

Computational Discovery of Pharmacological Chaperones to Rectify Protein Misfolding Using a Novel Support Vector Machine Classifier

By Hajira Fuad

Section 1: The Personal

Ever since I was a kid, I've been both fascinated and frightened by that full range of malicious, deadly human diseases that have no cure. The possibility of one's body turning against itself, waging war on its own cells and disrupting the complex biological processes that keep us healthy terrified me. I found the best way to confront my fears was to simply learn what causes debilitating diseases like cancer and Alzheimer's disease—what goes wrong in our body to cause these horrible maladies, and why. Somehow, learning about the purely technical, scientific aspects behind the pathogenesis of these diseases helped to erode my sense of powerlessness. I became hopeful and naturally progressed to thinking about cures. I indulged my growing curiosity by getting my hands dirty and reading esoteric abstracts, which, with lots of help from Google, I gradually became able to understand.

I found this process of research to be, dare I say it, fun. I'm someone who's always loved a good puzzle, and the more I got involved with cancer research, the more I realized that when it comes down to it, finding treatments and cures for diseases is kind of like solving one giant, complicated puzzle. But sometimes, the solution to the puzzle is much simpler and clearer than one would think. I became intrigued by the possibility of finding simple solutions for infinitely complicated diseases through the process of

connecting the dots, finding recurring motifs and patterns in human pathology. So, naturally, when I learned about protein misfolding, I was hooked.

Protein misfolding is involved in up to half of all human diseases, including Alzheimer's disease, Parkinson's disease, cystic fibrosis and cancer. I was amazed to learn how the simple occurrence of a misfolded protein can have such vast and deadly complications. As I learned more about the fascinating process of protein misfolding, I decided that this was the topic I wanted to focus on. I resolved to tackle this complex field of research with the powerful tool of machine learning, a subfield of artificial intelligence that has made possible the emerging field of bioinformatics, and that can greatly expedite drug research.

To complete this project, I did have to learn additional mathematics. Having never taken AP Statistics, I had to teach myself what one-sample t-tests and p-Values are and how to calculate them. I also had to wrap my head around the various distance-based calculations of molecular graphs that I ended up using in my project. However, learning these mathematical concepts was much less dull than it would've been had I learned it all in a classroom; somehow, the fact that I was going to apply these concepts to analyze the structures of real-life molecules and test the validity of a machine learning classifier I had created on my own made it all much more exciting.

I performed all my research on my laptop, in the comfort of my own home, after learning how to program in Java from a textbook. I hope this fact encourages any student thinking about embarking on their own research endeavor, but who does not have all the resources of a university lab—you can perform some really powerful research with just your laptop and a basic knowledge of coding. I'd also give the following basic advice to

students who want to undertake a project combining science and mathematics, but aren't quite sure where to begin: we're lucky to be living in an age where so many research papers and journals are free and accessible online, so take advantage of that! Browse abstracts about a general topic of interest, write down your thoughts and ideas in a notebook, and go from there. And don't be afraid to have some fun with it!

Section 2: The Research

Abstract

Protein misfolding is a simple phenomenon that is involved in the pathogenesis of a number of human diseases such as Alzheimer's disease, Parkinson's disease, Cystic Fibrosis, and cancer. Pharmacological chaperones are orally administered small molecules that, when bound to a misfolded protein, revert the misfolding process. To discover pharmacological chaperones for specific protein targets, knowledge of the 3D structure of the protein is required to identify exosites for the chaperone to bind to. Even then, most misfolded proteins do not possess natural binding sites. This project aims to find the structural analogues of ligands of misfolded proteins that can function as pharmacological chaperones. Using Java, I developed a classifier based on the support vector machine learning model to predict the structural similarity between two molecules using 2D molecular descriptors that function as similarity metrics. The classifier achieved accuracy greater than 90% and was used to find FDA approved drugs with high structural similarity to Curcumin, an amyloid beta ligand. According to the similar property principle, these drugs have a higher binding affinity towards amyloid beta and can therefore function as pharmacological chaperones. As a proof of concept, the drugs with highest-predicted structural similarity were entered into the SwissDock docking

simulation with amyloid beta, producing favorable approximated Full Fitness values. An FDA-approved drug predicted to be a structural analogue of Curcumin by the classifier, Salsalate, was recently discovered to reverse protein plaque accumulation in an animal model of dementia. Therefore, the binary classifier is capable of finding pharmacological chaperones through drug repurposing and computational methods, streamlining the costly and often protracted drug development process.

Protein Misfolding and Disease

Proteins are structurally complex biological molecules that play critical roles in almost every biological process [1]. For a protein to be fully functional, it first must acquire a certain structural conformation. The misfolding of a protein will cause a loss of function. Proteinopathies, diseases associated with misfolded proteins, can be initiated by very subtle errors in folding, which could include alterations in the primary structure caused by mutations, partial unfolding during thermal and oxidative stress, or RNA modification [6]. An effective strategy for treating proteinopathies is to restore proper protein function by inducing proper three-dimensional structure through the binding of molecules known as pharmacological chaperones.

Pharmacological chaperones are small molecules that, when bound to misfolded proteins, stabilize the protein against factors that could lead to misfolding and induce proper three-dimensional structure [3, 4]. The firm, specific binding of the pharmacological chaperone to any exposed, hydrophobic area on the surface of an unstable protein will initiate stabilization [2]. Additionally, as small molecules, pharmacological chaperones can cross the blood-brain barrier, which is integral for their application to neurodegenerative diseases such as Alzheimer's disease, Parkinson's

disease, and cystic fibrosis [6]. Such diseases are often related to the accumulation of protein deposits or “plaques” that are created when misfolded proteins clump together, leading to cell death and loss of function [3]. These plaques are known as “amyloids.”

A method for finding pharmacological chaperones is to search for the structural analogues of known ligands, substrates or even inhibitors of the affected protein— basically, molecules that are already known to be able to bind tightly to the protein. That way, the 3D structure of the protein does not have to be determined and the tight, secure binding of the pharmacological chaperone to the protein will be ensured due to its high structural similarity to ligands, based off of the similar property principle, which states that structurally similar molecules will exhibit similar physicochemical and biological properties [11]. This can all be accomplished through drug repurposing, which is the development of novel uses and applications for existing drugs, in that the molecules investigated for structural similarity to ligands are drugs that are already FDA-approved.

Small Molecule Similarity Searching

Structural similarity searching is a data mining application that serves to identify structural analogues of a query molecule from a database of thousands of molecules. Such compound databases are inexpensive and publically available. The computational methods behind similarity searching vary, though the majority of similarity search engines employ 2D molecular descriptors as the basis of their similarity calculations [8]. 2D descriptors are essentially numerical values associated with chemical constitution that correlate with physical properties [10]. Many 2D descriptors are distance-based molecular structure descriptors that can model physical, pharmacological, and biological properties of molecules [16].

Machine Learning

Machine learning is a branch of artificial intelligence and method for data analysis that allows computers to independently make predictions about previously unseen data through the employment of algorithms that learn iteratively and intuitively by detecting patterns in previous data. These algorithms are given inputs that are described by a certain number of attributes or “features.” The computer will give an output based on these features. Using machine learning, I created a classifier that can predict structural similarity between molecules with high accuracy and used the classifier to find pharmacological chaperones for a specific disease-causing misfolded protein by finding drugs with high structural similarity to ligands of that same protein.

Methodology

To create a classifier that can accurately predict structural similarity between molecules, a list of molecules that are structurally similar to Ibuprofen was generated using ChemMapper, a web server for computational drug discovery that uses 3D superpositioning of molecules to rank molecules from a selected database in terms of their structural similarity to a query molecule. The 500 compounds most similar to Ibuprofen were taken and labeled as the “positive” instances of the training data set of the classifier. Similarly, a list of compounds with structural similarity to Caffeine from the ZINC Traditional Chinese Medicine database of 142,148 compounds was generated. With the significant structural differences between Caffeine and Ibuprofen taken into account, the 500 compounds most similar to Caffeine were taken and labeled as the “negative” instances of the training data set of the classifier. Next, in Java, I implemented the Support Vector Machine (SVM) classifier, a machine learning algorithm and

classification method that is widely used in bioinformatics due to its high accuracy and flexibility in modeling diverse sources of data [13], using the Waikato Environment for Knowledge Analysis application programming interface (API) to classify molecules as structural analogues or non-analogues of Ibuprofen.

The features used to train the SVM algorithm were similarity metrics between 2D molecular descriptors of the molecules and Ibuprofen. I recorded the Simplified Molecular Input Line Entry System (SMILES), which encodes the chemical structure of a molecule into a single line of text, for every molecule in the dataset. I then input the SMILES for each molecule into programs I wrote in Java using the Chemical Development Kit (CDK), a library of Java classes for bioinformatics. These programs calculated certain 2D molecular descriptors of each molecule, along with the similarity between the 2D molecular descriptors of the molecules with Ibuprofen using similarity metrics. The 2D molecular descriptors used, which I won't describe in too much detail here, were all either topological indices (numerical values) or a list of bits unique to each molecule in which structural information was encoded.

Application to Alzheimer's Disease

Amyloid beta ($A\beta$) is a protein that plays a significant role in the pathogenesis of Alzheimer's disease. $A\beta$ misfolding leads to $A\beta$ aggregation, and thus the development of Alzheimer's disease [19]. Curcumin is an anti-inflammatory molecule in the turmeric root that binds tightly to $A\beta$; it has also shown to inhibit $A\beta$ aggregation [20]. One explanation for Curcumin's ability to inhibit $A\beta$ aggregation is that it functions similarly to a pharmacological chaperone when binding to $A\beta$ —through tight, specific binding to the protein, conformational integrity can be restored by the molecule. However,

Curcumin's poor absorption in the blood stream and rapid metabolization diminishes its candidacy as a pharmacological chaperone.

Due to Curcumin's ability to bind tightly to $A\beta$ and function as an $A\beta$ ligand, it was input as a query molecule into the SVM classifier. The SMILES of 1329 FDA approved drugs [21] were also input into the classifier and the probabilities of the drugs being structural analogues of Curcumin were determined. As proof of concept, the drug predicted by the classifier to be the most structurally similar to Curcumin was docked onto $A\beta$ using the Swissdock program and the binding affinity of the drug was calculated. Similarly, the drug predicted by the classifier to be the least structurally similar to Curcumin was docked onto $A\beta$ using the Swissdock program and its binding affinity was calculated as well.

Hesperetin, an FDA approved drug that had the highest probability of being a structural analogue of Curcumin out of the total list of 1329 FDA approved drugs, was docked onto amyloid beta using Swissdock to demonstrate its binding affinity. Similarly, Chlorotrianisene, an FDA approved drug that had the smallest probability of being structural analogous to Curcumin, was docked onto amyloid beta using Swissdock to demonstrate its binding affinity.

When given a ligand and target protein, Swissdock uses a similarity metric called Full Fitness which measured in kcal/mol to rank the several possible conformations of the drug. The most negative values are ranked first, as negative Full Fitness values represent an energetically favorable complex of a protein and bound ligand, since the ligand would release energy upon binding to the protein.

When bound to amyloid beta, the calculated Full Fitness for Hesperetin, which was given a probability of 0.998362 by the classifier of being a structural analogue of Curcumin, was -434.01 kcal/mol. The Hesperetin- $A\beta$ complex is shown below in Figure 3.2. Additionally, when bound to amyloid beta, the calculated Full Fitness for Chlorotrianisene, which had a probability of $1.67e-7$ in terms of being structurally analogous to Curcumin, was only -373.07 kcal/mol.



Figure 1: Docked conformation of Hesperetin with $A\beta$.

Results

5 fold cross validation was performed 50 times on the dataset and the accuracy of the Support Vector Machine classifier achieved was 91.08% on the training set. The Receiver Operating Characteristic (ROC) curve for a single run of cross validation is

shown in Figure 2. The ROC curve demonstrates the true positive rate versus the false positive rate for the algorithm as the discrimination threshold is varied. The area under the ROC curve measures the discrimination of the classifier, which in this case is the ability of the classifier to correctly classify analogues and non-analogues. On average for the fifty trials of cross validation the area of the ROC curve was 0.9501 for the training set. An area of 1 represents a perfect classifier whereas a random algorithm would have an area of 0.5. A one-sample t-test was conducted on the null hypothesis that the classifier was not significantly more accurate than a random classifier. The p-value for this test was $p < 0.001$. At any reasonable alpha level, I reject the null hypothesis that the classifier is not significantly more accurate than a random classifier.

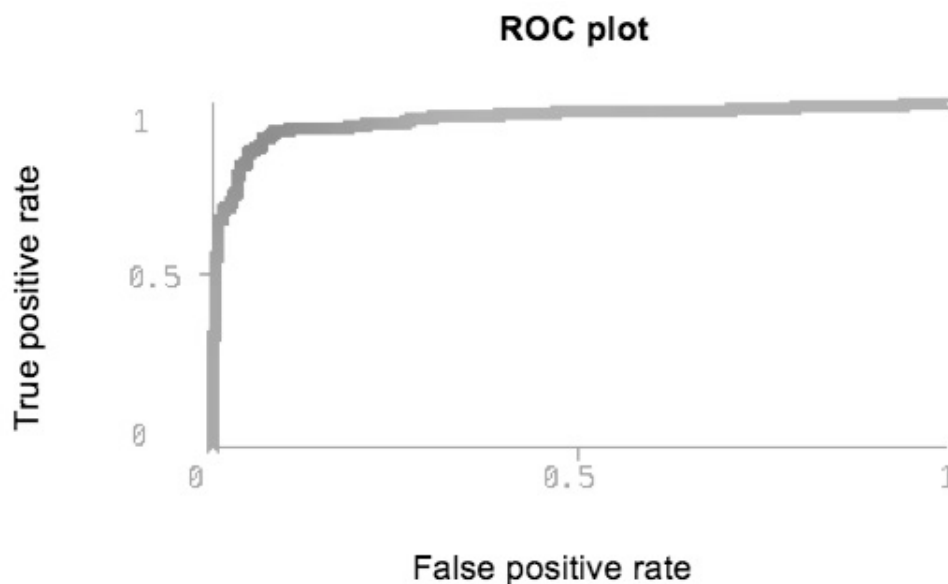


Figure 2: The ROC curve for the training set. The area under the ROC curve is 0.9501 where an area of 1 would represent a perfect classifier.

Of the drugs with the top twenty highest probabilities of being structural analogues of Curcumin, six drugs demonstrated significant potential of functioning as

pharmacological chaperones of amyloid beta, either through the current uses of the drug or experimental evidence from other studies. These drugs are shown in Figure 4.1.

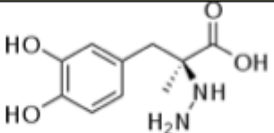
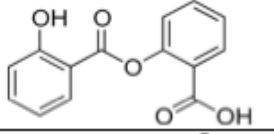
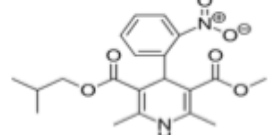
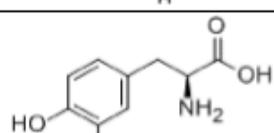
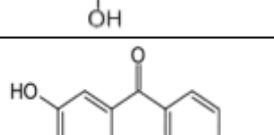
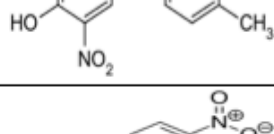
Name	Structure	Function	Probability
Carbidopa		Treats Parkinson's disease	0.9971503502659321
Salsalate		Nonsteroidal anti-inflammatory drug used to treat pain or arthritis	0.9897523952545544
Nisoldipine		Treats high blood pressure	0.9913267920604899
Levodopa		Used alone or in combination with Carbidopa to treat Parkinson's disease	0.9862308956634223
Tolcapone		Treats symptoms of Parkinson's disease	0.9857283574509320
Nitrendipine		Treats primary hypertension to decrease blood pressure	0.9847525015113828

Figure 4.1: Six drugs predicted to be structural analogues of Curcumin, along with their structure, function, and probabilities.

Carbidopa and Levodopa, which had probabilities of 0.997 and 0.986 respectively, are FDA approved drugs used in combination to treat Parkinson's disease. Salsalate, a drug used to treat rheumatoid arthritis and given a probability of 0.990 of being structurally analogous to Curcumin, reduced the accumulation of tau in an animal

model of frontotemporal dementia; tau is a protein that, when misfolded, is involved in the pathogenesis of Alzheimer's disease [22]. Nisoldilpine, which was given a probability of 0.991 of being structurally analogous to Curcumin, was shown in a study to mitigate $A\beta$ production in whole cells and reduce $A\beta$ plaque in a mouse model of Alzheimer's disease [23]. There is also *in vitro* evidence that Nitrendipine, an FDA approved drug that is used to treat primary hypertension and was given a probability of 0.985 of being structurally analogous to Curcumin, has led to the reduction of $A\beta$ pathology and improved cell survival [24].

Conclusion

These results signify the first time that a machine learning algorithm has been utilized in the discovery of pharmacological chaperones and offers a promising future for the discovery of pharmacological chaperones that does not solely depend on time-consuming and expensive screening methods. The classifier not only allows one to find structural analogues of a query molecule through a computationally inexpensive manner based solely on the atomic composition and topological features of molecules, but also provides a unique method for finding pharmacological chaperones without the added complexity of obtaining detailed knowledge of a protein's structure.

I present a new method for discovering pharmacological chaperones through the use of a machine learning classifier to discover structural analogues of known ligands of misfolded proteins. This classifier can also be applied to other proteinopathies such as Cystic Fibrosis, Parkinson's disease and cancer. The classifier I created was used to find FDA approved drugs that are structural analogues of Curcumin and therefore candidates for pharmacological chaperones that can be utilized to restore the conformational

integrity of amyloid beta, a misfolded protein involved in the pathogenesis of Alzheimer's disease. The classifier exposes new relationships between proteins and small molecules through the utilization of the similar property principle and demonstrates how drug repurposing is one way to take advantage of this principle.

References

1. Hartl U., Bracher A., Hayer-Hartl M., Molecular chaperones in protein folding and proteostasis, *Nature* 475, 324-332 (2011)
2. Singh R. L., Dar A. T., Parvaiz A., Proteostasis and Chaperone Surveillance, 172-174, 2015
3. Mahley W. R., Huang Y., *Journal of Medicinal Chemistry* 2012 55 (21), 8997-9008
4. Oh M, Lee. J, Wang W., Lee H., Lee W., Burlak C., Im W., Hoang Q., Lim H., Potential pharmacological chaperones target cancer-associated MCL-1 and Parkinson disease-associated alpha-synuclein. *Proceedings of the National Academy of Sciences of the United States of America*, 111(30), 11007-11012.
5. Small A. S., *Proceedings of the National Academy of Sciences of the United States of America*, 111(34), 12274-12275
6. Chaudhuri, T. K., and Subhankar P. "Protein-misfolding Diseases and Chaperone-based Therapeutic Approaches." *FEBS Journal* 273.7 (2006): 1331-349.
7. Stefani M., Protein misfolding and aggregation: new examples in medicine and biology of the dark side of the protein world, 1739(1), 5-25
8. Kriegel H., Schmidt T., Seidl T., 3d similarity search by shape approximation, *Proceedings of the 5th International Symposium on Advances in Spatial Databases*, 11-28
9. Thimm M., Goede A., Hougardy S., Preibner R., Comparison of 2D Similarity and 3D Superposition. Application to searching a conformational drug database, *Journal of Chemical Information and Computer Sciences* 44 (2004), 1816-1822
10. Ghorbani M., Hosseinzadeh A. M., A new version of Zagreb indices, *Filomat* 26:1 (2012), 93-100

11. Hentabli H., Saeed F., Abdo A., Salim N. A New Graph-Based Molecular Descriptor Using the Canonical Representation of the Molecule *The Scientific World Journal*, vol. 2014.

12. Screening we can believe in, *Nature Chemical Biology* 5, 127 (2009)

13. Ben-Hur A., Weston J., A User's Guide to Support Vector Machines, *Methods Mol Biol.* 2010;609:223-39.

14. Platt, J. Fast Training of Support Vector Machines using Sequential Minimal Optimization, in *Advances in Kernel Methods – Support Vector Learning*, B. Scholkopf, C. Burges, A. Smola, eds., MIT Press (1998).

15. Ertl P., Rohde B., Selzer P., Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties
Journal of Medicinal Chemistry 2000 43 (20), 3714-3717

16. Ilić, A. & Ilić, M. *Graphs and Combinatorics* (2013) 29: 1403.

17. Nikolic S., Kovacevic G., Milicevic A., Trinajstic N., *CROATICA CHEMICA ACTA CCACAA* 76 (2) 113-124 (2003).

18. Sadigh-Eteghad S., Sabermarouf B., Majdi A., Talebi M., Farhoudi M., Mahmoudi J., Amyloid-Beta: A Crucial Factor in Alzheimer's Disease. *Med Princ Pract* 2015;24:1-10
19. Kim B., Lyubchenko L. Y., Misfolding and Aggregation of Amyloid Beta Peptide: Single Molecule AFM Force Spectroscopy, *Biophysical Journal*, 98(3), Supplement 1, p190a

20. Reinke AA, Gestwicki JE, Structure-activity relationships of amyloid beta-aggregation inhibitors based on curcumin: influence of linker length and flexibility, *Chem Biol Drug Des.* 2007 Sep;70(3):206-15.

21. Minikel E., List of FDA-approved drugs and CNS drugs with SMILES, 2013, <www.cureffi.org/2013/10/04/list-of-fda-approved-drugs-and-cns-drugs-with-smiles>.

22. Gladstone Institutes. "Old drug offers new hope to treat Alzheimer's disease: By repurposing a prescription drug used to treat rheumatoid arthritis, researchers successfully reversed tau-related symptoms in an animal model of dementia." *ScienceDaily*. ScienceDaily, 21 September 2015. <www.sciencedaily.com/releases/2015/09/150921133646.htm>.

23. Bachmeier C, Beaulieu-Abdelahad D, Mullan M, Paris D. Selective dihydropyridine compounds facilitate the clearance of β -amyloid across the blood-brain barrier. *Eur J Pharmacol.* 2011 Jun 1;659(2-3):124-9.

24. Drug repositioning for Alzheimer's disease

Corbett A, Pickett J, Burns A, Corcoran J, Dunnett B. S, Edison P, Hagan J. J, Holmes C, Jones

E, Katona C, Kearns I, Kehoe P, Mudher A, Passmore A, Shepherd N, Walsh F, & Ballard C. *Nature Reviews Drug Discovery* 11, 833-846 (November 2012)