Evaluation of Models to Predict Parkinson's Disease Progression from Peptide Abundance Levels

Maria Stephenson
msteph11@student.ubc.ca

Avery Quan
aquan01@student.ubc.ca

Partheeban Bharati
bparthee@student.ubc.ca

Abstract

Heterogeneity in Parkinson's disease (PD) progression is poorly understood, negatively impacting patient outcomes and clinical trial design. Recent research indicates that abnormalities in the protein and peptide abundances are key drivers of disease progression, and therefore could have use in predicting progression. In this study, we developed several machine learning models to predict PD progression from protein and peptide abundance values attained from mass spectrometry readings of cerebrospinal fluid samples from several hundred patients. Boosted trees and lasso regression models performed best, achieving test SMAPE+1 scores of 57.6 and 61.5 respectively. APLP2 and CHL1 protein and peptide abundances were amongst those with greatest influence over model predictions, and were more strongly correlated with PD progression than other proteins, indicating that they could being useful in the generation of future prediction models.

1 Introduction

Parkinson's disease (PD) is a degenerative movement disorder originating in the central nervous system. There is currently no cure for it, although treatments exist which can improve symptoms. Due to populations around the world growing older, the disease burden is expected to increase in the coming decades.

The motor symptoms of PD are caused by the death of dopamine-producing neurons in the part of the brain called the substantia nigra (Triarhou (2013)). Dopamine is a neurotransmitter that, among other things, is involved in motor control. The dead neuronal cells have been found to contain Lewy bodies (abnormal protein aggregations) of the α-synuclein protein (Karayel O (2022)). Dysregulated expression of other proteins has been reported in PD as well. In mouse models of PD, a total of 518 proteins were found to be dysregulated in different areas of the brain (Zhang X (2010)). In cerebral spinal fluid (CSF) samples of PD patients, several proteins were found to be differentially present in PD compared to healthy controls. Proteins upregulated in PD were enriched for lysosomal-related terms, supporting the theory that α-synuclein in PD is caused by lysosomal dysregulation (Karayel O (2022)). Additionally, abundances of several proteins, such as CHST6, MIF and APLP1 were found to correlate with Unified Parkinson's Disease Rating Scale (UPDRS) scores (Karayel O (2022)). Together this shows that abnormal protein abundances may play a key role in PD development and severity.

PD is a heterogeneous disease, with patients not only exhibiting variation in symptoms, but also in the rate of disease progression (Greenland JC (2019)), complicating patient treatment and clinical trial design. Previous studies utilizing longitudinal data sets containing clinical data and abundance levels for certain proteins (ex. chemokines), have attempted to create machine learning models capable of predicting PD progression (Ahmadi Rastegar D (2019) and Dadu A (2022)). While resulting models hold promise, they are far from perfect, and currently there are no established reliable biomarkers or models for predicting PD progression. As stated previously, the expression of many proteins appear to be dysregulated in PD, thus use of proteomics data sets generated from techniques, such

Course project for UBC CPSC 440/540, April 2023.

as mass spectrometry, that attempt to quantify abundances of all proteins in a sample, could allow for the identification of promising new PD progression biomarkers. Previous studies using mass spectrometry data were not longitudinal however, preventing such analysis.

This study utilizes a new data set containing mass spectrometry data of cerebrospinal fluid samples in Parkinson's patients, taken over the course of several months. Various models were trained on this data set to predict UPDRS scores for current and future visits. The purpose of this study was to identify models with the best performance for this prediction task, as they could be useful in the development of future models. Additionally, we attempted to identify specific proteins and peptides that could aid in predicting UPDRS scores. Most of the proteins and peptides in the dataset did not appear to have predictive value. Protein and peptide abundances of APLP1 and CHLI1 were among those with the most influence over model predictions, and were two of the proteins that were most correlated with UPDRS scores. Both related to neuronal function, they may also be relevant to PD, and could be useful in the generation of future models for predicting PD progression.

2 Related Work

1. **Prediction with serum cytokines**: As immune dysfunction has been implicated in PD, researchers have attempted to predict PD progression using serum cytokine levels (Ahmadi Rastegar D (2019)). This was done by training two different models on patient baseline serum abundance levels of 27 different cytokines and evaluating their accuracy at predicting 2-year follow up disease severity scores. While related to our study, as both attempt to predict PD progression from protein levels, their analysis focused only on cytokine expression, limiting their ability to identify new biomarkers. In contrast, the data set we are working with contains over 200 unique proteins, giving us greater ability to identify proteins with predictive power, as we have more to examine.
2. **Stratification into progression risk groups**: Other work has focused on stratifying patients into risk groups based on progression rate, and developing models to predict the risk group a patient belongs to (Dadu A (2022)). While models perform well at this task (AUC $\geq$ 88), there is still variation in progression rate within each group, and accurate prediction of individual disease severity scores would be more informative. In our study we train models to predict individual scores, providing the research community with information regarding the models that show more potential than others for this task.
3. **Proteomics studies**: Previous studies have compared protein abundance levels in PD patients versus healthy controls, and associated abundance levels with disease severity scores (Karayel O (2022)). However, the data sets used in these studies were not longitudinal, so they could not relate protein abundances to disease progression, only to disease severity at the time of sample collection. In contrast our study uses a longitudinal data set, allowing us to address this research problem.

3 Materials and Methods

3.1 Data set

The data used in this project was obtained from the Kaggle challenge: AMP®-Parkinson's Disease Progression Prediction. The training data is publicly available but the test data is currently hidden.

The training data consists of 3 tables. The first contains clinical records for 248 patients over a total of 2615 visits. It contains their patient ID, visit ID, UPDRS scores divided into 4 categories (UPDRS I - IV), medication status, and time since their first visit in months. There are 10 missing UPDRS scores in category 3 and 499 in 4. Each clinical record is associated with measurements of protein and peptide abundance taken from the patient's CSF stored in their own tables. There are 227 unique proteins and 968 unique peptides, although not all patients have measurements of all proteins or peptides.

We merged the clinical records table with the protein and peptide tables so that each row, with a visit ID key, contains all the clinical, protein, and peptide data for a visit. We then dropped the medication statuses (since they are not provided in the test data), visit IDs, and patient IDs.

We then duplicated each row for each of the timesteps we are predicting scores for (0, 6, 12, 24 months) and updated the UPDRS scores accordingly.

After splitting the data set into labels (UPDRS scores) and features, we normalized the features using scikit-learn's MinMaxScaler with a minimum of 0 and a maximum of 1. We imputed missing protein and peptides counts using a KNNImputer with 5 neighbors. We imputed missing UPDRS scores using the median strategy.

3.2 Models

Our choice of models depended primarily on the size of our data set. The data set was small and had many missing values in important columns. For this reason, our focus was on simple machine learning models rather than deep learning. We started with strong "out of the box" models: linear regression, random forest, naive Bayes and boosted trees. Then MLP was added as our only neural network.

We used scikit-learn's implementations for models. All models were trained using 3-fold cross-validation and grid search to optimize hyper-parameters. Once optimal hyper-parameters were found, models were retrained on the full data set and submitted.

3.3 Metric

The Kaggle competition uses SMAPE+1 as its evaluation metric so we used it to train our models as well. It ranges from 0 to 200, where 0 means the target and prediction values are the same. It is defined as follows.

$$SMAPE + 1 = \frac{100}{n} \sum_{t=1}^{n} \frac{|F_t - A_t|}{(|A_t + 1| + |F_t + 1|)/2}$$

4 Results

Model	Validation	Test
Linear Regression	108.94	151.2
Ridge Regression	84.82	138
Lasso Regression	87.11	61.5
Support Vector Regression	84.86	72.1
Gaussian Naive Bayes	118.08	110.1
Multi-layer Percepton Classifier	82.18	79.5
Boosted Trees	68.7	57.6
Random Forest	77.07	70.3

Table 1. Model test and validation errors

Boosted trees and lasso regression had the the best performance, achieving test SMAPE+1 scores of 57.6 and 61.5 respectfully. The linear regression, ridge regression and naive Bayes models performed the worst, achieving SMAPE+1 test scores of 151.2, 138 and 110.1. The reason for the poor performance of naive Bayes could be because it assumes peptides/proteins are independent of each other given UPDRS score, which is not true. Also, naive Bayes relies on probabilities and has does not have inherent feature selection like lasso regression does. This results in naive Bayes being unable to filter out uncorrelated features, which could have resulted in its worse score. Linear regression and ridge regression also do not have inherent feature selection. Ridge regression performs better than linear regression due to the L2 regularization, but still performs badly due to its inability to set weights to zero.

As boosted trees and lasso regression had similar scores, we chose to further investigate the lasso regression model, as it was more interpretable. All the protein and peptide feature weights for each lasso model were zero, except for the model to predict UPDRS III scores. Predictions for UPDRS I, II and IV scores were made solely using the bias and visit month. Peptide and protein abundances in general correlated poorly with UPDRS scores, the highest correlation being 0.25 (Fig. 1). The zero

feature weights in the lasso regression models reinforce the notion that the vast majority of proteins and peptide abundances do not correlate with UPDRS scores, and do not have predictive power.

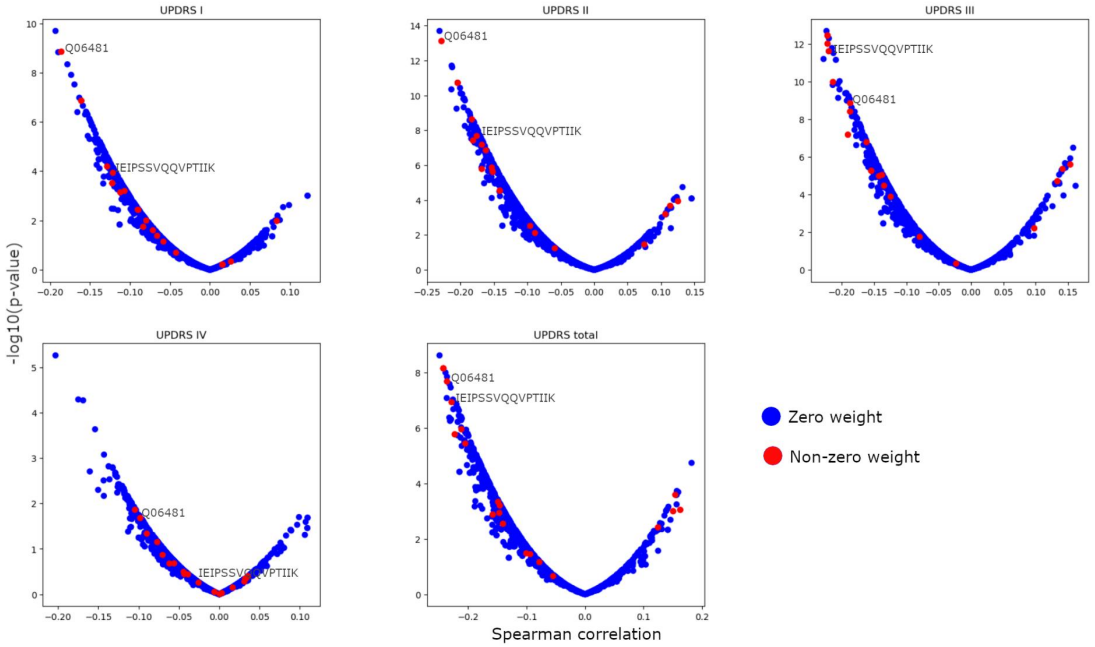

Figure 1. Volcano plots of Spearman correlations of protein and peptide abundances with UPDRS scores. Features are coloured according to whether or not they had a non-zero weight in the lasso model for predicting UPDRS III score. UPDRS total is the sum of all UPDRS scores.

Interestingly, when examining the proteins and peptides with non-zero weights in the UPDRS III lasso model, only a minority were found to be amongst those that were most correlated with UPDRS scores. As lasso regression weights consider interactions between features, this could explain why many of the peptides and proteins do not individually correlate with UPDRS score. Two of the proteins/peptides that were most correlated with UPDRS scores were Q06481 and IEIPSSVQQVPTIIK. Q06481 is also referred to as APLP2 and is a amyloid precursor- like protein involved in neuromuscular transmission, spatial learning and synaptic plasticity (Bethesda (2023)). Additionally it is implicated in Alzheimer's disease (Bethesda (2023)), and is related to the protein APLP1, which was shown to correlate with UPDRS scores in a previous study (Karayel O (2022)). Similarly, the peptide IEIPSSVQQVPTIIK is found in the protein CHL1, which is a neural recognition molecule suggested to be involved in signal transduction pathways (Bethesda (2023)). Given their apparent predictive power and involvement in neuronal function and neurodegenerative disorders, these two proteins/peptides could be of interest for the development of future predictive models for PD progression, and may also potentially play a causal role in progression.

5 Discussion and Future Work

In conclusion, in this study we trained 8 machine learning models to predict the progression of PD using measurements of protein and peptide abundance in patients' CSF. Boosted trees and lasso regression performed the best, with test SMAPE+1 scores of 57.6 and 61.5 respectively. Upon further inspection of the weights in the lasso regression model, we discovered that the model did not necessarily assign non-zero weights to the features that were most correlated to UPDRS scores. Of proteins and peptides that had non-zero weights and comparatively high correlation, we identified APLP2 and CHL1 protein and peptide abundances as being of interest for future model development, as in addition to their predictive power, both proteins are involved in neuronal processes, which affected in PD.

Even the best models we trained did not perform very well but there are several ways to make improvements. Feature engineering and selection could improve performance. It is possible that some proteins may have more of an impact on UPDRS scores when they co-occur with other proteins, and thus it could be useful to include a feature indicating their co-occurance. The difficulty here would be in identifying such interactions. Our model currently only uses the current measurements of protein and peptide abundance to make predictions but it is possible to make use of data from previous measurements. Models suited for time-series predictions, such as RNNs, could prove effective here.

Additionally, more high-quality data would be useful. Data with more patients, proteins, peptides, or even easy to collect data like age, medication state, and sex may improve predictions. A control group containing the protein and peptide data of people without PD but in similar demographics can be used as a baseline to identify abnormalities.

References

Ahmadi Rastegar D Ho N, et al (2019). "Parkinson's progression prediction using machine learning and serum cytokines." *npj Parkinsons Dis*.

Bethesda (Apr. 2023). *National Library of Medicine (US), National Center for Biotechnology Information*.

Dadu A Satone V, et al (2022). "Identification and prediction of Parkinson's disease subtypes and progression using machine learning in two cohorts." *NPJ Parkinsons Dis*.

Greenland JC Williams-Gray CH, Barker RA (2019). "The clinical heterogeneity of Parkinson's disease and its therapeutic implications." *Eur J Neurosci*.

Karayel O, et al. (2022). "Proteome profiling of cerebrospinal fluid reveals biomarker candidates for Parkinson's disease." *Cell reports. Medicine*.

Triarhou, LC (2013). *Dopamine and Parkinson's Disease*.

Zhang X, et al. (2010). "Region-specific protein abundance changes in the brain of MPTP-induced Parkinson's disease mouse model." *Journal of proteome research*.

A Supplementary material

Code for training and evaluating models, and calculating correlations between protein/peptide abundances and UPDRS scores can be found in the following Jupyter notebooks:

https://www.kaggle.com/code/mariastephenson/peptide-and-protein-correlations

https://www.kaggle.com/code/averyquan/evaluation-of-models-to-predict-parkinson-s-diseas