Combining Human Disease Genetics and Mouse Model Phenotypes towards Drug Repositioning for Parkinson’s disease

Yang Chen, MS¹, Xiaoshu Cai, BS¹, Rong Xu, PhD²
¹Department of Electrical Engineering and Computer Science, School of Engineering, ²Department of Epidemiology and Biostatistics, School of Medicine, Case Western Reserve University, Cleveland, Ohio, USA

Abstract
Parkinson’s disease (PD) is a severe neurodegenerative disorder without effective treatments. Here, we present a novel drug repositioning approach to predict new drugs for PD leveraging both disease genetics and large amounts of mouse model phenotypes. First, we identified PD-specific mouse phenotypes using well-studied human disease genes. Then we searched all FDA-approved drugs for candidates that share similar mouse phenotype profiles with PD. We demonstrated the validity of our approach using drugs that have been approved for PD: 10 approved PD drugs were ranked within top 10% among 1197 candidates. In predicting novel PD drugs, our approach achieved a mean average precision of 0.24, which is significantly higher (p<e-11) than 0.16 for a state-of-art drug discovery approach based on mouse phenotype data. Comparison of gene expression profiles between PD and top-ranked drug candidates indicates that quetiapine has the potential to treat PD.

Introduction
Parkinson’s disease (PD) is the second most common neurodegenerative disorder and affects 5 million people throughout the world¹. The prevalence of PD is expected to grow significantly with the aging of population². Current pharmacological treatment options are primarily dopamine replacement strategies³,⁴. Among them, levodopa remains the most effective agent in treating PD⁵, but may cause serious adverse events, such as motor and neuropsychiatric complications⁶. In addition, the dopaminergic therapies usually fail in reversing the progressive neuronal loss and controlling non-dopaminergic features of PD, such as sleep impairment, dementia, and sensory disturbances, which have become a major source of disability for most patients²,³,⁴. Therefore, new drugs are desired to improve the treatment of PD.

Computational drug repositioning approaches lead to rapid drug discovery. Most current repositioning strategies can be categorized as drug-based, disease-based, and profile-based (Figure 1). Drug-based (Figure 1(a)) and disease-based (Figure 1(b)) approaches exploit drug-drug or disease-disease similarity and existing drug-treatment knowledge to infer new disease-drug associations⁶,⁷,⁸. Drug similarities are based on pharmacological data, such as drug chemical structure⁹ and drug side effects¹⁰. Disease similarities are calculated using disease phenotype⁷, disease genetic and genomic data¹¹. Recently, profile-based repositioning strategies (Figure 1(c)) have successfully found new drugs for inflammatory bowel disease¹², small cell lung cancer¹³, and lung adenocarcinoma¹⁴. These approaches do not rely on the knowledge of existing drug treatments, and have increased ability to discover new drug-disease pairs compared with drug-based and disease-based strategies.

Current profile-based drug repositioning approaches mainly exploit the gene expression profiles of drugs and diseases¹²,¹³,¹⁴, and have an inherent challenge: the lower-level genomics profile similarities between drugs and diseases do not necessarily translate into higher-level drug treatment efficacy in diseases. Recently, the Mouse Genome Informatics (MGI) has made available large amounts of phenotypic descriptions for mouse genetic mutations based on systematic gene knockouts¹⁵, which are impossible on human. These causal gene-phenotype associations in mice have been demonstrated useful in discovering of new disease genes¹⁶,¹⁷ and drug targets¹⁸, and also have the potential to overcome the challenge in profile-based drug repositioning approaches. Hoehndorf and colleagues have used mouse phenotypes in their drug repositioning approach¹⁹. They first predict disease genes through comparing human and animal phenotype ontologies, and then link the predicted disease-gene associations with the drug-target data to suggest new drug indications. In this study, we present a novel profile-based drug repositioning strategy to exploit both lower-level disease and drug genetics and higher-level mouse phenotypes. We apply the approach for PD drug discovery and compare with Hoehndorf’s approach.

1851
Figure 1. Computational drug repositioning strategies: (a) disease-based methods, (b) drug-based methods, and (c) profile-based methods

Our approach leverages disease-gene associations, drug-target relationships, and mouse phenotype data to predict new drugs for PD. We first identified PD-specific mouse phenotypes using human disease genes in OMIM, and then screened all the FDA-approved drugs for candidates that share similar mouse phenotype profiles with PD. Recent studies have demonstrated that disease-associated genes have potential to guide drug discovery. Mendelian genetics in Online Mendelian Inheritance in Man (OMIM) has been underappreciated in drug target identification, but in fact provides invaluable opportunities for drug repurposing. To date, OMIM has included 15 high-penetrance PD genes that are likely to cause the PD symptoms in the mice carrying their mutations. Though these genes are mostly associated with familial PD, clinical studies have shown that the familial and sporadic forms of PD usually share the same molecular pathways. We compared our approach with the pure genetics-based approaches and demonstrated the importance of mouse phenotypes in improving the performance of PD drug detection. We also compared with Hoehndorf’s approach and show that combining disease genetics with mouse phenotypes using our novel approach achieves significantly better precision. Finally, we analyzed the gene expression profiles for top-ranked candidate drugs and the result suggests that quetiapine has the potential treat PD.

Figure 2. Drug discovery approach for Parkinson’s disease combining human disease genetics and mouse mutation phenotypes
Methods

Our hypothesis is that a drug has the potential to treat PD if the drug target genes are associated with PD phenotypes in mice. Gene-phenotype associations based on systematic mouse gene knockouts provide rich information to link drugs and their new indications. Figure 2 shows that our drug repositioning approach based on mouse phenotypes contains two steps. In the first step, we searched for the mouse phenotypes associated with PD using the well-studied disease genes. In the second step, we extracted a set of mouse phenotype features for each candidate drug and systematically calculated the semantic similarities (using mammalian phenotype ontology) of the phenotype profiles between PD and candidate drugs. Using the mouse phenotype similarity between the drugs and disease, we predicted how likely the drugs can be used to treat PD.

A. Identify mouse model phenotypes for PD using disease genetics in OMIM

We searched for mouse model phenotypes for PD using 15 genes associated with 20 subtypes of PD in OMIM. The mutations of these genes highly increase the risk for PD and are likely to cause PD phenotypes. All these human genes have homologies among mice. We downloaded the phenotype annotations for mouse genes from MGI, and extracted 358 phenotypes that are linked to the 15 PD genes. Different PD genes may share common phenotype annotations. For example, 7 out of 15 PD genes point to the phenotype of neurodegeneration. We weighted each phenotype with the number of its associated PD genes. The weights intuitively represent the confidence that the phenotype is related with PD.

We ranked the PD-specific mouse model phenotypes by their weights, and investigated the category of the top-ranked phenotypes. The mammalian phenotype ontology classifies mouse phenotypes into 30 categories. We first mapped each PD phenotype to its categories by tracing the isa relationship in the mammalian phenotype ontology. The 358 phenotypes were mapped into 24 categories. Then we calculated a score for each category by summing the weights of all the phenotypes in it. We ranked the categories based on these scores and examined the top five categories.

B. Prioritize candidate PD drugs based on the similarities of mouse phenotype profiles between disease and drugs

We collected a set of candidate drugs from DrugBank. The drug-target database in DrugBank contains information for 1427 FDA-approved (for any indication) drugs. We extracted 1197 drugs that target on human/mouse orthologous genes, and included them into the candidate drug set. Then we combined the drug-target relationships and phenotype annotations for the target genes to link each candidate drug to a set of mouse model phenotypes through the drug target genes. We constructed a vector of mouse phenotypes for each drug, and weighted each phenotype by the number of its associated target genes.

We calculated the semantic similarity between the vector of mouse phenotypes associated with PD and each candidate drug to determine how likely the drug can be used to treat PD. A similar calculation of semantic similarity between ontology concept vectors was used in a previous study. We first quantified the information content for each phenotype term \( t \) as \(-\log p(t)\), in which \( p(t) \) represents the frequency among phenotype annotations to all the 7568 mouse genes. In calculating the information content, if a gene is annotated by one phenotype term, we assumed that it is also annotated by the ancestors of this term in the hierarchy of mammalian phenotype ontology. Hence, a phenotype term has higher information content than its ancestors, which lie on higher levels in the ontology. Then we defined the semantic distance \( \text{sim}(t_1,t_2) \) between phenotype terms \( t_1 \) and \( t_2 \) as:

\[
\text{sim}(t_1,t_2) = \max_{a \in A(t_1,t_2)} -\log p(a),
\]

where \( A(t_1,t_2) \) is the set of common ancestors for \( t_1 \) and \( t_2 \). To calculate the distance from the phenotype vector \( p_1 \) to \( p_2 \), we matched each phenotype term in \( p_1 \) to the most similar term in \( p_2 \) and took the average:

\[
\text{sim}(p_1 \rightarrow p_2) = \text{avg}(\Sigma_{t_1 \in p_1} \max_{t_2 \in p_2} \text{sim}(t_1,t_2)).
\]

The matching similarity was weighted by the product of weights for phenotype term \( t_1 \) and \( t_2 \). The similarity between \( p_1 \) and \( p_2 \) was defined as the average of semantic similarities in both directions:

\[
\text{sim}(p_1,p_2) = \frac{1}{2} \text{sim}(p_1 \rightarrow p_2) + \frac{1}{2} \text{sim}(p_2 \rightarrow p_1).
\]

C. De novo evaluation in prioritizing FDA-approved PD drugs

We investigated if our method can prioritize approved PD drugs. We ranked the 1197 candidate drugs using the semantic similarities of the mouse phenotype profiles between the drugs and PD. Then we extracted FDA-approved...
PD drugs from FDA drug labels. Our drug ranking algorithm does not use any information of the approved PD drugs. In the de novo evaluation, we calculated the distribution of approved PD drugs among our ranks by plotting a 10-bin histogram. Specifically, we divided the ranks into 10 ranges, and counted the number of approved PD drugs within each range. In addition, we investigated the target genes for the top 10% candidate drugs. We ranked these drug target genes by the number of drugs (ranked within top 10%) that target on each gene. We also calculated the distribution of genes targeted by the FDA-approved drugs among all the drug target genes using histogram.

We demonstrated the importance of using mouse phenotypes to predict drugs for PD. Recent studies have shown that disease-associated genes can guide the detection of existing drug therapies and promising candidate drugs\(^{22,23}\). We compared our approach with two genetics-based drug discovery methods (figure 3). The first method\(^{22}\) directly matches the disease genes in OMIM with the drug target genes to repositioning existing drugs for new indications. The second method\(^{23}\) extends the disease-associated genes with their functionally related genes based on protein-protein interactions (PPIs), and matched the extended gene set with the drug target genes for drug repositioning. We downloaded the PPIs from the STRING database\(^{31}\), and used the experiment data source, which contains PPI databases such as HPRD, BIND, and GRID. We evaluated if the two methods have the ability to identify approved PD drugs without using mouse phenotypes, and compared the result with our approach.

Figure 3. Our approach compared the mouse phenotype features between disease and drugs. We compared with genetics-based drug discovery methods, which directly match the disease genes and their interacting genes with the drug target genes. The comparison aims at demonstrating the importance of using mouse phenotypes.

**D. Evaluation in ranking novel PD drugs and comparison with an existing drug repositioning approach**

We investigated if our approach has the ability to prioritize novel PD drugs. In our recent studies, we constructed large-scale drug-disease treatment knowledge bases from multiple data resources using techniques including natural language processing, text mining and data mining\(^{22,33,34}\). The databases included 9,216 drug-disease treatment pairs extracted from FDA drug labels, 34,306 pairs extracted from 22 million published biomedical literature abstracts, and 69,724 pairs extracted from 171,805 clinical trials. Based on these knowledge bases, we constructed two evaluation sets as the proxy of novel PD drugs: the first set consists of the drugs that have been tested for PD in clinical trials and the second set consists of PD drugs extracted from literature abstracts in Medline. We removed the FDA-approved PD drugs from both sets. We used histogram to investigate the distribution of drugs in each set among our rank. We also generated a precision-recall curve and calculated the mean average precision to evaluate the ranking of drugs in the union of the two sets.

We compared the performance of our approach with a recent drug discovery approach proposed by Hoehndorf\(^{49}\). In their approach, the human diseases were linked to mouse phenotypes through phenotype ontology comparison, and then associated with orthologous genes based on the gene-phenotype relationships in animal models. After that, they linked the predicted disease genes with the drug-target data to suggest candidate drugs for a given disease. We compared the histograms that represent the distributions of evaluation drugs as well as the precision-recall curves for the two methods.

**E. Test the top-ranked drugs using gene expression data analysis**

We further examined the top-ranked drugs by comparing their gene expression profiles in Gene Expression Omnibus (GEO)\(^{35}\) with that of PD. For the drugs, we extracted data sets that contain gene expression levels before and after adding the drugs to human or animal brain tissues. For PD, we downloaded the data sets that compared the
PD patients and healthy controls. We used the GEO2R software\textsuperscript{36} to identify the significantly differential expressed genes (adjusted p value <0.05) for the disease and drugs, respectively. Then we investigated if common significant genes exist between PD and the drug, and if these common genes have opposite directions of regulation.

**Results**

**A. Our disease genetics-based phenotype prioritization algorithm identified PD-specific mouse model phenotypes**

We ranked and classified the mouse model phenotypes detected using PD genes in OMIM. The top ranked phenotype categories are nervous system and behavior/neurological phenotypes as expected (Table 1). Examples of nervous system phenotypes with the highest weights include neurodegeneration and alpha-synuclein inclusion body, which characterize the pathology of PD. In addition, top-ranked behavior/neurological phenotypes, such as impaired coordination and abnormal gait, mostly include typical motor symptoms of PD. Interestingly, the rest top-ranked phenotype categories show that the pathology of PD is complex and involves not only the nervous system, but also immune system, homeostasis and other aspects.

**Table 1.** The top-ranked categories of mouse phenotypes extracted using PD genes in OMIM.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Phenotype Category</th>
<th>Example top-ranked phenotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>nervous system phenotype</td>
<td>Neurodegeneration</td>
</tr>
<tr>
<td>2</td>
<td>behavior/neurological</td>
<td>Impaired coordination</td>
</tr>
<tr>
<td>3</td>
<td>immune system phenotype</td>
<td>Decreased double-positive T cell number</td>
</tr>
<tr>
<td>4</td>
<td>homeostasis/metabolism</td>
<td>Decreased dopamine level</td>
</tr>
<tr>
<td>5</td>
<td>hematopoietic system</td>
<td>Decreased hemoglobin content</td>
</tr>
</tbody>
</table>

**B. Our approach prioritized FDA-approved PD drugs**

We extracted 22 FDA-approved drugs for PD and 474 genes targeted by these drugs. The median rank of the 22 drugs is 125 (top 10\% among 1197 drugs). The histogram in figure 4 shows that our approach prioritized 10 approved PD drugs within top 10\%. The table in figure 4 shows the rank and percentile of the top 10 approved PD drugs. Among them, the most effective dopamine replacement agent, levodopa, was ranked within top 5\%. Figure 5 shows that the drugs prioritized by our approach frequently target on the drug target genes for approved PD drugs. In figure 5(a), nine in the top ten drug target genes (except GABRA1) are target genes for approved PD drugs. Figure 5(b) shows that half of the top 10\% genes have been targeted by approved PD drugs, while the other half are new drug targets and may lead to novel candidate PD drugs.

**Figure 4.** Our approach ranked the approved PD drugs in the top. A total of 10 among 22 approved PD drugs were ranked within top 10\% among all the 1197 drugs.
Approved PD drugs and their target genes cannot be easily detected through matching disease genes and drug target genes. We compared the performance in identifying approved PD drugs with two genetics-based drug discovery methods. Using the first method, none of the 15 PD genes directly matches the target genes for approved PD drugs and we detected zero approved drug. Using the second method, we detected one approved PD drug, rasagiline, through its target gene BCL2, which interacts with the PD gene PARK2. Though the disease genes for PD and their interacting genes do not directly provide information on the drug target genes, our approach prioritized 10 out of 22 approved PD drugs by exploiting the gene-phenotype associations in mouse models.

C. Our approach outperformed an existing approach in prioritizing novel PD drugs

The top ranked drugs generated by our approach are enriched for the novel PD drugs in the two evaluation sets (figure 6). We extracted 81 drugs from clinical trials to construct the first set, and the candidate drugs in our approach contain 69 of them. Our approach ranked a total of 22 drugs in the top 10% (3.2-fold enrichment comparing to the random cases, \( p < 10^{-4} \)), and this number is 450% higher than 4 drugs in the bottom 10%. Most testing drugs (68%) in the clinical trial set were ranked within top 30%. The evaluation set based on Medline contains 102 drugs, and our candidate drugs included 85 among them. We ranked 26 within top 10% (3.0-fold enrichment comparing to the random cases, \( p < 10^{-4} \)), which is a 760% increase comparing with 3 drugs in the bottom 10%. In contrast, figure 7 shows that the evaluation drugs spread out in different rank ranges when using the existing drug discovery approach based on mouse model phenotypes. Comparison between figure 6 and 7 show that our approach performed better than Hoehndorf’s approach in ranking novel PD drugs in the two evaluation sets.
Figure 7. The distribution of evaluation sets based on clinical trials and Medline texts among the ranks generated by the baseline approach based on mouse phenotypes.

The precision-recall curves in figure 8 further show that our performance is significantly better than the previous approach. The mean average precision for our approach is 0.24, which is significantly higher than 0.16 for the Hoehndorf’s approach (p<e-11). The result means that our approach achieved higher precision averagely at all recall levels, and mostly ranked the novel PD drugs higher than the previous approach.

Figure 8. Precision-recall curves in ranking the novel PD drugs for our approach and Hoehndorf’s approach based on PhenomeNet.

D. Gene expression analysis suggests quetiapine as a potential PD drug

Among the top 10 candidate drugs, we found a set of gene expression samples available for quetiapine in GEO. We identified 61 significant genes for quetiapine from GEO series GSE45229 and 1650 significant genes for PD from GSE8397. Table 2 lists the common significantly differential genes between PD and quetiapine, as well as the direction of regulation for each gene and the logarithm of fold change. Among these genes, MAOA regulates the metabolism of neurotransmitters such as dopamine and is closely associated with PD. In addition, MAOA is not a drug target gene for quetiapine based on the drug-target data in DrugBank. The gene expression analysis suggests that quetiapine, one of the top ranked drugs, has the potential to treat PD.

Table 2. Common significantly differential genes for PD and quetiapine as well as their directions of regulation and fold change.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Quetiapine regulation</th>
<th>Log(FC)</th>
<th>PD regulation</th>
<th>Log(FC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HSPB1</td>
<td>Up</td>
<td>1.5</td>
<td>Down</td>
<td>-1.4</td>
</tr>
<tr>
<td>CHORDC1</td>
<td>Up</td>
<td>0.6</td>
<td>Down</td>
<td>-1.1</td>
</tr>
<tr>
<td>MAOA</td>
<td>Down</td>
<td>-0.6</td>
<td>Up</td>
<td>0.8</td>
</tr>
<tr>
<td>MRPL15</td>
<td>Down</td>
<td>-0.6</td>
<td>Up</td>
<td>0.8</td>
</tr>
</tbody>
</table>
SPEN  Up  0.3  Down  -0.5
EIF5  Down  -0.2  Up  0.4

Discussion

Our study has several limitations that may be addressed in the future. Currently, we used the disease genetics knowledge in OMIM as the seeds to detect PD mouse phenotypes. We have demonstrated in several recent works that disease genes predicted by analyzing human disease phenotype networks and genetic functional relationship networks also have the translational potential in drug discovery. In the future, we will develop approaches to integrate disease associated genes in OMIM, GWAS and prediction results from computational approaches in the drug repositioning approach.

In addition, other information including human disease phenotypes, disease similarities and drug similarities may also contribute in further prioritize strong candidate drugs. In the future, we will develop approaches to seamlessly integrate these data, and investigate the candidate drugs from different angles.

Finally, our experiments tested the ranks of drugs in three evaluation sets: approved PD drugs, drugs that are tested for PD in clinical trials, and PD drug treatments extracted from biomedical literature. These evaluations tested the performance of our approach in retrieving the most successful PD drugs, and approximated its ability of detecting novel PD drugs. In the future, biomedical experiments and clinical studies are required to further demonstrate the effectiveness of the candidate drugs in treating PD.

Conclusions

In this study, we developed a novel drug repositioning approach to predict new drugs for Parkinson’s disease using both disease genetics knowledge and mouse model phenotypes. Our approach can identify FDA-approved PD drugs and prioritize novel PD drugs. A comparison with pure genetics-based drug repositioning approaches shows the importance of using mouse model phenotypes in identifying PD drugs. In addition, our approach outperformed a recently proposed mouse phenotype based drug discovery method through combining disease genetics with mouse model phenotypes using a novel computational approach. Further gene expression analysis on top-ranked candidate drugs suggested quetiapine as a potential PD therapy.

Acknowledgement

RX conceived the study. YC performed algorithm design, data analysis, and evaluation, and wrote the manuscript. All authors have participated in the study discussion and final manuscript review.

Funding

The project was supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under the NIH Director’s New Innovator Award number DP2HD084068.

References


