PrionX - Workshop Final Report

Maggi Plutov (209223155), Shani Daniel (315231902), Eyal Blyachman (208273672), Almog Boaron (313119265)

Abstract

In this study, we explored the prion phenomenon utilizing convolutional neural networks (CNNs) to predict prion proteins with higher accuracy. Utilizing data from PrionHome Database (Harbi, 2012 [1]) and employing negative sets composed of disordered ([2]) and globular proteins ([3]), we developed two models: a basic neural network model and a CNN model. Our comparative analysis revealed that the CNN model significantly outperformed its basic equivalent, identifying the abundance of Asparagine and Glutamine as the critical feature for prion classification. Despite its superior performance in initial tests, the CNN model demonstrated limitations in accurately predicting prions within test genomes when compared to standard tools from previous research, such as PrionScan. Furthering our investigation, we applied the CNN model to the human genome to identify known prion-forming proteins, providing valuable insights and affirming the model's potential utility in genomic studies. This research underscores the potential of CNNs in prion protein classification while highlighting areas for further improvement and investigation.

Introduction

Motivation

Prions, mysterious and often misunderstood proteins, have captured both scientific and popular interest due to their unique properties and the deadly neurodegenerative diseases they are associated with. Unlike typical proteins, prions can misfold and induce other proteins to misfold similarly, leading to a cascade of detrimental effects within the brain after reaching a critical threshold. This phenomenon, associated with diseases such as Alzheimer's, Parkinson's, Creutzfeldt-Jakob, and Lou Gehrig's, among others, underscores the need for deeper understanding and innovative approaches to prion research. This highlights the urgency of advancing our methods for their study and classification. Furthermore, the expanding universe of prion and prion-like phenomena (Harbi, 2014, [5]) poses significant challenges in classification and prediction, necessitating a comprehensive understanding and experimental validation.

Problem Formulation

Prion proteins, primarily existing as an isoform of the major prion protein (PrP^C), play crucial roles in cellular communication and the preservation of myelin. Their physiological roles, however, are overshadowed by their potential to misfold into a pathological form (PrP^{Sc}) under certain conditions (Colby, 2011, [11]). This misfolding acts as a catalyst, triggering a domino effect that leads to the misfolding of other proteins.

The term "prion" spans a range of meanings, including "prion-like" and "prion-related" proteins, each reflecting different level of association with the core prion phenomena. Here we concentrated on prion-forming proteins –proteins that has the potential to misfold in a manner that leads to neurodegenerative

diseases, like prions.
These prion-forming proteins, which bear resemblance to amyloid fibrils, underscore a deep link to a wider spectrum of protein misfolding diseases, while identifying these proteins is crucial for early detection, understanding disease mechanisms, and developing targeted therapies.
The prediction of which proteins may adopt prionogenic properties is a significant challenge, intensified by the complexity of protein folding, the absence of universal sequence motifs, and the context-dependent nature of prion formation.

The complex nature of prions, characterized by their rich Asparagine (N) and Glutamine (Q) domains, plays a crucial role in their structural conversion and propagation.
The Q/N rich regions in prion proteins are crucial due to their structural and biochemical properties, significantly influencing the misfolding and aggregation associated with prion diseases. These regions are prone to forming stable β-sheet structures through hydrogen bonding, differing from the α-helical structure in normal prion proteins (PrP^C). The flexibility and bonding capabilities of these Q/N rich regions not only facilitates the misfolding process but also promotes the aggregation of prion proteins into β-sheet-rich amyloid fibrils, indicative of prion diseases (Colby, 2011, [11]).
Thus, understanding Q/N rich regions is key to advancing our knowledge of prion diseases and investigating novel treatment options.

This project aims to leverage convolutional neural networks (CNNs) to address the described challenges. By focusing on the distinctive properties of prions, such as their N and Q domain richness and the disordered composition of these domains, we seek to:

- Distinguish prions from non-prions with greater accuracy.
- Compare the effectiveness of different negative datasets in improving classification performance.
- Enhance our understanding of prion characteristics and their implications for neurodegenerative diseases.
- Identify new prion-forming proteins within the human genome, thereby contributing to the early detection and potential treatment of related diseases.

Methods

Data

Our Data was composed from 3 different datasets:

1. Dataset of prions from PrionHome Database (Harbi, 2012 [1]).
2. Dataset of disordered proteins (disport [2]).
3. Dataset of globular proteins (cathdb [3]).

Preprocessing

Our preprocessing pipeline followed the following steps:

1) **Data filtering by protein size**. Proteins with fewer than 40 amino acids were excluded. The primary aim was to remove proteins that did not possess enough amino acids required to form a globular hydrophobic core, which is crucial for prion behavior.
2) **Removal of faulty data and duplicates.** Random faulty data instances were removed from the dataset to minimize noise and inaccuracies. Additionally, duplicate sequences were identified and removed to ensure data integrity and prevent redundancy in the training set.
3) **Data clustering for grouping homologs:**
    a. We used MMSEQ2 ([4]) to cluster sequences with a 70% identity threshold, forming clusters of sequences with significant sequence similarity. This clustering approach ensured that sequences sharing common characteristics were grouped together, facilitating subsequent data partitioning, and preventing data leakage.
    b. A graph was constructed, where each node represented a sequence, and edges were established based on the clusters formed by mmseq2.
    c. The networkX (Hagberg, 2008, [12]) connected component's function was employed to group sequences into connected components based on the graph representation. This step ensured that sequences within the same cluster remained grouped together in the same set at training and testing, maintaining the structural integrity of the data and preventing information leakage between different subsets.
4) **Data Partition:**
    The dataset was partitioned into three subsets: 60% for training, 20% for validation, and 20% for testing. Importantly, as we mentioned, the integrity of sequence clusters was preserved during partitioning to prevent bias and ensure representative subsets for model training and evaluation.

Models

In our quest to develop an effective neural network for prion classification, we designed a sequential model capable of handling variable-length sequences and extracting relevant features. We added the following layers to the model using Keras ([13]):

1. **Masking Layer**: The masking layer identifies and masks padded elements (tokens) in the input sequences, indicating to subsequent layers that these tokens should be ignored during computation. This ensures that sequences of varying lengths are processed uniformly, maintaining the integrity of the input data, and facilitating effective feature extraction.
   We padded the sequences with zeros to ensure they were of uniform size and masked the value of 0 to handle the sequences length variability.
2. **Average Pooling (1D) Layer**: The average pooling layer applies a sliding window mechanism to the input sequences, computing the average occurrence of each amino acid within the window. By aggregating information across neighboring amino acids, this layer captures local sequence patterns and generates feature representations that are robust to variations in sequence length.
   In our models, we employed a window size of 30, as it represents enough amino acids to be considered a domain.
3. **Dense Layer with Sigmoid Activation**: The dense layer receives the feature representations generated by the previous layers and applies a sigmoid activation function to produce probability scores for binary classification. The sigmoid function maps the input values to a range between 0 and 1, representing the likelihood of each window belonging to the positive class (e.g., prion-like) or negative class.
4. **Max Pooling (1D) Layer**: The max pooling layer takes the maximum probability score across all windows, effectively summarizing the information extracted from the entire sequence. By selecting the maximum score, the layer identifies the most salient features indicative of prion-like behavior, enabling robust classification. This layer serves as the final step in the model's decision-making process, producing a single prediction for each input sequence.

After attempting the NN model with these layers, we added a convolution layer to analyze how it impacts performance.

5. **Convolutional Layer**: a convolutional layer is introduced after the masking layer to perform feature extraction through convolutional operations. This layer applies a set of learnable filters to the input sequences, convolving them with sliding windows to detect spatial patterns and structural motifs. By incorporating convolution, the model gains the ability to capture higher-order sequence features and exploit spatial dependencies, potentially improving classification performance.
   In our CNN models we used 256 filters, a kernel size of 25, and the kernel regularization l1 with strength 0.001, to control the amount of non-zero weights in the kernel.
   These hyper parameters were optimized to minimize validation loss.
   Additionally, we utilized the "relu" activation function.

Both models (the NN and the CNN) were trained using the Adam optimizer with a learning rate of 0.001. Adam optimizer was chosen for its effectiveness in training deep neural networks and handling sparse gradients. The learning rate of 0.001 was empirically determined to balance training speed and convergence stability.

The binary cross entropy loss function was utilized to measure classification errors. Early stopping with a high number of epochs was employed to monitor validation loss and halt training when no improvement was observed, preventing overfitting, and ensuring optimal model generalization.

This approach yielded four trained models:

1. Basic NN model, with only disordered proteins as a negative set.
2. Basic NN model, with disordered and globular proteins as a negative set.
3. CNN model, with only disordered proteins as a negative set.
4. CNN model, with disordered and globular proteins as a negative set.

Results

Models Metrics

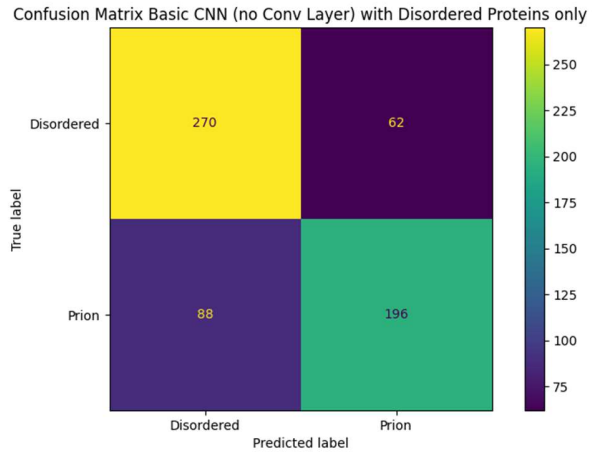
The accuracies and losses of each model regarding the test are summarized in Table 1:

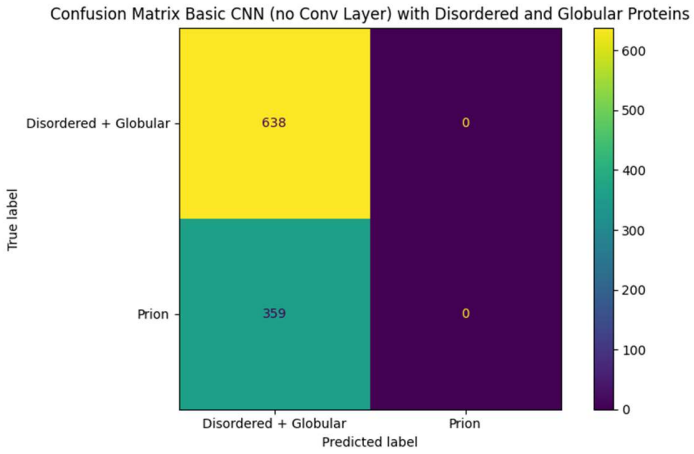
Model	Accuracy	Loss (BCE)
Basic NN model (without conv layer), only disordered proteins in negative set	0.786033	0.641325
Basic NN model (without conv layer), globular & disordered proteins in negative set	0.720701	0.560538
CNN, only disordered proteins in negative set	0.832095	0.459483
CNN, globular & disordered proteins in negative set	0.851041	0.434653

Table 1: Accuracy and loss of all models

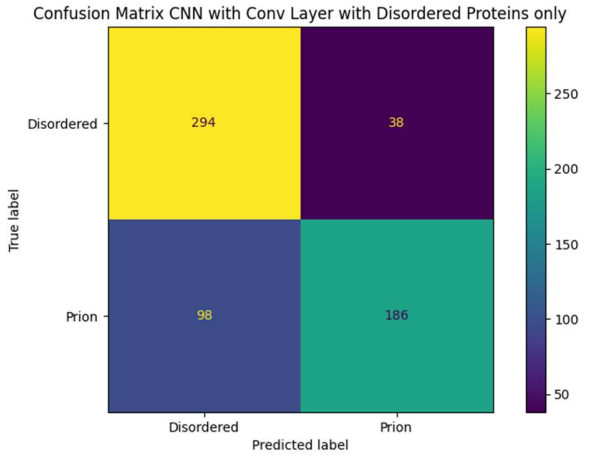
It is evident that the CNN models, incorporating convolution layers, generally outperformed their basic alternatives across these metrics. They exhibit higher accuracies and lower losses.
Additionally, we calculated the confusion matrix for each model, along with their true positive rates (TPR), false positive rates (FPR), true negative rates (TNR) and false negative rates (FNR).

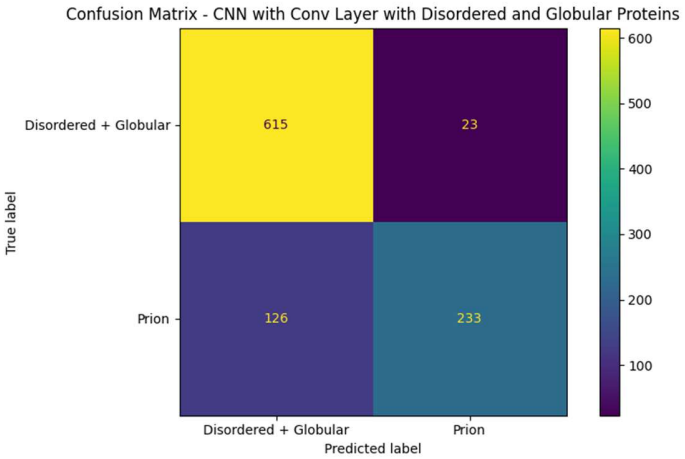
Basic Neural Network models





CNN models





	Basic NN only disordered	Basic NN disordered and globular	CNN only disordered	CNN disordered and globular
TPR	0.690	0.0	0.655	0.649
FPR	0.187	0.0	0.114	0.036
TNR	0.813	1.0	0.886	0.964
FNR	0.310	1.0	0.345	0.351

Table 2: Rates calculated for each model

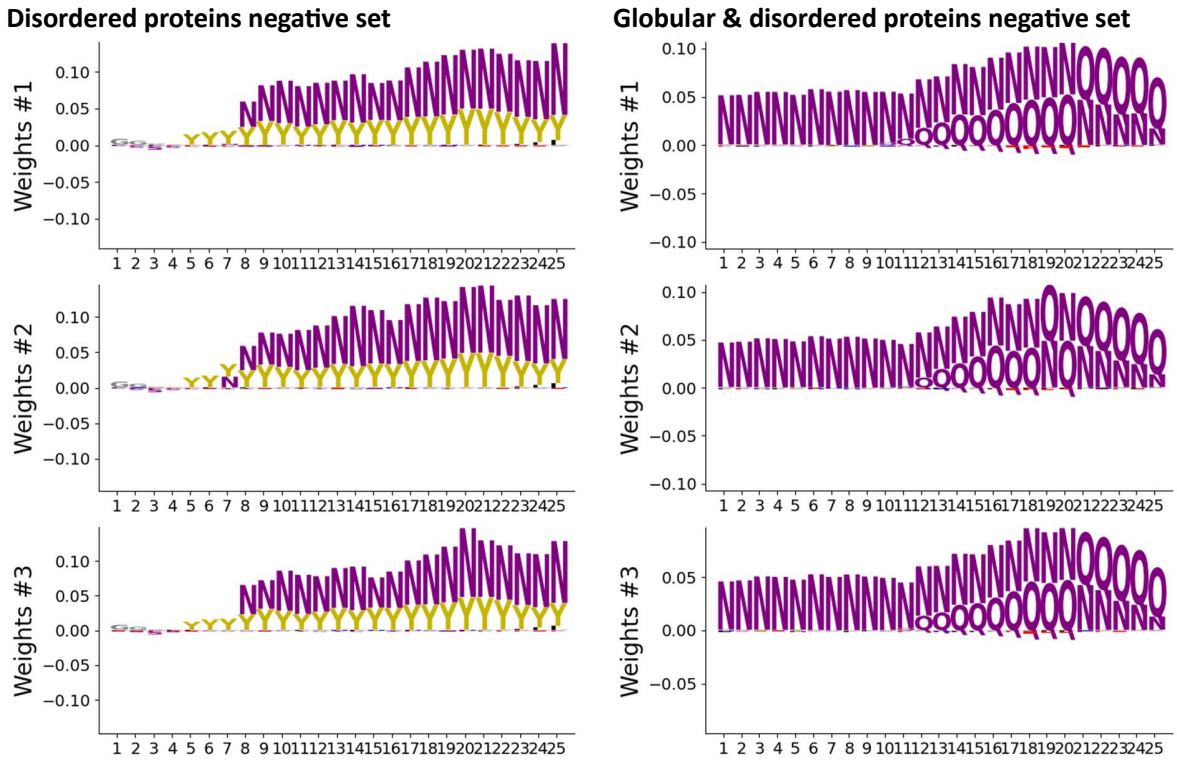
The confusion matrices indicate that the CNN models performed better at detecting negative samples, meaning non-prion proteins, while they also demonstrated effective detection of prions. The CNN model trained on disordered and globular proteins detected negative samples more accurately than the CNN trained only on disordered proteins, but less at detecting positive samples.

We observed unexpected results with the basic neural network model that included both disordered and globular proteins in the negative set. Surprisingly, this model classified all samples in the test set as non-prions, indicating very poor performance in practice. However, incorporating a convolutional layer significantly enhanced the accuracy and performance of this model.

When considering only disordered proteins in the negative set using the basic NN model, we observed a reasonable detection of prions (TPR of 0.69). However, this model still achieved a lower True Negative Rate (TNR) compared to the CNN models.

Amino Acids Weights Visualization

Furthermore, we aimed to visualize the importance of each amino acid in the CNN models (using PGM visualization library ([8])). The weight assigned to each amino acid corresponds to its significance in the classification process. In the visualization, the x-axis denotes the kernel size (utilizing a kernel size of 25), with each feature (filter) represented by a separate plot. The 256 features (filters) were arranged based on weight, and the following show the top three features with the highest weights:



Initially, it's evident that the plots of the top three features for each negative set exhibit similarities. This observation suggests that the significance of amino acid weights is not a random phenomenon. Significantly, asparagine (N) stands out as a key factor in prion classification, evident from its notable presence in both negative datasets.
This finding aligns with existing literature linking prions to asparagine (N) and glutamine (Q) rich domains. Moreover, adding globular proteins in the negative dataset, glutamine (Q) emerges as a notably significant amino acid, as anticipated.

The significance of tyrosine (a hydrophobic amino acid) is highlighted when considering only disordered proteins, as hydrophobic domains may be absent in such proteins due to their disordered nature and lack of a hydrophobic core. This implies that regions rich in hydrophobic amino acids may serve as an additional discriminative factor between disordered proteins and prions.

In summary, these visualizations are largely consistent with existing literature on prion sequences, providing valuable validation for our models.

Classifying The Human Genome

We aimed to assess the predictive performance of our models in identifying prions within the human genome (2024 [7]). For simplicity, we limited our focus to genes with 500 amino acids or less. This approach yielded a dataset of 12,281 genes sequences on which we applied our models.

We focused on known proteins associated with prions in the human genome and assessed their probability rankings as potential prions according to our models. We excluded these proteins, if present, from the original datasets to prevent data leakage and bias.

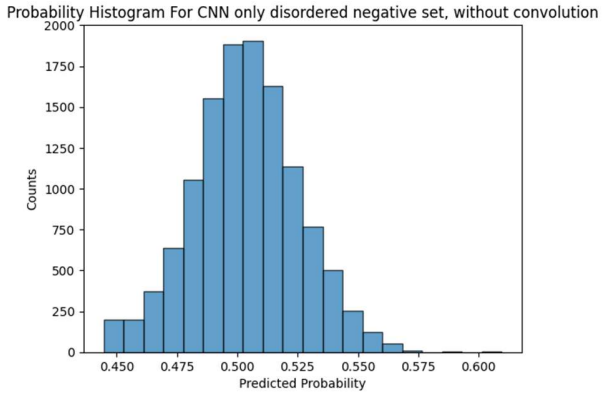
The proteins we chose to focus on are:

1) Human protein 1 - PRIO_HUMAN Major prion protein. This protein is known as the main prion protein that has a direct effect on CJD (Creutzfeldt-Jakob disease). CJD is caused by variants affecting the gene PRIO_HUMAN. Thus, we would expect it to rank lower, meaning likely to be a prion.
2) Human protein 2 - PRND_HUMAN Prion-like protein doppel. This protein exhibits certain properties associated with prions, and CJD is also associated with mutations in this protein.
3) Human protein 3 - SPRN_HUMAN Shadow of prion protein. Prion-like protein that has PrP(C)-like neuroprotective activity. May act as a modulator for the biological actions of normal and abnormal PrP (By similarity). This implies that this protein actively regulates prion activity.
4) Human protein 4 - APRIO_HUMAN Alternative prion protein. This proteins' sequence has a repeated region. Mutations in this gene have been associated with several diseases, such as Creutzfeldt-Jakob disease, Huntington disease-like 1, kuru, etc.

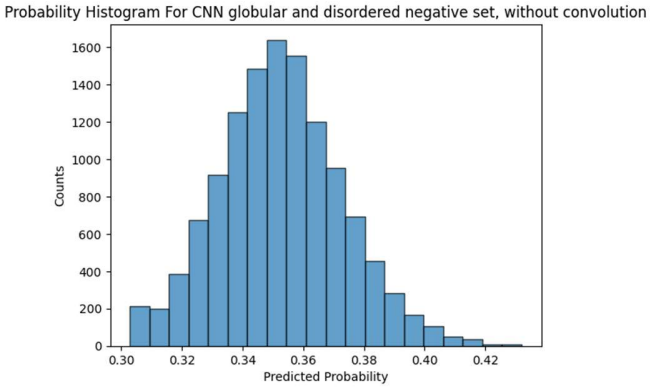
5) Human protein 5 - PRNT_HUMAN Putative testis-specific prion protein. This protein appears to share structural similarities with prions, yet it apparently lacks prion activity.

The following graphs illustrate the probability distributions of being a prion for each of the four models on the human genome. We present the rankings of the five proteins mentioned earlier, where higher probabilities correspond to lower ranks.

Basic NN models

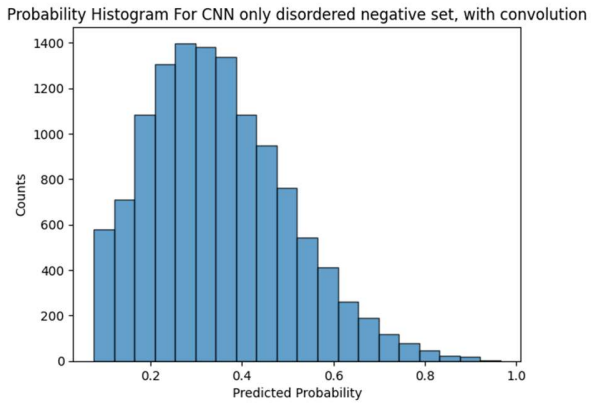


rank for “human protein 1” - 665
rank for “human protein 2” – 5,203
rank for “human protein 3” – 12,052
rank for “human protein 4” – 12,041
rank for “human protein 5” – 1,462

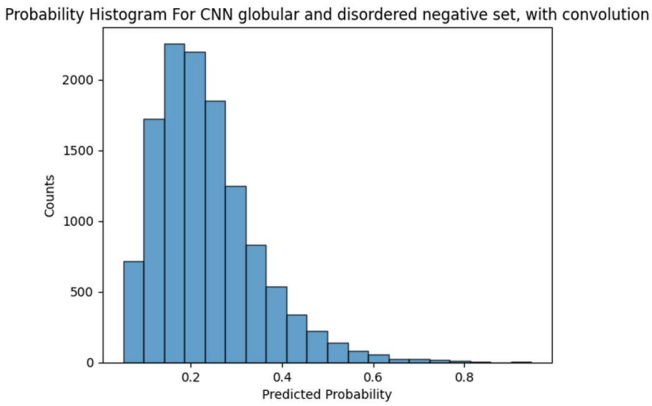


rank for “human protein 1” – 8,993
rank for “human protein 2” – 3,277
rank for “human protein 3” – 9,279
rank for “human protein 4” – 4,533
rank for “human protein 5” – 4,103

CNN models



rank for “human protein 1” - 12
rank for “human protein 2” - 630
rank for “human protein 3” – 11,114
rank for “human protein 4” – 12,265
rank for “human protein 5” – 11,387



rank for “human protein 1” - 350
rank for “human protein 2” - 556
rank for “human protein 3” – 11,984
rank for “human protein 4” – 12,108
rank for “human protein 5” – 11,477

Basic NN models

The basic NN models exhibit notably poor performance.
The probability distribution associated with the negative set consisting of disordered proteins tend to mean around 0.5. This is not ideal, considering that prion proteins in the human genome are expected to be relatively rare.
Interestingly, "human protein 1" received the lowest rank among the five proteins in all the models, as anticipated. However, its rank is not significantly low, and the ranks of the remaining four proteins appear mostly arbitrary.

With globular and disordered proteins in the negative set, we see again that the basic model predicts overall low probabilities. All the predicted probabilities are below 0.5, meaning this model generally hasn't predicted any prions in the human genome. We also observe that the rankings for all five proteins mentioned appear to be visibly random, which further supports the conclusion that this model is not very effective.

CNN models

The CNN models show improved performance. The probabilities are concentrated around 0.2 to 0.3 (equivalent to 20% to 30%) with a right tail that nearly reaches 1, which is more plausible.
Regarding the model where only disordered proteins are included in the negative set, human protein 1 received an exceptionally low rank of 12. Following this, human protein 2 received the next lowest rank, aligning with expectations. The remaining three proteins obtained high ranks, indicating lower probabilities of being prions. Considering our existing knowledge on them as described previously, this can be reasonable.
The CNN with globular proteins in the negative set shows similar results.

Discussion

Important Features and Motifs

Our analysis revealed that the CNN model successfully captured essential features indicative of prion-like behavior, such as regions enriched in asparagine (N) and glutamine (Q) residues. These findings are consistent with previous studies in the literature, which validate our model's predictions ([1]).

Comparing Models Performance

In this study, we aimed to develop a convolutional neural network (CNN) model for the detection of prions. We evaluated our two models on both Human and Arabidopsis genomes and compared their performance with other algorithms results reported in the literature (Marcos 2021 [6]). It's important to note that while our comparison may not be entirely equivalent due to genome simplicity (filtering part of the genome beforehand) we believe it provides valuable insights.
We observed different results between our two CNN models. Specifically, the CNN trained on both

disordered and globular proteins demonstrated better results in prion prediction on Human and Arabidopsis genomes. It predicted approximately 3% of prions in the human genome and 6.7% in Arabidopsis genome compared to the second CNN which got 16.6% and 25.1% respectively.
Prior models achieved predictions of 1% or less, with PrionScan demonstrating the most promising results, at 0.29% for human data and 0.09% for Arabidopsis. In conclusion, employing CNNs for predicting prions in genomes requires further refinement to achieve performance comparable to established benchmark tools.

Organism	Proteome enrichment (%) by PLAAC, PAPA, LPS	Proteome enrichment (%) by PrionScan	Proteome enrichment (%) by CNN trained on disordered	Proteome enrichment (%) by CNN trained on disordered and globular
H. Sapiens	1	0.29	16.6	3.0
A. Thaliana	1	0.09	25.1	6.7

Table 3: Percent of human genome classified as prions by different tools

We found that the CNN model which incorporated both disordered and globular proteins in the negative set, exhibited a lower prediction rate for prions, indicating higher accuracy. This trend is illustrated in the graph below, where the cutoff value of 0.5 represents a prediction of a prion.

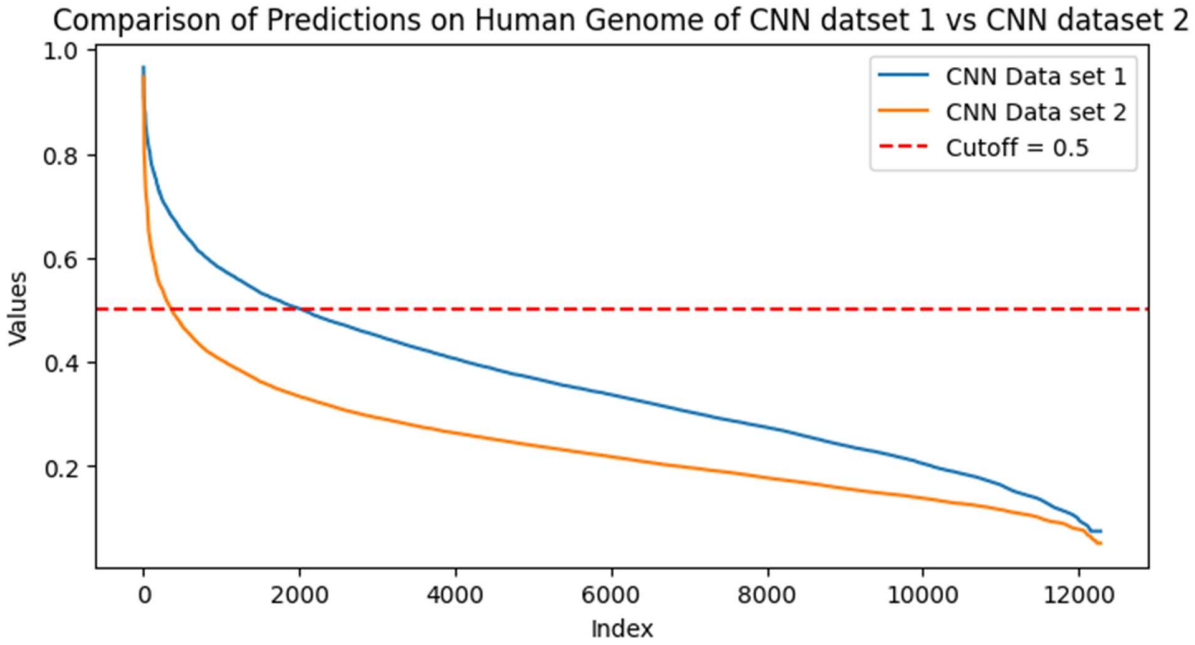


Figure 1: The amount of proteins classified as prions is lower using CNN with globular proteins (orange curve)

The comparison between the two models revealed both differences and similarities in their predictions for new sequences. While some sequences received high scores from both models, indicating some confidence in their classification as prions related, others displayed variations, receiving low scores from one model but high scores from the other.

Protein seq	name	CNN disordered predictions	CNN disordered and globular predictions
MKFTIVFAGLLGVFLAPALANYNINV NDDNNNAGSGQQSVSVNNEH…	Gastrokine-1	0.965796	0.920569
MPAENSPAPAYKVSSHGGDSGLD GLGGPGVQLGSPDKKKRKANTQG…	Pygopus homolog 1	0.939643	0.903039
MNWHMIISGLIVVVLKVVGMT LFLLYFPQIFNKSNDGFTTTRSYGT…	C-type lectin domain family 5 member	0.910259	0.790942
MDQNNSLPPYAQGLASPQGAM TPGIPIFSPMMPYGTGLTPQPIQNT…	TATA-box-binding protein	0.289791	0.947320
MQQQQQQQQQQQQQQQQQQQ QQQQQQQQQQQQQQQQQQQQ…	Ataxin-8	0.075199	0.936216

Table 4: New Predictions by the CNN models, the first 3 rows show high predictions by both models, the last 2 rows show the difference between predictions.

While most of the proteins above appear unrelated to prions, two stand out.
Gastrokine, known as an anti-amyloidogenic protein, plays a role in preventing the formation of amyloid fibers (Overstreet, 2021, [9]). Given the similarity between amyloid aggregation and prion phenomena, it raises the possibility of viewing Gastrokine as anti-prion and suggesting potential parallels in their mechanisms.
Ataxin-8 is a protein primarily consisting of polyglutamine expansion, which is implicated in neurodegenerative disorders (NCBI [10]).

Conclusions

In conclusion, our study demonstrates the feasibility of utilizing a basic CNN architecture for prion detection. While our model achieved relatively high accuracy and successfully identified key features associated with prion-like regions, further refinement and optimization are warranted to enhance its performance. Future research efforts should focus on incorporating additional genomic features, exploring complex CNNs, and integrating structural data to develop more robust and accurate prion detection models.

Bibliography

1. PrionHome: A Database of Prions and Other Sequences Relevant to Prion Phenomena
   https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0031785
   * Parsed prions data – prion_db.csv
2. Disport.fasta
   https://disprot.org/api/search?release=2022_06&show_ambiguous=true&show_obsolete=false&format=fasta&namespace=all&get_consensus=false
3. Cath.fasta
   http://download.cathdb.info/cath/releases/all-releases/v4_3_0/sequence-data/cath-domain-seqs-S35-v4_3_0.fa
4. https://github.com/soedinglab/mmseqs2
5. Classifying prion and prion-like phenomena
   https://www.tandfonline.com/doi/full/10.4161/pri.27960
6. Prion-like proteins: from computational approaches to proteome-wide analysis
   https://febs.onlinelibrary.wiley.com/doi/full/10.1002/2211-5463.13213
7. Human genome from Uniport
   https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/reference_proteomes/Eukaryota/UP000005640/UP000005640_9606.fasta.gz
8. Probabilistic Graphical Models (PGM)
   https://github.com/jertubiana/PGM
9. Gastrokine-1, an anti-amyloidogenic protein secreted by the stomach, regulates diet-induced obesity
   https://www.nature.com/articles/s41598-021-88928-8
10. ATXN8 ataxin 8
    https://www.ncbi.nlm.nih.gov/gene/724066
11. Prions
    https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3003464/
12. NetwrokX
    https://networkx.org/documentation/stable/index.html#citing
13. Keras
    https://keras.io/