OPEN

Experimental & Molecular Medicine (2013) 45, e48; doi:10.1038/emm.2013.100

www.nature.com/emm

ORIGINAL ARTICLE

Discriminant analysis of prion sequences for prediction of susceptibility

Ji-Hae Lee[1,2,7], Se-Eun Bae[1,3,7], Sunghoon Jung[1,4], Insung Ahn[5] and Hyeon Seok Son[1,2,3,6]

Prion diseases, including ovine scrapie, bovine spongiform encephalopathy (BSE), human kuru and Creutzfeldt–Jakob disease (CJD), originate from a conformational change of the normal cellular prion protein (PrP^{C}) into abnormal protease-resistant prion protein (PrP^{Sc}). There is concern regarding these prion diseases because of the possibility of their zoonotic infections across species. Mutations and polymorphisms of prion sequences may influence prion-disease susceptibility through the modified expression and conformation of proteins. Rapid determination of susceptibility based on prion-sequence polymorphism information without complex structural and molecular biological analyses may be possible. Information regarding the effects of mutations and polymorphisms on prion-disease susceptibility was collected based on previous studies to classify the susceptibilities of sequences, whereas the BLOSUM62 scoring matrix and the position-specific scoring matrix were utilised to determine the distance of target sequences. The k-nearest neighbour analysis was validated with cross-validation methods. The results indicated that the number of polymorphisms did not influence prion-disease susceptibility, and three and four k-objects showed the best accuracy in identifying the susceptible group. Although sequences with negative polymorphisms showed relatively high accuracy for determination, polymorphisms may still not be an appropriate factor for estimating variation in susceptibility. Discriminant analysis of prion sequences with scoring matrices was attempted as a possible means of determining susceptibility to prion diseases. Further research is required to improve the utility of this method.
Experimental & Molecular Medicine (2013) **45,** e48; doi:10.1038/emm.2013.100; published online 11 October 2013

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

INTRODUCTION

Transmissible spongiform encephalopathy (TSE) is a mammalian prion disease associated with neurodegenerative disorders including ovine scrapie, bovine spongiform encephalopathy (BSE), cervine chronic wasting disease, feline spongiform encephalopathy, mink transmissible mink encephalopathy, human kuru, Gerstmann–Sträussler–Scheinker syndrome, Creutzfeldt–Jakob disease (CJD) and fatal familial insomnia.[1] BSE, whose characteristics are similar to scrapie, was first identified in cows with abnormal progressive neurological symptoms in 1986. As it was shown that variants of CJD can be transmitted through the consumption of meat from infected cattle, there has been increasing concern regarding prion disease.[2] Prion disease can cause dysfunctions in motor activity, cognition and brain function, as well as mortality; in addition, it has long incubation periods, although there is variation among strains and host species.[1,3] Prion disease is caused by the accumulation of protease-resistant prion protein (PrP^{Sc}), which is derived from conformational conversion of normal cellular PrP (PrP^{C}) and has neuropathological features involving the formation of amyloid plaques that develop spongiform brain tissue.[4] The structure of PrP^{Sc} contains a high proportion of β-sheets because of conformational conversion of PrP^{C}, which is composed mainly of α-helical structures.[5] Prion protein (PrP) is encoded by the PrP gene (*PRNP*) and is strongly conserved among mammals.[6] PrP^{C} is a protein of 210 amino acids in length that resides mainly on the surface of the

[1]Laboratory of Computational Biology and Bioinformatics, Graduate School of Public Health, Seoul National University, Gwanak-gu, Korea; [2]Interdisciplinary Graduate Program in Bioinformatics, College of Natural Science, Seoul National University, Gwanak-gu, Korea; [3]Institute of Public Health and Environment, Seoul National University, Gwanak-gu, Korea; [4]Molecular Recognition Research Center, Korea Institute of Science and Technology, Seongbuk-gu, Korea; [5]High-performance Biocomputing Team, Supercomputing R&D Center, National Institute of Supercomputing and Networking, Korea Institute of Science and Technology Information, Yuseong-gu, Korea and [6]SNU Bioinformatics Institute, Seoul National University, Gwanak-gu, Korea
[7]These authors contributed equally to this work.
Correspondence: Professor HS Son, Laboratory of Computational Biology and Bioinformatics, Graduate School of Public Health, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 151-742, Korea.
E-mail: hss2003@snu.ac.kr
Received 9 April 2013; revised 22 July 2013; accepted 29 July 2013

mammalian neuronal cells through C-terminal glycosyl-phosphatidyl inositol anchors, although it is also expressed in other cell types.[7] Nuclear magnetic resonance spectroscopy has indicated that the C-terminal region of PrP^{C} forms a highly structured globular domain and the N-terminal region forms a non-structured flexible tail.[8] The C-terminal globular domain of mammalian PrP^{C} is composed of two antiparallel β-sheets and three α-helices.[9] Sequence variation in the *PrP* gene appears to be important, as it changes the susceptibility to prion disease.[6,10] Moreover, the M129V polymorphism of human PrP affects resistance to sporadic CJD.[11] The E219K polymorphism is not found in sCJD patients, suggesting resistance to sCJD.[12] The P102L polymorphism is seen in Gerstmann–Sträussler–Scheinker syndrome families with rapid development of dementia and cortical damage similar to CJD,[13] whereas Gerstmann–Sträussler–Scheinker syndrome patients with spastic paraparesis show the P105L polymorphism.[14] D178N is related to fatal familial insomnia,[15] whereas V180I and E200K mutations are also known to cause familial CJD.[16,17] Table 1 shows the polymorphisms and mutations that are relevant to prion diseases in humans, cattle, sheep, goats and mice. Polymorphisms at positions 136, 154 and 171 are relevant to the onset of scrapie in sheep. Sheep with the ARR/AHQ genotype are resistant to scrapie infection, whereas those with the ARQ/ARH/VRQ genotype are highly susceptible to scrapie infection.[18] Among the observed polymorphism regions that affect bovine PrP expression, a 23-bp indel in the promoter region upstream of the transcriptional start site and a 12-bp indel polymorphism in intron 1 are known to be related to BSE susceptibility.[19,20] The E211K mutation of the *PrP* gene in cattle has also been shown to be related to BSE.[21]

PrP^{Sc} is difficult to crystallise and has insoluble physico-chemical properties, making it difficult to perform experimental structural analysis, including X-ray crystallography and nuclear magnetic resonance spectroscopy.[6] Various studies of the relevance of conformational changes associated with amino-acid substitutions, amyloid propensity and *PRNP* genetic variability with prion-disease susceptibility have been conducted using bioinformatics methods, including molecular dynamics and other computational science algorithms.[10,22,23] There have been a number of studies regarding PrP polymorphisms that influence the susceptibility to prion disease, leading to the possibility of discrimination of prion-disease susceptibility based only on the primary structure information of PrP without requiring complicated structural analysis. In the present study, information on the effects of polymorphisms on prion-disease susceptibility was collected from the literature and used to predict the level of susceptibility of PrP sequences. BLOSUM62 and position-specific scoring matrices (PSSM) were used to calculate

Table 1 Prion sequence polymorphisms associated with prion diseases

Residues	*Host*	*Res*[a]	*Suspt*[b]	*Residues*	*Host*	*Res*[a]	*Suspt*[b]
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	+	

L109F,[35] M109T,[36] G127V,[37] A133V,[18] M134T,[38] L138F,[39] I139K,[38] R151H,[18] Q168R,[18] M129V,[11] N171S,[40] N173K,[38] T190V,[35] and E219K[12] polymorphisms influence susceptibility or resistance of prion disease. V180I,[16,41] T183A,[42] E196K,[43] E200K,[17] V203I,[43] R208H,[44] V210I,[45] E211Q,[43] and M232R[46] mutations cause Creutzfeldt–Jakob disease (CJD). P102L,[13] P105L,[14] A117V,[47] G131V,[48] F198S,[47] D202N,[49] Q212P,[49] and Q217R[49] mutations associated to Gerstmann–Sträussler–Scheinker syndrome (GSS). D178N mutation related to Fatal familial insomnia (FFI) or CJD.[15] E211K mutation related to Bovine spongiform encephalopathy (BSE).[21] Q168E, Q172R, V215I and Q219E mutations prevented PrP^{Sc} formation.[50] I139M[51] and Q219K[52] associated with scrapie resistance.
[a]Res: resistant effect
[b]Suspt: susceptibility effect.

distance scores of target sequences considering amino-acid substitutions, and susceptibility predictions were validated using k-nearest neighbour discriminant analysis.

MATERIALS AND METHODS

Data sets

Prion sequences of 23 mammalian species were collected from the NCBI protein database and a sequence homology search was performed with BLASTP 2.2.27.[24] A total of 222 PrP sequences were collected by choosing non-partial sequences with coverage of >50% from a preliminary search with default parameters and a search set was chosen that incorporated non-redundant (nr) protein databases and relevant species. Multiple sequence alignment was performed on collected mammalian PrP sequences with ClustalW using the default parameters.[25] The region of residues 98–227 (numbered according to the human sequence) with relatively few gaps and good alignment was selected, discarding the remaining regions with many gaps using BioEdit 7.1.3.[26] This region is a globular domain composed of two β-sheets and three α-helices and shows many prion disease-relevant mutations and polymorphisms. Sequence variation in PrP is relevant in the conformational conversion of PrP^{C} to PrP^{Sc} and in protein expression, which may strongly influence prion-disease susceptibility.[6,19] The effects of mutations and polymorphisms at different sites on susceptibility to prion disease were investigated (Table 1) and used to classify groups of susceptibility from a training set of 170 sequences. A test data set was constructed with new mutant sequences changed by substituting mouse, sheep, human, goat and cattle reference PrP protein sequences with new residues either negatively or positively influencing prion-disease susceptibility. Reference sequences of NP_000302.1 (human), NP_035300.1 (mouse), P52113.1 (goat), P10279.2 (cattle) and NP_001009481.1 (sheep) were used. A total of 177 sequences consisting of the 170 training sequences and 7 new mutant sequences were used for discriminant analysis. A total of four groups were classified based on the sequence information: Group 1 (sequences with polymorphisms that increase susceptibility), Group 2 (sequences with polymorphisms that decrease susceptibility), Group 3 (sequences with both types of polymorphisms) and Group 4 (sequences with polymorphisms that neither increase nor decrease susceptibility). The frequencies of the groups were 31, 15, 7 and 117, respectively.

The k-nearest neighbour discriminant analysis

The SAS 9.3 statistical package was used for k-nearest neighbour discriminant analysis to classify susceptibility to prion disease based on sequence variation because of polymorphisms. Discriminant analysis is a type of multivariate analysis, in which data are classified according to predetermined groups and a model is built from variables and numerous objects to identify pertinent groups for novel target sequences.[27] The k-nearest neighbour discriminant analysis method classifies an object as belonging to the group with the most nearest k-objects, calculated with the Mahalanobis distance from pooled covariance matrices of non-normally distributed sets.[27]

Distance with a BLOSUM62 matrix and the PSSM method

Classification of prion sequence members into susceptibility groups was performed using sequence similarity based on sequence alignment. A scoring matrix was used to represent the genealogical distance of a sequence from the consensus sequence. A BLOSUM62 matrix[28] was used to calculate the distance. The distance of a residue of a sequence was measured as the substitution score from the amino acid of the relevant column in the consensus sequence. The consensus sequence was built as the sequence with the most frequent amino acids for each column of the alignment. The results obtained using the BLOSUM62 matrix were compared with those obtained using PSSM.[29] PSSM is a scoring matrix that represents position-dependent substitution scores from multiple sequence alignment. The matrix showed different substitution scores for the same type of substitution considering the column-wise status of amino-acid types. The column-wise frequency-based log odds score was calculated. A positive score signified more frequent occurrences than random. Pseudocounts were added for non-occurring amino acids to avoid the impossible calculation of log(0). The column-wise distance scores of prion sequence residues were derived as follows. Relative frequency (rf) was calculated as the ratio of the frequency of a specific amino acid i in a residue (N_i) to the frequency of any amino acid (N_t).

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

The odds ratio was calculated as the ratio of the relative frequency (rf) to the background frequency of amino acid i. Background frequency was calculated using amino-acid propensities observed from 170 prion sequences.

$$\text{Odds ratio} = \frac{\mathrm{rf}}{\text{background frequency}}$$

PSSM was calculated by applying the logarithm of base 2 to the odds ratio.

$$\text{PSSM score} = \log_2(\text{odds ratio})$$

The calculation of distance scores using BLOSUM62 and PSSM was performed with the JAVA programming language. An amino acid 'z' of the AFM91139.1 sequence had the same score, with a consensus amino acid at the same residue. Thus, it was excluded from the prediction. BLOSUM62 and PSSM scores of collected sequences were stored in a MySQL database.

Accuracy evaluation

Leave-one-out cross-validation estimation was performed to analyse the accuracy of discriminant analyses.[30] The rates of erroneous classification were compared. A single subject member was first omitted from training and the discriminant model was built in leave-one-out cross-validation. The omitted subject was classified after training based on the built model. A total of 170 rounds of class predictions were performed. Then the rate of misclassification was assessed. The k-nearest neighbour discriminant analysis method with k-object numbers 3, 4, 5 and 6 were compared where lower misclassification rate signified better accuracy. The performance of kNN can be determined by calculating the sensitivity, specificity, error rate and total classification accuracy. The sensitivity, specificity, error rate and total classification accuracy are defined as follows:

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

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

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

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

The accuracies of test data sets were presented by the ratio of the number of correctly predicted sequences to the total number of test sequences.

RESULTS

Analysis of polymorphisms in collected PrP sequences

Sequence analyses of 12 mammalian families (*Cervidae*, *Bovidae*, *Canidae*, *Equidae*, *Felidae*, *Hominidae*, *Cercopithecidae*, *Cricetidae*, *Muridae*, *Mustelidae*, *Leporidae* and *Suidae*) were performed. A total of 222 sequences from 23 species were collected. These sequences contained the region encompassing residues 98–227, which incorporates three α-helices and two β-sheets. Table 2 shows the taxonomical information and number (n) of analysed sequences. The type and frequency of polymorphisms were analysed through multiple sequence alignment. *Cervus elaphus scoticus*, *Mesocricetus auratus*, *Mustela putorius furo*, *Neovison vison*, *Oryctolagus cuniculus*, *Pan troglodytes*, *Rattus norvegicus* and *Sus scrofa* showed no polymorphisms with identical amino acids among species on multiple sequence analysis. Polymorphism analysis was performed with residues of 201 sequences from 15 species with the presence of sequence polymorphisms as determined by multiple sequence alignment. Figure 1 shows the frequencies and sites of polymorphisms for each species from consensus sequence. The polymorphisms appeared on different residues, whereas similar polymorphisms could appear as polymorphisms on residue 129 based on analysis. The largest numbers of polymorphic residues (m) were found in *Ovis aries* ($m = 46$), whereas *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 showed comparatively frequent amino-acid substitutions. Prion disease may be transmitted among different species with generally lengthened incubation times, which is also referred to as a 'species barrier'.[31] *Canis lupus familiaris* represents a species barrier to TSE transmission.[32] The lack of significant differences in the number of polymorphisms between TSE-resistant and -susceptible species suggests that the number of polymorphisms and susceptibility of prion diseases are not related.

Table 2 Taxonomy of mammalian species considered in this study

Family	*Genus*	*Species*	*Subspecies*	*Common name*	*Freq*
Cervidae	*Alces*	*Alces*	*Gigas*	Alaskan moose	2
				Red deer	5
	Cervus	*Elaphus*	*Canadensis*	Wapiti	2
			Nelsoni	American elk	3
			Scoticus	Scottish red deer	1
	Odocoileus	*Hemionus*		Mule deer	4
		Virginianus		White-tailed deer	5
Bovidae	*Bos*	*Taurus*		Cattle[a]	15
	Capra	*Hircus*		Goat[a]	23
	Ovis	*Aries*		Sheep[a]	45
Canidae	*Canis*	*Lupus*	*Familiaris*	Dog	6
Equidae	*Equus*	*Caballus*		Horse	10
Felidae	*Felis*	*Catus*		Cat	4
Hominidae	*Homo*	*Sapiens*		Human[a]	16
	Pan	*Troglodytes*		Chimpanzee	1
Cercopithecidae	*Macaca*	*Mulatta*		Rhesus monkey	4
Cricetidae	*Mesocricetus*	*Auratus*		Golden hamster	5
Muridae	*Mus*	*Musculus*		House mouse[a]	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

Freq: the number of sequences of each species that were used to construct the training set for discriminant analysis.
[a]Species for which effects of polymorphisms and mutations were revealed.

Table 1 shows the known mutations and genetic variation in prion sequences that affect susceptibility to mammalian prion diseases. The mutations and polymorphisms were identified from the literature. A total of 170 sequences, excluding those with deletion mutations and amino acids other than the typical 20 types, were used for analysis. Table 3 shows the frequencies and groups of mutations and polymorphisms in the collected sequences. Increases and decreases in susceptibility were considered negative and positive, respectively. No negative polymorphisms were found in *Bos taurus*, whereas only positive polymorphisms were found in *Capra hircus* and *Mus musculus* in the training data set for discriminant analysis.

Susceptibility prediction using k-nearest neighbour discriminant analysis

Discriminant analysis using the k-nearest neighbour method was used to predict prion-disease susceptibility based on sequence and polymorphism data. A total of 42 polymorphisms, including 17 positive polymorphisms and 25 negative polymorphisms, were utilised for the predictions. A data set of seven sequences was generated by mutating reference sequences from *Bos taurus*, *Capra hircus*, *Homo sapiens*, *Mus musculus* and *Ovis aries*, which have no polymorphisms. Four sequences with polymorphisms consisting of two sequences from each of the reference sequences of *Homo sapiens* and *Ovis aries* were generated. A sequence with only negative polymorphisms was generated from the *Bos taurus* reference sequence, whereas two sequences with only positive polymorphisms were generated from reference sequences of *Capra hircus* and *Mus musculus*. Table 4 shows the classification results from k-nearest neighbour discriminant analysis with the distance score calculated through substitution scores based on the BLOSUM62 matrix and PSSM method. Group 1 of the

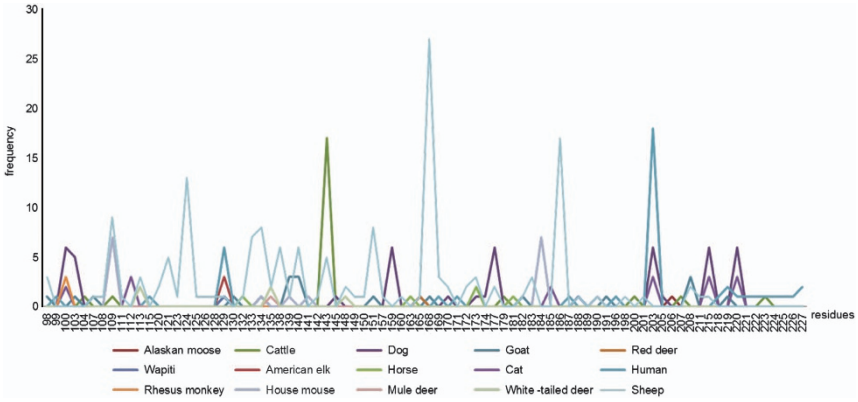

Figure 1 Prion polymorphism information of mammal species. The polymorphisms of 201 sequences from 15 species are shown. These diverse polymorphisms show that there are differences among species.

Table 3 Frequencies of amino-acid polymorphism

Species	*Group*	*Residue*	*Amino acid*	*Frequency*	*Amino acid*	*Frequency*
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

negative polymorphism sequences had higher sensitivity than Group 2 and Group 3. Table 5 shows the total classification accuracies of the k-nearest neighbour discriminant analysis using BLOSUM62 and PSSM matrices. The total classification accuracy peaked when the BLOSUM62 matrix with three and four k-objects were used. The results of k-nearest neighbour discriminant analysis with BLOSUM62 had accuracies of 57.14%, 57.14%, 42.86% and 42.86% in test data sets, respectively. Using k-nearest neighbour discriminant analysis with PSSM, the accuracies were 28.57%, 42.86%, 42.86% and 42.86%, respectively. The low accuracy of negative polymorphism sequences of *Bos taurus*, *Homo sapiens* and of positive polymorphism sequence of *Ovis aries* may have been because of the small training set of 170 sequences, which may have been insufficient to show the effects of prion-sequence polymorphisms on disease susceptibility.

DISCUSSION

Prion sequences of 23 mammalian species were collected and the relevant polymorphisms were examined. Mutant sequences were generated according to previous studies on amino-acid substitutions that affect prion-disease susceptibility and were used to construct a test data set for the discriminant analysis. Polymorphisms of PrP are known to significantly affect susceptibility to prion disease. In the present study, we evaluated the genetic risk for prion disease based on individual polymorphisms in the PrP sequence. The lack of significant differences in the numbers of polymorphisms between PrP sequences associated with high susceptibility and resistance suggests that the number of polymorphisms may not affect susceptibility to TSE. BLOSUM62 and PSSM were constructed to predict susceptibility groups from target sequences, and the accuracies of the classifications with the two matrices were compared. Accurate classifications were observed using the k-nearest neighbour method with three and four k-objects. Group 1 showed relatively accurate classification with higher sensitivity. Incorrect results from the test data set may have been because of the incomplete coverage of possible polymorphisms that affect prion-disease susceptibility. Thus, it is still difficult to accurately predict prion-disease susceptibility based on polymorphism information. However, the significant accuracy of polymorphisms that negatively influence prion-disease susceptibility suggests that this approach could be improved by using more data on prion sequences and polymorphisms. In this study, the results from BLOSUM62 were more precise than PSSM, which differed from other studies.[33] PSSM consisted of pair-wise sequence substitution scores calculated from the ratio of the frequency of amino acids within each column of aligned sequences, with the background frequency leading to more specific matrices for

Table 4 Cross-validation results of discrimination analyses for prediction of prion-disease susceptibility by scoring matrix

	True-positive	*False-positive*	*True-negative*	*False-negative*	*Sensitivity (%)*	*Specificity (%)*	*Error rate (%)*
BLOSUM62							
k=3							
Group 1	23	19	120	8	74.19	86.33	25.81
Group 2	10	9	146	5	66.67	94.19	33.33
Group 3	3	7	156	4	42.86	95.71	57.14
Group 4	96	1	52	21	82.05	98.11	17.95
k=4							
Group 1	25	21	118	6	80.65	84.89	19.35
Group 2	10	9	146	5	66.67	94.19	33.33
Group 3	3	7	156	4	42.86	95.71	57.14
Group 4	94	1	52	23	80.34	98.11	19.66
k=5							
Group 1	24	24	115	7	77.42	82.73	22.58
Group 2	9	10	145	6	60.00	93.55	40.00
Group 3	3	7	156	4	42.86	95.71	57.14
Group 4	92	1	52	25	78.63	98.11	21.37
k=6							
Group 1	28	26	113	3	90.32	81.29	9.68
Group 2	8	5	150	7	53.33	96.77	46.67
Group 3	3	7	156	4	42.86	95.71	57.14
Group 4	92	1	52	25	78.63	98.11	21.37
PSSM							
k=3							
Group 1	25	25	114	6	80.65	82.01	19.35
Group 2	7	9	146	8	46.67	94.19	53.33
Group 3	2	12	151	5	28.57	92.64	71.43
Group 4	87	3	50	30	74.36	94.34	25.64
k=4							
Group 1	28	26	113	3	90.32	81.29	9.68
Group 2	7	4	151	8	46.67	97.42	53.33
Group 3	2	12	151	5	28.57	92.64	71.43
Group 4	89	2	51	28	76.07	96.23	23.93
k=5							
Group 1	27	26	113	4	87.10	81.29	12.90
Group 2	7	6	149	8	46.67	96.13	53.33
Group 3	2	12	151	5	28.57	92.64	71.43
Group 4	89	1	52	28	76.07	98.11	23.93
k=6							
Group 1	27	28	111	4	87.10	79.86	12.90
Group 2	5	7	148	10	33.33	95.48	66.67
Group 3	1	12	151	6	14.29	92.64	85.71
Group 4	89	1	52	28	76.07	98.11	23.93

alignments. The lowest classification accuracy using PSSM may have originated from overoptimisation of the matrices to the specific groups from the limited information of column-wise frequencies of amino-acid types. However, the BLOSUM62 matrix with consensus sequences was not affected by this issue, as it was built for general alignments. Moreover, it is possible that diverse polymorphisms formed irregular profiles, which lost important information from each column. This is partly

Table 5 Total classification accuracy (%) of discrimination analyses

k-nearest	*k* = 3	*k* = 4	*k* = 5	*k* = 6
BLOSUM62	77.65	77.65	75.29	77.06
PSSM	71.18	74.12	73.53	71.76

shown as the smaller range of score values from −3.8 to 6.9 for PSSM compared with −4 to 11 for BLOSUM62. Therefore, the BLOSUM62 matrix may better differentiate between amino acids.

Clustering (an unsupervised learning method) makes data point clusters based on their similarities using Euclidean distance.[34] k-means clustering using the same data set was unsuccessful because of different groups gathered in the same cluster. Nevertheless, BLOSUM62 improved discrimination between the four groups on the two-dimensional plot compared with PSSM when the canonical analysis (which is a type of discriminant analysis) was employed using the same data set.

The possibility of rapid confirmation of prion strains without requiring complex structural and molecular biological analyses suggests the importance of this method. Further research is necessary for the development of a classification method based on other types of polymorphisms. Consideration of the effects of protein secondary structure, hydrophobicity and charge on prion-disease susceptibility may improve this method.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (No. 2012008344). This research was also supported by the National Research Foundation (NRF) funded by the Ministry of Science, ICT and Future Planning (No. 2012M3A9D1054622). Support from Korea Institute of Science and Technology Information (KISTI, K-13-L01-C02-S04) is gratefully acknowledged.

1 Prusiner SB. Prions. *Proc Natl Acad Sci USA* 1998; **95:** 13363–13383.
2 Britton TC, al-Sarraj S, Shaw C, Campbell T, Collinge J. Sporadic Creutzfeldt-Jakob disease in a 16-year-old in the UK. *Lancet* 1995; **346:** 1155.
3 Sakudo A, Ikuta K. Prion protein functions and dysfunction in prion diseases. *Curr Med Chem* 2009; **16:** 380–389.
4 Klamt F, Dal-Pizzol F, Conte da Frota ML Jr, Walz R, Andrades ME, da Silva EG *et al.* Imbalance of antioxidant defense in mice lacking cellular prion protein. *Free Radic Biol Med* 2001; **30:** 1137–1144.
5 Pan KM, Baldwin M, Nguyen J, Gasset M, Serban A, Groth D *et al.* Conversion of alpha-helices into beta-sheets features in the formation of the scrapie prion proteins. *Proc Natl Acad Sci USA* 1993; **90:** 10962–10966.
6 Sakudo A, Xue G, Kawashita N, Ano Y, Takagi T, Shintani H *et al.* Structure of the prion protein and its gene: an analysis using bioinformatics and computer simulation. *Curr Protein Pept Sci* 2010; **11:** 166–179.
7 Colby DW, Prusiner SB. Prions. *Cold Spring Harb Perspect Biol* 2011; **3:** a006833.
8 Wüthrich K, Riek R. Three-dimensional structures of prion proteins. *Adv Protein Chem* 2001; **57:** 55–82.
9 Riek R, Hornemann S, Wider G, Billeter M, Glockshuber R, Wüthrich K. NMR structure of the mouse prion protein domain PrP(121-231). *Nature* 1996; **382:** 180–182.
10 Stewart P, Campbell L, Skogtvedt S, Griffin KA, Arnemo JM, Tryland M *et al.* Genetic predictions of prion disease susceptibility in carnivore species based on variability of the prion gene coding region. *PLoS One* 2012; **7:** e50623.
11 Palmer MS, Dryden AJ, Hughes JT, Collinge J. Homozygous prion protein genotype predisposes to sporadic Creutzfeldt-Jakob disease. *Nature* 1991; **352:** 340–342.
12 Shibuya S, Higuchi J, Shin RW, Tateishi J, Kitamoto T. Codon 219 Lys allele of PRNP is not found in sporadic Creutzfeldt-Jakob disease. *Ann Neurol* 1998; **43:** 826–828.
13 Hainfellner JA, Brantner-Inthaler S, Cervenáková L, Brown P, Kitamoto T, Tateishi J *et al.* The original Gerstmann-Sträussler-Scheinker family of Austria: divergent clinicopathological phenotypes but constant PrP genotype. *Brain Pathol* 1995; **5:** 201–211.
14 Yamada M, Itoh Y, Inaba A, Wada Y, Takashima M, Satoh S *et al.* An inherited prion disease with a PrP P105L mutation: clinicopathologic and PrP heterogeneity. *Neurology* 1999; **53:** 181–188.
15 Parchi P, Capellari S, Chin S, Schwarz HB, Schecter NP, Butts JD *et al.* A subtype of sporadic prion disease mimicking fatal familial insomnia. *Neurology* 1999; **52:** 1757–1763.
16 Mutsukura K, Satoh K, Shirabe S, Tomita I, Fukutome T, Morikawa M *et al.* Familial Creutzfeldt-Jakob disease with a V180I mutation: comparative analysis with pathological findings and diffusion-weighted images. *Dement Geriatr Cogn Disord* 2009; **28:** 550–557.
17 Hsiao K, Meiner Z, Kahana E, Cass C, Kahana I, Avrahami D *et al.* Mutation of the prion protein in Libyan Jews with Creutzfeldt-Jakob disease. *N Engl J Med* 1991; **324:** 1091–1097.
18 Baylis M, Goldmann W. The genetics of scrapie in sheep and goats. *Curr Mol Med* 2004; **4:** 385–396.
19 Sander P, Hamann H, Drögemüller C, Kashkevich K, Schiebel K, Leeb T. Bovine prion protein gene (PRNP) promoter polymorphisms modulate PRNP expression and may be responsible for differences in bovine spongiform encephalopathy susceptibility. *J Biol Chem* 2005; **280:** 37408–37414.
20 Juling K, Schwarzenbacher H, Williams JL, Fries R. A major genetic component of BSE susceptibility. *BMC Biol* 2006; **4:** 33.
21 Heaton MP, Keele JW, Harhay GP, Richt JA, Koohmaraie M, Wheeler TL *et al.* Prevalence of the prion protein gene E211K variant in U.S. cattle. *BMC Vet Res* 2008; **4:** 25.
22 López de la Paz M, de Mori GM, Serrano L, Colombo G. Sequence dependence of amyloid fibril formation: insights from molecular dynamics simulations. *J Mol Biol* 2005; **349:** 583–596.
23 Conchillo-Sole O, de Groot NS, Aviles FX, Vendrell J, Daura X, Ventura S. AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. *BMC Bioinformatics* 2007; **8:** 65.
24 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990; **215:** 403–410.
25 Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of prgressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Res* 1994; **22:** 4673–4680.
26 Hall TA. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl Acids Symp Ser* 1999; **41:** 95–98.
27 Fernandez GCJ. Discriminant Analysis, A Powerful Classification Technique in Data Mining. *Proceeding of the SAS User International conference*. Orlando, FL. SAS Institute: Cary, NC. 2002, pp 247–250.
28 Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* 1992; **89:** 10915–10919.
29 Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997; **25:** 3389–3402.
30 Lachenbruch PA, Mickey MA. Estimation of error rates in discriminant analysis. *Technometrics* 1968; **10:** 1–11.

31 Sweeting B, Khan MQ, Chakrabartty A, Pai EF. Structural factors underlying the species barrier and susceptibility to infection in prion disease. *Biochem Cell Biol* 2010; **88:** 195–202.
32 Kirkwood JK, Cunningham AA. Epidemiological observations on spongiform encephalopathies in captive wild animals in the British Isles. *Vet Rec* 1994; **135:** 296–303.
33 Ou YY, Chen SA, Wu SC. ETMB-RBF: discrimination of metal-binding sites in electron transporters based on RBF networks with PSSM profiles and significant amino acid pairs. *PLoS One* 2013; **8:** e46572.
34 Ubeyli ED, Doğdu E. Automatic detection of erythemato-squamous diseases using k-means clustering. *J Med Sys* 2010; **34:** 179–184.
35 Westaway D, Goodman PA, Mirenda CA, McKinley MP, Carlson GA, Prusiner SB. Distinct prion proteins in short and long scrapie incubation period mice. *Cell* 1987; **51:** 651–662.
36 Saunders GC, Lantier I, Cawthraw S, Berthon P, Moore SJ, Arnold ME *et al.* Protective effect of the T112 PrP variant in sheep challenged with bovine spongiform encephalopathy. *J Gen Virol* 2009; **90:** 2569–2574.
37 Mead S, Whitfield J, Poulter M, Shah P, Uphill J, Campbell T *et al.* A novel protective prion protein variant that colocalizes with kuru exposure. *N Engl J Med* 2009; **361:** 2056–2065.
38 Vaccari G, D'Agostino C, Nonno R, Rosone F, Conte M, Di Bari MA *et al.* Prion protein alleles showing a protective effect on the susceptibility of sheep to scrapie and bovine spongiform encephalopathy. *J Virol* 2007; **81:** 7306–7309.
39 Saunders GC, Cawthraw S, Mountjoy SJ, Hope J, Windl O. PrP genotypes of atypical scrapie cases in Great Britain. *J Gen Virol* 2006; **87:** 3141–3149.
40 Appleby BS, Appleby KK, Hall RC, Wallin MT. D178N, 129Val and N171S, 129Val genotype in a family with Creutzfeldt-Jakob disease. *Dement Geriatr Cogn Disord* 2010; **30:** 424–431.
41 Chasseigneaux S, Haïk S, Laffont-Proust I, De Marco O, Lenne M, Brandel JP *et al.* V180I mutation of the prion protein gene associated with atypical PrPSc glycosylation. *Neurosci Lett* 2006; **408:** 165–169.
42 Grasbon-Frodl E, Lorenz H, Mann U, Nitsch RM, Windl O, Kretzschmar HA. Loss of glycosylation associated with the T183A mutation in human prion disease. *Acta Neuropathol* 2004; **108:** 476–484.
43 Peoc'h K, Manivet P, Beaudry P, Attane F, Besson G, Hannequin D *et al.* 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* 2000; **15:** 482.
44 Capellari S, Cardone F, Notari S, Schininà ME, Maras B, Sità D *et al.* Creutzfeldt-Jakob disease associated with the R208H mutation in the prion protein gene. *Neurology* 2005; **64:** 905–907.
45 Biljan I, Ilc G, Giachin G, Raspadori A, Zhukov I, Plavec J *et al.* Toward the molecular basis of inherited prion diseases: NMR structure of the human prion protein with V210I mutation. *J Mol Biol* 2011; **412:** 660–673.
46 Shiga Y, Satoh K, Kitamoto T, Kanno S, Nakashima I, Sato S *et al.* Two different clinical phenotypes of Creutzfeldt-Jakob disease with a M232R substitution. *J Neurol* 2007; **254:** 1509–1517.
47 Piccardo P, Liepnieks JJ, William A, Dlouhy SR, Farlow MR, Young K *et al.* Prion proteins with different conformations accumulate in Gerstmann-Sträussler-Scheinker disease caused by A117V and F198S mutations. *Am J Pathol* 2001; **158:** 2201–2207.
48 Panegyres PK, Toufexis K, Kakulas BA, Cernevakova L, Brown P, Ghetti B *et al.* A new PRNP mutation (G131V) associated with Gerstmann-Sträussler-Scheinker disease. *Arch Neurol* 2001; **58:** 1899–1902.
49 Piccardo P, Dlouhy SR, Lievens PM, Young K, Bird TD, Nochlin D *et al.* Phenotypic variability of Gerstmann-Sträussler-Scheinker disease is associated with prion protein heterogeneity. *J Neuropathol Exp Neurol* 1998; **57:** 979–988.
50 Kaneko K, Zulianello L, Scott M, Cooper CM, Wallace AC, James TL *et al.* Evidence for protein X binding to a discontinuous epitope on the cellular prion protein during scrapie prion propagation. *Proc Natl Acad Sci USA* 1997; **94:** 10069–10074.
51 Goldmann W, Martin T, Foster J, Hughes S, Smith G, Hughes K *et al.* Novel polymorphisms in the caprine PrP gene: a codon 142 mutation associated with scrapie incubation period. *J Gen Virol* 1996; **77:** 2885–2891.
52 Vaccari G, Di Bari MA, Morelli L, Nonno R, Chiappini B, Antonucci G *et al.* Identification of an allelic variant of the goat PrP gene associated with resistance to scrapie. *J Gen Virol* 2006; **87:** 1395–1402.