Identification of Diseases caused by non-Synonymous Single Nucleotide Polymorphism using Random Forest and Linear Regression Algorithms

Muhammad Junaid Anjum
COMSATS University Islamabad
Fatima Tariq
Lahore College for Women University
Khadeeja Anjum
CMH Medical College
Momina Shaheen
University of Roehampton
Faizan Ahmad (fahmad@cardiffmet.ac.uk)
Cardiff School of Technologies, Cardiff Metropolitan University

Article

Keywords:

Posted Date: June 26th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3001745/v1

Abstract

The analysis of different types of diseases is an extremal vital task which would help in producing vaccines for that particular type of disease. However, this is a very costly process as to test every disease it would mean to analyze every gene related to that specific disease. This issue of genic analysis is further elevated when different variations of each disease is considered. As such the use of different computational methods is taken into consideration to tackle the task of genic variation identification. This research makes use of Machine Learning algorithms to help in the identification and prediction of Single Nucleotide Polymorphism or more specifically Single Amino Acid Polymorphism. Taking into consideration ten different types of diseases, this research makes use of Random Forest and Linear Regression algorithms to identify and predict different genic variations of these diseases. From the extensive research, this article concludes that Random Forest algorithm performs better in comparison to Linear regression in genic variation predictions.

Introduction

The investigation of generic variations is currently being done through different practices such as Gene Panel which basically provides identification of variants in more than one gene. Through this test it is possible to pinpoint a disease and its different symptoms. However, this process can take a relatively long time. Because of this, recently, researchers have shifted their focus towards the identification of gene variants using computational methods or more specifically through the use of Machine Learning (ML). Over several years, different ML algorithms have been developed that aim to speed up the investigation of variants in genes. The main goal behind the use of ML is to speed up the identification process as well as lower the cost involved in the testing process. This sped up process of identification would allow researchers to gather information about the different variants of a gene and thus help in diagnosing and creating a medicine against the investigated variant.

This research focuses on the identification of single nucleotide polymorphism (SNP) or, more specifically, its sub-type of nonsynonymous SNP (nsSNP), also known as single amino acid polymorphism (SAP). Using the concept of ML, this research provides the identification of different diseases that are caused by substitutions, insertions, or deletions in an amino acid sequence of a particular protein. The identification is done using a simple random forest ML algorithm and a linear regression ML algorithm. Both identification results are calculated and compared with each other.

This research is divided into several different sections which are as follows:

- Section II provides background information about the concept of SNP and its different categories.
- Section III provides a comprehensive look into the research work conducted by different researchers regarding the identification of SNPs.
- Section IV details the different materials and methods used to evaluate the performance of the ML algorithm used in the identification of different disease specific SAPs.
- Section V discusses about the evaluated datasets and their results.
- Finally, in Section VI, the conclusion and future work is presented.

Background

To better understand the concept of nsSNP, one needs to understand the basic structure of a protein. A protein is basically composed of multiple peptide subunits, where peptides refer to a short chain of amino acids, which are also known as polypeptides. If one was to differentiate between proteins and peptides, then the biggest distinguishing factor would be that peptides are molecules consisting of 2–50 amino acids whereas proteins are made up of more than 50 amino acids.

The sequence of amino acids in a protein determines the function of that protein in the human body. The Lawrence Livermore National Laboratory and researchers from the Protein-Based Identification Technologies, LLC, collaborated with each other to develop a method for the biological identification that basically extracts the information present in a protein. The main focus of this research was to identify a person by just analysing the proteins found from a single human hair (Girard 2021). This method or forensic technique identified the different genetic mutations by analysing amino acids in the protein of the hair.

However, there are situations where over time, a mutation occurs in the DNA sequence which results in a synthesized protein witnessing a different amino acid sequence from the normal sequence. This mutation is known as single nucleotide polymorphism (SNP). However, it should be noted that the mutation of a DNA sequence is known as single nucleotide variation (SNV), while the mutation in the protein amino acid sequence is referred to as SNP.

A single nucleotide polymorphism is categorized into two distinct categories, as shown in Fig. 1, which is synonymous and nonsynonymous. A synonymous SNP refers to the concept where the amino acid sequence of a protein is not mutated; however, the function of a protein may be affected. Nonsynonymous SNP on the other hand refers to the mutation of amino acid sequence in a protein. Nonsynonymous SNP or nsSNP also has two different sub-categories that being missense and nonsense. In relation to nonsense mutation, it refers to the changes in the amino acid to a STOP codon, while missense mutation refers to the polymorphism of one amino acid to another amino acid in the sequence of a protein amino acid.

Identification of SNP or more specifically nsSNP is very important as this would result in researchers in the identification of different diseases that are caused by mutations in the amino acids. However, the normal methods of using biological weapons are too expensive and can take a long period of time. As such, researchers are now focusing their attention on in-silico methods which basically incorporate computational methods, machine learning (ML), in the predication or identification of mutations in proteins. There are different ML algorithms that can be used for this type of prediction such as Random Forest (RF), Support Vector Machines (SVM), Linear Regression (LR) or Decision Trees (DTs).

Related Work

Over the past few years, different types of research have been conducted in identification of nsSNPs in proteins. Some of these research works have been discussed below.

The authors in (Zhang et al. 2020) presented their study in which the authors predicted 57 of 88 high risk pathogenic nsSNPs that were involved in the pathogenesis of congenital cataracts. The authors collected data for the Gap junction protein alpha 3 (GJA3) which is an important pathogenic gene of congenital cataracts. The prediction of the mutation in GJA3 were made using different in silico web tools which included SIFT, PROVEAN, PolyPhen and others. The authors also investigated the secondary prediction of the amino acid structure produced from the GJA3 gene.

The authors (Choudhury et al. 2020) provided the analysis of different nsSNPs of the CTC1 gene, specifically the C-Terminal OB-fold region. Through their research, the researchers identified 75 out of 126 mutations present in the C-Terminal OB-fold region that were destabilized and deleterious. Out of the identified 75, the researcher concluded that 11 of those mutations could be considered pathogenic. The prediction of the nsSNP of the CTC1 gene was made using different prediction tools such as SIFT, PolyPhen, and Mutation assessor.

In (Akhtar et al. 2019), the authors presented their findings that SNP in the CCR6 gene have been the possible cause of different diseases such as psoriasis, lupus nephritis, rheumatoid arthritis, and systemic sclerosis. Through their study the authors also discussed that the gene-gene interaction of CCR6 with other genes such as CCL21 and CCL20, presented an association with different diseases related to CCR6. The authors also discussed that though they used different in-silico

methods to review the SNPs of CCR6, they pressed that a detailed investigation is required to be done to study the protein structure and function.

The authors in (Emadi et al. 2020) review that the HLA-G (Human Leukocyte Antigen G) protein is an immune tolerogenic molecule consisting of 7 isoforms. The mutations in the HLA-G have resulted in different diseases. As such, the authors have presented a study which reviews and predicts the most pathogenic missense nsSNPs in HLA-G isoforms through different in-silico methods. The authors also examined the functional and structural effects of predicted nsSNPs on HLA-G isoforms. For their prediction of different nsSNPs, the authors made use of different web prediction software such as SIFT, PROVEAN, PhD-SNP and SNAP2.

The research (Song et al. 2021) discusses that nsSNPs can result in different pathogenic mutations that can cause various human diseases. Though there are different prediction tools present, there are still improvements that can be made in these in-silco based prediction tools. As such, the authors have proposed a new sequence-based predictor known as DMBS which provides accurate predictions of deleterious nsSNPs. The proposed DMBS makes predictions through the combined efforts of two components, which are the predicted ligand-binding residues and sequence conservation. The authors test their proposed DMBS against different benchmark datasets which result in an Area under the Curve (AUC) ranging between 0.95 and 0.98.

The authors (Lira and Ahammad 2021) reviewed the DRD2 gene, which is a neuronal cell surface protein that is involved in brain function and development. There is a great clinical significance in the variations in the DRD2 gene as DRD2 is a pharmacotherapeutic target for the treatment of psychiatric disorders such as schizophrenia and ADHD. The authors have presented a study in which they make use of different bioinformatic tools such as SIFT, PROVEAN, PhD-SNP, and SNPs & GO to investigate the impact of nsSNP on the functionality and structure of the protein. From the 260 nsSNP collected from the dbSNP database, the authors found that 9 were considered deleterious. Upon further investigation, the authors found that the mutant variation F389V was considered as the most impactful variant.

The research (Khoruddin et al. 2021) review that the most common genetic variations are caused through SNPs for various diseases, one of which is cancer. The authors present that genome-wide association studies (GWASs) are often conducted, which are used to identify SNPs which can increase the risk of cancer and leukemia. However, the authors discuss that most GWAS don’t include the population of Orang Asli and Malays. Considering this, the authors have developed a comprehensive bioinformatic pipeline that would mine the entire genome sequence database of the Orang Asli and the Malays to identify the presence of pathogenic SNPs that could ultimately increase the risk of cancer. For this the authors made use of different in-silico prediction tools which include SIFT, PolyPhen, Condel and PANTHER to predict and identify the functional impact of the SNP. Out of the 80 SNPs in the GWAS dataset, the authors found 52 SNPs that are prominent in the Orang Asli and Malays.

In their research (Bhatnagar and Dang 2018), the authors review that an important regulator of collagen metabolism is Prolidase. Although there are several studies present on the prolidase deficiency, which is a rare autosomal recessive disorder, there is still a lack of studies related to prolidase at the molecular level. The authors have also discussed that a number of SNPs are still uncategorized for Prolidase. As such, the authors have presented a study, which is the first of its kind to predict the structure and function of nsSNP on the structure and function. For their prediction the authors used different prediction tools such as SIFT, PROVEAN, Mutation Assessor and nsSNP Analyzer.

The authors in (Arifuzzaman et al. 2020) discussed that due to mutations occurring in the SMPX gene, the regular activity of the SMPX protein can be disrupted. This disruption can result in the occurrence of the hearing process. The authors provide that recent studies have shown a connection between SNP and SMPX and hearing loss. As such, the authors have provided a study from damaging nsSNPs of SMPX using 13 different bioinformatics tools. Using the prediction tools, the impact of nsSNPs in the SMPX gene was evaluated, and the different deleterious convergent changes were recorded. The

authors also reviewed the different pathogenic effects of mutations in SMPX-meditated protein-protein interactions that were characterized through binding energy calculations and structural modelling.

The research (Lim et al. 2021) reviewed that MYB proteins are classified as highly conserved DNA binding domains (DBD). Mutations in MYB oncoproteins have been reported to cause augmented and aberrant cancer progression. Through the identification of MYP molecular biomarkers prediction of cancer progression can be made use to improve the management of cancer. As such, the authors have made use of a biomarker discovery pipeline for the identification of deleterious nsSNPs. For their investigation, the family of MYB protein was extracted from the NCBI database. For the prediction of the nsSNP, different in-silico tools were used. From their investigation, the authors found a total of 45 nsSNPs that were damaging and high-risk.

In their research the authors (Desai and Chauhan 2019) present that the encoding of methionine synthase by the MTR gene is one of the key enzymes involved in the S-Adenosyl Methionine (SAM) cycle that catalyzes the conversion of homocysteine to methionine. The authors in their study predicted the functional consequences of nsSNP in the human MTR gene using different in silico prediction tools such as Poly-Phen2, SNAP2, PROVEAN and PMut. The authors also predicted the PTM sites within a protein as well as generating the 3D protein structure. The authors presented that in the human MTR gene, D621G, G682D, V744L, V766E, and R1027W were found to be structurally and functionally significant nsSNPs.

Some other research works that have incorporated the use of in-silico methods to predict or identify diseases caused by nsSNPs include like that of (Havranel and Islam 2020) where the authors have made use of different available in-silico methods to investigate the pathogenic effect of 14 nsSNPs that may have a relationship with Neurofibromatosis type 2 (NF2), a deadly disease which is caused by nsSNPs in the NF2 gene. Similarly, another research work that uses in-silico based methods to identify a severe disease caused by nsSNP is (Quan et al. 2019) where the authors make use of CONSURF web server to identify the most pathogenic variations of the STXBP1 gene, which is a gene that is associated with early infantile epileptic Encephalopathy (also known as Ohtahara Syndrome). Another research work is conducted by (Saxena et al. 2021) where the authors make use of in-silico methods to identify the disease associated SNPs of human GOT1 (Glutamic-Oxaloacetic Transaminase 1). The human GOT1 is a gene that can be associated with several neurodegenerative diseases and also different types of cancer. Another research work done is where the authors in (Al Mehdi et al. 2019) proposed a computational algorithm DAMpred that could be used to identify nsSNPs that cause different type of diseases. This is done through the coupling of protein structure predictions and protein-protein interactions with evolutionary profiles. This entire process is then trained through a novel Bayes guided artificial neural network algorithm.

Materials and Methods

For the collection from data of different nsSNPs, different databases were used, including UniProtKB (also known as Swiss-Prot) (Uniprot 2021), dbSNP of NCBI (Sherry et al. 1999) and VariBench (Nasir and Vihinen 2013). As this research is focused on the non-synonymous of SNPs which results in the connection between polymorphism of amino acid sequencing which would ultimately lead to some rare or pathogenic diseases, the datasets collected were related to different diseases. The diseases considered are shown in Table 1.

Table 1
Disease categorization

Sr#	Disease Names
1	Hemophilia (Hemo)
2	Epileptic Encephalopathy (Epil)
3	Mucopolysaccharidosis (Muco)
4	Cardiomyopathy (Card)
5	Charcot-Marie-Tooth Disease (Char)
6	Retinitis Pigmentosa (Reti)
7	Marfan Syndrome
8	Glycogen Storage Disease (Glyc)
9	Niemann-Pick Disease (Niem)
10	Osteogenesis Imperfecta (Oste)

From these databases, positive datasets which included pathogenic proteins and negative datasets which included neutral datasets were collected belonging to the organism of Humans [homo sapiens]. After the collection of both positive and negative datasets, feature extraction was performed using the web server of Feature Extraction App (Hussain 2020).

The dataset was then trained and tested with an 80 – 20 split with both the RF and LR algorithms.

Random Forest:

Random forest classification, one of the finest learning algorithms in machine learning (Breiman 2001), uses unpruned classification trees produced by bootstrap sampling and random features (Svetnik et al. 2003). This algorithm assigns a class or label to each input data set. Many fields, from proteomics (Izmirlian 2004) to ecological studies (Mursalin et al. 2017; Jian et al. 2018), use the random forest classifier. Moreover, character recognition (Mursalin et al. 2017), sentiment analysis (Shaheen et al. 2019), malware detection (Joshi et al. 2018), traffic accident detection (Dogru and Subasi 2018), and medical imaging and diagnosis have all employed it.

Linear Regression:

A variable's value can be predicted using linear regression (Su et al. 2012) analysis based on the value of another variable. The dependent variable is the one you want to be able to forecast. The independent variable is the one you're using to make a prediction about the value of the other variable (Montgomery et al. 2021). With the help of one or more independent variables that can most accurately predict the value of the dependent variable, this type of analysis calculates the coefficients of the linear equation. The differences between expected and actual output values are minimized by linear regression by fitting a straight line or surface. The best-fit line for a set of paired data can be found using straightforward linear regression calculators that employ the "least squares" technique.

After training and testing, different effective performance which include Accuracy, Precision, Sensitivity, Specificity, F-score and MCC, were calculated using the formulas presented below:

$$Accuracy\,(Acc) = \frac{TP + TN}{N}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{FP + TN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F - Score = F1 = \frac{Precision \times Sensitivity}{Precision + Sensitivity}$$

$$Matthews Correlation Coefficient\,(MCC) = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

Here TP, TN, FP and FN represent True Positive, True Negative, False Positive and False Negative, respectively. Beside these performance metrics the use of Receiver Operating Characteristics (ROC) was used to verify the performance of the used ML algorithms. Basically, the ROC is a curved based graph which represents the Area Under the Curve (AUC) which is the representation of TP rate and FP rate. The larger the AUC, the better the performance of the ML algorithm being used.

Results and Discussion

Through training and testing of the obtained datasets, the performance evaluations of both RF and LR were conducted which in general demonstrates that RF algorithm provides a better prediction of pathogenic nsSNP compared to LR algorithm. The results of the performance evaluation metrics of accuracy, sensitivity, specificity, precision, F score, and MCC are shown in Table 2.

Table 2
Performance Evaluation Metric Results

Algorithm	Accuracy	Sensitivity	Specificity	Precision	F1	MCC	Self-Consistency
RF	100	100	100	100	100	1	100
LR	100	100	100	100	100	1	83.5

After the performance evaluation of the collected datasets, using independent testing of both RF and LR, the test accuracy of both these algorithms are done, which showed than RF performed better that LR with a score of 86.6 while the LR presented a test accuracy score of 60.0.

After the independent test, the ROC was generated which again shows that FR performs better that LR as RF had an AUC of 0.777 while LR had an AUC of 0.643. The plotting of ROC of both these ML algorithms are shown in Fig. 2.

Another test that was conducted was that of MCC where the score for Splitting testing was done for both RF and LR which showed that RF had an MCC score of 73.214 while LR had an MCC score of 20.01.

After this, the cross validation of both RF and LR were done using 10-fold validation was done which represents the accuracy of both these algorithms. The accuracy results of the 10-Fold validation are shown in Table 3.

Table 3
10-Fold Validation Results

Sr#	Accuracy	Folds	Algorithm
0	93.969397	1	RF
1	93.699370	2	RF
2	93.609361	3	RF
3	93.249325	4	RF
4	92.619262	5	RF
5	93.429343	6	RF
6	92.979298	7	RF
7	92.619262	8	RF
8	93.243243	9	RF
9	93.963964	10	RF
10	93.185000	Mean	RF
0	93.699370	1	LogReg
1	93.609361	2	LogReg
2	93.519352	3	LogReg
3	93.249325	4	LogReg
4	92.799280	5	LogReg
5	93.339334	6	LogReg
6	92.529253	7	LogReg
7	92.709271	8	LogReg
8	92.882883	9	LogReg
9	93.513514	10	LogReg
10	93.230139	Mean	LogReg

The accuracy of the 10-fold validation is further represented in the form of a plotted violin graph as shown in Fig. 3.

Conclusion and Future Work

From the above research, it can be concluded that on the diseases that contain nsSNP variants, the detection of such variants can be refined and processed faster using computational methods. Using random forest and linear regression ML algorithms, this research verifies that the algorithm of Random Forest performs better in comparison to Linear Regression when used for the detection of disease specific nsSNP or SAPs. This is verified and validated through different performance metrics, as well as ROC graphs, and 10-fold validation.

In future, this research aims to include more disease to check the performance of Random Forest as well as the comparison of other algorithms which can be both supervised and unsupervised learning algorithms.

Declarations

Data Availability

The data that support the findings of this study are available from the author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

1. Akhtar M, Jamal T, Jamal H, Din JU, Jamal M, Arif M, Arshad M, Jalil F. 2019. Identification of most damaging nsSNPs in human CCR6 gene: In silico analyses. Int. J. Immunogenet. 46(6):459–471.
2. Al Mehdi K, Fouad B, Zouhair E, Boutaina B, Yassine N, Chaimaa AEC, Najat S, Hassan R, Rachida R, Abdelhamid B et al. 2019. Molecular Modelling and dynamics study of nsSNP in STXBP1 gene in early infantile epileptic encephalopathy disease. Biomed Res. Int.
3. Arifuzzaman M, Mitra S, Das R, Hamza A, Absar N, Dash R. 2020. In silico analysis of nonsynonymous single-nucleotide polymorphisms (nsSNPs) of the SMPX gene. Ann. Hum. Genet. 84(1):54–71.
4. Bhatnagar R, Dang AS. 2018. Comprehensive in-silico prediction of damage associated SNPs in Human Prolidase gene. Sci. Rep. 8(1):1–14.
5. Breiman L. 2001. Random Forests. Mach. Learn. 45(1):5–32.
6. Choudhury A, Mohammad T, Samarth N, Hussain A, Rehman M, Islam A, Alajmi MF, Singh S, Hassan M. 2021. Structural genomics approach to investigate deleterious impact of nsSNPs in conserved telomere maintenance component 1. Sci. Rep. 11(1):1–13
7. Desai M, Chauhan JB. 2019. Predicting the functional and structural consequences of nsSNPs in human methionine synthase gene using computational tools. Syst Biol Reprod Med. 65(4):288–300.
8. Dogru N, Subasi A. 2018. Traffic accident detection using random forest classifier. 15th Learning and Technology Conference (L&T). p. 40–45.
9. Emadi E, Akhoundi F, Kalantar SM, Emadi-Baygi M. 2020. Predicting the most deleterious missense nsSNPs of the protein isoforms of the human HLA-G gene and in silico evaluation of their structural and functional consequences. BMC Genet. 21(1):1–27.
10. Girard JE. 2021. Criminalistics: Forensic science, crime, and terrorism. Jones & Bartlett Learning
11. Havranel B, Islam SM. 2020. Prediction and evaluation of deleterious and disease causing non-synonymous SNPs (nsSNPs) in human NF2 gene responsible for neurofibromatosis type 2 (NF2). J. Biomol. Struct. 39(18):7044–7055.
12. Hussain W. 2020. Fea_Protein. GitHub Repository. [accessed 2022]. https://github.com/WaqarHusain/FEA_Protein.
13. Izmirlian G. 2004. Application of the Random Forest Classification Algorithm to a SELDI-TOF Proteomics Study in the Setting of a Cancer Prevention Trial. Ann. N. Y. Acad. Sci. 1020(1):154–174.
14. Jian CW, Ibrahim MZ, Thum W, Seong T, Ei W, Khatun S. 2018. Embedded Character Recognition System using Random Forest Algorithm for IC Inspection System. Electr. Comp. Eng. 10(1–3):121–125.
15. Joshi S, Upadhyay H, Lagos L, Akkipeddi NS, Guerra V. 2018. Machine Learning Approach for Malware Detection Using Random Forest Classifier on Process List Data Structure. Proceedings of the 2nd International Conference on Information System and Data Mining – ICISDM. p. 98–102.

16. Khoruddin NA, Noorizhab MN, Teh LK, Mohd Yusof FZ, Salleh MZ. 2021. Pathogenic nsSNPs that increase the risks of cancers among the Orang Asli and Malays. Sci. Rep. 11(1):1–22.
17. Lim SW, Tan KJ, Azuraidi OM, Sathiya M, Lim EC, Lai KS, Yap WS, Afizan NARNM. 2021. Functional and structural analysis of non-synonymous single nucleotide polymorphisms (nsSNPs) in the MYB oncoproteins associated with human cancer. Sci. Rep. 11(1):1–14.
18. Lira SS, Ahammad I. 2021. A comprehensive in silico investigation into the nsSNPs of Drd2 gene predicts significant functional consequences in dopamine signaling and pharmacotherapy. Sci. Rep. 11(1):1–16.
19. Montgomery DC, Peck EA, Vining GG. 2021. Introduction to linear regression analysis. John Wiley & Sons.
20. Mursalin M, Zhang Y, Chen Y, Chawla NV. 2017. Automated epileptic seizure detection using improved correlation-based feature selection with random forest classifier. Neurocomputing. 241:204–214.
21. Nasir PS, Vihinen M. 2013. VariBench: A benchmark database for variations. Hum. Mutat. 34(1):42–49.
22. Quan L, Wu H, Lyu Q, Zhang Y. 2019. DAMpred: Recognizing Disease-Associated nsSNPs through Bayes-Guided Neural-Network Model Built on Low-Resolution Structure Prediction of Proteins and Protein–Protein Interactions. JMB. 431(13):2449–2459.
23. Saxena S, Murthy TK, Chandramohan V, Yadav AK, Singh TR. 2021. Structural and functional analysis of disease-associated mutations in GOT1 gene: An in-silico study. Comput. Biol. Med. 136:104695.
24. Shaheen M, Awan SM, Hussain N, Gondal ZA. 2019. Sentiment analysis on mobile phone reviews using supervised learning techniques. Int. j. mod. educ. comput. sci. 7:32–43.
25. Sherry ST, Ward M, Sirotkin K. 1999. dbSNP—database for single nucleotide polymorphisms and other classes of minor genetic variation. Genome Res. 9(8):677–679.
26. Song R, Cao B, Peng Z, Oldfield CJ, Kurgan L, Wong KC, Jang Y. 2021. Accurate Sequence-Based Prediction of Deleterious nsSNPs with Multiple Sequence Profiles and Putative Binding Residues. Biomolecules. 11(9):1337.
27. Su X, Yan X, Tsai CL. 2012. Linear regression. Wiley Interdiscip. Rev. Comput. Stat. 4(3):275–294.
28. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. 2003. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. J. Chem. Inf. Comput. 43(6):1947–1958.
29. UniProt: the universal protein knowledgebase in 2021. 2021. Nucleic Acids Res. Spec. Publ. 49(D1): D480-D489.
30. Zhang M, Huang C, Wang Z, Lv H, Li X. 2020. In silico analysis of non-synonymous single nucleotide polymorphisms (nsSNPs) in the human GJA3 gene associated with congenital cataract. BMC Mol. Cell Biol. 21(1):1–13.

Figures

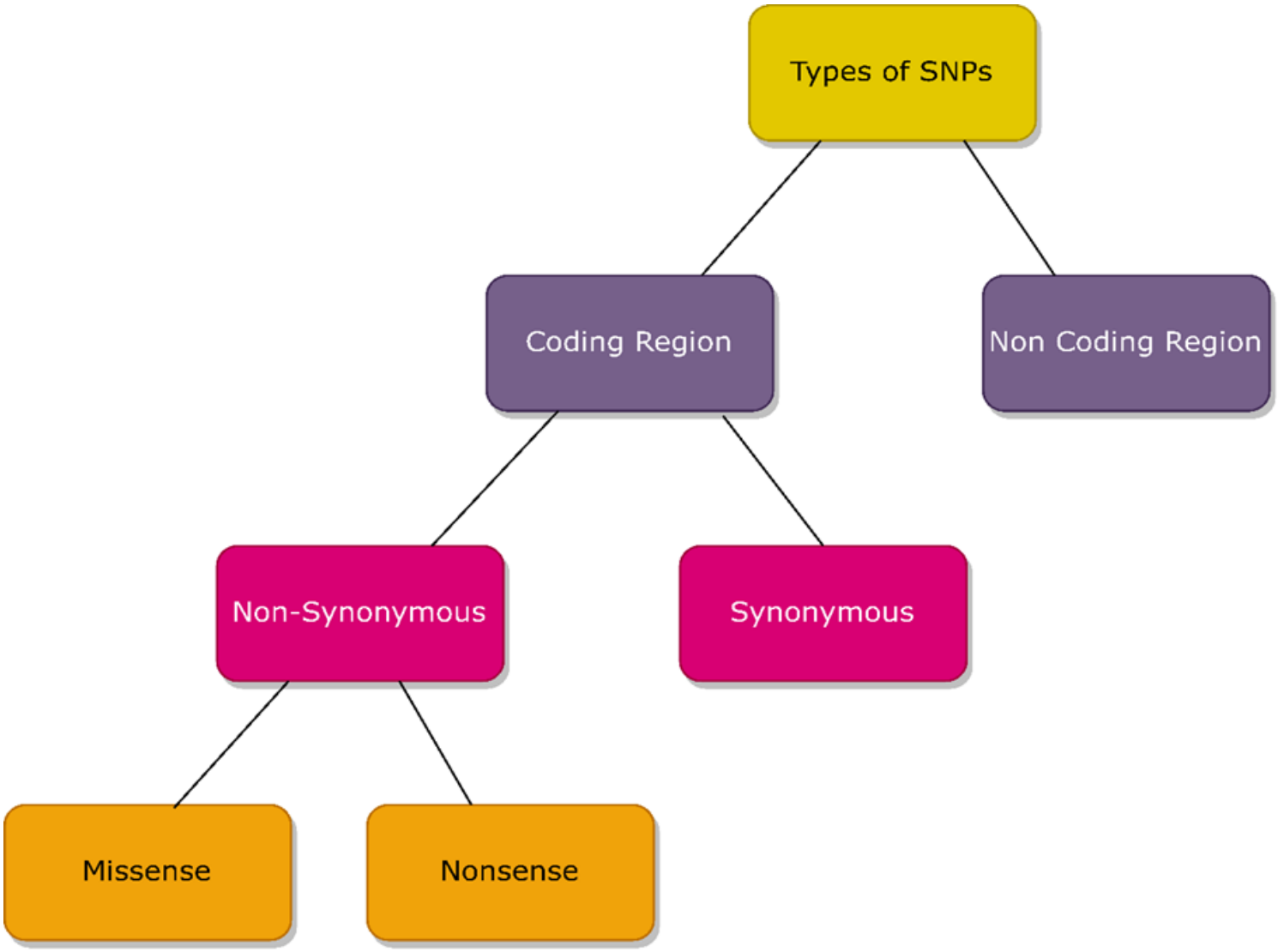

Figure 1

Categories of single nucleotide polymorphism (SNP)

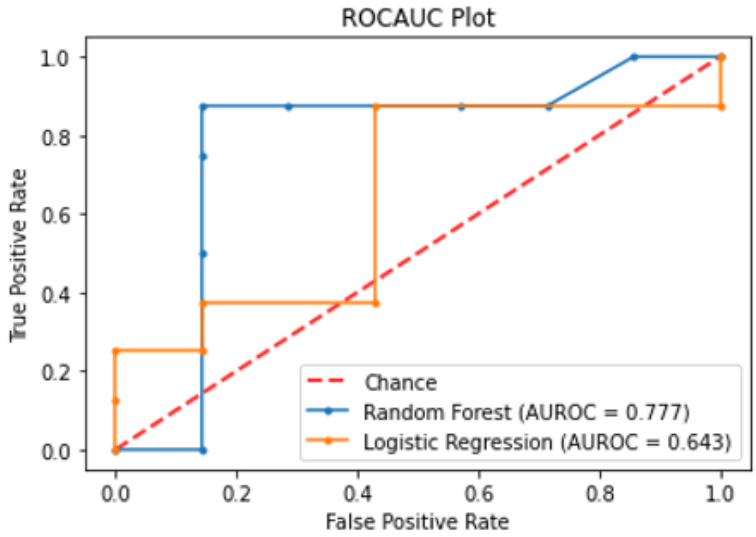

Figure 2

ROC of RF and LR

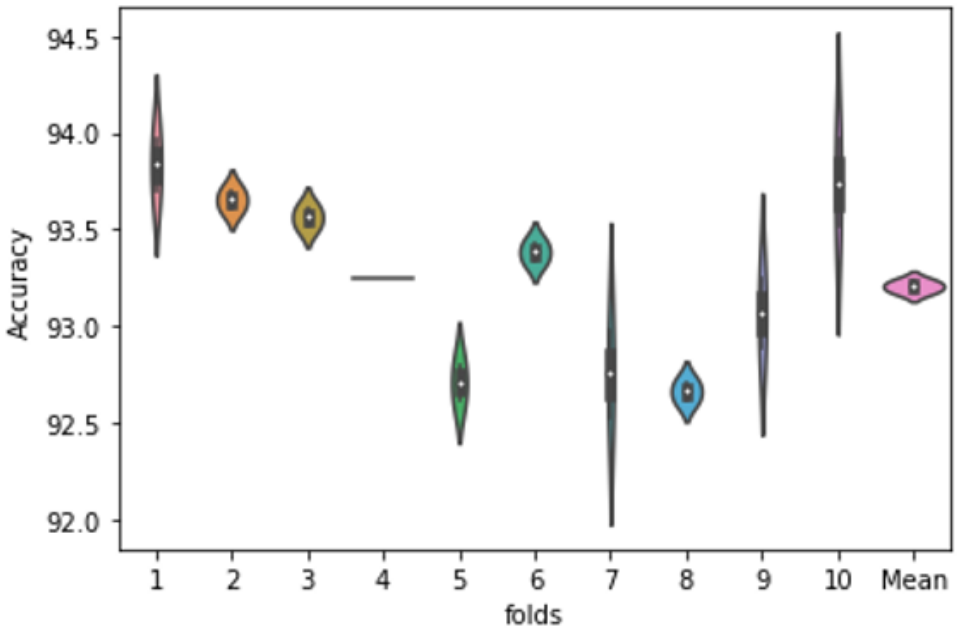

Figure 3

10-Fold Validation Violin Plot Graph