5_1	1.5_1.5	1.5_2	2_1	2_1.5	2_2
# of Peptides	242	190	171	250	185	171	226	161	163
% Matching Peptides	93.8	93.6	93.5	94.4	94.6	93.5	94.7	95.1	95.7

The same sample was run through the Orbitrap at double the concentration.

Optimization searches were run to confirm the results; however, a fragment tolerance of

1 amu was used (Table 4). More peptides were identified in all cases due to the higher concentration of sample being run through the mass spectrometer. There was not a single set of parameters that yielded the best results in all cases, so the optimized parameters from the original sample, C-terminal tryptic digest with a peptide and fragment tolerance of 1 amu, were chosen for all subsequent tryptic DH5α library samples. For the 50x tryptic DH5α library cut with GluC, 279 peptides were identified 97.1% of which had E or D in P1 once again consistent with the preference of GluC , confirming the validity of the method. A second set of parameters gave similar results; the non-specific enzyme specification with fixed modifications and peptide and protein tolerance of 1 amu were also considered to be optimal parameters for this sample.

Table 4: A tryptic DH5α library was cut with GluC as a control run at double the concentration compared to Table 2. SEQUEST search parameters were optimized. The most peptides were found using a fragment tolerance of 1.

SEQUEST Data Optimization for Tryptic DH5α Library cut with GluC with Sample 3-50x			
SEQUEST Search Specification	**Partial Tryptic digest with variable modifications**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	273	275	271
% Matching Peptides	97.9	97.5	97
SEQUEST Search Specification	**C-terminal Tryptic digest with variable modifications**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	277	275	270
% Matching Peptides	97.5	97.5	97.1
SEQUEST Search Specification	**Non-specific digest with variable modification**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	270	266	255
% Matching Peptides	94.8	96.6	96.5
SEQUEST Search Specification	**Partial Tryptic digest with fixed modifications**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	272	274	270
% Matching Peptides	97.8	97.1	97.1
SEQUEST Search Specification	**C-terminal Tryptic digest with fixed modifications**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	279	275	268
% Matching Peptides	97.1	97.4	97.7
SEQUEST Search Specification	**Non-specific digest with fixed modification**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	281	279	264
% Matching Peptides	96.8	96.4	97.3

Once the optimization for SEQUEST search parameters was completed, TPP parameters were optimized using the second set of mass spectrometric data with double the amount of sample loaded. The best search results from SEQUEST, C-terminal tryptic digest with a peptide and fragment tolerance of 1 amu was used to optimize the enzyme specification, peptide length, probability cut-off, and Peptide Prophet parameters. Tryptic and semi-tryptic searches resulted in the same peptide list (Table 5). There was no difference between the two searches; therefore, the semi-tryptic was the parameter setting used for all subsequent searches. Probability represents the percent error allowed.

While values of 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, and 0.1 were evaluated, only three of these values are shown in Table 5. The optimal value was 0.08, which allowed for an error of 8%. The shortest peptide of 7 resulted in the best data. Finally, three different Peptide Prophet parameters were optimized: choosing to not use isotope-coded affinity tagged (icat) data, not using the number of missed cleavages (NMC), and using accurate mass binning. These settings did not result in a significantly different number of peptides that matched the expected amino acids of E or D.

Table 5: A tryptic DH5α library cut with GluC was used as a control to optimize parameters for SEQUEST and TPP. The optimal parameters were semi-tryptic enzyme specification, 0.08 probability cut-off, peptide length of 7, and default Peptide Prophet settings.

TPP Parameter Optimization for Sample 3_50x			
TPP Protease Search	**Tryptic**	**Semi -Tryptic**	**Non-specific**
# of Peptides	182	182	192
% Match	87.9	87.9	85.4
TPP Probability	**0.06**	**0.08**	**0.1**
# of Peptides	182	181	180
% Match	87.9	88.4	88.4
TPP Peptide length	**6**	**7**	**8**
# of Peptides	201	181	159
% Match	87	88.4	88
Peptide Prophet	**no icat**	**Accurate mass**	**no NMC**
# of Peptides	182	187	187
% Match	88.5	87.1	87.7

After the TPP parameters had been optimized, SEQUEST searches were run through TPP using the optimized parameters to refine the SEQUEST parameters used in TPP for the original mass spectrometric data (Table 6). A fragment tolerance of 1 amu is shown. Several searches were run with other fragment tolerances, but they resulted in fewer peptides and lower percentages of matching peptides. The SEQUEST parameters

resulting in the highest number of peptides with E or D in P1 were non-specific enzyme specification with a peptide tolerance of 1.5 amu (Table 6). However, using a peptide tolerance of 1 amu, the difference between the two sets of parameters was only 4 peptides or 0.1%. Both fixed and variable modifications gave similar results, but fixed modifications had a slightly higher percent of matching peptides. The set of parameters used for all further samples was: peptide tolerance of 1.5 amu, fixed modifications, and non-specific enzyme specification (Table 6). This set of parameters yielded 226 peptides, 92.5% containing E or D in the P1 position.

Table 6: A tryptic DH5α library cut with GluC was used as a control to optimize parameters for SEQUEST and TPP. The optimal parameters were non-specific enzyme specification with a peptide tolerance of 1.5 and a fragment tolerance of 1.

TPP Data Optimization for Tryptic DH5α Library cut with GluC for Sample 3			
SEQUEST Search Specification	**Partial Tryptic digest with variable modifications**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	183	199	196
% Matching Peptides	90.2	86	86.9
SEQUEST Search Specification	**C-terminal Tryptic digest with variable modifications**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	176	219	217
% Matching Peptides	90.9	80.4	82
SEQUEST Search Specification	**Non-specific digest with variable modification**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	219	230	221
% Matching Peptides	91.8	90	91.4
SEQUEST Search Specification	**Partial Tryptic digest with fixed modifications**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	178	196	199
% Matching Peptides	92.2	89.1	87
SEQUEST Search Specification	**C-terminal Tryptic digest with fixed modifications**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	225	215	222
% Matching Peptides	79.1	71.9	79.3
SEQUEST Search Specification	**Non-specific digest with fixed modification**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	222	226	285
% Matching Peptides	92.4	92.5	88

The TPP parameters were optimized using the same sample run through the Orbitrap at double the concentration to confirm the results. The results were similar, with non-specific enzyme parameters yielding the best results; however, variable modifications resulted in a significantly larger number of peptides than for any other set of parameters (Table 7). There was not one set of parameters that was significantly better

than the others, though non-specific enzyme specification, fragment tolerance of 1.5 amu, and variable modifications for K and C yielded the highest number of peptides (287) with greater than 90% of those peptides matching the specificity of GluC.

Table 7: A tryptic DH5α library cut with GluC was used as a control to optimize parameters for SEQUEST and TPP. The optimal parameters were non-specific enzyme specification with a peptide tolerance of 1 and a fragment tolerance of 1.

TPP Data Optimization for Tryptic DH5α Library cut with GluC for Sample 3_50x			
SEQUEST Search Specification	**Partial Tryptic digest with variable modifications**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	175	172	173
% Matching Peptides	90.3	91.8	91.4
SEQUEST Search Specification	**C-terminal Tryptic digest with variable modifications**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	181	178	178
% Matching Peptides	88.4	91	91
SEQUEST Search Specification	**Non-specific digest with variable modification**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	281	287	209
% Matching Peptides	90	91	89.5
SEQUEST Search Specification	**Partial Tryptic digest with fixed modifications**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	169	167	166
% Matching Peptides	91.7	93.4	93.4
SEQUEST Search Specification	**C-terminal Tryptic digest with fixed modifications**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	183	173	174
% Matching Peptides	88.6	91.9	92.5
SEQUEST Search Specification	**Non-specific digest with fixed modification**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	202	201	265
% Matching Peptides	94.1	92.7	89.8

Once all of the search methods had been optimized, the results were compared to determine the best search method for identifying peptides within our data sets (Table 8). The SEQUEST/Scaffold optimized method was used to analyze tryptic DH5α libraries

cut with the protease of interest, due to the larger percentage of peptides matching GluC cleavage specificity.

Table 8: A tryptic DH5α library cut with GluC was used to optimize parameters for several data analysis methods. The method that yielded the most peptides with the highest percent match for GluC specificity was using the SEQUEST search program combined with the Scaffold viewing program.

Optimal Search Parameters for Sample 3_50x		
	TPP	Scaffold
# of Peptides	287	279
% Match	91	97.1

The CLIP-PICS website designed by Overall and Schilling was used to re-validate and create heat maps to visualize the data. Heat maps for each of the 2 methods of analysis were created (Figure 17). The only amino acids for which there is a preference in P1 are E and D. These samples all display the specificity of GluC, once again confirming that the PICS method can be confidently used to obtain information about the amino acid sequence preferentially cleaved by the protease of interest.

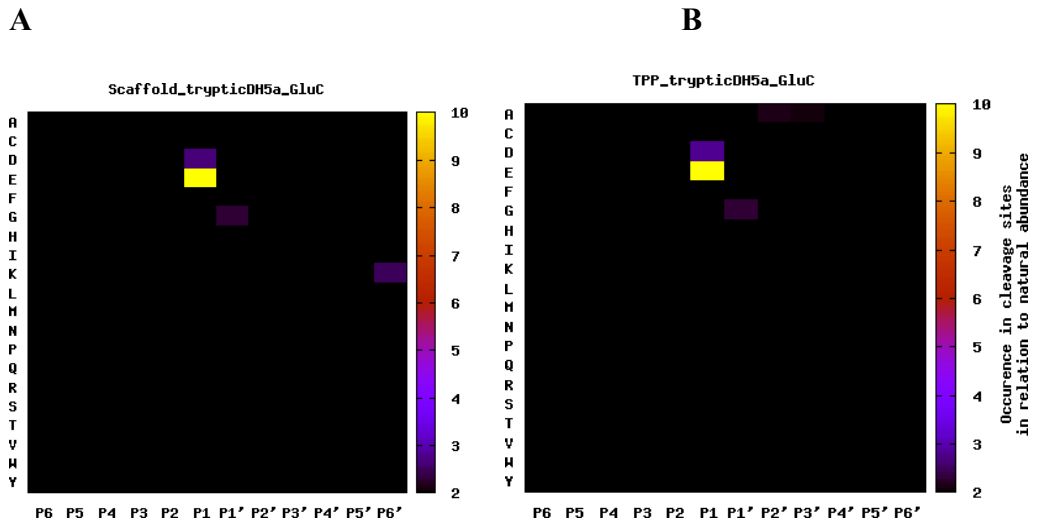


Figure 17: The CLIP-PICS website was used to create heat maps to visualize the analysis of the tryptic DH5α library cut with GluC in a 1:100 ratio of protease to Library. Heat maps for each of the two methods used are shown here. All three methods yielded over 250 peptides. A) The Sequest/Scaffold method had greater than 97% selectivity for E and D in P1. B) The Sequest/TPP method had greater than 90% selectivity for E and D in P1.

The group who wrote the PICS method also analyzed a tryptic library cut with GluC as a control.[56] They found that a 1:100 ratio of GluC : Library yielded more than 290 peptide sequences with greater than 94% matching the expected specificity for GluC. Also, a 1:1000 ratio of GluC : Library yielded over 400 peptides with greater than 94% matching the expected specificity for GluC. Our 1:100 sample yielded a similar number of peptides with a higher percent matching the expected values. However, our sample with a 1:1000 ratio of GluC:Library yielded very few peptides and is not shown here.

Control optimization: Tryptic DH5α library cut with chymotrypsin

Two different protease-to-library ratios were used to cut tryptic DH5α libraries with chymotrypsin. The ratios were 1:100 and 1:1000 for chymotrypsin to library (μg/μg). Parameters were optimized using both of these control samples (Tables 9-15). Chymotrypsin is a less specific protease than GluC. It cleaves amino acid sequences

following large hydrophobic amino acids: phenylalanine, tyrosine, tryptophan, and leucine.[58] Using this protease as a control, we can optimize the parameters that lead to the identification of the most peptides with a significant portion (~65%) of the amino acids matching chymotrypsin specificity.

Multiple methods using SEQUEST as the database search program were used to gain the most information from our samples. The first method used SEQUEST as the database search engine, and then used Scaffold to view the peptide list at a certain probability threshold. The Scaffold probability threshold was optimized using SEQUEST parameters of non-specific enzyme search, fixed modifications on lysine and cysteine, and protein and peptide tolerances of 1 amu (Table 9). The optimal probability threshold was 80% protein tolerance and 90% peptide tolerance. These are large tolerances; however, there is a subsequent validation that increases the level of confidence. These tolerance levels were used for all of the SEQUEST optimization tests.

Table 9: A tryptic DH5α library cut with chymotrypsin was used to optimize the Scaffold probability threshold. The protein and peptide probability thresholds yielding the most peptides with the highest percent of matching peptides were 80% and 90%, respectively.

Scaffold optimization for tryptic DH5α cut with chymotrypsin in a 1:1000 ratio			
Probability threshold Protein_Peptide	**80_80**	**80_90**	**80_95**
# of Peptides	292	261	231
% Matching Peptides	67.8	68.9	71.4
Probability threshold Protein_Peptide	**90_90**	**90_95**	**95_95**
# of Peptides	255	231	137
% Matching Peptides	69.8	71.4	73.7

Once the Scaffold parameters were optimized, multiple SEQUEST parameters were tested and optimized (Table 10 and 11). The enzyme search specification options

were: partial tryptic digestion, C-terminal tryptic digestion, and non-specific digestion. These options were tested with fixed and variable modifications as well as differing peptide and protein tolerance levels. The optimal values for a protease to library ratio of 1:100 were non-specific digestion with fixed modifications and protein and peptide tolerance of 1 amu (Table 10). The optimal parameters for SEQUEST resulted in 118 peptides, with 54.2% of those peptides having F, W or Y in the P1 position, consistent with the cleavage specificity of chymotrypsin.

Table 10: A tryptic DH5α library cut with chymotrypsin was used as a control to optimize parameters for SEQUEST and Scaffold. The optimal parameters were non-specific enzyme specification, fixed modifications, and a peptide and fragment tolerance of 1 amu.

SEQUEST Data Optimization for Tryptic DH5α Library cut with Chymotrypsin with a 1:100 ratio			
Enzyme Search Specification	**Partial Tryptic digest with variable modifications**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	72	102	89
% Matching Peptides	47.2	8.9	49.4
Enzyme Search Specification	**C-terminal Tryptic digest with variable modifications**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	147	120	120
% Matching Peptides	29.9	39.1	39.1
Enzyme Search Specification	**Non-specific digest with variable modification**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	110	103	95
% Matching Peptides	44.6	54.4	56.9
Enzyme Search Specification	**Partial Tryptic digest with fixed modifications**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	83	86	85
% Matching Peptides	53	55.8	53
Enzyme Search Specification	**C-terminal Tryptic digest with fixed modifications**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	130	119	118
% Matching Peptides	37.7	41.1	41.6
Enzyme Search Specification	**Non-specific digest with fixed modification**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	118	107	107
% Matching Peptides	54.2	52.3	53.3

The SEQUEST parameters were also optimized using a protease to library ratio of 1:1000. The optimal parameters for this control were again non-specific enzyme search specifications, fixed modifications and a protein and peptide tolerance of 1 amu (Table 11). This set of parameters matched the optimized parameters for the 1:100 protease:library ratio. It also matched the second set of previously optimized parameters for the tryptic DH5α library cut with GluC (Table 3 and 4). The optimal parameters resulted in 261 peptides, 68.9% of which had F, W or Y in the P1 position corresponding to the known chymotrypsin cleavage preference.

Table 11: A tryptic DH5α library was cut with chymotrypsin as a control to optimize the SEQUEST parameters with Scaffold. The optimal parameters were non-specific digestion, fixed modifications, and a peptide and fragment tolerance of 1 amu.

SEQUEST Data Optimization for Tryptic DH5α Library cut with Chymotrypsin with a 1:1000 ratio			
Enzyme Search Specification	**Partial Tryptic digest with variable modifications**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	210	98	209
% Matching Peptides	67.2	20.4	69.9
Enzyme Search Specification	**C-terminal Tryptic digest with variable modifications**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	260	240	240
% Matching Peptides	56.6	64.1	64.1
Enzyme Search Specification	**Non-specific digest with variable modification**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	222	233	236
% Matching Peptides	68.5	68.2	69.9
Enzyme Search Specification	**Partial Tryptic digest with fixed modifications**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	213	209	211
% Matching Peptides	70	72.2	71.1
Enzyme Search Specification	**C-terminal Tryptic digest with fixed modifications**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	271	262	255
% Matching Peptides	60.5	62.6	63.2
Enzyme Search Specification	**Non-specific digest with fixed modification**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	261	251	252
% Matching Peptides	68.9	69.4	70.2

Once the controls had been optimized for SEQUEST/Scaffold parameters, the TPP method was optimized. The control with a ratio of 1:100 was used to optimize TPP parameters for a SEQUEST search with fixed modifications, non-specific enzymatic specifications, and protein and peptide tolerances of 1 amu (Table 12). The optimal TPP parameters were a semi-tryptic enzyme specification with a probability threshold of 0.08. These parameters were the same as those identified for the tryptic GluC library (Table 5).

Table 12: A tryptic DH5α library was cut with chymotrypsin as a control to optimize the TPP parameters. The optimal parameters were semi-tryptic cleavage with a probability of 0.08.

Optimization of TPP parameters with a ratio of 1:100			
Enzyme specification	**Tryptic**	**Semi tryptic**	**Nonspecific**
# of peptides	112	112	53
% matching peptides	57.1	57.1	69.7
TPP probability threshold	**0.03**	**0.05**	**0.08**
# of peptides	139	112	101
% matching peptides	49.7	57.1	61.4

After the TPP parameters were optimized, SEQUEST searches were performed using the best parameters. As seen for the tryptic DH5α GluC control, the optimal enzymatic specification was non-specific. The optimal parameters for the 1:100 of protease to library control were also fixed modifications with a peptide and fragment tolerance of 1 amu (Table 13). These parameters agreed with those for sample 3 (Table 6). They also agreed with those for sample 3 run at double the concentration (Table 7). The best set of parameters resulted in 101 peptides, 61.4% of which had F, W, or Y in the P1 position, which is consistent with the cleavage preference of chymotrypsin.

Table 13: A tryptic DH5α library cut with chymotrypsin was used to optimize TPP parameters. The optimal parameters were non-specific enzyme constraints with fixed modifications and a peptide and fragment tolerance of 1 amu.

TPP Data Optimization for Tryptic DH5α Library cut with Chymotrypsin with a ratio of 1:100			
Search Specification	**Partial Tryptic digest with variable modifications**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	54	62	71
% Matching Peptides	64.7	9.7	57.7
Search Specification	**C-terminal Tryptic digest with variable modifications**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	80	77	94
% Matching Peptides	46.3	54.6	47.8
Search Specification	**Non-specific digest with variable modification**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	79	51	125
% Matching Peptides	62.1	64.7	48
Search Specification	**Partial Tryptic digest with fixed modifications**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	65	62	72
% Matching Peptides	60	64.5	59.7
Search Specification	**C-terminal Tryptic digest with fixed modifications**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	67	75	99
% Matching Peptides	61.3	54.7	47.5
Search Specification	**Non-specific digest with fixed modification**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	101	113	125
% Matching Peptides	61.4	59.3	48

TPP parameters were also optimized using a 1:1000 ratio of chymotrypsin to tryptic DH5α library. The optimal parameters were non-specific enzyme specification with fixed modifications and a peptide and fragment tolerance of 1 amu (Table 14); the same as those for the 1:100 ratio (Table 13). The optimal parameters resulted in 327 peptides, 64.9% of which displayed chymotrypsin cleavage preference.

Table 14: A tryptic DH5α library cut with chymotrypsin was used to optimize the TPP parameters. The optimal parameters were non-specific enzyme specification with fixed modifications and a peptide and fragment tolerance of 1 amu.

TPP Data Optimization for Tryptic DH5α Library cut with Chymotrypsin with a ratio of 1:1000			
Search Specification	**Partial Tryptic digest with variable modifications**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	179	59	237
% Matching Peptides	68.7	23.8	66.3
Search Specification	**C-terminal Tryptic digest with variable modifications**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	201	297	318
% Matching Peptides	61.8	54.2	51.9
Search Specification	**Non-specific digest with variable modification**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	298	315	324
% Matching Peptides	64.2	65.2	65.4
Search Specification	**Partial Tryptic digest with fixed modifications**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	204	236	238
% Matching Peptides	69.6	65.2	66.4
Search Specification	**C-terminal Tryptic digest with fixed modifications**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	236	298	321
% Matching Peptides	61.1	54.6	51.1
Search Specification	**Non-specific digest with fixed modification**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	327	344	322
% Matching Peptides	64.9	62.5	61.9

Once the 2 controls cleaved with chymotrypsin had been optimized, the different methods were compared to determine the best method for analyzing tryptic DH5α sample cleaved by our protease of interest (Table 15). Analysis with TPP resulted in similar percentages of matched peptides for both samples. Further samples were analyzed using SEQUEST/Scaffold, as this method was determined to be better for the tryptic DH5α

libraries cut with GluC, and neither method showed a significant difference for these tryptic DH5α libraries cut with chymotrypsin.

Table 15: A tryptic DH5α library cut with chymotrypsin was used to optimize parameters for several data analysis methods.

Optimal Search Parameters				
	TPP		Scaffold	
	26	27	26	27
# of Peptides	101	327	118	261
% Matching Peptides	61.4	64.9	54.2	68.9

Following optimization of parameters for the tryptic DH5α library cut with GluC, heat maps were generated to visualize the data. All of the heat maps show that amino acids matching chymotrypsin cleavage predominate in P1 (Figure 18).

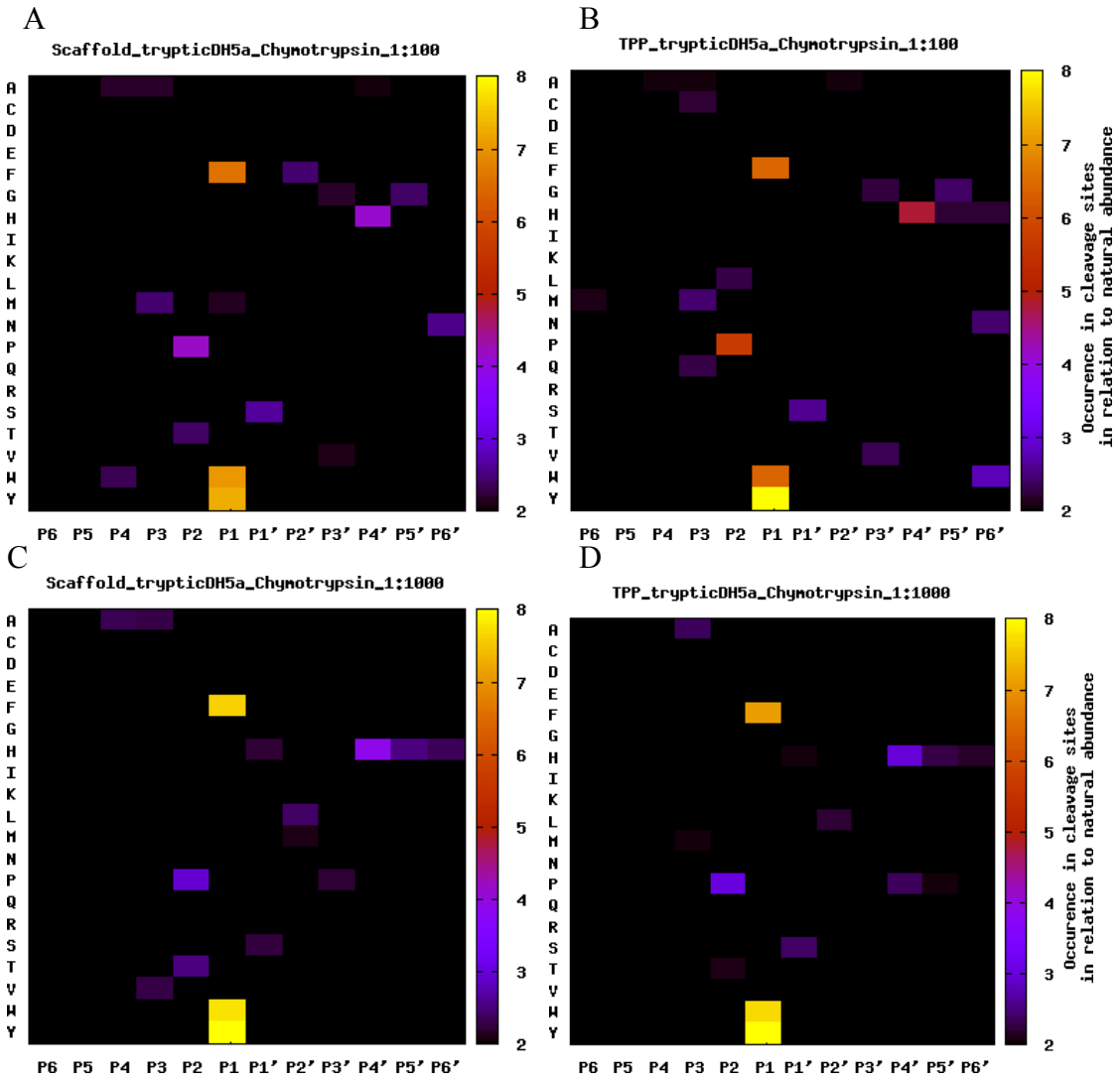


Figure 18: A DH5α library cut with chymotrypsin was used as a control to validate the PICS method. Multiple database search methods were tested to determine the best method for identifying peptides cleaved by the protease of interest and heat maps were generated to visualize the data. For all heat maps corresponding to the analysis of the chymotrypsin control samples, the observed amino acid preferences are consistent with the established chymotrypsin cleavage preferences. A) A 1:100 chymotrypsin to DH5α library was optimized using SEQUEST and Scaffold. B) A 1:100 chymotrypsin to DH5α library was optimized using SEQUEST and TPP. C) A 1:1000 chymotrypsin to DH5α library was optimized using SEQUEST and Scaffold. D) A 1:1000 chymotrypsin to DH5α library was optimized using SEQUEST and TPP.

Control optimization: Tryptic S2 library cut with GluC

Two different protease-to-library ratios were used to cut tryptic S2 libraries with GluC. The ratios were 1:100 and 1:1000 for chymotrypsin to library. Parameters were optimized using both of these controls (Tables 16-22). Again, the fact that GluC is a very specific protease was used to optimize search parameters such that only greater than 95% of the correct peptides were identified.[56] Multiple methods testing different database search programs were used, so that we could be sure to gain the most from our samples. The first method used SEQUEST as the database search engine, and then used Scaffold to view the peptide list at a certain probability threshold. The Scaffold probability threshold was optimized using SEQUEST parameters of non-specific enzyme search, fixed modifications on the amino acids, lysine and cysteine, protein tolerance of 1.5 amu and peptide tolerances of 1 amu (Table 16). The optimal probability threshold was 80% protein tolerance and 80% peptide tolerance.

Table 16: A tryptic S2 library cut with GluC was used to optimize the parameters for Scaffold. The optimal parameters resulting in the most peptides with the highest percentage matching GluC specificity were 80% probability thresholds for both protein and peptide tolerance.

Scaffold optimization for Sample 25			
Probability threshold Protein_Peptide	**80_80**	**80_90**	**80_95**
# of Peptides	870	835	793
% Matching Peptides	94.7	94.6	94.8
Probability threshold Protein_Peptide	**90_90**	**90_95**	**95_95**
# of Peptides	687	646	646
% Matching Peptides	94.8	95	95

The optimized Scaffold probability thresholds were then used to find the best SEQUEST parameters for the tryptic S2 controls cut with GluC. Non-specific enzymatic

constraints resulted in the most peptides and the highest number of peptides matching GluC specificity. Fixed modifications with a peptide tolerance of 2 amu increased the number of peptides identified (Table 17). These parameters resulted in 1183 peptides, with 91.3% of the peptides consistent with the expected GluC cleavage preference.

Table 17: A tryptic S2 library cut with GluC was used to optimize the parameters for SEQUEST. The optimal parameters for a SEQUEST search were non-specific enzyme specifications with fixed modifications and a peptide tolerance of 2amu.

SEQUEST Data Optimization for Tryptic S2 Library cut with GluC for 1:100 ratio			
Enzyme Search Specification	**Partial Tryptic digest with variable modifications**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	90	89	89
% Matching Peptides	78.9	79.8	83.1
Enzyme Search Specification	**C-terminal Tryptic digest with variable modifications**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	114	104	97
% Matching Peptides	66.6	74.1	81.5
Enzyme Search Specification	**Non-specific digest with variable modification**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	768	1121	1168
% Matching Peptides	91.1	91.4	91.9
Enzyme Search Specification	**Partial Tryptic digest with fixed modifications**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	93	92	89
% Matching Peptides	83.9	82.6	85.4
Enzyme Search Specification	**C-terminal Tryptic digest with fixed modifications**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	115	113	110
% Matching Peptides	68.7	69	72.7
Enzyme Search Specification	**Non-specific digest with fixed modification**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	1173	1175	1183
% Matching Peptides	91.1	91.3	91.3

Once the 1:100 ratio of GluC to tryptic S2 library was optimized, the 1:1000 ratio was used to confirm the best parameters. Again, non-specific search constraints with

fixed modifications and a peptide tolerance of 2 amu yielded the best results (Table 18). These parameters resulted in 880 peptides, 94.6% of which had E or D in the P1 position.

Table 18: A tryptic S2 library cut with GluC was used to optimize the parameters for SEQUEST. The optimal parameters for a SEQUEST search were non-specific enzyme specifications with fixed modifications and a peptide tolerance of 2 amu.

SEQUEST Data Optimization for Tryptic S2 Library cut with GluC for 1:1000 ratio			
Enzyme Search Specification	**Partial Tryptic digest with variable modifications**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	65	72	67
% Matching Peptides	81.6	79.2	80.6
Enzyme Search Specification	**C-terminal Tryptic digest with variable modifications**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	81	86	81
% Matching Peptides	70.4	72.1	75.3
Enzyme Search Specification	**Non-specific digest with variable modification**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	542	720	753
% Matching Peptides	93.4	94.1	94.1
Enzyme Search Specification	**Partial Tryptic digest with fixed modifications**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	71	76	75
% Matching Peptides	80.3	78.9	78.7
Enzyme Search Specification	**C-terminal Tryptic digest with fixed modifications**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	89	100	96
% Matching Peptides	68.6	66	66.7
Enzyme Search Specification	**Non-specific digest with fixed modification**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	851	870	880
% Matching Peptides	94.5	94.7	94.6

After the parameters for the first database search method were optimized, parameters for the second method were optimized. The TPP parameters were optimized using non-specific enzyme constraints with variable modifications and a peptide and

fragment tolerance of 1 for the SEQUEST parameters. The optimal parameters for TPP were non-specific enzyme specification with a probability threshold of 0.08 (Table 19).

Table 19: A tryptic S2 library cut with GluC was used to optimize parameters for TPP. The optimal parameters were non-specific enzyme specification and a probability threshold of 0.08.

Optimization of TPP parameters Sample 24			
Enzyme specification	**Tryptic**	**Semi tryptic**	**Nonspecific**
# of Peptides	2037	2037	1735
% Matching Peptides	80.5	80.5	86.4
TPP probability threshold	**0.03**	**0.05**	**0.08**
# of Peptides	2172	2037	1917
% Matching Peptides	77.3	80.5	83.4

The optimal TPP parameters were used to optimize the SEQUEST parameters. The best parameters were non-specific enzyme specification, fixed modifications, and a peptide and fragment tolerance of 1 amu (Table 20). These parameters resulted in 1624 peptides, 88.3% of which had E or D in the P1 position.

Table 20: A tryptic S2 library cut with GluC was used to optimize SEQUEST and TPP parameters. The optimal parameters were non-specific enzyme constraints, fixed parameters, and a peptide and fragment tolerance of 1 amu.

TPP Data Optimization for Tryptic DH5α Library cut with Chymotrypsin Sample 24			
Enzyme Search Specification	**Partial Tryptic digest with variable modifications**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	85	87	88
% Matching Peptides	82.3	85	87.5
Enzyme Search Specification	**C-terminal Tryptic digest with variable modifications**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	87	79	90
% Matching Peptides	71.3	88.6	83.4
Enzyme Search Specification	**Non-specific digest with variable modification**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	938	1579	1727
% Matching Peptides	88.4	84	82.5
Enzyme Search Specification	**Partial Tryptic digest with fixed modifications**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	78	84	90
% Matching Peptides	89.8	85.7	84.5
Enzyme Search Specification	**C-terminal Tryptic digest with fixed modifications**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	74	85	100
% Matching Peptides	83.8	85.8	78
Enzyme Search Specification	**Non-specific digest with fixed modification**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	1624	1831	1878
% Matching Peptides	88.3	84.7	84.1

After optimizing TPP and SEQUEST parameters for the 1:100 ratio of GluC to tryptic S2 library, the same parameters were optimized using a 1:1000 ratio control. The optimal parameters were the same as those for the 1:100 ratio: non-specific enzyme constraints, fixed modifications, and a peptide tolerance of 1 amu (Table 21). These parameters resulted in 974 peptides, with 92.1% of those peptides consistent with the cleavage preference of GluC.

Table 21: A tryptic S2 library cut with GluC was used to optimize SEQUEST and TPP parameters. The optimal parameters were non-specific enzyme constraints, fixed parameters, and a peptide and fragment tolerance of 1 amu.

TPP Data Optimization for Tryptic DH5α Library cut with Chymotrypsin Sample 25			
Enzyme Search Specification	**Partial Tryptic digest with variable modifications**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	47	57	55
% Matching Peptides	93.6	89.5	92.7
Enzyme Search Specification	**C-terminal Tryptic digest with variable modifications**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	45	55	53
% Matching Peptides	84.5	83.7	86.8
Enzyme Search Specification	**Non-specific digest with variable modification**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	567	970	1109
% Matching Peptides	92.9	86.1	83.3
Enzyme Search Specification	**Partial Tryptic digest with fixed modifications**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	48	52	57
% Matching Peptides	93.8	94.2	94.8
Enzyme Search Specification	**C-terminal Tryptic digest with fixed modifications**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	43	48	53
% Matching Peptides	95.4	93.8	92.5
Enzyme Search Specification	**Non-specific digest with fixed modification**		
Peptide/Fragment Tolerance (amu)	1_1	1.5_1	2_1
# of Peptides	974	1185	1254
% Matching Peptides	92.1	86.2	84.4

Once the parameters had been optimized for both database searches, the methods were compared to determine the best set of parameters to use for S2 libraries cut with our protease of interest, BACE1. Scaffold identified fewer peptides for both S2 controls; however, the percentage of peptides matching GluC cleavage was higher for both S2 controls (Table 22). Therefore, the SEQUEST/Scaffold search method was used for all further S2 samples.

Table 22: The optimal search parameters for a tryptic S2 library cut with GluC were compared to determine the best database search method.

Optimal Search Parameters				
	TPP		Scaffold	
	24	25	24	25
# of Peptides	1624	974	1183	880
% Matching Peptides	88.3	92.1	91.3	94.6

Heat maps for each of the S2 controls and each of the database search methods were generated. In these heat maps, the only amino acids that apppear in P1 are E and D (Figure 19). These peptides are the only ones that should appear after cleavage by GluC.

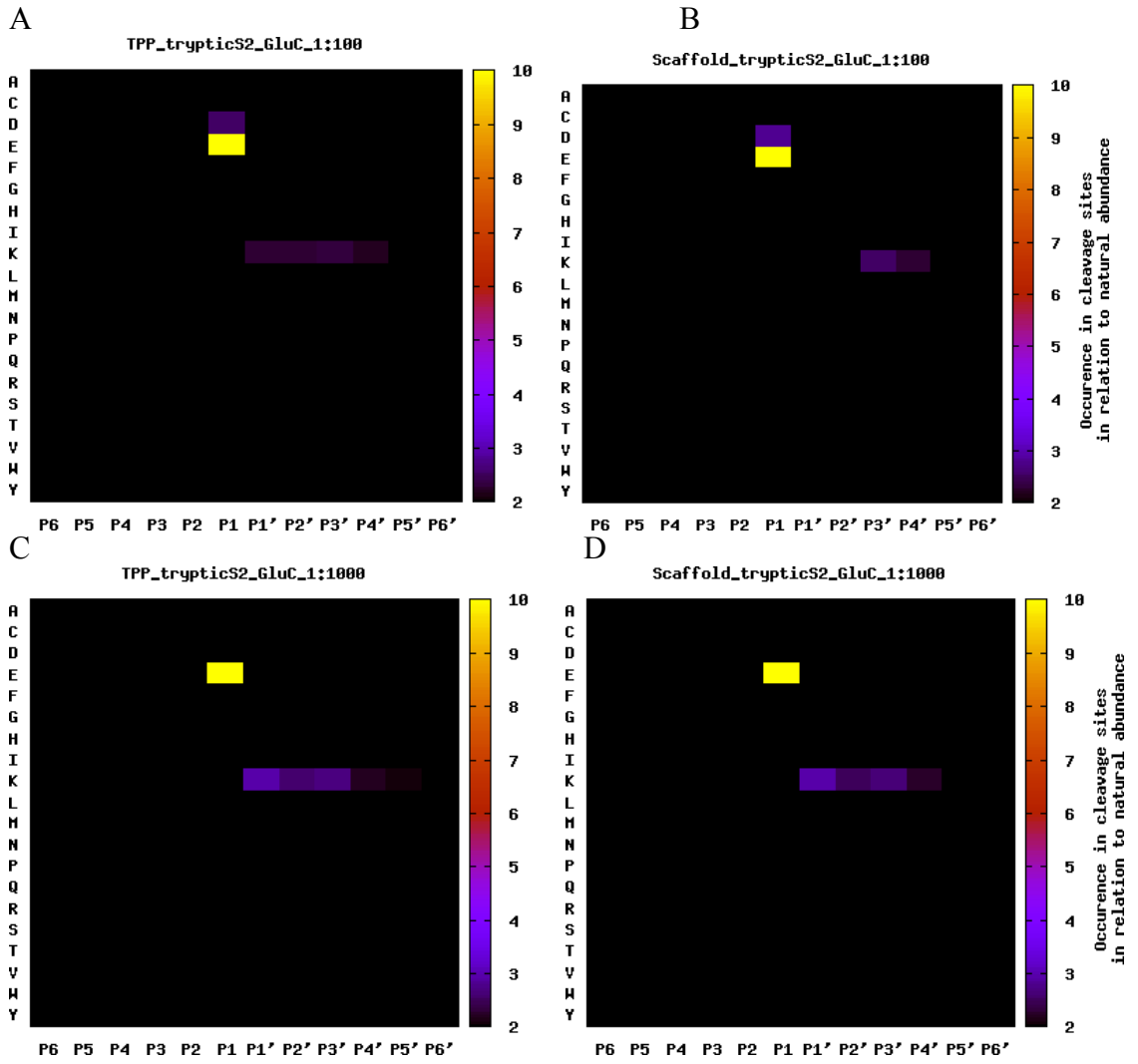


Figure 19: An S2 library cut with GluC was used as a control to validate the PICS method. Multiple database search methods were tested to determine the best method for identifying peptides cleaved by the protease of interest and heat maps were generated to visualize the data. For all of the panels, the only amino acids appearing in this heat map are those corresponding to GluC cleavage. A) A 1:100 GluC to S2 library was optimized using SEQUEST and Scaffold. B) A 1:100 GluC to S2 library was optimized using SEQUEST and TPP. C) A 1:1000 GluC to S2 library was optimized using SEQUEST and Scaffold. D) A 1:1000 GluC to S2 library was optimized using SEQUEST and TPP.

Tryptic DH5α library cut with BACE1A

A tryptic DH5α library was cut with BACE1A to determine the sequence of amino acids preferentially cut by this protease. Multiple ratios of protease to library were

used to determine the ratio resulting in the most peptides (Table 23). A ratio of approximately 1:50 yielded the highest number of peptides.

Table 23: A tryptic DH5α library was cut with BACE1A. Different ratios of protease to library were used to determine the best ratio. A ratio of 1:50 resulted in the highest number of peptides.

Sample Id #	Protease:Library Ratio	# of Scaffold Peptides	# of TPP Peptides
1	125	29	0
2	36	116	151
19	17	18	2
20	8	24	12
21	4	31	0
22	50	132	171
23	200	24	0
34	100	25	0
41	50	71	4
42	50	55	8
43	151	34	15
44	300	38	14
45	56	57	85
47	47	67	8
48	47	13	0
46.5	50	76	63
47.5	110	48	28
48.5	200	45	0

Heat maps were generated by pooling the peptide results for the best samples: 2, 22, 41 and 45. 311 peptides were used to generate the heat map. Cleavage site preferences for BACE1A were aspartate (D) in P4, isoleucine (I) in P3, aromatic residues (Y and F) or hydrophobic residues (L) in P1, methionine (M) in P1' and valine (V) in P2' (Fig. 20).

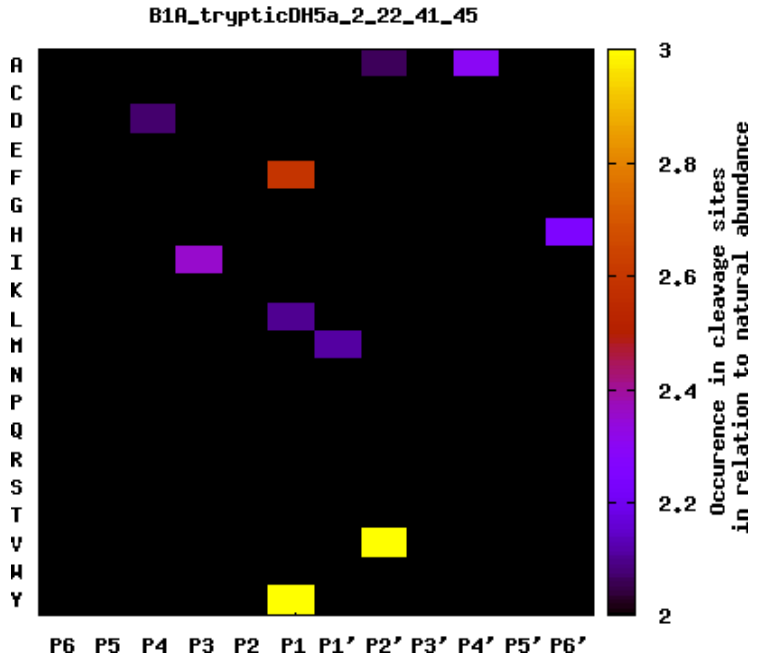


Figure 20: A tryptic library was used to create this heat map. The heat map shows the amino acids that preferentially comprise the cleavage site for BACE1A. There are not any amino acids with high specificity outside of P3 or P3'. 311 peptides were analyzed to generate this heat map. The most important amino acids here appear to be the aromatic residues in P1 and valine in P2'.

Neighbor effects for BACE1A cleavage sites were analyzed (Table 24, Figure 21, and Figure 22). The largest degree of cooperativity existed between P3 and P1.

Table 24: Potential subsite cooperativity is shown for BACE1A. Neighbor effects do not appear to a large extent. The largest cooperativity influence is between P3 valine and P1 phenylalanine. P1 refers to the N-terminal side of cleavage starting with the amino acid closest to the cleavage site. P1' refers to the C-terminal side of cleavage starting with the amino acid closest to the cleavage site (Fig. 1).

Fixed residue	Affected residue	Change (%)	% Occurrence Fixed Residue
P3_L	P1_N	11.6	13.2
P3_V	P1_F	21.2	10.3
P3_V	P2_E	14.7	10.3
P1 Y	P2' V	13.9	9
P1_F	P3' Q	20	10

When tyrosine is in P1, there was a marked preference for valine in P2'. This was seen with phenylalanine in P1, but it was not as intense (Fig. 21). There was also a preference for serine in P2. With phenylalanine in P1, there was a strong preference for valine in P3 and glutamine in P3'.

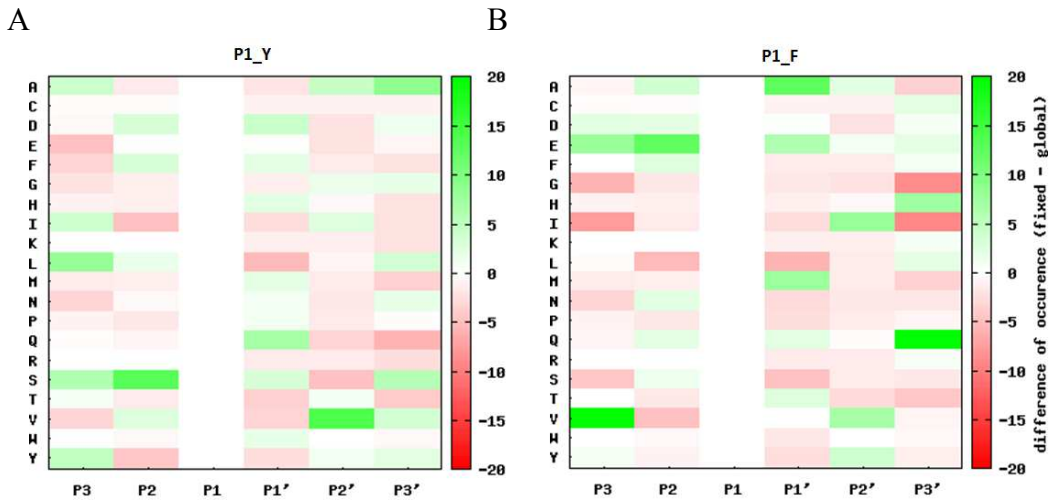


Figure 21: Heat maps depicting neighbor effects for fixed amino acids in P1. A) With tyrosine in P1, there was an increase in the preference for valine in P2'. B) The strong preference for valine was not seen with phenylalanine in P1, but an increase in the preference for valine in P3 was observed.

With valine in P3, there was a marked preference for phenylalanine in P1. This preference was not seen with isoleucine or leucine in P3. There was also a preference for glutamate in P2 (Fig. 22). When leucine was in P3, there was a slight preference for asparagine in P1. No strong neighbor effects were observed with isoleucine in the P3 position.

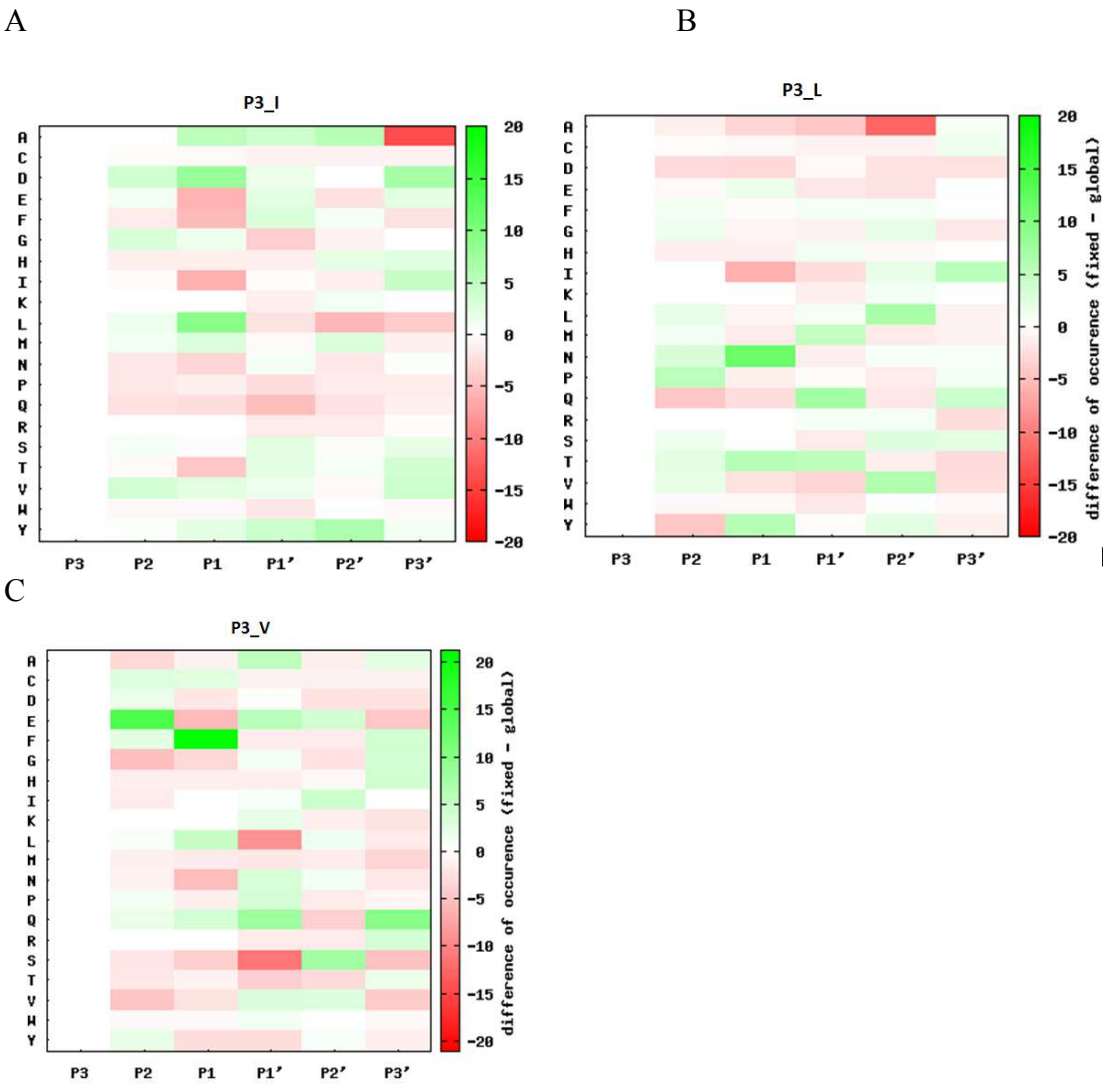


Figure 22: Heat maps depicting neighbor effects for fixed amino acids in P3. The largest neighbor effect occurs when valine was in P3. A) With isoleucine in P3, there were no noticeable neighbor effects. B) With leucine in P3, there was an increase in the preference for asparagine in the P1 position. C) With valine in P3, there was an increase in the preference for phenylalanine in P1.

Chymotrypsin DH5α library cut with BACE1A

A chymotrypsin DH5α library was cut with BACE1A to reinforce the tryptic library data. Multiple ratios of protease to library were used to determine the ratio resulting in the most peptides (Table 25). The optimal ratio appeared to be between 1:50 and 1:100.

Table 25: A chymotrypsin DH5α library was cut with BACE1A. Different ratios of protease to library were used to determine the best ratio. A ratio of between 1:50 and 1:100 resulted in the highest number of peptides.

Sample Id #	Protease : Library Ratio	# of TPP Peptides	# of Scaffold Peptides
5	1:250	22	32
28	1:90	54	38
29	1:40	29	34
35	1:100	8	16
36	1:125	9	21

A heat map was generated by pooling the identified peptide sequences from the best 2 samples, 28 and 29 (Fig. 23). 72 peptides were analyzed to create the heat map. The amino acids that were most prominent were methionine (M) in P4, glutamine (Q) in P2, methionine (M) or cysteine (C) in P1, glycine (G) in P2', and tryptophan (W) in P4'.

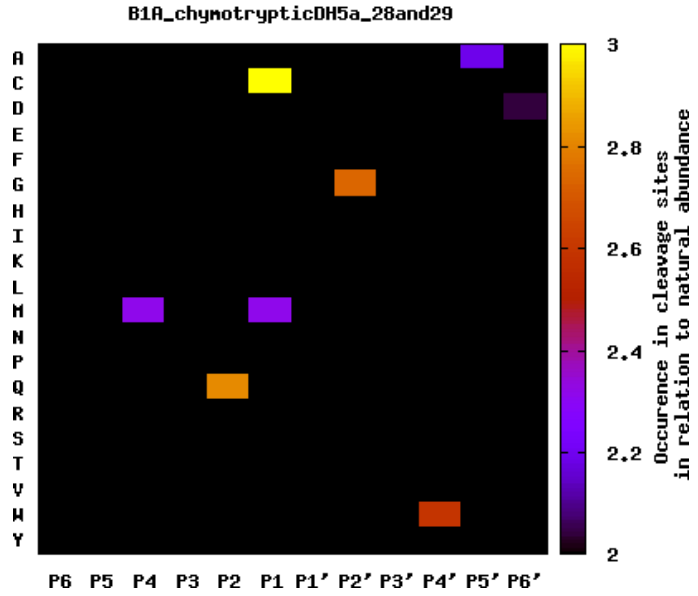


Figure 23: A heat map was created for chymotrypsin DH5α libraries cut with BACE1A. 72 unique peptides were analyzed to determine that M was preferred in P4, Q was preferred in P2, M or C were preferred in P1, G was preferred in P2', and W was preferred in P4'.

GluC DH5α library cut with BACE1A

A GluC DH5α library was cut with BACE1A to determine the sequence of amino acids preferentially cut by BACE1A. Multiple ratios of protease to library were used to determine the ratio resulting in the most peptides (Table 26). The optimal ratio appeared to be greater than 1:20.

Table 26: GluC DH5α libraries were cut with BACE1A using multiple ratios of protease to library to determine the optimal ratio for each sample set. A ratio of greater than 1:20 appeared to be optimal.

Sample Id #	Ratio of Protease : Library	# of Scaffold Peptides	# of TPP Peptides
16	1:9	4	0
17	1:9	3	0
30	1:20	48	16
31	1:10	38	59

A heat map was generated using the 2 best samples. 84 unique peptides were analyzed to create the heat map, which identified tryptophan (W) in P4, W in P3, histidine (H) in P2, W or serine (S) in P1, glycine (G) or W in P1', M or I in P2', and tyrosine (Y) in P4' (Fig. 24).

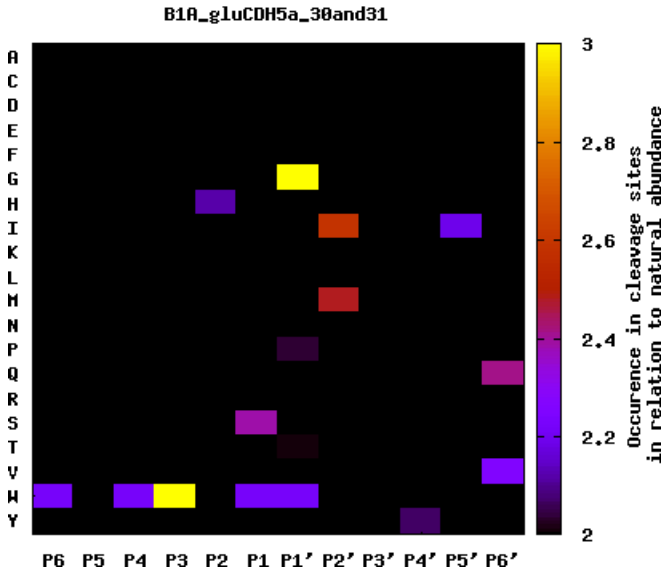


Figure 24: A heat map was generated for GluC libraries cut with BACE1A. 84 peptides were analyzed to determine the noticeable amino acids: W in P4, W in P3, H in P2, W or S in P1, G or W in P1', and M or I in P2'.

Tryptic S2 cut with BACE1A

Multiple tryptic S2 libraries were cleaved with BACE1A at different protease to library ratios (Table 27). Several ratios resulted in higher numbers of peptides. The best ratios appeared to be between 1:150 and 1:300; however, ratios of 1:50 and 1:25 also yielded similar numbers of peptides.

Table 27: Tryptic S2 libraries were cut with BACE1A at several protease to library ratios. The best ratios appeared to be between 1:150 and 1:300; however ratios of 1:25 and 1:50 resulted in similar numbers of peptides.

Sample Id #	Ratio of Protease : Library	# of Scaffold Peptides	# of TPP Peptides
15	1:14	9	0
32	1:50	21	10
33	1:25	21	18
37	1:100	11	6
38	1:125	13	0
39	1:150	36	3
40	1:300	41	4

A heat map was generated using the 4 best tryptic S2 libraries cut with BACE1A (Fig. 25). 110 unique peptides were analyzed. The amino acids that occurred at higher than natural abundance were W or Y in P2, S in P1, S in P1', and M in P4'.

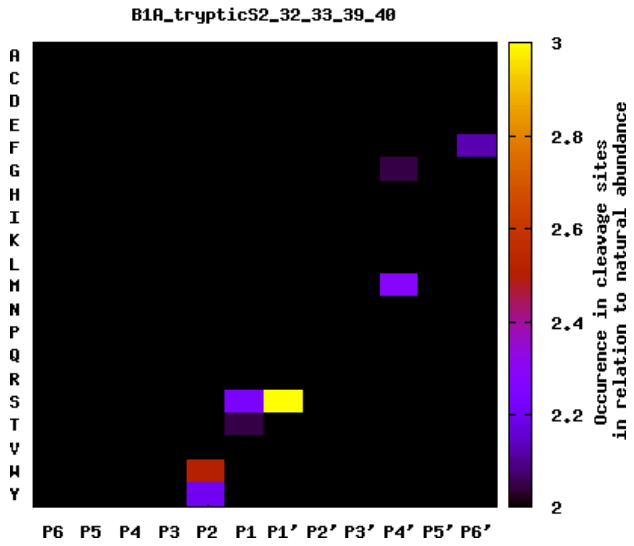


Figure 25: A heat map was generated for 4 tryptic S2 libraries cut with BACE1A. Amino acids occurring at higher than natural abundance were W or Y in P2, W in P1, S in P1', and M in P4'. 110 peptides were used in the analysis.

Tryptic DH5a cut with BACE2A

Two tryptic DH5α libraries were cut with BACE2A at different ratios of protease to library (Table 28). The two ratios resulted in a similar number of peptides, but there were not many peptides for either.

Table 28: Tryptic DH5α libraries were cleaved with BACE2A. Neither ratio yielded many peptides.

Sample Id #	Ratio of Protease : Library	# of Scaffold Peptides	# of TPP Peptides
12	1:75	23	8
13	1:150	24	8

A heat map was generated using both tryptic DH5α libraries cut with BACE2A. 57 unique peptides were analyzed (Fig. 26). The peptides occurring at higher than natural abundance were M in P4, W in P2, F or D in P1, P or M in P2', P in P3', and Y in P4'.

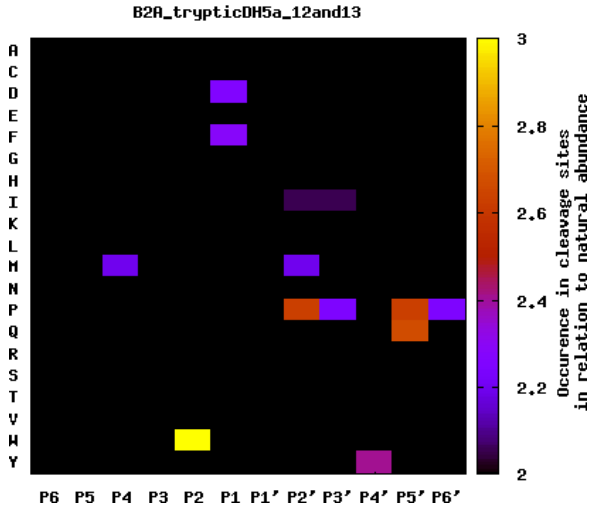


Figure 26: A heat map was created for tryptic DH5α libraries cut with BACE2A. The amino acids that were most noticeable in the 57 peptides analyzed were M in P4, W in P2, F or D in P1, P or M in P2', P in P3, and Y in P4'.

Chymotrypsin DH5a cut with BACE2A

A single chymotrypsin DH5α library was cut with BACE2A (Table 29). This sample yielded only 30 peptides with a ratio of 1:575

Table 29: A single chymotrypsin DH5α library was cleaved with BACE2A. Only 30 peptides resulted from a ratio of 1:575.

Sample Id #	Ratio of Protease : Library	# of Scaffold Peptides	# of TPP Peptides
6	1:575	30	27

A heat map was generated for the chymotrypsin DH5α library cut with BACE2A (Fig. 27). This heat map highlighted the amino acids occurring at higher than natural abundance for the 30 peptides analyzed. The amino acids were W or D in P4, I or W in P3, W or G in P2, T or E in P1, V or P in P1', L or P in P2', F or A in P3', and glutamine (Q) or I in P4'.

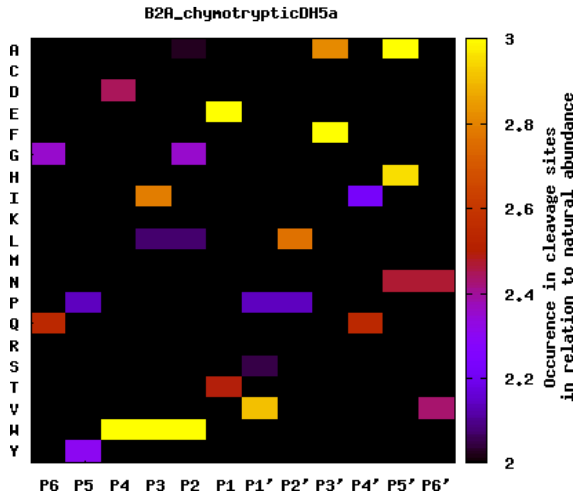


Figure 27: A chymotrypsin DH5α library was cleaved with BACE2A. A heat map displaying the amino acids occurring at higher than natural abundance was generated using 30 unique peptides. The amino acids were W or D in P4, I or W in P3, W or G in P2, T or E in P1, V or P in P1', L or P in P2', F or A in P3', and Q or I in P4'.

Discussion

Control optimization

Three different sets of controls were used to validate the PICS procedure in our lab: a tryptic DH5α library that was cut with GluC, a tryptic DH5α library cut with chymotrypsin, and a tryptic S2 library cut with GluC. Each of these controls was used to optimize the database search parameters to match the sequenced C-terminal peptide with their N-terminal counterpart (Fig. 7). Many different database search programs are available for use with mass spectrometric peptide sequence data. The programs used here were free software or online sources recommended by the staff at the University of Minnesota Center for Mass Spectrometry and Proteomics.

The two search methods used here both began by running the raw mass spectrometric data through SEQUEST. SEQUEST is a search program that matches the mass spectrometric peptide sequence data with the corresponding protein from the organism's proteome. This software generates a list of the C-terminal portion of the peptides; however, the list needs to be converted to a readable format. The two methods differ in the analysis of the list of peptides. One method involves using Scaffold, a program used to view SEQUEST data. This program also had the ability to limit the peptides by threshold tolerance for protein and peptide, so only peptides with a low rate of error were used. The second method used a different program to revalidate the peptide list, the trans-proteomic pipeline (TPP). TPP converts SEQUEST search data into a Peptide Prophet compatible format, which matches the protein sequence data to the known proteome of the library organism. The final step for each method was the same.

The C-terminal peptide list generated was uploaded into the website created by the Overall lab.[57] This website compares individual peptide sequences against the corresponding known proteome in order to determine the corresponding N-terminal peptide sequence. This final list of peptides was used to generate a heat map showing the amino acid preferences for the different subsites normalized according to their natural abundance. The two methods were optimized and then compared to determine the best method for database searches.

Multiple parameters for each of these methods were optimized to confirm the validity of the data as well as assuring that the highest number of peptides was analyzed. The parameters that were optimized were: level of error tolerated, enzyme cleavage specified for database searches, and fixed or variable modifications. Other parameters were tested, but did not have a significant impact on the results. Parameters that resulted in both the highest number of peptides and the highest percent of peptides matching the expected values were used for analyzing the samples to determine the amino acid sequence cleaved by the protease of interest (Table 30).

Table 30: The optimal parameters identified using the controls. The parameters that were optimized were: level of error, enzyme cleavage specified in the database search, and fixed versus variable modification on K and C.

Optimal Parameters for Database Analysis				
DH5α				
Sequest	Peptide Tolerance (amu)	Fragment Tolerance(amu)	Cleavage Enzyme	Fixed Modifications
	1	1	Non-specific	K, C, N-terminal
Scaffold	% Protein Tolerance	% Peptide Tolerance	# Peptides/Protein	
	80	90	1	
S2				
Sequest	Peptide Tolerance	Fragment Tolerance (amu)	Cleavage Enzyme	Fixed Modifications
	2	1	Non-specific	K, C, N-terminal
Scaffold	% Protein Tolerance	% Peptide Tolerance	# Peptides/Protein	
	80	80	1	

The optimal parameters for the control samples returned the highest number of peptide sequences along with the highest percentage of peptides matching the expected amino acids in P1. Peptide, fragment, and protein tolerance define the level of error allowed. The goal was to select the lowest possible level of error while avoiding false negatives. This level can be increased as long as subsequent programs are used to revalidate the peptides. 'Protease specificity' allowed the search program to set up a hypothetical digestion, which was used to match the mass spectrometric data entered. The 'protease selected' decreased the time of the search by limiting the possibilities for peptides; however, it could also result in missed peptides if the definition was too narrow. The possibilities for protease choices relevant to this data were partial tryptic digestion, C-terminal tryptic digestion, or non-specific cleavage. Each of the controls was a tryptic library, but they were also cut with another protease. Though using 'tryptic digestion' would limit and decrease the amount of time for a search, it could result in missing some of the peptides. The final parameter optimized was the level of amino acid modification.

Three different modifications were added in our libraries: thioacylation of the N-terminal peptide, carbamidomethylation of cysteine, and dimethylation of lysine. The N-terminal modification had to be fixed, because that was how peptides cut by the protease of interest were separated from other peptides. The other two modifications were fixed in the original protocol; however, assigning them as variable might have allowed for more peptides to be identified if some of the intended amino acid modifications did not go to completion.[55] The optimal parameters identified and used to identify the peptide sequences cleaved by BACE1A were shown in Table 30.

Sequence preferences for BACE1A

Trypsin, chymotrypsin, and GluC were the three different proteases used in making the libraries for this experiment. Since each of these proteases cleaves proteins after different classes of amino acids, the independent libraries should serve to minimize bias for the samples and to increase the number and variety of peptides. The trypsin library resulted in the largest number of peptides and thus the most information. However, due to their differing specificity, chymotrypsin and GluC libraries provided independent data as well as confirming some of the sequences obtained from the tryptic libraries.

Proteomes from DH5α (*E. coli*) and S2 cells (*Drosophila melanogaster*) were each used in conjunction with the proteases above to generate six independent peptide libraries. Both of these organisms have sequenced proteomes enabling the N-terminal sequences corresponding to the identified peptides to be identified through database searching (Fig. 7). Using both S2 and DH5α protease libraries increased the variety of

proteins and thus the number of different peptides in each library allowing us to gain more information about the preferred cleavage sequence of BACE1.

Compiling the data for all of the BACE1A samples gave a more complete understanding of the preferred cleavage sequence. Multiple amino acids are tolerated in each of the positions of the cleaved sequence; however, there are discernible preferences for certain amino acids or types of amino acids (Table 31). Hydrophobic amino acids, particularly isoleucine, leucine, or valine, are seen in most positions.

Table 31: Compilation of the amino acids comprising subsite positions P4-P4' that were preferred for BACE1A cleavage. The preferred amino acids identified by the PICS procedure were compared to known sequences cleaved by BACE1A as well as the preferred sequence identified by Turner, et al. in 2001.[33,42,60] *Bold amino acids indicate a preference that is only seen with PICS data. The outlined amino acid in P2' indicated the mutation that leads to decreased BACE1 cleavage.*

Subsite	P4			P3	P2			P1			P1'				P2'		P3'			P4'
APP Sequence	E			V	K			M			D				A		E			F
Swedish APP mutation	E			V	N			L			D				A		E			F
Protective APP mutation	E			V	K			M			D				T		E			F
Tryptic DH5α Library	D			I				F	L	Y	M				V					**A**
Chymotryptic DH5α Library	M				Q			M	C						G					W
GluC DH5α Library	**W**			**W**	H			W	**S**		G			**W**	M	I				Y
Tryptic S2 Library					**W**	**Y**		**S**			S									M
Neighbor Cooperativity				V	E			F									Q			
				L				**N**												
					S			Y							V					
Turner/Tang Data	E	Q	D	I/V/L	D	N	M	L	F/Y	M	M	E	Q	A	V/I	A	L	W	A	

In the P4 position, there were multiple types of amino acids identified by PICS. Three different amino acids were identified in three different libraries: aspartate, methionine, and tryptophan (Table 31). Published studies and the known APP sequences cleaved by BACE1A have glutamate, aspartate and glutamine in P4.[42] The wide variety of amino acids identified suggests that there is no strong preference for any particular amino acid in this position.

For P3, there was a preference for the hydrophobic isoleucine with a slight preference for the aromatic tyrosine (Table 31). Published studies have indicated a preference for valine and leucine in this site.[42] We showed that valine and leucine affect the neighboring amino acids in P2 and P1. With valine in P3, there is a cooperative affect between glutamate and phenylalanine in P2 and P1, respectively. This cooperativity was not observed with isoleucine or valine in P3. In fact, there was a slight negative cooperativity between isoleucine in P3 and phenylalanine in P1. With isoleucine in P3, there were no strong cooperativity effects. However, having leucine in P3 led to a preference for asparagine in P1.

There was no strong preference for any particular amino acid in P2. In our PICS data, we saw histidine and glutamine, larger, hydrophilic amino acids, as well as the aromatic amino acids tryptophan and tyrosine (Table 31). Amino acids identified by other groups and the amino acids in the APP and mutated APP sequences were all large and hydrophilic. No noticeable neighbor affects were seen for amino acids in P2 which supports the lack of specificity in P2.

There was a strong preference for aromatic or hydrophobic amino acids in P1 (Table 31). Surprisingly, cysteine was seen for the chymotrypsin DH5α library. However, for all peptides identified by PICS, cysteine was modified by carbamidomethylation, making it much larger. Serine was seen in both of the GluC DH5α and tryptic S2 libraries, but it is near the cut-off level for amino acids identified once normalized by their natural abundance. Data from additional GluC DH5α and tryptic S2 libraries would be necessary to confirm with statistical confidence that serine is

indeed preferred in P1. Neighbor affects were seen for phenylalanine and tyrosine in P1. With tyrosine in P1, there was a positive correlation between serine in P2 and valine in P2', but these preferences were not seen with phenylalanine in P1. With phenylalanine in P1, the reverse cooperativity with valine in P3 was observed. There was also a positive correlation for glutamate in P2 and glutamine in P3'.

There was no strong preference for any amino acid in P1'. Methionine, glycine, serine and tryptophan were all identified using PICS (Table 31). Data from other studies identified glutamine, glutamate, methionine, and alanine in P1'.[42] Again, this indicated that there were no strong preferences for a particular type of amino acid in P1', even tryptophan was identified in the GluC DH5α library.

There was a strong preference for hydrophobic or small amino acids in P2'. The hydrophobic amino acids valine, isoleucine, and methionine were strongly preferred in P2', though glycine and alanine were also seen here (Table 31). Our PICS data indicated that valine was the most preferred amino acid in this position, a trend seen in other studies as well.[42] A genome-wide study in Iceland revealed that a A673T mutation in the P2' site of APP was protective against Alzheimer's disease and age-related cognitive decline. Our PICS data also was consistent with this decrease in BACE1 activity when threonine is in P2'[33]. Threonine is a small, hydrophilic amino acid. This confirmed the preference for hydrophobic amino acids. No particularly significant neighbor effects were observed for P2'.

There was no strong preference for any amino acid in P3'as identified with the PICS method (Table 31). Different studies identified both large and small amino acids in

this position, indicating that the amino acid in this position does not determine cleavage by BACE1A.

There were both aromatic and small amino acids identified in P4' (Table 31) as well indicating no particular amino acid preference in this position for BACE1A.

In conclusion, only the positions P3, P1, and P2' showed strong preferences for certain types of amino acids in the PICS samples. In P3, the amino acids isoleucine, valine, and leucine were favored. In P1, aromatic or large amino acids were favored. In P2', valine was strongly preferred, but other amino acids such as glycine and isoleucine were also seen. Only hydrophilic amino acids were identified in this position. Neighbor affects indicate the following sets of amino acids: P3: V, P2: E, P1: F, P3': Q and P3: L, P1: N and P2: S, P1: Y, P2': V (Table 31). The optimal sequences identified using PICS and the cooperative neighbor affects were P3: I, P2: S, P1: Y, P1': S, P2': V or P3: V, P2: E, P1: F, P1': M, P2': V/I, P3': Q.

The BACE1A active site with an inhibitor bound is shown in Fig. 28. There are no strong constraints on the size of the peptide, except in P1'. This confirmed the amino acids identified with PICS, except for tryptophan seen in the GluC DH5α library.

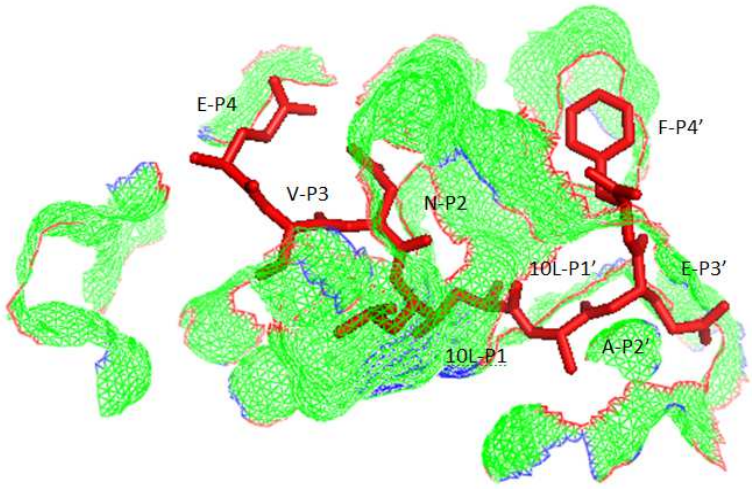


Figure 28: BACE1A active site generated using PyMOL. The inhibitor, modified at P1 and P1', is shown in red. The protein's active site is green. There were no large constraints except for P1'.

Sequence preferences for BACE2A

Compiling the data for all of the BACE2A (a close homolog of BACE1A) samples gives a more complete understanding of its preferred cleavage sequence. The cleaved sequence tolerated many amino acids in all positions; however, there are marked preferences for certain amino acids or types of amino acids (Table 32). Based on their homology, we predicted that there would be significant similarities between the sequence specificity for BACE1A and for BACE2A. However, we were also looking for the differences between the cleavage sequence preferences for the two proteases. For BACE2A as well as BACE1A, hydrophobic amino acids were seen in most positions. Neighbor effects were not seen, because there were too few peptides identified to gain a complete understanding of the cooperativity between neighboring amino acids. In general, the results for BACE2A should be considered preliminary and interpreted

cautiously due to the small number of peptides analyzed in the PICS procedure for this enzyme.

Table 32: Compilation of the amino acids comprising subsite positions P4-P4' that were preferred by BACE2A. The amino acids identified by the PICS procedure were compared to known sequences cleaved by BACE1A as well as the preferred sequence for BACE2 identified by Turner et al. in 2002.[33,41,60]

Subsite	P4	P3		P2			P1		P1'		P2'		P3'		P4'
APP Sequence	E	V		K			M		D		A		E		F
Swedish APP mutation	E	V		N			L		D		A		E		F
Protective APP mutation	E	V		K			M		D		T		E		F
Tryptic DH5α Library	M			W			F	D			P/M		P		Y
Chymotryptic DH5α Library	W/D	W	I	W		G	T	E	P	V	P	L	F	A	Q/I
Turner/Tang Data	E/Q	L/I/V		N	D		F/Y	L	M/L		I/V	E	F/Y/W	D	W/F

In P4, there were multiple types of amino acids identified (Table 32) indicating no strong preference for any amino acid in this position. In P3, hydrophobic and aromatic residues were preferred, specifically isoleucine and tryptophan (Table 32). These amino acids were similar to the preference found for BACE1A.No strong preference was seen for residues in P2 (Table 32) as small and large, aromatic amino acids were seen in this position. Though the number of peptides that were analyzed was small, the lack of any specific type of amino acids indicated that this position may not lead to selective cleavage by BACE2A consistent with what was seen for BACE1A.

In P1, there was no strong preference for any type of amino acid (Table 32). However, phenylalanine was seen, which was also seen for BACE1A. Interestingly, acidic amino acids were also seen in P1 for both the tryptic and chymotrypsin DH5α libraries. This did not coincide with the PICS data for BACE1A or with other research groups' studies.[41]

In P1', there was a preference for hydrophobic or small amino acids, which agreed with both the PICS data for BACE1A and other research groups' studies (Table 32).[41]

In P2', the only amino acids seen were hydrophobic (Table 32). This preference was seen in BACE1A and other research groups' studies as well.[41] There was a very strong preference for hydrophobic amino acids in both BACE1A and BACE2A. This position appeared most influential in determining cleavage by any BACE protease.

In P3', there was not a strong preference for any amino acid. Aromatic, hydrophobic, and small amino acids were all seen here (Table 32).

In P4', there was not a strong preference for any particular amino acid either. Glutamine, isoleucine, and tyrosine were all seen here (Table 32) indicating that this position does not have a strong effect on cleavage by BACE2A.

Conclusions

The PICS procedure can be used to gain an understanding of peptide sequences preferentially cleaved by any protease. Here, the PICS data analysis method was optimized with three different sets of controls to validate the data obtained for the protease of interest. The parameters of two different techniques used to sequence the N-terminal peptides and match them to the proteome of the organism to determine the C-terminal sequence were optimized. The best technique with optimized parameters was then used to determine the amino acid sequence cleaved by BACE1A and BACE2A. Previous studies have revealed information about the sequence cleaved by BACE1A, but these studies have not been able to identify the influence of the neighboring amino acid

cooperativity. This is one of the advantages of using PICS: neighbor affects can be investigated with this method, gaining novel information about the sequence cleaved by the protease of interest.

BACE1A is an important protease involved in Alzheimer's disease pathogenesis. A deeper understanding of the sequence cleaved by BACE1A would be used to understand the biological role of BACE1 as well as to guide the creation of inhibitors in the treatment of AD. The PICS procedure identified amino acid preferences for positions P4-P4', however only the positions P3, P1, and P2' showed strong preferences. In P3, the amino acids isoleucine, valine, and leucine were favored. In P1, the aromatic amino acids, tyrosine and phenylalanine were favored. In P2', valine was strongly preferred, but other amino acids such as glycine and isoleucine were also seen. Neighbor effects could be seen for P3 and P1, indicating that the following sets of amino acids are preferentially cleaved by BACE1A: P3: V, P2: E, P1: F, P3': Q and P3: L, P1: N and P2: S, P1: Y, P2': V. The optimal sequences identified using PICS and the cooperative neighbor affects were P3: I, P2: S, P1: Y, P1': S, P2': V or P3: V, P2: E, P1: F, P1': M, P2': V/I, P3': Q.

BACE2A is a homolog of BACE1A. It prefers to cleave a similar sequence, but there are some indications that its sequence preference is distinct. Initial data has been obtained for this protease using PICS. However, more samples need to be analyzed to gain a full understanding of the preferred sequence cleaved by this protease. Knowledge of the preferred sequence cleaved by the BACE family of proteases could lead to the

identification of novel native substrates, advancing our understanding of the function of this unique and interesting family of membrane-associated aspartyl proteases.

Works Cited

1. Barrett, A. J., Rawlings, N. D. & Woessner, J. F. *Handbook of Proteolytic Enzymes*. 1125–1127, 1483 (Elsevier Academic press: 2004).

2. Puente, X. S., Sánchez, L. M., Overall, C. M. & López-Otín, C. Human and mouse proteases: a comparative genomic approach. *Nature reviews. Genetics* **4**, 544–58 (2003).

3. Turk, B. Targeting proteases: successes, failures and future prospects. *Nature reviews. Drug discovery* **5**, 785–99 (2006).

4. McDonald, J. K. An overview of protease specificity and catalytic mechanisms: aspects related to nomenclature and classification. *The Histochemical journal* **17**, 773–85 (1985).

5. Schechter, I. & Berger, A. On the size of the active site in proteases: pronase. *Biochemical and biophysical research communications* **46**, 1956–60 (1972).

6. Hartley, B. B. S. Proteolytic Enzymes. *Annual reviews of biochemistryochemistry* **29**, 45–72 (1960).

7. Barman, A. & Prabhakar, R. Elucidating the catalytic mechanism of b--secretase (BACE1): A quantum mechanics / molecular mechanics (QM / MM) approach. *Journal of Molecular Graphics and Modelling* **40**, 1–9 (2013).

8. Polgar, L. The mechanism of action of aspartic proteases involves “push-pull” catalysis. *Federation of European Biochemical Societies* **219**, 1–4 (1987).

9. Becker-Pauly, C. *et al.* Proteomic analyses reveal an acidic prime side specificity for the astacin metalloprotease family reflected by physiological substrates. *Molecular & cellular proteomics : MCP* **10**, M111.009233 (2011).

10. Southan, C. A genomic perspective on human proteases as drug targets. *Drug discovery today* **6**, 681–688 (2001).

11. Abbenante, G. & Fairlie, D. P. Protease inhibitors in the clinic. *Medicinal chemistry* **1**, 71–104 (2005).

12. Yan, R. *et al.* Membrane-anchored aspartyl protease with Alzheimer’s disease b-secretase activity. *Nature* **402**, 533–537 (1999).

13. Lin, X. *et al.* Human aspartic protease memapsin 2 cleaves the beta-secretase site of beta-amyloid precursor protein. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 1456–60 (2000).

14. Hussain, I. *et al.* Identification of a novel aspartic protease (Asp 2) as beta-secretase. *Molecular and cellular neurosciences* **14**, 419–27 (1999).

15. Vassar, R. Beta-Secretase Cleavage of Alzheimer's Amyloid Precursor Protein by the Transmembrane Aspartic Protease BACE. *Science* **286**, 735–741 (1999).

16. Vassar, R., Kovacs, D. M., Yan, R. & Wong, P. C. The beta-secretase enzyme BACE in health and Alzheimer's disease: regulation, cell biology, function, and therapeutic potential. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **29**, 12787–94 (2009).

17. Ehehalt, R. *et al.* Splice variants of the beta-site APP-cleaving enzyme BACE1 in human brain and pancreas. *Biochemical and biophysical research communications* **293**, 30–7 (2002).

18. Tanahashi, H. & Tabira, T. Three novel alternatively spliced isoforms of the human beta-site amyloid precursor protein cleaving enzyme (BACE) and their effect on amyloid beta-peptide production. *Neuroscience letters* **307**, 9–12 (2001).

19. Mowrer, K. R. & Wolfe, M. S. Promotion of BACE1 mRNA alternative splicing reduces amyloid beta-peptide production. *The Journal of biological chemistry* **283**, 18694–701 (2008).

20. Zohar, O., Cavallaro, S., D'Agata, V. & Alkon, D. L. Quantification and distribution of beta-secretase alternative splice variants in the rat and human brain. *Brain research. Molecular brain research* **115**, 63–8 (2003).

21. Sinha, S. *et al.* Purification and cloning of amyloid precursor protein B-secretase from human brain. *Nature* **705**, 537–540 (1999).

22. Willem, M., Lammich, S. & Haass, C. Function, regulation and therapeutic properties of beta-secretase (BACE1). *Seminars in cell & developmental biology* **20**, 175–82 (2009).

23. Nussbaum, J. M. *et al.* Prion-like behaviour and tau-dependent cytotoxicity of pyroglutamylated amyloid-β. *Nature* **485**, 651–5 (2012).

24. Holler, C. J. *et al.* BACE2 expression increases in human neurodegenerative disease. *The American journal of pathology* **180**, 337–50 (2012).

25. Yang, L.-B. *et al.* Elevated β-secretase expression and enzymatic activity detected in sporadic Alzheimer disease. *Nature Medicine* **9**, 3–4 (2003).

26. Wang, H., Li, R. & Shen, Y. b-Secretase : its biology as a therapeutic target in diseases. *Trends in Pharmacological Sciences* **34**, 215–225 (2013).

27. Hu, X., He, W., Luo, X., Tsubota, K. E. & Yan, R. Report BACE1 Regulates Hippocampal Astrogenesis via the Jagged1-Notch Pathway. *CellReports* **4**, 40–49 (2013).

28. Cole, S. L. & Vassar, R. The Role of Amyloid Precursor Protein Processing by BACE1 , the B -Secretase , in Alzheimer Disease Pathophysiology. *The Journal of biological chemistry* **283**, 29621–29625 (2008).

29. Ishii, K. *et al.* Increased Ab 42 (43) -plaque deposition in early-onset familial Alzheimer ' s disease brains with the deletion of exon 9 and the missense point mutation (H163R) in the PS-1 gene. **228**, 17–20 (1997).

30. Feng, X., Zhao, P., He, Y. & Zuo, Z. Allele-specific silencing of Alzheimer's disease genes: the amyloid precursor protein genes with Swedish or London mutations. *Gene* **371**, 68–74 (2006).

31. Kumar-Singh, S. *et al.* Dense-core senile plaques in the Flemish variant of Alzheimer's disease are vasocentric. *The American journal of pathology* **161**, 507–20 (2002).

32. Kumar-Singh, S. *et al.* Behavioral disturbances without amyloid deposits in mice overexpressing human amyloid precursor protein with Flemish (A692G) or Dutch (E693Q) mutation. *Neurobiology of disease* **7**, 9–22 (2000).

33. Jonsson, T. *et al.* A mutation in APP protects against Alzheimer's disease and age-related cognitive decline. *Nature* **488**, 96–9 (2012).

34. Sun, X. *et al.* Distinct transcriptional regulation and function of the human BACE2 and BACE1 genes. 739–749doi:10.1096/fj.04-3426com

35. Barbiero, L. *et al.* BACE-2 is overexpressed in Down ' s syndrome. **182**, 335–345 (2003).

36. Gilmour, L. *et al.* ASP1 (BACE2) Cleaves the Amyloid Precursor Protein at the B-Secretase Site. *Molecular and cellular neurosciences* **619**, 609–619 (2000).

37. Ahmed, R. R., Holler, C. J., Webb, R. L., Li, Feng Beckett L., T. & Murphy, M. P. BACE1 and BACE2 enzymatic activities in Alzheimer's disease. *Journal of Neurochemistry* **112**, 1045–1053 (2010).

38. Casas, S. *et al.* BACE2 plays a role in the insulin receptor trafficking in pancreatic β -cells. *American Journal of Physical Endocrinal Metabolism* **299**, 1087–1095 (2010).

39. Hemming, M. L., Elias, J. E., Gygi, S. P. & Selkoe, D. J. Identification of beta-secretase (BACE1) substrates using quantitative proteomics. *PloS one* **4**, e8477 (2009).

40. Dominguez, D. *et al.* Phenotypic and biochemical analyses of BACE1- and BACE2-deficient mice. *The Journal of biological chemistry* **280**, 30797–806 (2005).

41. Turner, R. T., Loy, J. A., Nguyen, C., Devasamudram, T. & Ghosh, A. K. Specificity of Memapsin 1 and Its Implications on the Design of Memapsin 2. **2**, 8742–8746 (2002).

42. Turner, R. T. *et al.* Subsite Specificity of Memapsin 2 (B-Secretase): Implications for Inhibitor Design. *Biochemistry* **40**, 10001–10006 (2001).

43. Edman, P., Hoegfelt, E., Sillen, L. G. & Kinnel, P.-O. Method for determination of the amino acid sequence in peptides. *Acta Chem. Scand.* **4**, 283–293 (1950).

44. Becker, C. H. & Bern, M. Recent developments in quantitative proteomics. *Mutation research* **722**, 171–82 (2011).

45. Steen, H. & Mann, M. The ABC's (and XYZ's) of peptide sequencing. *Nature reviews. Molecular cell biology* **5**, 699–711 (2004).

46. Wilm, M. Femtomole sequencing of proteins from polyacrylamide gels by nano electrospray masss spectrometry. *Nature* **379**, 699–711 (2004).

47. Anderson, L. & Hunter, C. L. Quantitative mass spectrometric multiple reaction monitoring assays for major plasma proteins. *Molecular & cellular proteomics : MCP* **5**, 573–88 (2006).

48. Roepstorff, P. & Fohlman, J. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomedical Mass Spectrometry* **11**, 601 (1984).

49. Ziady, A. G. & Kinter, M. Protein sequencing with tandem mass spectrometry. *Methods in molecular biology (Clifton, N.J.)* **544**, 325–41 (2009).

50. Paizs, B. & Suhai, S. Fragmentation pathways of protonated peptides. *Mass spectrometry reviews* **24**, 508–48 (2005).

51. Zhang, Z. Prediction of low-energy collision-induced dissociation spectra of peptides. *Analytical chemistry* **76**, 3908–22 (2004).

52. Wysocki, V., Tsaprailis, G., Smith, L. & Breci, L. Mobile and localized protons: A framework for understanding peptide dissociation. *Journal of Mass Spectrometry* **35**, 1399–1406 (2000).

53. Perry, R. H., Cooks, R. G. & Noll, R. J. ORBITRAP MASS SPECTROMETRY : INSTRUMENTATION , ION MOTION AND APPLICATIONS. 661–699 (2008).doi:10.1002/mas

54. Han, X., Aslanian, A. & Yates, J. R. Mass spectrometry for proteomics. *Current opinion in chemical biology* **12**, 483–90 (2008).

55. Schilling, O., Huesgen, P. F., Barré, O., Auf dem Keller, U. & Overall, C. M. Characterization of the prime and non-prime active site specificities of proteases by proteome-derived peptide libraries and tandem mass spectrometry. *Nature protocols* **6**, 111–20 (2011).

56. Schilling, O. & Overall, C. M. Proteome-derived, database-searchable peptide libraries for identifying protease cleavage sites. *Nature biotechnology* **26**, 685–94 (2008).

57. Overall Lab CLIP-PICS. at <http://clipserve.clip.ubc.ca/pics/>

58. Neil D. Rawlings, G. S. S. *Handbook of Proteolytic Enzymes*. 2626 (Academic Press, 2012: 2012).

59. Humbolt University of Berlin. at <http://www.physik.hu-berlin.de/nano-en/forschung-en/np>

60. Johnstonat, J. A. *et al.* Increased B-amyloid release and levels of amyloid precursor protein (APP) in fibroblast cell lines from family members with the Swedish Alzheimer ' s disease APP670 / 67 1 mutation. *FEBS Letters* **354**, 274–278 (1994).