Comparative Proteomics of the Conservation of BACE1 Substrates Across Species

Shannon King, Meghan Fritz and Dr. Joseph L. Johnson

University of Minnesota Duluth, Swenson College of Science and Engineering, Department of Chemistry and Biochemistry

In a world of modern medicine where people are living longer, Alzheimer's disease (AD) is becoming increasingly relevant and problematic. As the number of people living with AD in the United States is greater than five million and increasing with each passing year, a better understanding of the disease mechanism is necessary to properly care for and treat these individuals. The beta-site amyloid precursor protein (APP) cleaving enzyme 1 (BACE1) is a membrane-bound aspartyl protease. It is responsible for cutting APP to form short β-amyloid (Aβ) peptides. The accumulation and aggregation of these peptides is a defining attribute of AD. Therefore, BACE1 has become a promising therapeutic target for the treatment and prevention of the disease. However, APP is not the only substrate cut by BACE1. Due to the importance of BACE1 as a potential target for AD treatment, it is imperative to fully understand its native function by identifying all possible substrates. In identifying such substrates, we hypothesize that native substrates (and the specific sequences) that are cleaved by BACE1 will be conserved throughout all species with BACE1 orthologs. In order to identify potential BACE1 substrates, the bioinformatics tools SignalP and TMHMM were used to identify all Type I membrane proteins (c-terminus of the protein is cytoplasmic) within a species' proteome that have a single transmembrane domain since all known BACE1 substrates possess these characteristics. We describe our comparative analysis of the preferred BACE1 cut sites within these proteomes using published data about the enzymes preferred cut sequences. Analysis of the data from a variety of organisms showed the protein sequences cut by BACE1 were conserved. Having a better understanding of the identity of BACE1 substrates will be beneficial in identifying and reducing the side effects that may be present due to AD treatments that target this enzyme.

INTRODUCTION

Alzheimer's disease (AD) is a neurodegenerative disease that affects memory and behavior. According to the Alzheimer's Association, 5 million Americans currently have AD, and that number is expected to increase to 14 million by the year 2050 (Alzheimer's Association). This drastic increase is due to improvements in modern medicine leading to increased longevity, meaning that the natural effects of aging are being seen more prominently. Currently there are no known cures for AD; however, there are a few medications that primarily treat its various symptoms. Researchers are looking for new and improved treatments that target specific aspects of the disease pathway. In order to achieve this, it is important to understand all aspects of the disease mechanism to create an effective treatment and avoid causing further damage. Specifically, this includes characterization of all proteins involved, what the roles and functions of these proteins are, as well as the identities of the substrates and the products or outcomes they may lead to.

Alzheimer's Disease Mechanism

In order to determine viable methods to treat AD, we must understand the disease mechanism and identify areas to target that would slow or reverse the progression of the disease or prevent it altogether. However, it is important to remember that the brain is extremely complex and involves many interconnected pathways, meaning that there can be many factors that may lead to the development of AD including genetics and environmental influences. The prevailing hypothesis for the development of AD is called the amyloid cascade hypothesis. The main components in this pathway are the amyloid precursor protein (APP) and microtubule associated protein Tau. Peptide fragments derived from APP (called β-amyloid or Aβ) aggregate

to form amyloid plaques and Tau aggregates to form neurofibrillary tangles (NFTs) which ultimately lead to neuronal loss, progressive impairment of short term memory, cognitive decline, and dementia (Lane et al., 2018). Our research focuses on the formation of APPs and looking to stop or slow AD by inhibiting their production.

β-secretase 1 (BACE1)

Amyloid plaques are composed of aggregated Aβ peptides derived from the APP and lead to the development of AD by disrupting neuronal function. β-secretase (BACE1), also called the beta-site amyloid precursor protein (APP) cleaving enzyme 1, is a membrane-bound aspartyl protease that cuts APP generating these Aβ peptides (Munro et al., 2016). These peptides aggregate to form the amyloid plaques that are a hallmark attribute of AD making them a promising target for AD treatment and prevention. BACE1 also has a homolog; the beta-site amyloid precursor protein (APP) cleaving enzyme 2 (BACE2) which is similar in structure and function to BACE1 and was analyzed simultaneously using the same procedure. Study of both proteins is critical in order to fully understand their role in the AD mechanism.

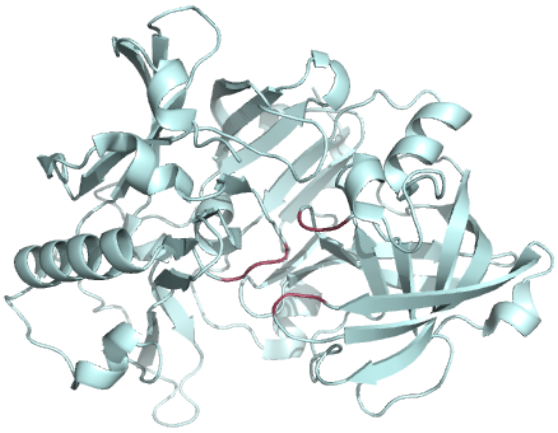

Figure 1. BACE1 structure. Active site is shown in red.

Research Goals

The purpose of performing this research was to work towards answering two main questions. What are the native substrates cut by BACE1 and, what happens if you block native substrates from being cut by BACE1 for prolonged periods of time? It is important to be able to answer these two questions in order to fully understand the effects of inhibiting this enzyme when designing new medications and treatments. Some of the specific goals of this summer project were to identify species with BACE1 orthologs, which are proteins with similar sequence, structure, and function in another species. Additionally, we aimed to identify possible BACE1 substrates. We hypothesized that the active sites of these orthologs would be conserved as well as the native substrates (and cut sites) across all of the species with BACE1 orthologs. Further, we hypothesized that species that are more closely related to humans would have orthologs and substrates that are more similar to that of a human. This is based on the theory of evolution and the idea that throughout time, as species evolved, important elements largely maintained their identity and species that evolved from each other or are more closely related would express more similar genetics. Ultimately, by identifying possible BACE1 substrates in these orthologs we can better understand the function of BACE1 and its role in the AD disease mechanism.

METHODS

Foundational Work

The entirety of this project was completed using bioinformatics which required the use of large databases of protein sequences and various softwares and programs to analyze the data. In

order to first identify species that contain human BACE1 orthologs, UniProt's BLAST feature was used to identify potential BACE1 orthologs (similar protein sequences in other organisms ranked by their percent similarity). Both MUSCLE and ClustalW multiple sequence alignments were used to compare the similarity of these sequences and to develop phylogenetic trees that show the homology of the sequence across the organisms. Jalview was also used to view these multiple sequence alignments and to identify regions that were conserved across all species. Another important tool that was used was PyMOL which enabled us to view the protein in three dimensions. Using a published BACE1 structure that includes an inhibitor, we were able to locate the active site of the protein and identify the amino acid residues responsible for forming the active site. By comparing the protein structure and the multiple sequence alignments, we could determine if the active sites were conserved as we hypothesized.

Preliminary Results

Using the aforementioned steps we compiled a list of over 100 species that have BACE1 orthologs. That list was narrowed down to about 50 species distributed among the six animal classes. Once the active site was located in PyMOL, each of these residues was examined in the multiple sequence alignment. The aspartate (Asp, D) residue was 100% conserved in each of the sequences in our list of 50 species. The other active site residues were also highly conserved. This was true for both the MUSCLE and ClustalW alignments which validated the approach of our study. Using the alignments of the amino acid sequences for BACE1 orthologs, we were able to develop a phylogenetic tree as shown in Figure 2. This tool allowed us to observe how closely related species were predicted to be based on the sequence of a single protein and compare it to the relationships that are commonly accepted in evolutionary theory. Our approach

for this project was rooted in the idea that species that are more closely related will have more similar genetics. Specifically, species that are known to be more closely related to humans will have more similar BACE1 orthologs as well as natural substrates that have a higher degree of conservation. By looking at how closely related specific species were, we were able to predict which ones would have natural BACE1 substrates that were most similar to that of a human. Using the phylogenetic tree the list was narrowed to 10 species, maintaining the representation of all six of the animal classes, for the sake of the short timeframe of this summer project. The animals that were chosen were chimpanzee, black bear, pig, mouse, turtle, chicken, African frog, northern mallard, zebrafish and cockroach. We focused on these ten species when comparing BACE1 substrates in organisms with BACE1 orthologs.

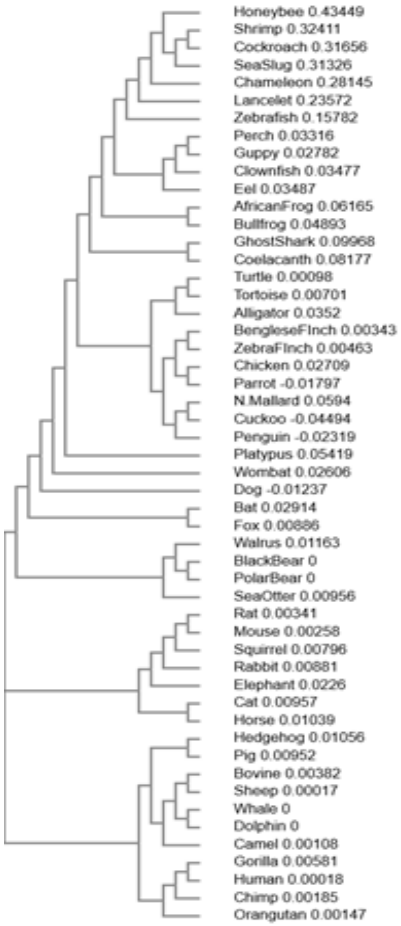

Figure 2. Phylogenetic tree of species with BACE1 orthologs. List of 50 species with highly conserved BACE1 orthologs. Ten species were selected to focus on for this study; chimpanzee, black bear, pig, mouse, turtle, chicken, African frog, northern mallard, zebrafish and cockroach.

Identifying BACE1 Substrates

Coming back to the question of how to identify specific BACE1 substrates, we used software called TMHMM and SignalP in order to identify proteins within each organism's proteome that could be potential substrates. We based our search for substrates on characteristics of known BACE1 substrates. Therefore, we were looking for proteins that contain a single Type I transmembrane domain as BACE1 is also a Type I integral membrane protein. Using UniProt we were able to download the entire proteome of each of the ten species we chose to focus on. Sections of the protein sequences were first run through the program TMHMM which scans each protein in the proteome for transmembrane domains by using an algorithm that identifies hydrophobic stretches of amino acids approximately 20 amino acids long that would likely form transmembrane helices. Based on the results from TMHMM, we selected the amino acid sequences that had one or more supposed transmembrane domains for further analysis. These protein sequences were then run through the program SignalP which scans the proteins or amino acid sequences for signal sequences. A signal sequence is a stretch of approximately 20 amino acids at the N-terminus of a protein that targets it to the secretory pathway. This additional step was necessary because the TMHMM algorithm frequently misidentifies signal peptides as transmembrane domains. Therefore, we used the mature SignalP outputs which removes the signal sequence from the amino acid sequence to then rerun TMHMM. This ensures that we are only looking at potential substrates that will fit in the active site of BACE that contain a single transmembrane domain. The outputs from the second run through TMHMM were sorted by the number of transmembrane helices and their orientation in the membrane. We proceeded by first

looking at the proteins with one transmembrane domain and then considered those with two transmembrane domains, but did not evaluate proteins with more than two transmembrane domains as they typically become too complex to be likely substrates.

Sequence Analysis

In order to analyze the group of Type I membrane proteins with one transmembrane domain, we used an Excel program written by Dr. Joseph Johnson. The program was designed based on research published by Turner et al. characterizing preferred cut sites for BACE1 which is shown in Figure 2. BACE1 is known to cut the middle peptide bond in octapeptide recognition sequences (between the P1 and P1' positions). They experimentally determined the preference of BACE1 to cut different octapeptide sequences with different amino acids at each of the eight positions by measuring the relative rate of catalysis. This sequence preference information was incorporated into the Excel algorithm by assigning a preference index each amino acid in each position. The preference for a known BACE1 substrate octapeptide sequence was used to determine the threshold cutoff values of $1x10^{-4}$. The full length amino acid sequence of each protein in the single transmembrane domain subset was loaded into the program, and the algorithm scanned the protein with a sliding frame looking at eight amino acids at a time. Each octapeptide was given a score. The potential octapeptide cut site sequences that exceeded the threshold were recorded. This was done for every protein that had a single Type I transmembrane domain in each of the ten chosen species as well as in humans.

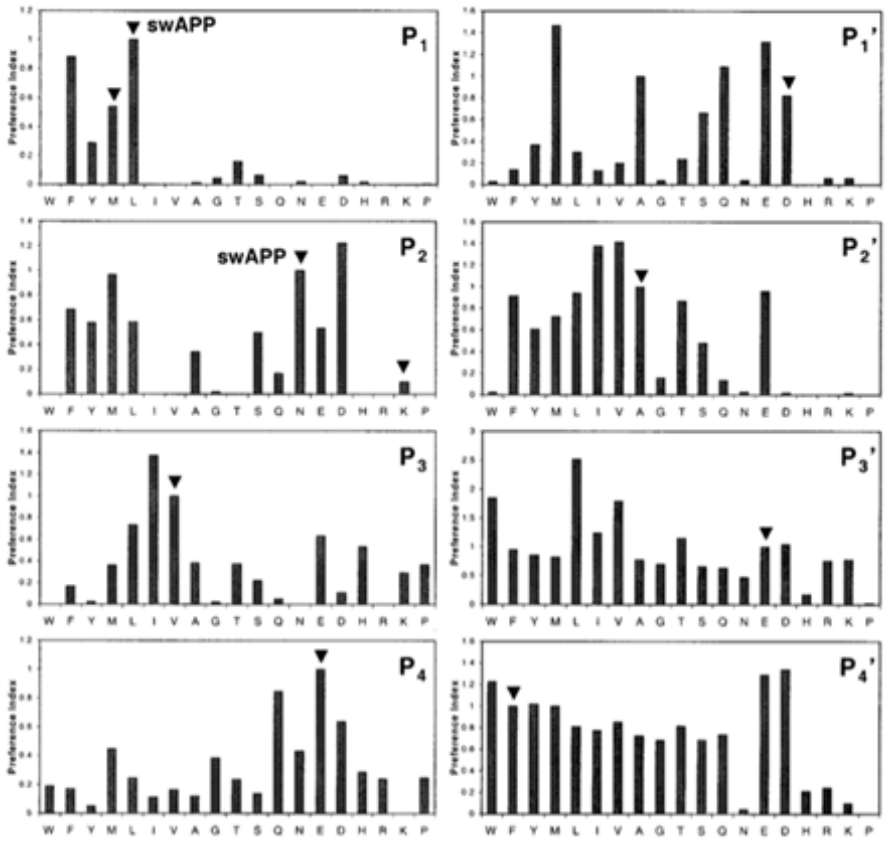

Figure 3. Preference of BACE1 for amino acids in each position in substrate octapeptide. BACE1 cuts octapeptides (P4, P3, P2, P1, P1', P2', P3', P4') between the P1 and P1' positions; however, the likelihood that it will cut an amino acid depends on which amino acid is in each of those eight positions. These graphs show the experimentally determined preference of BACE1 for a substrate with each amino acid in each position (a taller bar indicates greater preference). The arrows mark the natural amino acid sequence for APP which is a common substrate for BACE1. The arrows with swAPP indicate a Swedish mutation of APP that was more readily cleaved by BACE1 (meaning that it is a better substrate).

RESULTS

As each protein sequence in every species was analyzed 8 amino acids at time, each octapeptide was assigned a number value or score corresponding to how likely BACE1 was to cut that octapeptide based on which amino acids were in each position. This was determined by the preference table that was generated by Turner et al. For each protein, a list was compiled

with all of the octapeptides that met the threshold of $1x10^{-4}$, which corresponded to the score for the octapeptide cleavage site in native APP. The top twenty octapeptides that surpassed this threshold were considered to be likely cut sites and were listed in the final table produced by the program. The table for each protein listed the location of the transmembrane region residues within the amino acid sequence, what the octapeptide cut sequence was, as well as its location. This is important because we were specifically looking for a cut site that is within approximately 50 amino acids of the transmembrane region based on what we know about the structure of BACE1, the location of the active site, and the cleavage sites in known substrates. Using the information presented in the results table for each protein, we were able to select protein sequences that would likely be viable substrates based on their meeting the desired criteria.

amyloid precursor protein							
APP	1209	668	1.41E-03	0.740	EVKMDAEF	Human	
APP	933	594	1.41E-03	0.740	EVKMDAEF	Big-headed Turtle	
APP	463	649	1.41E-03	0.740	EVKMDAEF	Chicken	
APP	1292	668	1.41E-03	0.740	EVKMDAEF	Chimpanzee	
APP	1275	668	1.41E-03	0.740	EVKMDAEF	Mouse	
APP	782	615	1.41E-03	0.740	EVKMDAEF	N. Mallard	
APP	782	635	1.17E-02	0.829	LVFFAEDV	N. Mallard	*
APP	1260	668	1.41E-03	0.740	EVKMDAEF	Pig	
APP	2323	647	3.48E-04	0.665	EVKMDSEY	African Clawed Frog	*
APP	2323	655	1.34E-04	0.587	RHDTAYEV	African Clawed Frog	*
APP	958	653	1.17E-02	0.829	LVFFAEDV	Black Bear	*
APP	565	242	1.04E-03	0.742	DPETMEDE	German Cockroach	*
APP	759	619	9.23E-04	0.835	VPDLDLAT	Zebrafish	*
p-selectin glycoprotein ligand 1							
p-SGL1	616	256	1.12E-03	0.885	LAAMEALS	Human	
p-SGL1	634	272	1.12E-03	0.885	LAAMEALS	Chimpanzee	
p-SGL1	550	1691	2.59E-04	0.600	TLDTSTIQ	African Clawed Frog	*
p-SGL1	555	255	1.14E-04	0.682	PEATDALS	Black Bear	*
p-SGL1	465	191	1.69E-04	0.673	NVSSEAAV	Big-headed Turtle	*
p-SGL1	969	982	1.57E-04	0.714	ATELAEQF	Zebrafish	*

Figure 4. Comparative analysis of BACE1 cut sites in known BACE1 substrates. Compiled list of possible cut sites identified by the Excel program for two known substrates of BACE1. APP has a conserved cut site in each of the ten species we chose to research as well as additional cut sites in some species. Five species had possible cut sites in p-SGL1, which were less conserved.

We began by comparing the results produced by the program for two of the known substrates of human BACE1: amyloid precursor protein (APP) and p-selectin glycoprotein ligand 1 (p-SGL1). By beginning with two known substrates of BACE1, we were able to confirm that our process for selecting potential substrates substrates was effective. Additionally, it allowed us to analyze the conservation of predicted cut sites across the species and compare these differences to our phylogenetic tree.

Amyloid Precursor Protein (APP)

APP is the source of Aβ peptides and is the most well known substrate of BACE1. All of the species we studied had this protein and six of them had the cut sequence entirely conserved. As shown in Figure 3, the big-headed turtle, chicken, chimpanzee, mouse, northern mallard, and pig all had APP cut sequences that were identical to the cut site in human APP. The other four species had at least one viable cut site in their APP sequence that were not entirely conserved and varied in their degree of conservation. For example, the african frog had all but two amino acids conserved. However, these single amino acid differences may have made the cut site more favorable based on the preference table. Additionally, the northern mallard and african frog both have additional possible cut sites in the APP sequence that are downstream from the known cut sequence. When considering the use of BACE1 and APP as potential targets for AD treatment, it is important to know if these sequences will be cut along with the first possible octapeptide cut sequence.

P-selectin glycoprotein ligand 1 (p-SGL1)

Another known substrate of human BACE1 is p-SGL1. Using UniProt, we were able to determine that not all of the species that we were studying had p-SGL1 or an ortholog. However, those that did have this protein did produce results showing probable cut sites within the protein sequence. The octapeptide recognition sites for the possible cut sites in p-SGL1 were far less conserved across the five species that contained this protein sequence; however, chimpanzees did have a sequence that was entirely conserved. Based on the phylogenetic tree that was constructed and what is understood about evolution, this is logical because chimpanzees are one of humans' closest animal ancestors. It is important to note, however, that for both this ligand and APP, although the cut sequence may not be entirely conserved, based on the substrate preference index, BACE1 is still likely to cut these octapeptides.

DISCUSSION

From these results we can conclude that both APP and p-SGL1 are conserved although to different extents. APP is known to be a more common substrate for BACE1 which is confirmed in our results showing an ortholog in each of the ten species we studied as compared to p-SGL1 which is less common and less conserved. Additionally, it is important to note that although the cut sequence in a potential substrate is not entirely conserved, the BACE1 protein itself is not 100% conserved. The protein also has the ability to cut different sequences as long as they meet the threshold and orientation requirements. There are also other known substrates that we did not look at in our analysis because they are not naturally occurring substrates, do not have orthologs in the species we looked at or did not have acceptable cut sites based on the parameters we used. After confirming that the native BACE1 substrates do have cut sites that surpass the

threshold set by the program, we can now begin to look for other possible substrates that may be conserved across species. In order to do this we will look through about 2,500 proteins for each species to find cut sites that meet the requirements previously established. By comparing these cut sequences across species, we can identify other potential substrates for BACE1 based on which sequences are conserved. From there we can expand our search through the set of proteins in the proteome of each species to those that have two transmembrane domains to identify possible substrates among those proteins as well.

CONCLUSION

Throughout this process we were able to use programs such as BLAST and various multiple sequence alignment tools in order to identify species with BACE1 orthologs. After generating a list of all of the proteins with a single transmembrane domain in each of our ten selected species using TMHMM, SignalP, and the Excel scoring algorithm, we were able to conclude that, in species with BACE1 orthologs, known or native substrates are conserved. This shows the importance of both the gene and the substrate through their consistency throughout evolution. Additionally, understanding that BACE1 cuts multiple different combinations of amino acids in its substrates shows that the protein may be active in more or different ways than previously thought. This leads us to have a better understanding of BACE1 in the AD disease mechanism which will aid in the development of new treatments for the disease.

CONTEXTUALIZATION OF WORK

The work discussed in this paper is quite relevant to the current state of scientific research. The field of neuroscience is becoming increasingly popular and is one of the most prominent areas of scientific research of recent years. We are learning more and more about how our brains work everyday and as we develop a greater understanding of the brain and neuroscience, we are able to tackle more complex issues such as AD and other neurodegenerative diseases with intricate and interconnected disease mechanisms and pathways. Specifically, Alzheimer's Disease research is very important due to the increase in the prominence of the disease. Many people personally know someone with AD or at least a grandparent that developed dementia, so they understand the pain the disease causes not only the individual with AD but to their friends and family as well with the strains and difficulties they face second hand. The decline of cognitive function is a unique kind of suffering and one that many people wish could be cured or slowed.

An increasing concern is that people now who are currently young to middle age are becoming more likely to develop the disease because we are living longer and due to increased exposure to environmental risk factors that have been introduced to our daily lives. Some of these include poor air quality from nitrogen oxides and tobacco smoke, pesticides in our foods and in the air and water, prolonged exposure to electric and magnetic fields in machines and from technology like computers and phones, as well as increased exposure to potentially harmful metals and other elements in our water, the soil or at work (Killin et al., 2016). Additional research has also shown that sleep deprivation and brain injuries can also increase risk for AD and both of these have seen an increase in numbers in recent years (Sadeghmousavi et al, 2020 &

Ramos-Cejudo et al., 2018). However, it is also important to remember that AD is highly linked to specific genetic factors as well.

In an attempt to contribute to the growing amount of knowledge on AD, my research is advancing the field of study by taking a deeper look into a specific aspect of the known disease mechanism. We were able to characterize BACE1 by identifying which regions of the protein are conserved across species as well as identifying new potential substrates. Although this research is still at its beginning stages, the ultimate end result will lead to an overall better understanding of the AD mechanism and pathway and may help researchers develop a new treatment for the disease.

ACKNOWLEDGMENTS

I would like to thank the Swenson Family Foundation for sponsoring this summer research project. I am so grateful for this opportunity to develop new research skills and prepare for my future career. I would also like to thank Meghan Fritz who worked on this project with me all summer for her collaboration and Dr. Johnson for mentoring us. Lastly, I would like to thank the University Honors Program for providing me with so many opportunities to expand my knowledge outside my comfort zone and to grow to become a more well rounded student and person.

REFERENCES

"Facts and Figures." Alzheimer's Association, 2021, https://www.alz.org/alzheimers-dementia/facts-figures

Killin, Lewis O. J. et al. Environmental risk factors for dementia: a systematic review. *BMC geriatrics* vol. 16,1 175. Oct. 2016, doi:10.1186/s12877-016-0342-y

Lane, C. A. Hardy, J., Schott, J. M. Alzheimer's Disease. *European Journal of Neurology* 2018, 25: 59–70.

Munro, K. M., et al.Functions of the Alzheimer's Disease Protease BACE1 at the Synapse in the Central Nervous System. *J Mol Neurosci.* 2016.

Turner III et al. "Subsite Specificity of Memapsin 2 (β-Secretase): Implications for Inhibitor Design." *Biochemistry*, 2001.

Ramos-Cejudo, Jaime et al. "Traumatic Brain Injury and Alzheimer's Disease: The Cerebrovascular Link." *EBioMedicine* 2018, 28: 21-30. doi:10.1016/j.ebiom.2018.01.021

Sadeghmousavi, S., Eskian, M., Rahmani, F. et al. The effect of insomnia on development of Alzheimer's disease. *J Neuroinflammation* 2020, 17:289. doi:10.1186/s12974-020-01960-9