Discovery of Novel Subtypes in Parkinson’s Disease

Xiang GU1, Hongxu Ding1, Jinghao Sun1

1. contribute equally, names not listed in order

Abstract

Introduction

Parkinson’s disease (PD), a long-term neurodegenerative disease, mainly affects the motor neural system. Patients of PD bear with disabilities in obvious shaking, rigidity and slow movement. Unfortunately, there is no available therapeutics against the underlying neurodegenerative process currently. Therefore, it is of great importance to understand how the disease develops, how many subtypes within the disease and how to perform health care for different patients. This study aims to classify the patients based on their genetic data to identify novel subtyping mechanism for PD.

Genetic information has been successfully applied in subtyping and classifying other diseases including complex disease as cancer. Genetic mutations in the protein coding region imply potential alteration of protein structure and function which contribute to the disease development and progression. Therefore, patients can be classified the mutations they harbor which is in correlation with the cause of their disease. This theory has been proved successful in many cancer studies, including breast cancer[[1](#_ENREF_1)] and glioblastoma[[2](#_ENREF_2)]. In this study, we analyzed the mutation information of 645 patients obtained via Exome-seq which sequenced only the protein coding regions in current cohort, to identify novel subtyping scheme for PD.

Subtyping and identifying potential subtypes of a disease falls in the regime of unsupervised clustering problem. Numerous methods for unsupervised clustering have been developed with each’s own advantages and disadvantages depending on the format and features of unlabeled data. These methods have been widely used in other fields to understand both real-life problems such as pattern mining in customers and biomedical problems such as subtyping of breast cancer[[1](#_ENREF_1)]. Under the concept of social networks, analysis on bipartite graph formulated by the patients and their mutations proposes a novel dimension in studying subtyping of a disease[[3](#_ENREF_3)][[4](#_ENREF_4)]. Therefore, in this study, we apply both unsupervised machine learning algorithms and social networks to identify the subtypes of PD.

Method

Preprocessing of Exome-seq Data

Exome-seq of 645 patients was downloaded from Parkinson's Progression Markers Initiative (http://www.ppmi-info.org/). Variants were annotated using hg19 with ANNOVAR and Bioconductor package VariantAnnotation[[5](#_ENREF_5)] in R. Only the unique single nucleotide polymorphism (SNP) passing quality control filter, with amino acid sequence changed were selected. The 179,754 overlapping filtered SNPs from both packages were selected as final candidate features used for subtyping.

Dimension Reduction

Partitioning Around Medoids (PAM)

PAM is a way of implementing k-medoids clustering, a more robust version of k-means clustering[[6](#_ENREF_6)]. To get the best clustering scheme, we use silhouette score[[7](#_ENREF_7)] to evaluate the clustering performance under each k value (from 2 to 20). Result shows silhouette score increase as k goes up, reaches maximum value (0.22) when k ranges from 8 to 11and decreases as k is greater than 12, indicating the optimal classification scheme lies within k=8, 9, 10 and 11.

Unsupervised Clustering

Unsupervised clustering was performed using dimensionally reduced feature matrix. Bla bla bla bla bla

Consensus Clustering

Consensus clustering can 1) determine the number of clusters and 2) assess the stability of the discovered clusters by evaluating the consensus across multiple runs of a clustering algorithm (in our case PAM clustering)[[8](#_ENREF_8)]. Result shows a general trend that as k increases, the PAC (proportion of ambiguous clustering) score decreases. Also, no significant decrease of PAC is observed if k goes beyond 7, indicating the optimal classification scheme lies within k greater than 7.

Affinity Propagation

Affinity propagation determines heterogeneities within data by exchanging messages between data points. Such process is repeated until a high-quality set of exemplars and corresponding clusters gradually emerges[[9](#_ENREF_9)]. Affinity propagation gives clusters with few patients, and we consider those as non-representative. After removing clusters with less than 10 patients, we have 12 representative clusters (negative distance as pairwise similarity, clustering scheme slightly variates when different pairwise similarity measurements methods are used).

Bipartite Network Modularity

The relationship between SNPs and patients can be modeled with a bipartite network[[3](#_ENREF_3)]. It has been reported that the heterogeneity information with the data can be reflected by the network[[4](#_ENREF_4)]. Based on the constructed bipartite network, we measure modularity using method developed by Newman[[10](#_ENREF_10)]. This is an especially powerful method compared to the above mentioned ones, because the cluster specific SNPs are also highlighted.

Result

Discussion

Conclusion

**References**

1. Cancer Genome Atlas, N., *Comprehensive molecular portraits of human breast tumours.* Nature, 2012. **490**(7418): p. 61-70.

2. Verhaak, R.G., et al., *Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1.* Cancer Cell, 2010. **17**(1): p. 98-110.

3. Ciriello, G., et al., *Emerging landscape of oncogenic signatures across human cancers.* Nat Genet, 2013. **45**(10): p. 1127-1133.

4. Girvan, M. and M.E. Newman, *Community structure in social and biological networks.* Proc Natl Acad Sci U S A, 2002. **99**(12): p. 7821-6.

5. Obenchain, V., et al., *VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants.* Bioinformatics, 2014. **30**(14): p. 2076-8.

6. Park, H.-S. and C.-H. Jun, *A simple and fast algorithm for K-medoids clustering.* Expert Systems with Applications, 2009. **36**(2, Part 2): p. 3336-3341.

7. Rousseeuw, P.J., *Silhouettes: A graphical aid to the interpretation and validation of cluster analysis.* Journal of Computational and Applied Mathematics, 1987. **20**: p. 53-65.

8. Monti, S., et al., *Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data.* Machine Learning, 2003. **52**(1): p. 91-118.

9. Frey, B.J. and D. Dueck, *Clustering by Passing Messages Between Data Points.* Science, 2007. **315**(5814): p. 972-976.

10. Newman, M.E., *Fast algorithm for detecting community structure in networks.* Phys Rev E Stat Nonlin Soft Matter Phys, 2004. **69**(6 Pt 2): p. 066133.