Classification of Prion Strains with Polymorphism Dataset

August 2013

Laboratory of Computational Biology and Bioinformatics
Interdisciplinary Program in Bioinformatics
College of Natural Science
The Graduate School
Seoul National University

Ji-Hae Lee

A thesis submitted in fullfillment of the requirements for
the degree of Master of Science to
Seoul National University

다형성 데이터를 이용한 프라이온 서열의 분류

Classification of Prion Strains with Polymorphism Dataset

지도교수 손 현 석

이 논문을 이학석사 학위 논문으로 제출함

2013년 8월

서울대학교 대학원

협동과정 생물정보학

이 지 혜

이지혜의 석사 학위 논문을 인준함

2013년 6월

위 원 장 김 희 발 (인)

부 위 원 장 손 현 석 (인)

위 원 안 인 성 (인)

Abstract

Classification of Prion Strains with Polymorphism Dataset

Ji-Hae Lee

Interdisciplinary Program in Bioinformatics

The Graduate School

Seoul National University

Pathogenic prion which has undergone conformational change of normal PrP^{C} into abnormal PrP^{Sc} is known to be the causing material of Transmissible Spongiform Encephalopathy (TSE) including ovine scrapie, bovine spongiform encephalopathy (BSE), human kuru and Creutzfeld-Jacob Disease (CJD). These prion diseases are strongly concerned in public health context for their possibility of transmission to the human from other host species. PrP^{C} is mainly composed of α-helical structures while PrP^{Sc} is mainly composed of β-sheet structures. Numerous polymorphisms and mutations are found in the sequence of prion protein. These sequence difference might influence prion disease susceptibility through the modulation of protein conformational change and expressions. Though prion protein sequence polymorphisms which influence prion disease susceptibility was considered important, the database which is specific on theses polymorphisms has not been developed. Also, PrP^{Sc} is biochemically

difficult for experiments in the laboratory and computational simulations including molecular dynamics study need much computational resource. Thus, the prion disease susceptibility prediction method based on the polymorphism information of prion sequences without molecular biological experiments and detailed structural analysis was quite necessary. Following procedures were performed for the generation of prion polymorphism database. BLASTP application was exploited for the collection of mammalian prion protein sequences and ClustalW application was used for the multiple sequence alignment. Also, necessary information was parsed using JAVA programming language, and archived in the MySQL database tables. Following procedures were performed for the construction of computational application for the prion susceptibility prediction. The effect of polymorphisms and mutations on the prion disease susceptibility was investigated through literary search and training dataset was classified into four groups using polymorphism information. Discriminant analysis was performed based on training data, and they were classified as groups. In order to determine the accuracy of prediction in new strains, manually mutated sequences from reference sequence which does not appear polymorphism or mutation were generated and were used as test dataset of discriminant analysis. Position-specific scores were calculated using JAVA codes and the accuracy of discriminant analyses based on distance from either BLOSUM62 or PSSM were compared. Cross-validations were performed for the accuracy analysis of k-nearest neighbor and linear discriminant analysis. The classification was visualized in a 2D graph in canonical discriminant analysis. As a result, there was no association between the frequency of polymorphisms and susceptibility of prion disease. K-nearest neighbor method with k of three and four showed the most accurate susceptibility prediction among

the used discriminant analyses. The rate of misclassification was decreased and would more clearly discriminate in the 2D plot of canonical discriminant analysis when using BLOSUM62 than PSSM matrix. In addition, sequences with negative polymorphism have relatively higher accuracy of classification. But, the presence or absence of a specific polymorphism in prion sequence was difficult to accurately assess the risk of prion diseases. Through this research, polymorphism information was incorporated for the classifications and the discriminant analysis with distances from amino acid substitution matrix developed in this study. It might help prompt and correct prediction of prion disease susceptibility without experiments and also might be useful in public health context through additional research.

Keywords : prion, polymorphism, susceptibility, substitution score matrix, discriminant analysis

Student Number **:** 2011-20446

Table of Contents

Chapter 3. Results and Discussion

Chapter 4. Conclusions ····· 52

List of Figures

List of Tables

Chapter 1.
Introduction

1.1 Prion

Transmissible Spongiform Encephalopathies (TSE) is a mammalian neurodegenerative disorder which originates from the pathological functioning of abnormal form of prion protein (PrP^{Sc}) (Prusiner, 1991). This includes scrapie of sheep and goats, bovine spongiform encephalopathy (BSE) of cattle, chronic wasting disease (CWD) of deer, feline spongiform encephalopathy of cats, transmissible mink encephalopathy of minks, and kuru, Gerstmann-Sträussler-Scheinker syndrome (GSS), sporadic Creutzfeldt-Jakob disease (CJD), variant CJD, familial CJD and fatal familial insomnia (FFI) of human (Prusiner, 1998). Normal cellular prion (PrP^{C}) converts conformation into abnormal isoform (PrP^{Sc}) and accumulates in the brain tissue (Wadsworth *et al.*, 1999). The accumulation forms amyloid plaques and induces sponge-like vacuolation of brain tissues as neuropathological characters (Klamt *et al.*, 2001).

TSE was believed to be caused by a “slow virus” which does not provoke immunological reactions and only damages brain tissues (Gajdusek, 1972; Prusiner, 1998). The first remark of the slow virus which onsets the disease after the long incubation period from the infection of host appeared in 1954 by Sigurdsson (Sigurdsson, 1954). Gadjusek and colleagues later found that kuru, which is one of the prion diseases, is transmitted by cannibalism in Papua New Guinea (Gajdusek and Zigas, 1957). In 1959, Haldow suggested slow virus as the agent of the kuru disease of inhabitants of Papua New Guinea

which is similar to scrapie (Haldow, 1959). Gajdusek found that kuru also infects chimpanzees (Gajdusek, 1966). Griffith remarked that infectious agent of scrapie, which is one of the prion diseases, is a protein and explained the mechanism of self-replication of this protein (Griffith, 1967). Prusiner showed that infectious agent of scrapie is resistant to treatments that modify nucleic acids but sensitive to the protease treatments and named this non-virial, non-plasmidic, and non-viroidal proteinaceous infectious particles as "prion" (Prusiner, 1982). PrP gene knockout mice show increased resistance to the inoculation of mouse scrapie prion which implies the necessity of PrP gene for the susceptibility of scrapie (Büeler *et al*., 1993). TSE diseases are caused by prions through inheritance of pathogenic mutations, infections, and spontaneous processes (Prusiner, 1998). Human CJD generally displays progressive dementia and ovine scrapie and bovine spongiform encephalopathy usually develops ataxic illnesses (Wells *et al*., 1987). Pathogenic prion (PrP^{Sc}) has prevalent β-sheet structures which have been converted from α-helical structures of PrP^{C} (Pan *et al*., 1993).

Prion protein (PrP) is encoded by PrP gene (*PRNP*) and is well conserved in mammals (Sakudo *et al*., 2010). Open reading frame (ORF) of PrP gene of mouse is on the third exon of the second chromosome (Baybutt and Manson, 1997). PrP gene ORF of rat is on the third exon of the third chromosome, while PrP gene ORFs of sheep and cattle are on third exon of the thirteenth chromosome (Saeki *et al*., 1996; O'Neill *et al*., 2003; Inoue *et al*., 1997). Human PrP gene ORF is on the second exon of the 20th chromosome (Mahal *et al*., 2001). ORF of mammalian PrP genes usually exist on the last exon according to this information (Sakudo *et al*., 2010). 5' region of transcriptional initiating site of PrP gene shows short GC-rich feature which is generally shown in housekeeping genes (Puckett *et al*., 1991). 3'

-untranslated region (UTR) of mRNA contains functional sequence (ATTAAA) and intron 1 has putative binding sites for transcription factors (Kim *et al*., 2008). Specific protein 1 (Sp1)-binding site is related to the transcriptional control (Dynan and Tjian, 1983). Three Sp1-binding sites which are able to control the PrP gene promoter activity exist in the promoter region of bovine and human PrP genes (Sakudo *et al*. 2010). PrP expression level is related to the prion disease susceptibility. If polymorphism in promoter region hinders the binding of Sp1 transcription factor, expression level of PrP protein decreases and BSE susceptibility is subsequently reduced (Juling *et al*., 2006). On the contrary, BSE susceptibility increases when the expression of prion protein (PrP^{C}) increases which converts into PrP^{Sc} (Juling *et al*., 2006).

1.2 Prion protein structure

PrP^{C} is a protein with 210 amino acid residues which is attached to mammalian neuronal cell through glycosylphosphatidyl inositol (GPI) anchor of C-terminus (Colby and Prusiner, 2011). PrP is synthesized by removing N-terminal signal peptide (SP) and carboxyterminal peptide through post-translational processing which reduces the protein into 210 residues (Colby and Prusiner, 2011). PrP^{C} and PrP^{Sc} have the same primary structure (Stahl *et al.*, 1993). NMR spectroscopy analysis of human, recombinant bovine, mouse, and syrian hamster prion protein shows the well structured globular domain of C-terminal region and flexible “tail” structure of N-terminal region (Wüthrich and Riek, 2001). Glycine-rich octapeptide repeats (OR) regions of N-terminus show different frequency according to the host species and strains (Goldfarb

et al., 1991) and show binding ability to divalent cations of Zn^{2+}, Fe^{2+}, Ni^{2+}, Mn^{2+} (Choi *et al*., 2004), and specially to Cu^{2+} (Millhauser, 2004). C-terminal globular domain of mammalian PrP^{C} is highly conserved and consisted with two short anti-parallel β-sheets and three α-helices (Riek *et al*., 1996). <Figure 1.1> displays the secondary structure cartoon diagram of human prion protein of residues from 121 to 230 amino acids using protein structure viewer, Jmol 12.2.15 (Zahn *et al*., 2000; PDB ID: 1QM2). PrP also possesses disulfide bond bridge (S-S) between α_2 and α_3, hydrophobic regions (HRs) in central part (HR1) and C-terminal region (HR2), and two Asn(N)-linked glycosylation sites(CHO) on position 180 and 196 which are not relevant with PrP^{Sc} formation (Brown, 2001; Sakudo *et al*., 2006).

Though the normal cellular function of PrP^{C} is not clearly revealed, it is believed to have anti-apoptotic activity, anti-oxidative activity, copper ion homeostasis, transmembrane signaling, formation and maintenance of synapses, and cell adhesion (Westergard *et al*., 2007). PrP^{C} is engaged in copper ion metabolism with increased endocytosis in high concentrations and decreased one in the lower concentrations (Pauly and Harris, 1998).

Similar self-propagating proteins were revealed in yeast and fungi (King and Diaz-Avalos, 2004; Baxa *et al*., 2006). Yeast prion includes [URE3] and [PSI] which are alternative conformational states of Ure2p and Sup35 each (Wickner, 1994; Patino *et al*., 1996). Fungi species, *Podospora anserina*, has [HET-s] prion state of Het-s protein (Coustou *et al*., 1997). Yeast prions dose not induces diseases but serves necessary functions in the host which is different to the case of mammals.

1.3 Mechanism of the Conversion of Conformation

PrP^{Sc} shows difference in biophysical properties from PrP^{C}. It is from conformational conversion of PrP^{C} after post-translational processes in spite of the identical primary structure (Borchelt *et al*., 1990). The conformational change of PrP^{C} to PrP^{Sc} is highly relevant to the cause of prion disease. Cellular PrP^{C} is monomeric, soluble, and sensitive to proteinase, while PrP^{Sc} is multimeric, insoluble, and resistant to proteinase K (PK) (Caughey *et al*., 1991; Pan *et al*., 1993). Conversion to PrP^{Sc} is directly related to the formation of amyloid fibrils. PrP^{Sc} converts PrP^{C} into PrP^{Sc} isoform by working as a template as self-propagating protein (Kocisko *et al*., 1994; Cobb and Surewicz, 2009). Conformational conversion induces change of secondary structures as higher β-sheet content for the PrP^{Sc} than that of PrP^{C} (Pan *et al*., 1993). This conformational transition is thought to be important in prion propagation considering the fact that chemical modification has not been found (Pan *et al*., 1993). High β-sheet content of PrP^{Sc} leads to the aggregation and the formation of insoluble fibrils to gain resistance to proteinase K (PK) (Cohen and Prusiner, 1998). Nucleation-dependent polymerization process incorporates the conversion of PrP^{C} to PrP^{Sc} which has two steps of the binding of PrP^{C} to the PrP^{Sc} template and subsequent conformational change (Cobb and Surewicz, 2009). PrP^{Sc} monomers bind to make partially structured intermediate. This oligomer is increased in solvent exposure and hydrophobicity which leads to intermolecular interactions (Apetri *et al*., 2006). This changed property enables the oligomer to act as a template which induces conformational conversion of PrP^{C} into PrP^{Sc} (Apetri *et al*., 2006). When sufficient amount of oligomer accumulates to form stable nucleus in lag phase, monomeric PrP^{C} binds to this oligomeric nucleus and transformed to PrP^{Sc} (Jarrett ad Lansbury, 1993). The

length of the lag phase could be varied by the amount of the seeding aggregates on which amyloid fibrils grow (Jarrett and Lansbury, 1993). In the case of yeast prion, chaperone protein Hsp104 severs amyloid fibers into fragments to produce more seeds for polymerizations that induces increased replication of fungal abnormal prions (Shorter and Lindquist, 2008). Also, Hsp40 and Hsp70 are involved in the replication of yeast abnormal prions (Shorter and Lindquist, 2008). The growth phase is influenced by this fibril fragmentation (Xue *et al*., 2008).

Prion diseases display the deposition of amyloid-like fibrils as in other types of neurodegenerative diseases including Alzheimer's, Huntington's, Parkinson's disease (Chiti and Dobson, 2006). Amyloid fibrils generally show "cross-β" structure though different proteins aggregates in diverse diseases (Chiti and Dobson, 2006). β-strands are vertical to fibril axis and are parallel to hydrogen bonds in cross-β structure (Tycko, 2004). β-sheet pairs stack in parallel and side chains bind strongly to form "steric zipper" structure according to the analysis of microcrystals of short peptides of yeast Sup35 protein of seven residues which forms cross-β structure (Nelson *et al*., 2005).

1.4 Effect of Prion Polymorphism

Mutation and polymorphism of PrP gene is quite important because it changes the susceptibility of the disease (Sakudo *et al*., 2010). The polymorphism of PrP gene ORF in human is an important determining factor of the prion disease susceptibility. <Figure 1.2> displays PrP mutations and polymorphisms which are relevant to prion disease in human, mice, sheep, elk and cattle (Prusiner, 1998; Sakudo *et al*.,

2010). Effect of polymorphism or mutation in each species are as follows. In human PrP M129V polymorphism, for example, most sporadic CJD cases are homozygous at residue 129, but more than half of normal population are heterozygous at this site (Palmer *et al*., 1991). Therefore, this polymorphism influences the resistance of sporadic CJD. Four types of PrP^{Sc} with different physicochemical properties appear in the CJD patient's brain tissue (Collinge *et al*., 1996; Wadsworth *et al*., 1999; Hill *et al*., 2003). Type 1~3 appear in classical CJD (sporadic and iatrogenic) and type 4 appears in vCJD. Type 1 and 4 PrP^{Sc} are observed only in human of homozygous 129M, type 3 is observed in human with one or more 129V alleles, and type 2 is observed in all genotypes (Collinge *et al*., 1996; Wadsworth *et al*., 1999; Hill *et al*., 2003). According to this, polymorphism on the position 129 of PrP determines type and phenotype of CJD (Wadsworth *et al*., 2004). E219K polymorphism does not appear in sporadic CJD patient, suggesting that E219K might be resistant to sCJD (Shibuya *et al*., 1998). P102L polymorphism in GSS family induces fast dementia development and cortical damage which is similar to that of CJD (Hainfellner *et al*., 1995). P105L with 129V polymorphism was observed in GSS patient with spastic paraparesis (Yamada *et al*., 1999). D178N polymorphism is relevant to fatal familial insomnia (FFI) (Parchi *et al*., 1999). N171S with 129V polymorphism was confirmed in CJD family with a strong psychiatric clinical presentation (Appleby *et al*., 2010). V180I is a causative point mutation of CJD and is recognized as the most common cause of familial CJD in Japan (Chasseigneaux et al., 2006; Mutsukura *et al*., 2009). E200K mutations causes CJD (Hsiao *et al*., 1991). Kaneko *et al*. reported that PrP^{Sc} formation was suppressed when position 214 and 218 was changed to human PrP residues in neuroblastoma cell of scrapie-infected mouse

(Mo) with chimeric Hu/Mo PrP gene (Kaneko *et al*., 1997). Also, mutations and polymorphisms of T183A (Grasbon-Frodl *et al*., 2004), E196K (Peoc'h *et al*., 2000), F198S (Piccardo *et al*., 2001), D202N (Piccardo *et al*., 1998), V203I (Peoc'h *et al*., 2000), R208H (Capellari *et al*., 2005), V210I (Biljan *et al*., 2011), E211Q (Peoc'h *et al*., 2000), and R232M (Shiga *et al*., 2007) are known to be related with the susceptibility of prion diseases including CJD and GSS. <Table 1.1> displays the PrP polymorphisms and mutations and their effect on the susceptibility of prion diseases.

Ovine polymorphisms on position 136, 154, and 171 are related to scrapie. Sheep with ARR and AHQ genotype are resistant to scrapie, while sheep with ARQ, ARH and VRQ genotype are susceptible to scrapie (Baylis and Goldmann, 2004). Bovine polymorphisms which influences PrP expression includes Ins/Del on position 23 in promoter region which is upstream to the transcription start site and Ins/Del on position 12 in intron 1 (Sander *et al*., 2005). These 12bp/23bp Indel(Ins/Del)s are known to be related with BSE-susceptibility (Juling *et al*., 2006). 12bp allele is posed in Sp1 binding site and 23bp allele is posed in transcription factor RP58 binding site. 23bp polymorphism controls PrP gene promoter activity while 12bp deletion varies Sp1 binding affinity, affects promoter activity to lower the PrP gene expressions and subsequently lessens the BSE risk (Xue *et al*., 2008; Sakudo *et al*., 2010). Bovine PrP gene ORF also have E211K mutation which was revealed to be relevant to the atypical BSE (Heaton *et al*., 2008). Mice PrP has polymorphisms on 108 and 189 residue which are referred as $Prnp^{a}$ and $Prnp^{b}$ genotype each and affects TSE incubation time in strain-specific manners (Westaway *et al*., 1987).

1.5 Objective of study

Prion diseases generally have long incubation time and develop clinical symptoms including loss of motor functions, cognitive impairment, and brain dysfunctions which might cause the death of the infected subjects though variations are observed in different prion strains and species of hosts (Prusiner, 1998; Sakudo and Ikuta, 2009). More than 180,000 heads of cattle was infected with BSE in England after the first identification of BSE in cattle in 1986 which has similar characters to scrapie with abnormal progressive neurological disorders (Smith and Bradley, 2003). The interest in prion disease has been increased since the finding of the transmission of mad cow disease to human to cause vCJD through the consumption of infected meat (Britton *et al*., 1995). Protein Data Bank (PDB) contains about a score of 28 X-ray crystallography structures and about 73 nuclear magnetic resonance (NMR) structures with "prion" as a key word. The conversion mechanism of PrP^{C} into PrP^{Sc} using computer simulations based on this PDB structural information has been under study. Much research on the relationship between polymorphism and prion disease susceptibility has been performed using fungi and animal models. The results of the research are deposited in public databases including NCBI and EMBL. Prion specific database includes Prion Disease Database (PDDB), AMYPdb, and PrionHome. PDDB contains time-course expression profiles, tissue-specific expressions, and systems biological network data of genes (Gehlenborg *et al*., 2009). AMYPdb is a database which contains amyloid precursor proteins which are known to induce pathological propagation of proteins including prion by forming amyloid fibrils (Pawlicki *et al*., 2008). PrionHome contains prion and prion-related sequences with classifications of prionogenicity (Harbi *et al*., 2012). Both experimental and simulative methods could be applied

to prion disease research. However, PrP^{Sc} is insoluble and hard to crystallize to make NMR analysis and X-ray crystallography difficult. These points make the experimental analysis of PrP^{Sc} structure difficult (Sakudo *et al*., 2010). Research on conformational change, the relationship of PrP polymorphisms with prion disease susceptibility and others are currently performed using bioinformatics analysis methods. Computational analyses are based on improvements of the computational hardware, database techniques, and computational simulation techniques including molecular dynamics (MD). Though simulative approaches supply advantages for the study of the details of the prions, computer simulations need expensive hardware equipments and much time. Method that determines the susceptibility of prion diseases only referring the primary structure of prion protein without complex structural analysis would, thus, be necessary. Prion protein classification research includes the analysis of the disease-specific signature of BSE using mid-infrared spectroscopy of fluid serum and various classification algorithms (Martin *et al*., 2004). Also, there is a study of prion conformational stability using support vector machine (SVM) where Euclidean distances and 3D distance count descriptor from 3D pseudo-folding graph representation (Fernández *et al*., 2008). Though the prion protein polymorphism which influences the susceptibility of prion disease has been constantly concerned, relevant database has not been established. The effect of polymorphisms could be collected from experimental literatures and amino acid polymorphism information of mammalian prion could be collected to build comparative database. Predictive classification tools could be built from this information. Training dataset for discriminant analysis classified into four groups depending on the effect of polymorphisms on prion diseases. Also, Test dataset was generated by introducing a mutation at each site in the

reference sequences which does not appear polymorphisms. The species which are only investigated polymorphisms of specific direction have one mutant sequence and other species have two mutant sequences. Susceptibility of test dataset are predicted based on training data that are classified into each group. Here, amino acid substitutions of mammalian prion sequences were converted into score using BLOSUM62 and PSSM substitution matrices and diverse discriminant analysis was used to perform the prediction of classes to enable the anticipations of the possible dangers of a specific prion protein without experimental studies.

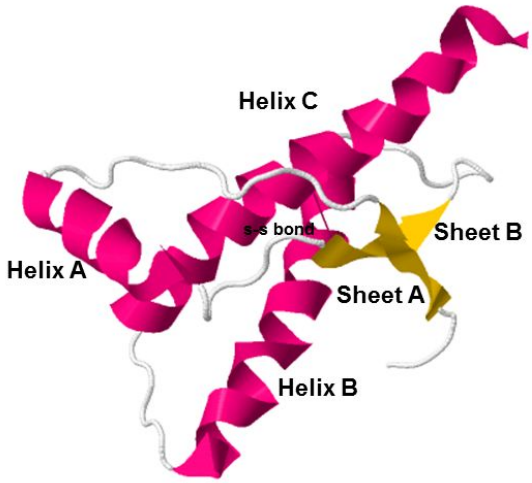

Figure 1.1 Human prion protein fragment 121-230 (PDB ID : 1QM2). C-terminal globular domain of mammalian PrP^{C} consisted with two short anti-parallel β-sheets and three α-helices (Zahn *et al*., 2000).

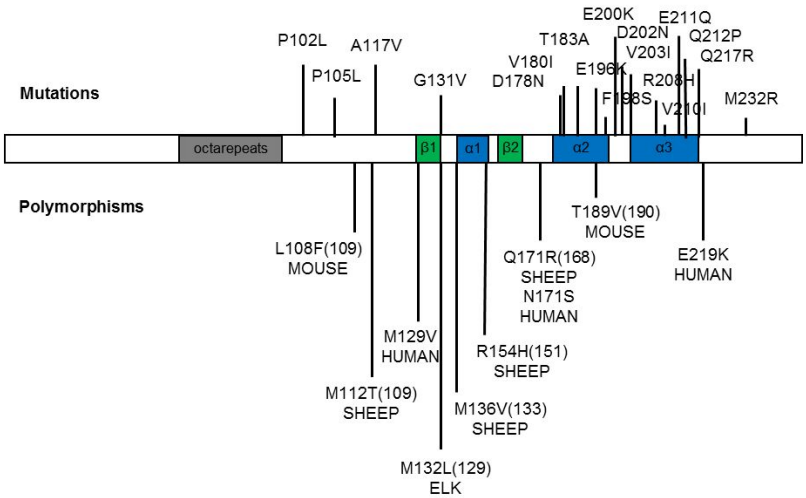

Figure 1.2 Mutations (above) and polymorphisms (bottom) associated with prion disease (Prusiner, 1998; Sakudo *et al*., 2010)**.**

Table 1.1 Prion sequence polymorphism associated with prion disease.

Residues	Host	Res	Sus	Residues	Host	Res	Sus
P102L	human		+	**V210I**	human		+
P105L	human		+	**E211Q**	human		+
L109F	mouse	+		**E211K**	cattle		+
A117V	human		+	**Q212P**	human		+
G127V	human	+		**V215I**	mouse	+	
M129V	human	+		**Q217R**	human		+
G131V	human		+	**E219K**	human	+	
I139M	goat	+		**Q219E**	mouse	+	
Q168E	mouse	+		**Q219K**	goat	+	
N171S	human		+	**M232R**	human		+
Q172R	mouse	+		**A133A+R151R+Q168K**	sheep		+
D178N	human		+	**A133A+R151R+Q168Q**	sheep		+
V180I	human		+	**A133V+R151R+Q168Q**	sheep		+
T183A	human		+	**A133A+R151R+Q168H**	sheep		+
T190V	mouse	+		**A133A+R151R+Q168R**	sheep	+	
E196K	human		+	**A133A+R151H+Q168Q**	sheep	+	
F198S	human		+	**M109T+A133A+R151R+Q168Q**	sheep	+	
E200K	human		+	**M134T+A133A+R151R+Q168Q**	sheep	+	
D202N	human		+	**L138F+A133V+R151R+Q168Q**	sheep		+
V203I	human		+	**I139K+A133A+R151R+Q168Q**	sheep	+	
R208H	human		+	**N173K+A133A+R151R+Q168Q**	sheep	+	

Res: resistant effect; Sus: susceptibility effect. L109F (Westaway *et al*., 1987), M109T (Saunders *et al*. 2009), G127V (Mead *et al*. 2009), A133V (Baylis and Goldmann, 2004), M134T (Vaccari *et al*. 2007), L138F (Saunders *et al*., 2006), I139K (Vaccari *et al*. 2007), R151H (Baylis and Goldmann, 2004), Q168R (Baylis and Goldmann, 2004), M129V (Palmer *et al*., 1991), N171S (Appleby *et al*., 2010), N173K (Vaccari *et al*. 2007), T190V (Westaway *et al*., 1987), and E219K (Shibuya *et al*., 1998) polymorphisms influence susceptibility or resistance of prion disease. V180I (Chasseigneaux *et al*., 2006; Mutsukura *et al*., 2009), T183A (Grasbon-Frodl *et al*., 2004), E196K (Peoc'h *et al*., 2000), E200K (Hsiao *et al*., 1991), V203I (Peoc'h *et al*., 2000), R208H (Capellari *et al*., 2005), V210I (Biljan *et al*., 2011), E211Q (Peoc'h *et al*., 2000), and M232R (Shiga *et al*., 2007) mutations cause Creutzfeldt-Jakob disease (CJD). P102L (Hainfellner *et al*., 1995), P105L (Yamada *et al*., 1999), A117V (Piccardo *et al*., 2001), G131V (Panegyres *et al*. 2001), F198S (Piccardo *et al*., 2001), D202N (Piccardo *et al*., 1998), Q212P (Piccardo *et al*., 1998), and Q217R (Piccardo *et al*., 1998) mutations associated to Gerstmann-Sträussler-Scheinker syndrome(GSS). D178N (Parchi *et al*., 1999) mutation related to Fatal familial insomnia (FFI) or CJD. E211K (Heaton *et al*., 2008) mutation related to Bovine spongiform encephalopathy (BSE). Q168E, Q172R, V215I and Q219E (Kaneko *et al*., 1997) mutations prevented PrP^{Sc} formation. I139M (Goldmann *et al*. 1996) and Q219K (Vaccari *et al*, 2006) associated with scrapie resistance.

Chapter 2.
Materials and Methods

2.1 BLAST

BLAST (Basic Local Alignment Search Tool) was employed for the collection of homologous prion protein sequences to build mammalian prion polymorphism database. BLAST is the most widely used application with heuristic algorithms for the analysis of sequence similarity (Altschul, 1990). Sequence alignment includes global alignment and local alignment. Global alignment compares the whole lengths of each of the two sequences to find the optimal alignment (Needleman and Wunsch, 1970). This method is generally used for the pair of sequences with similar lengths and high similarity. Local alignment finds the optimal regions of best similarity from a pair of sequences (Smith and Waterman, 1981). This method is usually applied to sequences of moderate similarity and different lengths. BLAST algorithm performs local sequence alignment. According to Baxevanis and Ouellette, BLAST method first generates “query words” of short subsequence of which length is set as three letters as default and “neighborhood words” of related words with conservative substitutions. Neighborhood words are derived from the calculated scores using BLOSUM62 scoring system as the words with higher scores than the threshold score T. BLAST algorithm extends the alignment of the matched words between query and target sequences and calculates the cumulative score which sums the scores of matches, mismatches and gaps until it reaches the maximal length. The maximal length is determined as the stop of the extension when the cumulative score

decrease by mismatches and gaps exceeds significance decay threshold, X. High-scoring segment pair (HSP) is obtained as the alignment with the maximal cumulative score by this method. The reliability of the alignment is measured by bit score and expectation value (E-value) which indicates statistical significance. Higher bit score and lower expectation value (E-value) signifies closer alignment; i.e. E-value of 0.05 signifies identical closeness would occur 5% in alignments by chance (Baxevanis and Ouellette, 2005; McEntyre and Ostell, 2002). Here, prion protein sequences from 23 species were obtained from protein database of NCBI. Sequence homology search was performed using BLASTP 2.2.27 with the collected prion sequences from 23 species. Default parameters were selected and non-redundant protein database with chosen organisms was searched. Total of 222 sequences were collected by choosing non-partial sequences of more than about 50% of coverage. Taxonomy information, common name, and frequency of sequences from dataset depending on each species displayed in <Table 2.1>.

2.2 Multiple Sequence Alignment (MSA)

Multiple sequence alignment was performed with collected mammalian prion protein sequences using ClustalW (Thompson *et al.*, 1994). Default parameters were used in the alignment. ClustalW is a widely used application for the multiple sequence alignment (MSA) of DNA and protein sequences. BioEdit 7.1.3 (Hall, 1999) application was used to selectively extract region of residues from 98 to 227 in human prion sequence numbering. This region showed few gaps and better aligned than other regions with frequent gaps. Region of residue 98-227 forms globular domain with two β-sheets and three α-helices and

contains many polymorphisms and mutations related with prion diseases. Polymorphisms and mutations have functional differences; some but not all of polymorphisms significantly affect the onset or phenotype of diseases while mutations cause diseases (Harris, 1969, Harris, 1971). Polymorphisms could be used as genetic markers for the human diseases (Johnson and Todd, 2000). Polymorphisms on prion protein have significant influence on the susceptibility of the prion disease and are known to be related with the conversion of PrP^{C} to PrP^{Sc}.

2.3 Prion Polymorphism Database

MySQL database was built with sequence of C-terminal region of residue 98-227 and annotation information. Annotations and sequence information were collected from public resources of protein database of NCBI. C-terminal region with low gap content was selected through the multiple sequence alignment. JAVA code was used for the extraction of the accession number, species and sequence information from alignment results in fasta format. JDBC-MySQL driver was downloaded and installed from MySQL web page (http://www.mysql.com) for the implementation of MySQL transaction functions into JAVA. Parsed information was stored into relevant fields of tables. Explanation of each table of the database is displayed in <Table 2.2>. Accession number, species information and sequence was parsed and stored into the “prion sequence table.” The character information of each residue of collected sequences were sorted into “prion character table” of 222 rows and 130 columns.

The frequency of 20 amino acids of each residue of the collected sequences were stored into “amino acid frequency table”. JAVA code was used for the calculation of PSSM scores and calculated values

were stored in the “relative frequency table”, “odds ratio table”, and “pssm table” of 130 rows and 20 columns. The PSSM scores of collected sequences were stored in the “pssm trans table” of 177 rows and 130 columns of MySQL database. The character information of “prion character table” was transformed into scores using BLOSUM62 matrix and stored into “blosum62 trans table” of 177 rows and 130 columns . <Table 2.3> illustrates the structure of the “pssm trans table” and “blosum62 trans table”. Sequences used in these tables were included in the discriminant analysis. Polymorphism table for 23 species was additionally built for the illustration of species-specific polymorphism information. This table has information of amino acid frequency of each residue of each species as shown in <Table 2.4>.

2.4 Discriminant Analysis

Discriminant analysis was applied to the classification of the susceptibility of the prion diseases referring differences by polymorphisms. SAS 9.3 statistical packages were used to perform K-nearest, canonical, and linear discriminant analysis and accuracy of the methods were compared. Discriminant analysis (DA) is a multivariate statistical method which is generally used for complex data with large members and variables (Fisher, 1936). The groups are previously determined and appropriate model for the prediction of groups are built using information of variables (Fernandez, 2002). Through discriminant analysis, this research conducted to examine whether there is difference between groups, to identify what residue is important to determine groups using variable selection, and to classify susceptibility group of new sequence. These analysis used three types of DA: Linear, Canonical, k-nearest-neighbor discriminant analysis.

When each group is assumed to follow multi-variate normal distribution, linear discriminant analysis can be used to as parametric method and performs classification using discriminant functions formed from the linear combination of variables to maximize distance between the groups (Fernandez, 2002; SAS Inst. Inc. 2008). Canonical discriminant analysis (CDA) which use analysis of variance when group is more than three reduces dimensions by introducing canonical variables which are linear sets of the most representative variables for the differences among groups when the discriminant variables are abundant (SAS Inst. Inc. 2008). Canonical variables are mutually independent and used for the visual representation of an object with much more variables as low-dimensional plots. When any distribution is not considered, K-nearest-neighbor method (kNN) is non-parametric discriminant analysis and classifies members into a group that has the most members in the nearest k members according to Mahalanobis distances from the discriminant variables of the groups that are not normally distributed (Fernandez, 2002; SAS Inst. Inc. 2008). Stepwise method was used to find important variables. SLE (Significant Level for Entry) of 0.25 was used for the inclusion of variables and SLS (Significant Level for Stay) of 0.15 was used for the elimination of variables. Iteration of inclusion and elimination of variables were performed to find significant variables. The accuracy of the results from all variables and significant variables were compared. <Figure 2.2> displays the processes of our study, used applications, scoring matrix and used computer programming language.

The significance of the discriminant function for the four groups was examined using the Wilk's Lambda and Pillai's Trace. When the null hypothesis was set up that there is no difference between the groups, the null hypothesis was rejected if p-value of Wilk's Lambda is

less than 0.05 of significance level (α). In other words, it can be seen that the difference between groups is statistically significant. Wilk's Lambda means variation within the group / total variation, and the smaller value is better. Pillai's Trace means variation between the group / total variation, and the higher value is better.

2.4.1 Datasets

C-terminal region with infrequent gaps of residues 98-227 which was selected from the alignment results of prions of 23 mammalian species was used for the analysis. This region contains three α-helices and two parallel β-sheets. The information of the effects of prion mutations and polymorphisms of each residue on the prion disease susceptibility was investigated <Table 1.1>. Collected mammalian prion sequences were categorized into groups according to their degree of susceptibility. Mutant sequences were generated from reference prion sequences which does not show polymorphism in hosts of mouse, sheep, human, goat, and cattle using the substitution of the residues to increase or decrease the susceptibility. Reference sequences of NP_000302.1 (human), NP_035300.1 (mouse), P52113.1 (goat), P10279.2 (cattle), and NP_001009481.1 (sheep) were utilized. In investigated reference sequence, the two mutant sequences generated by introducing positive or negative polymorphism in the reference sequence of the species *Homo sapiens* and *Ovis aries*. *Bos Taurus* with only negative polymorphism and *Capra hircus, Mus musculus* with only positive polymorphism had generated one mutant sequence depending on each species. Sequences with amino acids of other types rather than typical 20 amino acids were omitted. Total of 177 sequences from the collected 170 sequences and mutated 7 sequences were used as dataset

for classification analysis. Group 1 shows increased susceptibility to the prion diseases because it has polymorphisms that increase susceptibility. Group 2 shows increased resistance with lengthened incubation times from difficult conformational change from PrP^{C} to PrP^{Sc} because it has polymorphisms that reduce susceptibility. Group 3 has polymorphisms that both increase and decrease susceptibility. In group 4, polymorphisms that increase or decrease the degree of risk for infection of prion disease do not appear.

2.4.2 Distance Measurements

- BLOSUM62 Utilized Distance

Scoring matrix was used to represent difference of sequences according to the groups of the susceptibility of prion diseases as scores. The most widely used scoring matrix is BLOSUM62 matrix (Henikoff and Henikoff, 1992) which was derived from conserved motifs of protein families. According to Baxevanis and Ouellette, concept of block was exploited which is similar to protein motifs. Motif is a conserved amino acid sequence of a protein with specific function or structure while block is an ungapped alignment region of protein family. BLOSUM (Block Substitution Matrices) was built based on susbstitution patterns in conserved blocks. Numerous types of BLOSUM matrices exist according to the conservation level of the used sequences as BLOSUM45, BLOSUM62, BLOSUM80, etc. BLOSUM62 means that the matrix was calculated from the sequences with 62% identity with others. <Figure 2.1> illustrates the BLOSUM62 matrix. 20 amino acids are on the columns and rows. The scores are calculated as follows.

$$S_{i,j} = \log(\frac{q_{i,j}}{p_i p_j})$$

where p_i and p_j is the probability of the occurrence of amino acid i and j in random environments. $q_{i,j}$ is the probability of the co-occurrence of amino acids of i and j with genealogical relationship. In other words, log odds ratio of $S_{i,j}$ is the ratio of the odds between random and genealogical substitutions. BLOSUM matrix is composed of these log odds scores. Positive score infers that the genealogical substitutions between the residues would be more prone than random environments (Baxevanis and Ouellette, 2005). The application which calculates distance score of prion sequence from consensus sequence using BLOSUM62 score matrix was coded with JAVA programming language. Consensus sequence was built as the sequence with the most frequent amino acids. This distance score could be used for the classification of the susceptibility groups.

- PSSM(Position-Specific Scoring Matrix) Utilized Distance

The results using BLOSUM62 matrix was compared with the results using PSSM or PWM (position weight matrix; Altschul *et al.*, 1997) for the better discriminant analysis. PSSM is a different scoring matrix which represents position-dependent substitution scores from multiple sequence alignment. This matrix has different substitution scores for the same pair of amino acids according to the positions of amino acids. The frequency matrix of amino acids for each position is first calculated to build position specific scoring matrix. This frequency matrix is converted into score matrix by calculating log ratio between observed frequency and amino acid propensities in collected prion sequences. Relative frequency (*rf*) is calculated as the ratio of the frequency of a specific amino acid *i* in a residue (N_i) and the frequency of any amino acids (N_t).

$$rf = \frac{N_i}{N_t}$$

Odds ratio is calculated as the ratio of the relative frequency (*rf*) and the background frequency of amino acid *i*. Background frequency was used amino acid propensities which were observed from the total of 170 prion sequences <Table 2.5>.

$$odds\ ratio = \frac{rf}{background\,frequency}$$

PSSM (Position-specific Scoring Matrix) was calculated by applying logarithm of base 2 to the odds ratio.

$$PSSM\ score = \log_2(odds\ ratio)$$

Positive score means more frequent occurrences in the family than random occasions while negative score signifies the opposite. Pseudo-counts were added for the positions of zero frequencies for the impossibility of the calculation of log(0). Consensus sequence might be derived from this PSSM by selecting the most probable sequences. The column-wise scores of prion sequences were derived using the PSSM of prion sequences. An amino acid ‘z’ of 'AFM91139.1 sequence' has same score with consensus amino acid at the residue so that it is ignored in the prediction.

2.4.3 Accuracy evaluation

Leave one out cross-validation estimation was used to analyze the accuracy of the types of discriminant analysis by comparing the rate of misclassification (Lachenbruch and Mickey, 1968). Cross-validation method divides training and test sets using training set to derive discrimination function and test set to validate the derived function. In this validation method, a single subject is first omitted from training

and the discriminant model is built. The omitted subject is tested for the classification after the training. Every member of the training set is left once for the validations and others are trained for each omission. The rate of misclassification is assessed. K-nearest, canonical, and linear discriminant analysis methods were compared while lower misclassification rate signifies the better accuracy. Group-specific accuracy can be compared by comparing the error rate of each group. The performance of the classifiers is determined by the calculation of sensitivity, specificity, error rate and total classification accuracy. The sensitivity, specificity, error rate and total classification accuracy are defined as follow.

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{Error rate} = \frac{\text{The number of incorrectly classified sequences}}{\text{The number of sequences in a Group}}$$

$$\text{Total classification accuracy} = \frac{\text{The number of correctly predicted sequences}}{\text{The total number of sequences}}$$

The average misclassification rate was the mean error rate for four groups. The accuracies of test dataset were presented by the ratio of the number of correctly predicted sequences to the total number of test sequences.

Table 2.1 Taxonomy information of Collected Prion Sequences

Family	Genus	Species	Subspecies	Common Name	Freq
Cervidae	Alces	alces	gigas	**alaskan moose**	2
	Cervus	elaphus		**red deer**	5
			canadensis	**wapiti**	2
			nelsoni	**american elk**	4
			scoticus	**scottish red deer**	1
	Odocoileus	hemionus		**mule deer**	5
		virginianus		**white-tailed deer**	6
Bovidae	Bos	taurus		**cattle***	20
	Capra	hircus		**goat***	23
	Ovis	aries		**sheep***	85
Canidae	Canis lupus	familiaris		**dog**	7
Equidae	Equus	caballus		**horse**	10
Felidae	Felis	catus		**cat**	4
Hominidae	Homo	sapiens		**human**	19
	Pan	troglodytes		**chimpanzee**	1
Cercopithecidae	Macaca	mulatta		**rhesus monkey**	4
Cricetidae	Mesocricetus	auratus		**golden hamster**	5
Muridae	Mus	musculus		**house mouse***	7
	Rattus	norvegicus		**rat**	2
Mustelidae	Mustela	putorius	furo	**domestic ferret**	2
	Neovison	vison		**american mink**	2
Leporidae	Oryctolagus	cuniculus		**rabbit**	3
Suidae	Sus	scrofa		**pig**	3

*means to investigated the effect of susceptibility on prion disease in that species

Table 2.2 Database Table Information

Field name	Data Type	
prion sequence table		
accession	varchar	primary key
species	varchar	
sequence	text	
prion character table		
accession	varchar	primary key
species	varchar	
residue 98	char	
…		
residue 227	char	
amino acid frequency table		
residue	int	primary key
amino acid C	int	
…		
amino acid W	int	
amino acid relative frequency table		
residue	int	primary key
amino acid C	double	
…		
amino acid W	double	
amino acid odds ratio table		
residue	int	primary key
amino acid C	double	
…		
amino acid W	double	
pssm table		
residue	int	primary key
amino acid C	double	
…		
amino acid W	double	

Table 2.3 Table Structure of Discriminant Analysis Training Dataset

Field name	Data Type	
blosum62 trans table		
accession	varchar	primary key
species	varchar	
residue 98	int	
…		
residue 227	int	
pssm trans table		
accession	varchar	primary key
species	varchar	
residue 98	double	
…		
residue 227	double	

Table 2.4 Information of Polymorphism Table of 25 Species

Field name	Data Type	
species polymorphism table		
residue	int	primary key
amino acid C	int	
…		
amino acid W	int	

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	9																			
S	-1	4																		
T	-1	1	5																	
P	-3	-1	-1	7																
A	0	1	0	-1	4															
G	-3	0	-2	-2	0	6														
N	-3	1	0	-2	-2	0	6													
D	-3	0	-1	-1	-2	-1	1	6												
E	-4	0	-1	-1	-1	-2	0	2	5											
Q	-3	0	-1	-1	-1	-2	0	0	2	5										
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8									
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5								
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5							
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5						
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4					
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4				
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4			
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6		
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7	
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11

Figure 2.1 BLOSUM62 matrix (Henikoff S and Henikoff JG. 1992). Positive score infers that the substitutions between the residues would be more prone than random environments. Negative score infers that the substitutions between the residues would be more rare than random environments.

Table 2.5 Background Frequency

AA	Freq	%
Cys (C)	341	1.54
Ser (S)	864	3.91
Thr (T)	1,729	7.82
Pro (P)	852	3.86
Ala (A)	1,362	6.16
Gly (G)	1,542	6.98
Asn (N)	1,642	7.43
Asp (D)	860	3.89
Glu (E)	1,283	5.81
Gln (Q)	1,579	7.15
His (H)	716	3.24
Arg (R)	1,161	5.25
Lys (K)	1,240	5.61
Met (M)	1,139	5.15
Ile (I)	742	3.36
Leu (L)	502	2.27
Val (V)	1,866	8.44
Phe (F)	524	2.37
Tyr (Y)	1,972	8.92
Trp (W)	183	0.83

Amino acid propensities (%) of collected 170 prion sequences.

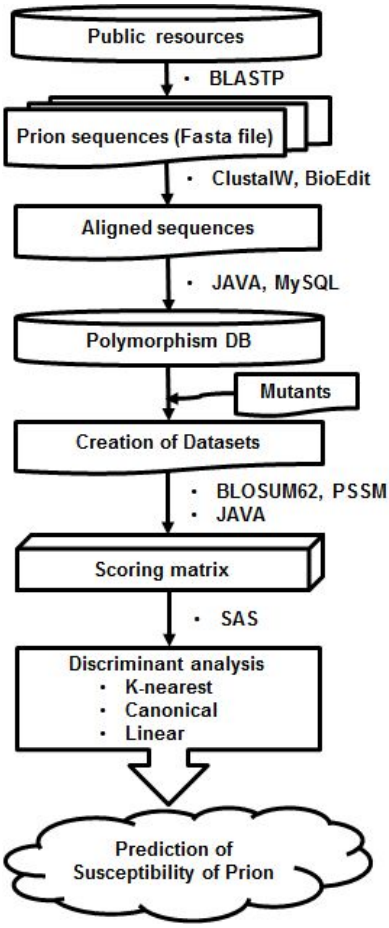

Figure 2.2 Flow diagram of the process of the study with used applications, analysis methods, and used computer language codes.

prion protein sequences from 25 species were obtained from protein database of NCBI. Sequence homology search was performed using BLASTP. Multiple sequence alignment was performed using ClustalW. BioEdit was used to selectively extract region of residues from 98 to 227. MySQL database was built with sequence and parsed annotation information using JAVA. JAVA code was used for the calculation of PSSM and BLOSUM62 scores. SAS 9.3 were used to perform discriminant analysis.

Chapter 3.

Results and Discussion

3.1 Overview of prion polymorphism DB

Prion sequences of 12 families of mammals were analyzed; *Cervidae, Bovidae, Canidae, Equidae, Felidae, Hominidae, Cercopithecidae, Cricetidae, Muridae, Mustelidae, Leporidae, Suidae.* All sequences include three α-helix and two β-sheet that contained the residue 98-227 fragments region, and a total of 222 sequence of 23 species were collected. Taxonomy information and the number of sequences (n) are presented in <Table 2.1>. Relatively, many sequences of *Ovis aries* (n = 85), *Capra hircus* (n = 23), *Bos Taurus* (n = 20), and *Homo sapiens* (n = 19) were collected. The number of collected sequences probably affect the accuracy of the analysis. Through multiple sequence alignment, the type and frequency of polymorphisms in collected PrP sequences of the 23 species was analyzed. *Cervus elaphus scoticus, Mesocricetus auratus, Mustela putorius furo, Neovison vison, Oryctolagus cuniculus, Pan troglodytes, Rattus norvegicus, Sus scrofa* species did not appear polymorphisms because collected sequences of each species have the same amino acid sequence. Polymorphism based on differences in the amino acid sequence shown in each species from a consensus sequence, <Table 3.1> shows the polymorphisms of 201 sequences from 15 species while sequences of no polymorphisms and of species which have only one sequence were excluded. Polymorphisms were shown on residue 206 in *Alces gigas* (Alaskan moose) and on residue 104, 109, 143, 179, 200, 207, and 223 in *Bos taurus* (cattle). Polymorphisms were shown on residue 100, 103,

129, 145, 159, 165, 170, 173, 174, 177, 203, 205, 215, and 220 in *Canis lupus familiaris* (dog) and on residue 98, 103, 107, 130, 134, 139, 140, 151, 168, 182, 191, 208, and 219 in *Capra hircus* (goat). *Cervus elaphus* (red deer) showed polymorphism on residue 165, *Cervus elaphus canadensis* (wapiti) showed polymorphisms in residue 129 and 188, *Cervus elaphus nelsoni* (american elk) showed polymorphism in residue 129. *Equus caballus* (horse) showed polymorphisms on residue 132, 163, 173, and 181. *Felis catus* (cat) showed polymorphisms on residue 100, 112, 185, 203, 215, and 220. *Homo sapiens* (human) showed polymorphisms in diverse residues of residue 99, 115, 129, 169, 171, 187, 196, 203, 218, 219, 220, 221, 222, 223, 224, 225, 226, and 227. *Macaca mulatta* (Rhesus monkey) showed polymorphism on residue 100 and *Mus musculus* (house mouse) showed polymorphisms on residue 109, 134, 139, 141, 184, and 190. *Odocoileus hemionus* (mule deer) showed polymorphism on residue 135 and *Odocoileus virginianus* (white-tailed deer) showed polymorphisms on residue 113, 135, and 148. Ovis aries (sheep) showed the most polymorphisms on residue 98, 107, 108, 109, 111, 113, 120, 121, 123, 124, 125, 126, 128, 129, 132, 133, 134, 135, 138, 139, 140, 142, 143, 148, 149, 150, 151, 157, 160, 165, 168, 169, 170, 172, 173, 177, 182, 183, 186, 188, 190, 198, 201, 208, 211, and 215. <Figure 3.1> shows residue which appeared polymorphism and the incidence (m) using line graph. *Ovis aries* (m = 46) have the most number of residues which appeared polymorphism, *Homo sapiens* (m = 18), *Canis lupus familiaris* (m = 14), *Capra hircus* (m = 13), *Bos Taurus* (m =7), *Felis catus* (m = 6), and *Mus musculus* (m = 6) also have relatively large number of amino acid substitutions than other species. Prion diseases can be transmitted between species, but transmission among distant species are relatively not well spread, and it is called the 'species barrier' (Sweeting

et al., 2010). *Canis lupus familiaris* is a representative species which can see species barrier to TSE transmission. TSE-infected feline species eating infected beef with BSE during the BSE epidemic occurs in many cases which have been reported, but the canine species was not found (Kirkwood and Cunningham, 1994). There is no distinct difference in the frequency of polymorphism between species known to be resistant to prion disease and species with high susceptibility to TSE. Thus, this study suggest that there is no association between the frequency of polymorphism and risk of prion disease. There are differences in the position of polymorphism occurs according to the species. Also, similar polymorphism occurs in various species. The polymorphism on residue 129 was the most frequent among species, while polymorphisms on 100, 109, 134, 135, 139, 165, 173, 203, 215 and 220 were the most frequent in order. <Table 3.2> shows the type and frequency of mutation or polymorphism on training dataset that affect the positive or negative susceptibility of prion disease. For discriminant analysis, negative polymorphism has only been found in *Bos Taurus* species and positive polymorphism was only discovered in *Capra hircus, Mus musculus*. The rarity of the events of the pathogenic mutations and polymorphisms related to the onset of the prion diseases were indicated that be able to significant influence to the disease susceptibility.

3.2 Validations of Classification

Discriminant analysis was carried out to predict genetic risk of prion disease based on the susceptibility effect of each polymorphism associated with prion disease. Mutation or polymorphism site have positive effect or negative effects on prion diseases. 17 changes having positive effects and 25 changes having negative effects as a total of 42

different amino acid substitutions were used. Collected prion sequences were categorized into four groups depending on whether polymorphism have any effect on susceptibility of prion disease. The frequency of each group is 31, 15, 7, and 117, respectively. Discriminant analysis of susceptibility groups was conducted based on the variables of distance scores of each residue. Cross-validation method was exploited to evaluate the accuracy of the discriminant analysis. Difference between the groups is significance through p-values of Wilk's Lambda (value=0.0442, p-value=<0.0001) and Pillai's Trace (value=1.9024, p-value=<0.0001) which are less than significance level (α =0.05) in statistical significance test of discriminant function generated by linear discriminant analysis with BLOSUM62 matrix. The misclassification rate of group 1, group 2, group 3 and group 4 were 0.2581, 0.600, 0.5714 and 0.2137 when BLOSUM62 matrix and variables of all residues were employed <Table 3.3>. The misclassification rates of group 2 and 3 are higher than group 1 and 4. Significant variables for the discriminant analysis were found through stepwise method as the variables of the residues were 109, 113, 120, 124, 126, 129, 134, 138, 139, 145, 151, 155, 168, 171, 173, 186, 190, 203, 205, 223, and 225. Linear discriminant analysis was also conducted from the selected residues. The misclassification rate of group 1 and group 4 were 0.4194 and 0.2735 in this case as displayed in <Table 3.4>. Wilks' lambda and Pillai's Trace are 0.0897 and 1.6302, respectively. The misclassification rate of group 1 and group 4 of the linear discriminant analysis of PSSM matrix with all residue variables were 0.2903 and 0.2393 <Table 3.5>. Difference between the groups is significance through p-values of Wilk's Lambda (value=0.0614, p-value=<0.0001) and Pillai's Trace (value=1.7720, p-value=<0.0001) which are less than significance level (α =0.05). As mentioned above, significant variables for the discriminant

analysis were found using stepwise method with residues of 109, 112, 113, 120, 124, 126, 129, 134, 138, 139, 151, 167, 168, 171, 173, 186, 190, 205, 215, 219, and 225 being selected. The misclassification rate of group 1 and group 4 were shown to be 0.1613 and 0.2735 from the linear discriminant analysis after variable selection <Table 3.6>. Wilks' lambda and Pillai's Trace are 0.1217 and 1.4902, respectively. K-nearest neighbor discriminant analysis was also conducted from the BLOSUM62 matrix. Cross-validation accuracies as misclassification rates of group 1 from the discriminant analyses with k of 3, 4, and 5 with all variables were 0.2581, 0.1935, and 0.2258 <Table 3.7>. Analysis with k of 4 showed the lowest misclassification rate. Misclassification rates of group 1 from cross-validations with 21 selected variables with k of 3, 4, and 5 were 0.1935, 0.2258, and 0.2258 <Table 3.8>. The rate of misclassification was decreased when k was 3 while variable selection with k of 4 was increased. The misclassification rates of group 1 which has the lowest value among groups with PSSM and all variables with k of 3, 4, and 5 were 0.1935, 0.0968, and 0.1290 <Table 3.9>. The cross-validation accuracies as misclassification rates of group 1 with 21 selected variables were 0.2258, 0.2258, and 0.2258 <Table 3.10>. The most accurate results were obtained from the BLOSUM62 matrix and kNN method with k of 4 with all variables <Table 3.11>. Error rate result for each group showed that error rate of Group 1 sequences having negative effects polymorphism which is relatively lower than other groups and were correctly classified in the most cases. Therefore, We were able to well predict exposure to the onset of prion when at least one negative effect polymorphism exists. In Group 3 with both positive and negative polymorphisms, classification error rate was relatively high.

Two dimensional plot was also built using two canonical variables

in the canonical discriminant analysis. In this case, selection of significant variables didn't helped deduct clear discriminations of the four groups on the plot. The classification results of the canonical discriminant analysis are shown in a two-dimensional graph in <Figure 3.2>. When using the BLOSUM62 matrix and variables of all residues, difference between the groups was found significant using Wilks' Lambda is 0.0442 (p-value=<0.0001). The groups are most clearly distinguished. When using only 21 selected variables, difference between the groups decrease through Wilks' Lambda is 0.0897 (p-value = <0.0001). When using PSSM matrix and variables of all residues, difference between the groups was found significant through Wilk's Lambda statistic is 0.0614 (p-value=<0.0001). When using only 21 selected variables, difference between the groups showed decrease through Wilk's Lambda statistic is 0.1217 (p-value=<0.0001).

3.3 Prediction of Susceptibility for New sequences

In this study, consensus sequence which does not appear polymorphism used as reference sequence in *Bos Taurus, Capra hircus, Homo sapiens, Mus musculus,* and *Ovis aries* species. Each test sequence was generated by introducing mutations at each sites in the reference sequences. The total of seven test dataset being generated. Through results of the validation, when using k-nearest neighbor method, misclassification rate was lower than linear discriminant analysis. <Table 3.11> is the result of classification analysis using the linear, k-nearest neighbor method with BLOSUM62 and PSSM which had differences in the way to the amino acid converted into a numerical score. Cross-validation results as average misclassification rates from the discriminant analyses with k of 3, 4, and 5 with all

variables of the case of BLOSUM62 matrix were 0.3356, 0.3237, and 0.3527. Average misclassification rates from discriminant analyses with k of 3, 4, and 5 with selected variables were 0.3513, 0.4260, and 0.5141. Average misclassification rates from the discriminant analyses with k of 3, 4, and 5 with all variables of the case of PSSM matrix were 0.4244, 0.3959, and 0.4040. Cross-validation accuracies from the discriminant analyses with k of 3, 4, and 5 with selected variables of the case of PSSM matrix were 0.4055, 0.3889, and 0.4603. Smallest misclassification rate was shown when k nearest neighbor with k of 4 and all variables with BLOSUM62 matrix was used. The results showed that we can get the most accurate prediction when we want to know the degree of prion disease genetic risk of any sequences.

<Table 3.12> shows the classification results from k-nearest neighbor discriminant analysis with the distance score calculated through substitution scores based on the BLOSUM62 matrix and PSSM method with all variables. Group 1 of the negative polymorphism sequences had the higher sensitivity than Group 2 and 3 when excluding Group 4. <Table 3.13> shows the total classification accuracies of the k-nearest neighbor, linear discriminant analysis using a BLOSUM62 matrix and PSSM matrix. The total classification accuracy was maximized when the BLOSUM62 matrix with three and four k-objects were used.

In order to determine the accuracy of prediction in new strains, analysis was performed on test dataset. Accuracy of prediction of susceptibility for test dataset was presented in <Table 3.14>. Accuracy on the test dataset from linear discriminant analysis with BLOSUM62 matrix and variables of all residues is 71.43% which means that groups accurately predicted five out of seven sequences. Accuracy on the test dataset from linear discriminant analysis with BLOSUM62 matrix and selected variables is 42.86%. Accuracy on the test dataset from linear

discriminant analysis with PSSM matrix and variables of all residues is 57.14% which means that groups accurately predicted four out of seven sequences. Accuracy on the test dataset from linear discriminant analysis with PSSM matrix and selected variables is 57.14%. K-nearest neighbor discriminant analysis were also conducted using BLOSUM62 and PSSM matrix. The accuracy on the test dataset with BLOSUM62 matrix and variables of all residues showed values as 57.14%, 57.14%, and 42.86%, respectively. The accuracy of the test dataset with BLOSUM62 matrix and selected variables showed all values of 42.86%. The accuracy on the test dataset with PSSM matrix and variables of all residues were low as 28.57%, 42.86%, and 42.86%. The accuracy on the test dataset with PSSM matrix and selected variables showed values of 57.14%, 71.43%, and 71.43%. There are difficulties in assessing accuracy of prediction of new sequence using the average misclassification rate. The groups predicted four out of seven sequences when k nearest neighbor used k of 4 and BLOSUM62 matrix with all variables. Despite the case of Group 1 has relatively low error rate, accuracies on test sequences were not high. High accuracy showed in sequences of *Ovis aries* which introduced negative effect mutation. Sequences of *Bos Taurus* and *Homo sapiens* which introduced negative effect mutation have low accuracy. For this reason, I think these low accuracy were caused by collected 170 sequences alone is difficult to reflect the effect of all polymorphisms to predict susceptibility. There are discrepancies in the accuracy of prediction among species depending on the number of collected sequences and polymorphism information. However, if more sequences are collected and analysed, it can be expected that the accuracy may increase.

Table 3.1 Prion Polymorphism Information of 15 host species

residue	Alaskan moose	Cattle	Dog	Goat	Red deer	Wapiti	American elk	Horse	Cat	Human	Rhesus monkey	House mouse	Mule deer	White -tailed deer	Sheep
	2	20	7	23	5	2	4	10	4	19	4	7	3	6	85
98				Q(22), R(1)											Q(82), X(2), R(1)
99										W(18), R(1)					
100			G(6), N(1)						G(2), N(2)		H(3), N(1)				
103			N(5), S(2)	S(22), R(1)											
104		K(19), R(1)													
107				T(22), P(1)											T(84), N(1)
108															N(84), H(1)
109		M(19), I(1)										L(6), F(1)			M(76), T(4), X(4), V(1)
111															H(84), X(1)
112									M(3), V(1)						
113														A(4), G(1), X(1)	A(82), X(2), P(1)
115										A(18) del(1)					
120															A(83), F(1), X(1)
121															V(80), L(4), X(1)
123															G(84), X(1)
124															G(72), X(6), S(5), V(2)
125															L(84), del(1)
126															G(84), S(1)
128															Y(84), X(1)
129			M(6), L(1)			L(1), M(1)	L(2), M(1), X(1)			M(13), V(5), X(1)					M(84), X(1)
130				L(22), Q(1)											
132								S(9), G(1)							S(84), X(1)
133															A(75), X(7), V(2), T(1)
134				M(22), I(1)								M(6), V(1)			M(77), X(7), T(1)
135													S(2), N(1)	S(4), N(1), X(1)	S(83), N(1), X(1)
138															L(79), F(3), X(3)
139				I(20), M(2), T(1)								I(6), V(1)			I(84), K(1)
140				H(20), R(3)											H(79), X(5), R(1)
141												F(6), L(1)			
142															G(84), X(1)
143		S(12), X(5), N(3)													N(80), X(3), S(2),
145			Y(6), C(1)												
148														R(5), X(1)	R(83), X(2)
149															Y(84), I(1)
150															Y(84), C(1)
151				R(22), H(1)											R(77), X(5), H(3)
157															Y(84), X(1)
159			D(4), E(2), N(1)												
160															Q(84), X(1)
163								Y(9), C(1)							
165			P(6),		P(4),										P(84),

		S(1)		S(1)						X(1)
168			Q(22), R(1)							Q(58), H(9), X(8), R(8), N(1), K(1)
169							Y(18), H(1)			Y(82), X(3)
170		S(6), N(1)								S(83), X(2)
171							N(18), S(1)			
172										Q(83), Z(2)
173		N(6), S(1)				N(8), K(2)				N(82), D(1), K(1), X(1)
174		N(6), T(1)								
177		R(6), H(1)								H(83), X(2)
179	C(19), R(1)									
181						N(9), D(1)				
182			I(22), F(1)							I(84), X(1)
183										T(82), X(3)
184									I(6), T(1)	
185							R(2), K(2)			
186										Q(68), L(11), X(5), R(1),
187								H(18), R(1)		
188					T(1), A(1)					T(84), R(1)
190									T(6), V(1)	T(84), X(1)
191			T(22), P(1)							
196								E(18), G(1)		
198										F(84), X(1)
200	E(19), X(1)									
201										T(84), X(1)
203		M(6), I(1)					M(3), I(1)	V(18), I(1)		
205		I(6), M(1)								
206	M(1), I(1)									
207	E(19), K(1)									
208			R(20), Q(2), G(1)							R(83), Q(1), X(1)
211										E(84), D(1)
215		V(6), I(1)					V(3), I(1)			I(84), P(1)
218								Y(18), L(1)		
219			Q(22), K(1)					E(17), K(1), I(1)		
220		K(6), R(1)					K(3), R(1)	R(18), N(1)		
221								E(18), T(1)		
222								S(18), L(1)		
223	Q(19), R(1)							Q(18), G(1)		
224								A(18), T(1)		
225								Y(18), D(1)		
226								Y(18), G(1)		
227								Q(17), H(1), K(1)		

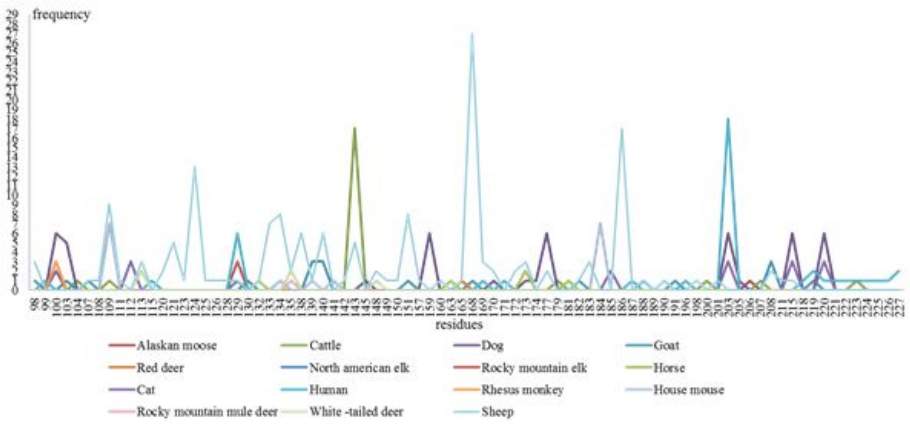

Figure 3.1 Prion Polymorphism Information of mammal species.

This result shows that the polymorphisms of 201 sequences from 15 species. Diverse polymorphisms have difference or was shared according o the species.

Table 3.2 Frequencies of amino acid polymorphism

Species	Group	Residue	AA	Freq	AA	Freq
Capra hircus	2	139	I	20	M	2
	2	219	Q	22	K	1
Ovis aries	1	133,151,168	ARQ	27		
	1				ARK	1
	1				ARH	7
	1				VRQ	1
	2				ARR	4
	2				AHQ	2
	2	109	M(ARQ)	23	T(ARQ)	3
	2	134	M(ARQ)	26	T(ARQ)	1
	2	139	I(ARQ)	26	K(ARQ)	1
	2	173	N(ARQ)	26	K(ARQ)	1
Homo sapiens	2	129	M	11	V	5
	1	171	N	15	S	1
	1	203	V	15	I	1
	2	219	E	15	K	1
Mus musculus	2	109	L	6	F	1
	2	190	T	6	V	1

Table 3.3 Cross validation results of the linear discriminant analyses considering of all possible variables with BLOSUM62 matrix based distance measurements

		Predicted classification group			
		Group 1	Group 2	Group 3	Group 4
Grouping	Group 1	23	1	2	5
	Group 2	4	6	1	4
	Group 3	1	2	3	1
	Group 4	19	2	4	92
	Total	47	11	10	102
	Error rate	0.2581	0.6000	0.5714	0.2137

Table 3.4 Cross validation results of the linear discriminant analyses after the selection of significant variables and with BLOSUM62 matrix based distance measurements

		Predicted classification group			
		Group 1	Group 2	Group 3	Group 4
Grouping	Group 1	18	0	2	11
	Group 2	5	7	0	3
	Group 3	1	2	2	2
	Group 4	19	5	8	85
	Total	43	14	12	101
	Error rate	0.4194	0.5333	0.7143	0.2735

Table 3.5 Cross validation results of the linear discriminant analyses considering of all possible variables with PSSM based distance measurements

		Predicted classification group			
		Group 1	Group 2	Group 3	Group 4
Grouping	Group 1	22	1	3	5
	Group 2	5	5	2	3
	Group 3	1	2	3	1
	Group 4	19	4	5	89
	Total	47	12	13	98
	Error rate	0.2903	0.6667	0.5714	0.2393

Table 3.6 Cross validation results of the linear discriminant analyses after the selection of significant variables and with PSSM based distance measurements

		Predicted classification group			
		Group 1	Group 2	Group 3	Group 4
Grouping	Group 1	26	1	3	1
	Group 2	4	7	0	4
	Group 3	2	2	2	1
	Group 4	19	7	6	85
	Total	51	17	11	91
	Error rate	0.1613	0.5333	0.7143	0.2735

Table 3.7 Cross validation results of the k-nearest discriminant analyses considering all possible variables with k values of 3, 4, and 5 and with BLOSUM62 matrix based distance measurements

		Predicted classification group			
		Group 1	Group 2	Group 3	Group 4
Grouping	**k=3**				
	Group 1	23	4	2	0
	Group 2	4	10	0	1
	Group 3	1	3	3	0
	Group 4	14	2	5	96
	Total	42	19	10	97
	Error rate	0.2581	0.3333	0.5714	0.1795
	k=4				
	Group 1	25	4	2	0
	Group 2	4	10	0	1
	Group 3	1	3	3	0
	Group 4	16	2	5	94
	Total	46	19	10	95
	Error rate	0.1935	0.3333	0.5714	0.1966
	k=5				
	Group 1	24	5	2	0
	Group 2	5	9	0	1
	Group 3	1	3	3	0
	Group 4	18	2	5	92
	Total	48	19	10	93
	Error rate	0.2258	0.4000	0.5714	0.2137

Table 3.8 Cross validation results of the k-nearest discriminant analyses with k values of 3, 4, and 5 after the selection of significant variables and with BLOSUM62 matrix based distance measurements

		Predicted classification group			
		Group 1	Group 2	Group 3	Group 4
Grouping	**k=3**				
	Group 1	25	3	2	1
	Group 2	4	11	0	0
	Group 3	1	4	2	0
	Group 4	18	3	6	90
	Total	48	21	10	91
	Error rate	0.1935	0.2667	0.7143	0.2308
	k=4				
	Group 1	24	3	2	1
	Group 2	8	7	0	0
	Group 3	2	3	2	0
	Group 4	19	2	6	90
	Total	53	15	10	91
	Error rate	0.2258	0.5333	0.7143	0.2308
	k=5				
	Group 1	24	3	2	1
	Group 2	9	6	0	0
	Group 3	5	2	0	0
	Group 4	19	2	6	90
	Total	57	13	8	91
	Error rate	0.2258	0.6000	1.000	0.2308

Table 3.9 Cross validation results of the k-nearest discriminant analyses considering of all possible variables with k values of 3, 4, and 5 and with PSSM based distance measurements

		Predicted classification group			
		Group 1	Group 2	Group 3	Group 4
Grouping	**k=3**				
	Group 1	25	3	2	1
	Group 2	5	7	2	1
	Group 3	3	1	2	1
	Group 4	17	5	8	87
	Total	50	16	14	90
	Error rate	0.1935	0.5333	0.7143	0.2564
	k=4				
	Group 1	28	1	2	0
	Group 2	5	7	2	1
	Group 3	3	1	2	1
	Group 4	18	2	8	89
	Total	54	11	14	91
	Error rate	0.0968	0.5333	0.7143	0.2393
	k=5				
	Group 1	27	2	2	0
	Group 2	5	7	2	1
	Group 3	3	2	2	0
	Group 4	18	2	8	89
	Total	53	13	14	90
	Error rate	0.1290	0.5333	0.7143	0.2393

Table 3.10 Cross validation results of the k-nearest discriminant analyses with k values of 3, 4, and 5 after the selection of significant variables and with PSSM based distance measurements

		Predicted classification group			
		Group 1	Group 2	Group 3	Group 4
Grouping	**k=3**				
	Group 1	24	3	2	1
	Group 2	3	9	2	1
	Group 3	3	1	2	1
	Group 4	21	5	7	84
	Total	51	18	13	87
	Error rate	0.2258	0.4000	0.7143	0.2821
	k=4				
	Group 1	24	4	2	1
	Group 2	4	10	0	1
	Group 3	3	1	2	1
	Group 4	21	5	7	84
	Total	52	20	11	87
	Error rate	0.2258	0.3333	0.7143	0.2821
	k=5				
	Group 1	24	4	2	1
	Group 2	4	10	0	1
	Group 3	6	1	0	0
	Group 4	21	5	7	84
	Total	55	20	9	86
	Error rate	0.2258	0.3333	1.000	0.2821

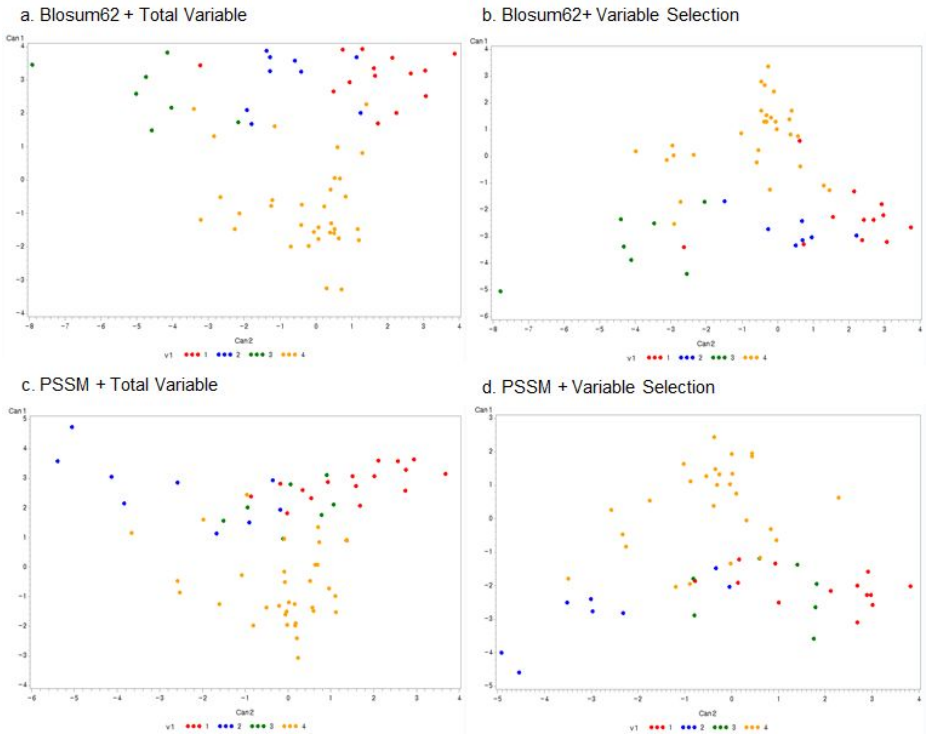

Figure 3.2 Canonical Discriminant Analysis Results in 2 by 2 Plots (Group 1- red, Group 2- blue, Group 3- green, and Group 4- orange). Two dimensional plot was also built using two canonical variables (y=can1, x=can2). **a**. Wilk's Lambda and Pillai's Trace are 0.0442 and 1.9025. The group is most clearly distinguished. **b**. Wilk's Lambda and Pillai's Trace are 0.0897 and 1.6302. **c**. Wilk's Lambda and Pillai's Trace are 0.0614 and 1.7720. **d.** Wilk's Lambda and Pillai's Trace are 0.1217 and 1.4902.

Table 3.11 The Average Misclassification Rates of Discrimination Analyses of the Prediction of Prion Disease Susceptibility Groups by Scoring Matrix

Score matrix	BLOSUM62		PSSM	
	all variables	variable selection	all variables	variable selection
LDA	0.4108	0.4851	0.4419	0.4206
k-nearest DA				
k=3	0.3356	0.3513	0.4244	0.4055
k=4	**0.3237**	0.4260	0.3959	0.3889
k=5	0.3527	0.5141	0.4040	0.4603

Table 3.12 Cross validation results of kNN discrimination analyses with all variables for prediction of prion disease susceptibility

		True Positive	False positive	True Negative	False Negative	Sensitivity	Specificity
BLOSUM62							
k = 3	Group 1	23	19	120	8	74.19%	86.33%
	Group 2	10	9	146	5	66.67%	94.19%
	Group 3	3	7	156	4	42.86%	95.71%
	Group 4	96	1	52	21	82.05%	98.11%
k = 4	Group 1	25	21	118	6	80.65%	84.89%
	Group 2	10	9	146	5	66.67%	94.19%
	Group 3	3	7	156	4	42.86%	95.71%
	Group 4	94	1	52	23	80.34%	98.11%
k = 5	Group 1	24	24	115	7	77.42%	82.73%
	Group 2	9	10	145	6	60.00%	93.55%
	Group 3	3	7	156	4	42.86%	95.71%
	Group 4	92	1	52	25	78.63%	98.11%
PSSM							
k = 3	Group 1	25	25	114	6	80.65%	82.01%
	Group 2	7	9	146	8	46.67%	94.19%
	Group 3	2	12	151	5	28.57%	92.64%
	Group 4	87	3	50	30	74.36%	94.34%
k = 4	Group 1	28	26	113	3	90.32%	81.29%
	Group 2	7	4	151	8	46.67%	97.42%
	Group 3	2	12	151	5	28.57%	92.64%
	Group 4	89	2	51	28	76.07%	96.23%
k = 5	Group 1	27	26	113	4	87.10%	81.29%
	Group 2	7	6	149	8	46.67%	96.13%
	Group 3	2	12	151	5	28.57%	92.64%
	Group 4	89	1	52	28	76.07%	98.11%

Table 3.13. Total classification accuracy of discrimination analyses

k-nearest	k=3	k=4	k=5	LD
all-variables				
BLOSUM62	**77.65%**	**77.65%**	75.29%	72.94%
PSSM	71.18%	74.12%	73.53%	70.00%
selected variables				
BLOSUM62	75.29%	72.35%	70.59%	65.88%
PSSM	72.94%	70.59%	69.41%	70.59%

Table 3.14 Accuracy of Prediction of Susceptibility for Test data

Score matrix	BLOSUM62		PSSM	
	all variables	variable selection	all variables	variable selection
LDA	71.43%	42.86%	57.14%	57.14%
k-nearest DA				
k=3	57.14%	42.86%	28.57%	57.14%
k=4	57.14%	42.86%	42.86%	71.43%
k=5	42.86%	42.86%	42.86%	71.43%

Chapter 4.
Conclusions

Mammalian prion sequence database was constructed for the analysis of polymorphisms in this study. Information of amino acid substitution which affects prion disease susceptibility was collected through literary search. Mutant sequences were generated through the information of substitutions and dataset of discriminant analysis was constructed by incorporating these sequences with previously collected prion sequences in the database. Polymorphisms on prion protein are known to be of important influence on the susceptibility of prion disease. However, database for the specific information of prion protein sequence polymorphisms has not been constructed. Therefore, this attempt to build prion polymorphism database will help set grounds of the research. There is no distinct difference in the frequency of polymorphism between species known to resistant to prion disease and species with high susceptibility to TSE. Thus, we suggest that there is no association between the frequency of polymorphism and risk of prion diseases. BLOSUM62 matrix is usually exploited in the homology search and multiple sequence alignment, though PSSM is recently frequently used for the search of distant protein and protein family. JAVA codes were used to build BLOSUM62 matrix and PSSM. Classification accuracy using each of the two matrices was compared. As a result, BLOSUM62 showed less misclassification than PSSM, and it is different from other studies (Ou et al., 2013). This is probably due to the fact that BLOSUM62 value has the range of -4 to 11, while PSSM values has the range of -3.8 to 6.9. Therefore, the

BLOSUM62 matrix may better present difference of amino acids. Generally, variable selection reduced accuracy of the classifications. When using the k-nearest neighbor method considering the three and four k objects, the most accurate susceptibility group was classified. Mainly, error rate of Group 1 is low. Inaccurate classification results can be attributed to training dataset which were used in discriminant analysis that do not contain all information about investigation on the effect of polymorphism. Probably caused by these reasons, the presence or absence of a specific polymorphism in prion sequence was difficult to accurately assess the risk of prion diseases. Eventually, the information of a specific amino acid alone does not yet explain the risk of prion diseases and does not show a solid foundation yet. However, because the sequence having a negative effect polymorphism in the prion disease has relatively high accuracy, the prion sequence collected more will be able to increase the accuracy of classification. Residue 129, 190, and 205 were selected as significant variables through logistic discriminant analysis. These residues were selected as significant variables through stepwise method in linear discriminant analysis. Thus, it is possible that these residues might have significant influences on discriminance of the susceptibility group. In fact, M129V, T190V substitutions increase resistance of prion disease. The concordance of the selected variables and previously known polymorphisms might suggest the correctness of the selection of significant variables. Clustering which is unsupervised learning method try to make the clusters of data points based on their similarities to each other using Euclidean distance (Ubeyli and Doğdu, 2010). When k-means clustering was performed using the same dataset, it is failed to get success results because of different groups are gathered in the same cluster.

The discriminant analysis method with training 177 prion sequences

with polymorphisms was shown to be reliable in this study. The effect of the use of PSSM instead of typical scoring matrix was also investigated. The research presented here has a significance of finding susceptibility information of mammalian prion with discriminant analysis and scoring matrix without the use of experimental methods in a short time. The study has limitations in the incorporated number of species and sequences for the discriminant analysis, but investigation of more diverse species and the effects of residue-wise susceptibility will support better accuracy. In the future, more prion sequence would be deposited into the prion polymorphism database to be possible to be accessed though internet. Through this research, polymorphism information from experimental researches was incorporated for the classifications. It was found that scoring matrix for sequence alignment and discriminant analysis could be efficient in the classification of the susceptibility of new prion strains. These methods might be valuable in the context of public health for the fast apprehension of the collected sequence without experimental procedures. In this study, it was investigated if the polymorphism affects the onset of prion diseases. As a result, it was confirmed that the existence of the negative polymorphism accurately predicted the risk of prion diseases, but simple polymorphism changing by itself had limitations to evaluate the impact of changes for susceptibility. Therefore, it seems that additional research is needed on the differences depending on species and the interaction with other factors such as secondary structure, hydrophobicity, and charge.

BIBLIOGRAPHY

Prusiner SB. 1991. Molecular biology of prion diseases. Science. 252(5012):1515-1522.

Prusiner SB. 1998. Prions. Proc Natl Acad Sci U S A. 95(23):13363-13383.

Wadsworth JD, Jackson GS, Hill AF, Collinge J. 1999. Molecular biology of prion propagation. Curr Opin Genet Dev. 9(3):338-345.

Klamt F, Dal-Pizzol F, Conte da Frota ML Jr, Walz R, Andrades ME, da Silva EG, Brentani RR, Izquierdo I, Fonseca Moreira JC. 2001. Imbalance of antioxidant defense in mice lacking cellular prion protein. Free Radic Biol Med. 30(10):1137-1144.

Gajdusek DC. 1972. Spongiform virus encephalopathies. J Clin Pathol Suppl (R Coll Pathol). 6:78-83.

Sigurdsson B. 1954. Rida, a chronic encephalitis of sheep: With general remarks on infections which develop slowly and some of their special characteristics. Br Vet J 110: 341–354.

Gajdusek DC, Zigas V. 1957. Degenerative disease of the central nervous system in New Guinea; the endemic occurrence of kuru in the native population. N Engl J Med. 257(20):974-978.

Hadlow WJ. 1959. Scrapie and kuru. Lancet ii, 289–290.

Gajdusek DC, Gibbs CJ Jr, Alpers, M. 1966. Experimental transmission of kuru-like syndrome to chimpanzee. Nature. 209(5025):794–796.

Griffith JS. 1967. Self-replication and scrapie. Nature. 215(5105):1043-1044.

Prusiner SB. 1982. Novel proteinaceous infectious particles cause scrapie. Science. 216(4542):136-144.

Büeler H, Aguzzi A, Sailer A, Greiner RA, Autenried P, Aguet M, Weissmann C. 1993. Mice devoid of PrP are resistant to scrapie. Cell. 73(7):1339-1347.

Wells GA, Scott AC, Johnson CT, Gunning RF, Hancock RD, Jeffrey M, Dawson M, Bradley R. 1987. A novel progressive spongiform encephalopathy in cattle. Vet Rec. 121(18):419–420.

Pan KM, Baldwin M, Nguyen J, Gasset M, Serban A, Groth D, Mehlhorn I, Huang Z, Fletterick RJ, Cohen FE. 1993. Conversion of alpha-helices into beta-sheets features in the formation of the scrapie prion proteins. Proc Natl Acad Sci U S A. 90(23):10962-10966.

Sakudo A, Xue G, Kawashita N, Ano Y, Takagi T, Shintani H, Tanaka Y, Onodera T, Ikuta K. 2010. Structure of the prion protein and its gene: an analysis using bioinformatics and computer simulation. Curr Protein Pept Sci. 11(2):166-179.

Baybutt H, Manson J. 1997. Characterisation of two promoters for prion protein (PrP) gene expression in neuronal cells. Gene.

184(1):125-131.

Saeki K, Matsumoto Y, Onodera T. 1996. Identification of a promoter region in the rat prion protein gene. Biochem Biophys Res Commun. 219(1): 47-52.

O'Neill GT, Donnelly K, Marshall E, Cairns D, Goldmann W, Hunter N. 2003. Characterization of ovine PrP gene promoter activity in N2a neuroblastoma and ovine foetal brain cell lines. J Anim Breed Genet. 120: 114-123.

Inoue S, Tanaka M, Horiuchi M, Ishiguro N, Shinagawa M. 1997. Characterization of the bovine prion protein gene: the expression requires interaction between the promoter and intron. J Vet Med Sci. 59(3):175-183.

Mahal SP, Asante EA, Antoniou M, Collinge J. 2001. Isolation and functional characterisation of the promoter region of the human prion protein gene. Gene, 268(1-2):105-114.

Puckett C, Concannon P, Casey C, Hood L. 1991. Genomic structure of the human prion protein gene. Am J Hum Genet. 49(2):320-329.

Kim Y, Lee J, Lee C. 2008. In silico comparative analysis of DNA and amino acid sequences for prion protein gene. Transbound Emerg Dis. 55(2):105-114.

Dynan WS, Tjian R. 1983. The promoter-specific transcription factor Sp1 binds to upstream sequences in the SV40 early promoter. Cell.

35(1):79-87.

Juling K, Schwarzenbacher H, Williams JL, Fries R. 2006. A major genetic component of BSE susceptibility. BMC Biol, 4:33.

Colby DW, Prusiner SB. 2011. Prions. Cold Spring Harb Perspect Biol. 3(1):a006833.

Stahl N, Baldwin MA, Teplow DB, Hood L, Gibson BW, Burlingame AL, Prusiner SB. 1993. Structural studies of the scrapie prion protein using mass spectrometry and amino acid sequencing. Biochemistry. 32(8):1991-2002.

Wüthrich K, Riek R. 2001. Three-dimensional structures of prion proteins. Adv Protein Chem. 57:55-82.

Goldfarb LG, Brown P, McCombie WR, Goldgaber D, Swergold GD, Wills PR, Cervenakova L, Baron H, Gibbs CJ Jr, Gajdusek DC. 1991. Transmissible familial Creutzfeldt-Jakob disease associated with five, seven, and eight extra octapeptide coding repeats in the PRNP gene. Proc Natl Acad Sci U S A. 88(23):10926-10930.

Choi CJ, Kanthasamy A, Anantharam V, Kanthasamy AG. 2006. Interaction of metals with prion protein: possible role of divalent cations in the pathogenesis of prion diseases. Neurotoxicology. 27(5):777-787.

Millhauser GL. 2004. Copper binding in the prion protein. Acc Chem Res. 37(2):79-85.

Riek R, Hornemann S, Wider G, Billeter M, Glockshuber R, Wüthrich K. 1996. NMR structure of the mouse prion protein domain PrP(121-231). Nature. 382(6587):180-182.

Zahn R, Liu A, Lùhrs T, Riek R, von Schroetter C, Lōpez Garcīa F, Billeter M, Calzolai L, Wider G, Wüthrich K. 2000. NMR solution structure of the human prion protein. Proc Natl Acad Sci U S A. 97(1):145-150.

Brown DR. 2001. Copper and prion disease. Brain Res Bull. 55(2):165-173.

Sakudo A, Onodera T, Suganuma Y, Kobayashi T, Saeki K, Ikuta K. 2006. Recent advances in clarifying prion protein functions using knockout mice and derived cell lines. Mini Rev Med Chem. 6(5), 589-601.

Westergard L, Christensen HM, Harris DA. 2007. The cellular prion protein (PrP(C)): its physiological function and role in disease. Biochim Biophys Acta. 1772(6):629-644.

Pauly PC, Harris DA. 1998. Copper stimulates endocytosis of the prion protein. J Biol Chem. 273(50), 33107-33110.

King CY, Diaz-Avalos R. 2004. Protein-only transmission of three yeast prion strains. Nature. 428(6980):319-323.

Baxa U, Cassese T, Kajava AV, Steven AC. 2006. Structure, function,

and amyloidogenesis of fungal prions: filament polymorphism and prion variants. Adv Protein Chem. 73:125-180.

Wickner RB. 1994. [URE3] as an altered URE2 protein: Evidence for a prion analog in Saccharomyces cerevisiae. Science 264(5158): 566–569.

Patino MM, Liu J-J, Glover JR, Lindquist S. 1996. Support for the prion hypothesis for inheritance of a phenotypic trait in yeast. Science 273(5275): 622–626.

Coustou V, Deleu C, Saupe S, Begueret J. 1997. The protein product of the het-s heterokaryon incompatibility gene of the fungus Podospora anserina behaves as a prion analog. Proc Natl Acad Sci 94(18): 9773–9778.

Borchelt DR, Scott M, Taraboulos A, Stahl N, Prusiner SB. 1990. Scrapie and cellular prion proteins differ in their kinetics of synthesis and topology in cultured cells. J Cell Biol. 110(3):743-752.

Caughey BW, Dong A, Bhat KS, Ernst D, Hayes SF, Caughey WS. 1991. Secondary structure analysis of the scrapie-associated protein PrP 27-30 in water by infrared spectroscopy. Biochemistry. 30(31):7672-7680.

Kocisko DA, Come JH, Priola SA, Chesebro B, Raymond GJ, Lansbury PT, Caughey B. 1994. Cell-free formation of protease-resistant prion protein. Nature. 370(6489):471-474.

Cobb NJ, Surewicz WK. 2009. Prion diseases and their biochemical

mechanisms. Biochemistry. 48(12):2574-2585.

Cohen FE, Prusiner SB. 1998. Pathologic conformations of prion proteins. Annu Rev Biochem. 67, 793-819.

Apetri AC, Maki K, Roder H, Surewicz WK. 2006. Early intermediate in human prion protein folding as evidenced by ultrarapid mixing experiments. J Am Chem Soc. 128(35):11673-11678.

Jarrett JT, Lansbury PT Jr. 1993. Seeding "one-dimensional crystallization" of amyloid: a pathogenic mechanism in Alzheimer's disease and scrapie? Cell. 73(6):1055-1058.

Shorter J, Lindquist S. 2008. Hsp104, Hsp70 and Hsp40 interplay regulates formation, growth and elimination of Sup35 prions. EMBO J 27(20): 2712–2724.

Xue WF, Homans SW, Radford SE. 2008. Systematic analysis of nucleation-dependent polymerization reveals new insights into the mechanism of amyloid self-assembly. Proc Natl Acad Sci U S A. 105(26): 8926–8931

Chiti F, Dobson CM. 2006. Protein misfolding, functional amyloid, and human disease. Annu Rev Biochem. 75:333-366.

Tycko R. 2004. Progress towards a molecular-level structural understanding of amyloid fibrils. Curr Opin Struct Biol. 14(1):96-103.

Nelson R, Sawaya MR, Balbirnie M, Madsen AØ, Riekel C, Grothe R,

Eisenberg D. 2005. Structure of the cross-beta spine of amyloid-like fibrils. Nature. 435(7043):773-778.

Palmer MS, Dryden AJ, Hughes JT, Collinge J. 1991. Homozygous prion protein genotype predisposes to sporadic Creutzfeldt-Jakob disease. Nature. 352(6333):340-342.

Collinge J, Sidle KC, Meads J, Ironside J, Hill AF. 1996. Molecular analysis of prion strain variation and the aetiology of 'new variant' CJD. Nature. 383(6602):685-690.

Wadsworth JD, Hill AF, Joiner S, Jackson GS, Clarke AR, Collinge J. 1999. Strain-specific prion-protein conformation determined by metal ions. Nat Cell Biol. 1(1):55-59.

Hill AF, Joiner S, Wadsworth JD, Sidle KC, Bell JE, Budka H, Ironside JW, Collinge J. 2003. Molecular classification of sporadic Creutzfeldt-Jakob disease. Brain. 126(Pt 6):1333-1346.

Wadsworth JD, Asante EA, Desbruslais M, Linehan JM, Joiner S, Gowland I, Welch J, Stone L, Lloyd SE, Hill AF, Brandner S, Collinge J. 2004. Human prion protein with valine 129 prevents expression of variant CJD phenotype. Science. 306(5702):1793-1796.

Shibuya S, Higuchi J, Shin RW, Tateishi J, Kitamoto T. 1998. Codon 219 Lys allele of PRNP is not found in sporadic Creutzfeldt-Jakob disease. Ann Neurol. 43(6):826-828.

Hainfellner JA, Brantner-Inthaler S, Cervenáková L, Brown P, Kitamoto

T, Tateishi J, Diringer H, Liberski PP, Regele H, Feucht M, et al. 1995. The original Gerstmann-Sträussler-Scheinker family of Austria: divergent clinicopathological phenotypes but constant PrP genotype. Brain Pathol. 5(3):201-211.

Yamada M, Itoh Y, Inaba A, Wada Y, Takashima M, Satoh S, Kamata T, Okeda R, Kayano T, Suematsu N, Kitamoto T, Otomo E, Matsushita M, Mizusawa H. 1999. An inherited prion disease with a PrP P105L mutation: clinicopathologic and PrP heterogeneity. Neurology. 53(1):181-188.

Parchi P, Capellari S, Chin S, Schwarz HB, Schecter NP, Butts JD, Hudkins P, Burns DK, Powers JM, Gambetti P. 1999. A subtype of sporadic prion disease mimicking fatal familial insomnia. Neurology. 52(9):1757-1763.

Appleby BS, Appleby KK, Hall RC, Wallin MT. 2010. D178N, 129Val and N171S, 129Val genotype in a family with Creutzfeldt-Jakob disease. Dement Geriatr Cogn Disord. 30(5):424-431.

Chasseigneaux S, Haïk S, Laffont-Proust I, De Marco O, Lenne M, Brandel JP, Hauw JJ, Laplanche JL, Peoc'h K. 2006. V180I mutation of the prion protein gene associated with atypical PrPSc glycosylation. Neurosci Lett. 408(3):165-169.

Mutsukura K, Satoh K, Shirabe S, Tomita I, Fukutome T, Morikawa M, Iseki M, Sasaki K, Shiaga Y, Kitamoto T, Eguchi K. 2009. Familial Creutzfeldt-Jakob disease with a V180I mutation: comparative analysis with pathological findings and diffusion-weighted images.

Dement Geriatr Cogn Disord. 28(6):550-557.

Hsiao K, Meiner Z, Kahana E, Cass C, Kahana I, Avrahami D, Scarlato G, Abramsky O, Prusiner SB, Gabizon R. 1991. Mutation of the prion protein in Libyan Jews with Creutzfeldt-Jakob disease. N Engl J Med. 324(16):1091-1097.

Kaneko K, Zulianello L, Scott M, Cooper CM, Wallace AC, James TL, Cohen FE, Prusiner SB. 1997. Evidence for protein X binding to a discontinuous epitope on the cellular prion protein during scrapie prion propagation. Proc Natl Acad Sci U S A. 94(19):10069-10074.

Grasbon-Frodl E, Lorenz H, Mann U, Nitsch RM, Windl O, Kretzschmar HA. 2004. Loss of glycosylation associated with the T183A mutation in human prion disease. Acta Neuropathol. 108(6):476-484.

Peoc'h K, Manivet P, Beaudry P, Attane F, Besson G, Hannequin D, Delasnerie-Lauprêtre N, Laplanche JL. 2000. Identification of three novel mutations (E196K, V203I, E211Q) in the prion protein gene (PRNP) in inherited prion diseases with Creutzfeldt-Jakob disease phenotype. Hum Mutat. 15(5):482.

Piccardo P, Liepnieks JJ, William A, Dlouhy SR, Farlow MR, Young K, Nochlin D, Bird TD, Nixon RR, Ball MJ, DeCarli C, Bugiani O, Tagliavini F, Benson MD, Ghetti B. 2001. Prion proteins with different conformations accumulate in Gerstmann-Sträussler-Scheinker disease caused by A117V and F198S mutations. Am J Pathol. 158(6):2201-2207.

Piccardo P, Dlouhy SR, Lievens PM, Young K, Bird TD, Nochlin D, Dickson DW, Vinters HV, Zimmerman TR, Mackenzie IR, Kish SJ, Ang LC, De Carli C, Pocchiari M, Brown P, Gibbs CJ Jr, Gajdusek DC, Bugiani O, Ironside J, Tagliavini F, Ghetti B. 1998. Phenotypic variability of Gerstmann-Sträussler-Scheinker disease is associated with prion protein heterogeneity. J Neuropathol Exp Neurol. 57(10):979-988.

Capellari S, Cardone F, Notari S, Schininà ME, Maras B, Sità D, Baruzzi A, Pocchiari M, Parchi P. 2005. Creutzfeldt-Jakob disease associated with the R208H mutation in the prion protein gene. Neurology. 64(5):905-907.

Biljan I, Ilc G, Giachin G, Raspadori A, Zhukov I, Plavec J, Legname G. 2011. Toward the molecular basis of inherited prion diseases: NMR structure of the human prion protein with V210I mutation. J Mol Biol. 412(4):660-673.

Shiga Y, Satoh K, Kitamoto T, Kanno S, Nakashima I, Sato S, Fujihara K, Takata H, Nobukuni K, Kuroda S, Takano H, Umeda Y, Konno H, Nagasato K, Satoh A, Matsuda Y, Hidaka M, Takahashi H, Sano Y, Kim K, Konishi T, Doh-ura K, Sato T, Sasaki K, Nakamura Y, Yamada M, Mizusawa H, Itoyama Y. 2007. Two different clinical phenotypes of Creutzfeldt-Jakob disease with a M232R substitution. J Neurol. 254(11):1509-1517.

Baylis M, Goldmann W. 2004. The genetics of scrapie in sheep and goats. Curr Mol Med. 4(4):385-396.

Sander P, Hamann H, Drögemüller C, Kashkevich K, Schiebel K, Leeb T. 2005. Bovine prion protein gene (PRNP) promoter polymorphisms modulate PRNP expression and may be responsible for differences in bovine spongiform encephalopathy susceptibility. J Biol Chem. 280(45):37408-37414.

Xue G, Sakudo A, Kim CK, Onodera T. 2008. Coordinate regulation of bovine prion protein gene promoter activity by two Sp1 binding site polymorphisms. Biochem Biophys Res Commun. 372(4):530-535.

Heaton MP, Keele JW, Harhay GP, Richt JA, Koohmaraie M, Wheeler TL, Shackelford SD, Casas E, King DA, Sonstegard TS, Van Tassell CP, Neibergs HL, Chase CC Jr, Kalbfleisch TS, Smith TP, Clawson ML, Laegreid WW. 2008. Prevalence of the prion protein gene E211K variant in U.S. cattle. BMC Vet Res. 4:25.

Westaway D, Goodman PA, Mirenda CA, McKinley MP, Carlson GA, Prusiner SB. 1987. Distinct prion proteins in short and long scrapie incubation period mice. Cell. 51(4):651-662.

Saunders GC, Lantier I, Cawthraw S, Berthon P, Moore SJ, Arnold ME, Windl O, Simmons MM, Andréoletti O, Bellworthy S, Lantier F. 2009. Protective effect of the T112 PrP variant in sheep challenged with bovine spongiform encephalopathy. J Gen Virol. 90(Pt 10):2569-2574.

Mead S, Whitfield J, Poulter M, Shah P, Uphill J, Campbell T, Al-Dujaily H, Hummerich H, Beck J, Mein CA, Verzilli C, Whittaker J, Alpers MP, Collinge J. 2009. A novel protective prion protein

variant that colocalizes with kuru exposure. N Engl J Med. 361(21):2056-5065.

Vaccari G, D'Agostino C, Nonno R, Rosone F, Conte M, Di Bari MA, Chiappini B, Esposito E, De Grossi L, Giordani F, Marcon S, Morelli L, Borroni R, Agrimi U. 2007. Prion protein alleles showing a protective effect on the susceptibility of sheep to scrapie and bovine spongiform encephalopathy. J Virol. 81(13):7306-9.

Saunders GC, Cawthraw S, Mountjoy SJ, Hope J, Windl O. 2006. PrP genotypes of atypical scrapie cases in Great Britain. J Gen Virol. 87(Pt 11):3141-3149.

Panegyres PK, Toufexis K, Kakulas BA, Cernevakova L, Brown P, Ghetti B, Piccardo P, Dlouhy SR. 2001. A new PRNP mutation (G131V) associated with Gerstmann-Sträussler-Scheinker disease. Arch Neurol. 58(11):1899-1902.

Goldmann W, Martin T, Foster J, Hughes S, Smith G, Hughes K, Dawson M, Hunter N. 1996. Novel polymorphisms in the caprine PrP gene: a codon 142 mutation associated with scrapie incubation period. J Gen Virol. 77 (Pt 11):2885-2891.

Vaccari G, Di Bari MA, Morelli L, Nonno R, Chiappini B, Antonucci G, Marcon S, Esposito E, Fazzi P, Palazzini N, Troiano P, Petrella A, Di Guardo G, Agrimi U. 2006. Identification of an allelic variant of the goat PrP gene associated with resistance to scrapie. J Gen Virol. 87(Pt 5):1395-1402.

Sakudo A, Ikuta K. 2009. Prion protein functions and dysfunction in prion diseases. Curr Med Chem. 16(3):380-389.

Smith PG, Bradley R. 2003. Bovine spongiform encephalopathy (BSE) and its epidemiology. Br Med Bull. 66:185-198.

Britton TC, al-Sarraj S, Shaw C, Campbell T, Collinge J. 1995. Sporadic Creutzfeldt-Jakob disease in a 16-year-old in the UK. Lancet. 346(8983):1155.

Gehlenborg N, Hwang D, Lee IY, Yoo H, Baxter D, Petritis B, Pitstick R, Marzolf B, Dearmond SJ, Carlson GA, Hood L. 2009. The Prion Disease Database: a comprehensive transcriptome resource for systems biology research in prion diseases. Database (Oxford). 2009:bap011.

Pawlicki S, Le Bechec A, Delamarche C. 2008. AMYPdb: a database dedicated to amyloid precursor proteins. BMC Bioinformatics 9: 273.

Harbi D, Parthiban M, Gendoo DM, Ehsani S, Kumar M, Schmitt-Ulms G, Sowdhamini R, Harrison PM. 2012. PrionHome: a database of prions and other sequences relevant to prion phenomena. PLoS One. 7(2):e31785.

Martin TC, Moecks J, Belooussov A, Cawthraw S, Dolenko B, Eiden M, Von Frese J, Kohler W, Schmitt J, Somorjai R, Udelhoven T, Verzakov S, Petrich W. 2004. Classification of signatures of Bovine Spongiform Encephalopathy in serum using infrared spectroscopy. Analyst. 129(10):897-901.

Fernández M, Caballero J, Fernāndez L, Abreu JI, Acosta G. 2008. Classification of conformational stability of protein mutants from 3D pseudo-folding graph representation of protein sequences using support vector machines. Proteins. 70(1):167-175.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215(3):403-410.

Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 48(3):443-453.

Smith TF, Waterman MS. 1981. Comparison of biosequences. Advan Appl Math. 2(4):482-489.

Baxevanis AD, Ouellette BFF. 2005. Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, Third Edition. Wiley, John & Sons. Chapter 11. 295-324.

McEntyre J, Ostell J. 2002. The NCBI Handbook. National Center for Biotechnology Information (US). Chapter 16 The BLAST Sequence Analysis Tool

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of prgressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acids Res. 22(22):4673-4680.

Hall TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl Acids Symp Ser. 41:95-98.

Harris H. 1969. Enzyme and protein polymorphism in human populations. Br Med Bull. 25(1):5-13.

Harris H. 1971. Polymorphism and protein evolution. The neutral mutation-random drift hypothesis. J Med Genet. 8(4): 444-452.

Johnson GC, Todd JA. 2000. Strategies in complex disease mapping. Curr Opin Genet Dev. 10(3):330-334.

Fisher RA. 1936. The use of multiple measurements in taxonomic problems. Ann of Eugenics. 7:179-188

Fernandez GCJ. 2002. Discriminant Analysis, A Powerful Classification Technique in Data Mining. Proceedings of SAS User Group International (SUGI27). 247-27

SAS Institute Inc. 2008. SAS/STAT 9.2 Users Guide. Cary NC: SAS Institute Inc. Chapter 10. 201-208.

Henikoff S, Henikoff JG. 1992. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A. 89(22):10915-10919.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res.

25(17):3389-3402.

Lachenbruch PA, Mickey MA. 1968. Estimation of error rates in discriminant analysis. Technometrics 10(1):1-11.

Sweeting B, Khan MQ, Chakrabartty A, Pai EF. 2010. Structural factors underlying the species barrier and susceptibility to infection in prion disease. Biochem Cell Biol. 88(2):195-202.

Kirkwood JK, Cunningham AA. 1994. Epidemiological observations on spongiform encephalopathies in captive wild animals in the British Isles. Vet Rec. 135(13):296-303.

Ou YY, Chen SA, Wu SC. 2013. ETMB-RBF: discrimination of metal-binding sites in electron transporters based on RBF networks with PSSM profiles and significant amino acid pairs. PLoS One. 8(2):e46572.

Ubeyli ED, Doğdu E. 2010. Automatic detection of erythemato-squamous diseases using k-means clustering. J Med Syst. 34:179-184.

국문초록

Prion은 정상 프라이온 단백질인 PrP^{C}가 전염성을 갖는 비정상 형태인 PrP^{Sc}로 구조적 변화가 일어난 것으로, 양의 스크래피(scrapie), 소의 광우병(BSE), 사람의 쿠루(kuru), 크로이츠 펠트 야콥병(CJD)와 같은 전염성 해면상 뇌증(TSE)의 원인 물질로 알려진 단백질이다. 이러한 프라이온 질환은 종을 뛰어넘어 인간에게도 다른 종의 질환이 전염될 수 있기 때문에 보건학적으로도 많은 관심을 받고 있다. PrP^{C}는 주로 α-helix 구조를 이루어져 있고, PrP^{Sc}는 주로 β-sheet이 우세한 구조로 구성되어 있다. 프라이온의 서열에서 다양한 돌연변이(mutation)와 다형성(polymorphism)이 발견되고 있는데, 이러한 서열상의 차이가, PrP^{Sc}로의 구조적인 변화와 단백질 발현을 조절해서 프라이온 관련 질환의 감수성(susceptibility)에 영향을 줄 수 있다. 그동안 실험 연구를 통해서 감수성에 영향을 주는 프라이온 단백질의 다형성에 대한 중요성이 확인되어 왔지만, 그에 특화된 데이터베이스가 아직 생성되지 못하고 있다. 또한 PrP^{Sc}의 생화학적 특징으로 인해서 실험 연구가 어렵고, 분자동역학과 같은 컴퓨터 시뮬레이션은 많은 시간이 요구된다는 어려움이 있다. 따라서 실험과 복잡한 구조 분석없이 프라이온 단백질의 일차구조에 나타난 다형성 정보만을 이용하여 빠르게 감수성을 판단할 수 있는 기법이 필요하다고 생각하였다. 프라이온 다형성 데이터베이스를 생성하기 위해서, BLASTP 프로그램을 이용해서 포유류의 프라이온 단백질 서열을 수집하였고, ClustalW 프로그램을 이용해서 다중 서열 정렬을 하였다. 그런 다음에, JAVA 프로그래밍 언어를 이용해서 필요한 정보를 파싱하였고, MySQL 데이터베이스에 테이블을 생성해서 수집한 정보들을 저장하였다. 감수성 예측 프로그램을 생성하기 위해서, 문헌조사를 통해서 포유류의 프라이온 단백질 서열에서 돌연변이와 다형성이 프라이온 질환의 감수성에 주는 효과를 조사하였고, 이러한 정보를 이용하여 훈련 데이

터 셋을 4개의 그룹으로 분류하였다. 판별분석은 그룹이 분류된 이러한 데이터를 이용해서 수행된다. 새로운 서열에서 그룹이 정확하게 예측되는 지를 알아보기 위해서 다형성이 나타나지 않는 참조 서열에 인위적인 변화를 준 서열을 생성하여 판별 분석을 위한 테스트 데이터 셋으로 사용하였다. JAVA 코딩으로 위치 특이 점수(position-specific score)를 계산하였으며, 점수 행렬(scoring matrix)인 BLOSUM62 행렬과 PSSM 행렬을 이용해서 서열에서의 아미노산 차이를 점수로 치환하여 점수 행렬에 따른 판별 분석의 정확성을 비교하였다. k-nearest neighbor(kNN)과 선형판별분석의 정확성을 교차검증(cross-validation)방법으로 비교하였으며, 정준판별분석을 통해서 2차원 그래프로 그룹의 분류를 시각화하였다. 그 결과, 다형성의 개수와 프라이온 질환의 감수성에는 연관성이 없었으며, 3 또는 4개의 k 개체를 고려한 k-nearest neighbor를 사용하였을 때 가장 정확하게 감수성 그룹이 판별되었다. PSSM 행렬보다 BLOSUM62 행렬을 사용하였을 때 오분류율이 감소하였으며, 2차원 그래프에서 더 명확히 그룹이 분류되었다. 또한, 질환에 대한 감수성을 높이는 다형성을 갖는 서열에서 비교적 판별의 정확성이 높았지만, 다형성 변화 자체만으로 감수성의 변화에 주는 영향을 평가하기에는 제한점이 있었다. 본 연구를 통해서, 다형성 정보와 점수행렬을 이용한 판별 분석의 가능성을 살펴보았으며, 이러한 방법은 실험 없이 프라이온 질환에 대한 감수성 정도를 쉽고 빠르게 판별하는 데에 있어서 도움을 줄 것이다. 또한 이러한 생명정보학 기법은 추가적인 연구 분석을 통해서 보건학적으로도 유용하게 사용될 수 있을 것이다.

주요어 : 프라이온, 다형성, 감수성, 치환 점수 행렬, 판별분석
학 번 : 2011-20446