Using Machine Learning to Link Genotype Data with Phenotypic Traits in Complex Genetic Disorders

By Keerthi Lakshmi Narayanan

Abstract

Genetic disorders are among the most devastating diseases that affect humans. Yet, several genetic disorders remain incomprehensible to scientists and doctors. Correlating genotype data with phenotypic manifestations has the potential to help scientists better understand complex genetic disorders, such as Fibrodysplasia Ossificans Progressiva (FOP), Wilson's disease, and Stiff Skin Syndrome. However, if such processes were done manually by humans, they might lead to severe errors or misconceptions and, quite simply, be a waste of time. Computer algorithms and machine learning techniques can be efficiently utilized to analyze genotype data along with phenotypic manifestations, providing scientists with the necessary data to reach accurate conclusions on misunderstood diseases.

Keywords: Genetics, Genotypes, Phenotypes, Genetic Disorders, Computer Algorithms

Introduction

Broadly defined, 'genotype' refers to the genetic makeup of an organism; it describes an organism's set of genes. The term can be used to refer to alleles, or variant forms of a gene carried by an organism. Humans are diploid organisms, meaning that humans carry two alleles for every genetic position. Alleles are present in two forms: either as dominant (represented by uppercase 'F') or as recessive (represented by lowercase 'f'). As long as one dominant allele is present, that gene will be reflected in the person's phenotypic manifestations. However, for recessive alleles to be reflected in a person's phenotypic traits, both alleles must be recessive. Alleles can code for everything, from simple physical characteristics like eye color to an individual's risk for developing a particular disease (Nature, n.d.).

This paper discusses the following disorders:

- **Wilson's Disease:** Wilson's disease (WD) is an autosomal recessive disorder caused by a mutation of the ATP7B gene, leading to abnormal copper metabolism. The major clinical features of WD include liver disease, neurological disorders, Kayser-Fleischer (K-F) rings, and osteoporosis. Early diagnosis and lifelong treatment lead to better outcomes. Drugs such as sodium dimercaptosuccinate (Na-DMPS), zinc, and Gandou Decoction can be used to treat WD (Liu et al., 2017).

- **Fibrodysplasia Ossificans Progressiva (FOP):** FOP is an extremely rare autosomal dominant disorder characterized by congenital malformations of the great toes and progressive heterotopic ossification that can induce a disabling second skeleton. Spontaneously occurring flare-ups can cause inflammatory soft tissue to swell, followed by progressive and disabling heterotopic endochondral ossification. There is no definitive treatment for FOP, but early diagnosis and appropriate prevention of flare-ups can extend patient longevity (Qi et al., 2017).

- **Stiff Skin Syndrome (SSS):** SSS is a rare disease characterized by thickened, indurated skin and limited joint movement. Multiple diverse phenotypes have been reported, and the correlation of severity with the clinical heterogeneity and histopathological findings of SSS needs to be refined (Zhao et al., 2024).

Methodology

This study involved a review of various research papers on rare, incomprehensible genetic disorders to gain a better understanding of the information that remains unknown. The study also included a review of several papers on machine learning algorithms to determine the specific type of algorithm required for classification problems. Furthermore, the study reviewed research papers on support vector machines (SVM) and hierarchical clustering to understand the strategies needed for their application in this field. Additionally, a case study on the use of SVM in cancer genomics was analyzed to understand its application (Cervantes et al., 2020; Huang et al., 2018; Patel et al., 2016).

Machine Learning Systems

Support Vector Machine (SVM)

One machine learning system that is particularly suited for this situation is the support vector machine (SVM). SVM is a type of machine learning algorithm used for classification and regression tasks. It helps computers learn how to classify data into different categories (Cervantes et al., 2020). An SVM plots all data as points on a graph, with every point belonging to one of two categories. The SVM attempts to draw a line or curve that best separates these points

into their correct categories. The primary task of an SVM is to find the line that separates the categories with the widest possible margin. Points closest to the line are labeled as "support vectors," and they help define where the line should be. SVMs are known for their accuracy in classifying data and can be used for both linear and non-linear data (data that cannot be separated by a straight line). They also work well in high dimensions when the data has multiple features (Cervantes et al., 2020). An SVM, when given the genotype data of an individual, can be trained to find the best line (or hyperplane) that separates the data based on the phenotypic labels (e.g., separating genetic data of healthy individuals from those with a genetic disorder). The genetic markers that the SVM identifies as most important for making its predictions can be reviewed to help strengthen the interpretation of the analysis. These markers could be key to understanding the genetic basis of the phenotypes (Huang et al., 2018).

Hierarchical Clustering

Another machine learning technique that would be extremely useful in this field is hierarchical clustering, a method for grouping items into clusters in a step-by-step process, forming a tree-like structure that shows how items are related. Hierarchical clustering begins by gathering items or data points to be grouped together (e.g., genetic information). The algorithm initially considers every individual data point as its own cluster. It then starts analyzing data points with the most similarities and merges them into a single cluster. This process continues, with the algorithm finding the next closest pair, and so on, until all points are combined into a single large cluster. The user can then divide the large, merged cluster into as many small and specific clusters as desired (Patel et al., 2016).

Hierarchical clustering does not require a pre-defined number of clusters, giving the user flexibility in making such decisions after viewing the complete cluster. This algorithm also handles various types of data, such as numerical or categorical (Patel et al., 2016).

By using hierarchical clustering to analyze genotypes and their phenotypic manifestations, patterns in the data can be identified, such as which genetic variations are associated with specific phenotypic traits. It also helps in grouping individuals with similar genetic profiles to better understand how differences in their phenotypic manifestations occur. Hierarchical clustering also provides insights into the relationships between different genetic markers and phenotypic traits, helping to understand the genetic basis of the disorder (Patel et al., 2016).

Conclusion

Genetic disorders are deadly diseases that affect millions of people every year, and many of these disorders are still unknown or not explored. Understanding the cause of a disorder and how it can be treated is crucial. Computer algorithms and machine learning techniques are extremely efficient tools for analyzing key data that may be pivotal to our understanding of genetic disorders.

References

Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. Neurocomputing, 408, 44-66. https://doi.org/10.1016/j.neucom.2020.04.064

Huang, S., Cai, N., Pacheco, P. P., Narandes, S., Wang, Y., & Xu, W. (2018). Applications of support vector machine (SVM) learning in cancer genomics. Cancer Genomics & Proteomics, 15(1), 41–51. https://doi.org/10.21873/cgp.20063

Liu, J., Luan, J., Zhou, X., Cui, Y., & Han, J. (2017). Epidemiology, diagnosis, and treatment of Wilson's disease. Intractable & Rare Diseases Research, 6(4), 249-255. https://doi.org/10.5582/irdr.2017.01057

Nature. (n.d.). Genotype. Nature Scitable. Retrieved August 4, 2024, from https://www.nature.com/scitable/definition/genotype-234/

Patel, S., Sihmar, S., & Jatain, A. (2016). A study of hierarchical clustering algorithms. 2016 IEEE International Conference on Computing, Analytics and Security Trends (CAST), 135-142. https://doi.org/10.1109/CAST.2016.7915073

Qi, Z., Luan, J., Zhou, X., Cui, Y., & Han, J. (2017). Fibrodysplasia ossificans progressiva: Basic understanding and experimental models. Intractable & Rare Diseases Research, 6(4), 242-248. https://doi.org/10.5582/irdr.2017.01055

Zhao, Q., Chu, Z., Li, L., Feng, C., Zhou, H., Hu, J., Zhao, L., Che, D., Zhang, X., Peng, B., Han, Y., & Geng, X. (2024). The clinical heterogeneity of stiff skin syndrome: A systematic review. Journal of Dermatology, 51(5), 487-495. https://doi.org/10.1111/1346-8138.16390