Human Genomics

RESEARCH

Open Access

Rapid discrimination between deleterious and benign missense mutations in the CAGI 6 experiment

Eshel Faraggi[1,2*], Robert L. Jernigan[3] and Andrzej Kloczkowski[4,5,6]

Abstract

We describe the machine learning tool that we applied in the CAGI 6 experiment to predict whether single residue mutations in proteins are deleterious or benign. This tool was trained using only single sequences, i.e., without multiple sequence alignments or structural information. Instead, we used global characterizations of the protein sequence. Training and testing data for human gene mutations was obtained from ClinVar (ncbi.nlm.nih.gov/pub/ClinVar/), and for non-human gene mutations from Uniprot (www.uniprot.org). Testing was done on post-training data from ClinVar. This testing yielded high AUC and Matthews correlation coefficient (MCC) for well trained examples but low generalizability. For genes with either sparse or unbalanced training data, the prediction accuracy is poor. The resulting prediction server is available online at http://www.mamiris.com/Shoni.cagi6.

Introduction

In recent years, the field of genetic interpretation is burgeoning. As of March 7, 2023, a Google Scholar search of the terms 'predict gene variant' gives 1,960,000 results. Valuable applications are emerging from these mutation studies. To mention a few examples: genetic variation and response to cancer treatment [1], mental health [2], geographic location of a specimen [3], educational attainment and longevity [4, 5], splicing [6], schizophrenia [7], non-alcoholic fatty liver disease [8], and obesity [9]. The Critical Assessment of Genome Interpretation (CAGI) [10, 11] experiment was developed to objectively assess computational methods for predicting the phenotypic outcomes of genomic variations, and to monitor the progress of research in this field. The work described here participated in the sixth round of the CAGI experiment.

Predicting the effects of genetic variations is a fundamental problem in biology and medicine. Machine learning based methods have proven to be the most successful approaches for protein structure prediction [12–14] and are promising contenders to tackle the problem of predicting the effects of gene variants as well. Distinguishing computationally, between variants that are associated with damage (deleterious) and those that are not (benign or neutral) is a major aim of this research. [9, 15–34]

Our participation in CAGI was restricted to Missense Mutations (MM): a change in a single codon that results in a different amino acid. MMs conserve the length of the protein and result in a single amino acid change in the expressed protein. Sometimes the terms Single Amino acid Variant (SAV) [35] or NonSynonymous Variant (NSV) [36] are used to describe such mutations. Our

*Correspondence:
Eshel Faraggi
efaraggi@gmail.com
[1] Research and Information Systems, LLC, 1620 E. 72nd ST., Indianapolis, IN 46240, USA
[2] Physics Department, Indiana University Purdue University Indianapolis, Indianapolis, IN 46202, USA
[3] Roy J. Carver Department of Biochemistry, Biophysics and Molecular Biology, Iowa State University, Ames, IA 50011, USA
[4] The Steve and Cindy Rasmussen Institute for Genomic Medicine, Columbus, OH 43205, USA
[5] Battelle Center for Mathematical Medicine, The Research Institute at Nationwide Children's Hospital, Columbus, OH 43205, USA
[6] Department of Pediatrics, The Ohio State University, Columbus, OH 43205, USA

participation in the CAGI6 experiment involved the prediction of the impact of MMs of the human gene Arylsulfatase A (ARSA) on its enzymatic activity. ARSA breaks down cerebroside 3-sulfate into cerebroside and sulfate. Cerebrosides are a group of lipids involved in animal muscle and nerve cell membranes. Variations in ARSA are implicated in Metachromatic Leukodystrophy, a disorder characterized by neuro-cognitive decline. In severe forms of the disease patients only survive to early childhood. [37]

Our approach for CAGI6 focused on predicting the phenotype of a MM purely from the amino acid sequence without additional input from structural information or evolutionary information from multiple sequence alignments. Instead, we tried to capture the local and global physical environment of a residue and the protein. As was shown in our earlier work [38], such an approach coupled with training machine learners on a large database, can compensate for some loss of information and provide robust predictions when alignments are unavailable. Additionally, due to the vast number of possible genetic variations, the time it takes to make predictions about their impact is crucial for large scale studies.

Materials and methods

We would like to numerically describe the local and global information from the amino acid sequence, as a substitute for information extracted from multiple sequence alignments. We would also like to find a representation of a given protein that is independent of its length. We do this by calculating projections of properties of the sequence on a set of preselected functions. Here we use discrete periodic functions as explained later. Reliance on global features instead of multiple sequence alignments follows the same approach that was used for ASAquick [38]. We will also use ASAquick as part of the input features. The Accessible Surface Area (ASA), sometimes called Solvent-Accessible Surface Area, is the surface area of a protein accessible to a solvent and is usually measured in $\mathrm{\AA}^2$.

Numerically capturing the local and non-local information from the amino acid sequence of a protein is crucial for any predictions from that sequence. We also wish to find a representation of a protein that is independent of its length. We would like to characterize an observable quantity, a, along the sequence. That is, characterize the set of values $\{a_i\}_{i=1}^{S}$, with S the number of residues in the sequence. For example, a_i could be the value of the ASA for the residue at position i.

We name the first method as *'joint'*, and express its value, for a given periodicity, n, as, $M_j(n)$. To calculate $M_j(n)$, we take the first $2n$ residues, sum up the observables a_i for $i = 1, .., n$, and subtract the observables a_i for $i = n + 1, ..., 2n$. We repeat this for the rest of the $2n$ blocks in the sequence. We ignore any remaining tail that is not covered by the $2n$ blocks. We then normalize this sum by dividing by $4\,S^{0.2}$. We chose this normalization constant after some testing, to obtain distributions visually in the interval $[-1, 1]$ independent of the length of the sequence.

The second method is termed *'disjoint'*, and its value is expressed by $M_d(n)$, for a given periodicity n. To obtain $M_d(n)$ for $n > 1$, we sequentially labeled each residue in the sequence as $\{1, 2, ..., n, 1, 2, ..., n, 1, 2, ...\}$. We will use l_i below to denote this label. We again ignore any remaining sequence tail that does not fit exactly within the n-blocks. We then calculate

$$M_d(n) = \frac{1}{3 \cdot S^{0.2}} \sum_{i=1}^{S} \left(-1 + \frac{2l_i}{n-1} \right) a_i \quad (\text{For } n > 1) \tag{1}$$

For $n = 1$, the sum in Eq. (1) is taken as four times the average value of a_i along the sequence. Note that $M_j(1) = M_d(2)$. We again selected the normalization to obtain distributions visually in the interval $[-1, 1]$ independent of the length of the sequence.

To represent a given sequence, we calculated for it $M_j(n)$ and $M_d(n)$ with $n = 1, .., 400$. For a given n we z-scored the values of $M_j(n)$ across all mutation data, and similarly for $M_d(n)$. Note that for a given mutation these values are calculated for the mutated sequence. By using this method we are able to represent a sequence by its patterns of discrete periodicity, irrespective of its length. Because of some similarity with moment integrals, we will use the term to refer to them here. Other efforts have been made to represent sequences in length invariant ways [39–45]. However, these approaches are derived from discretization of the continuum assumption.

ASAquick [38] is our single sequence ASA predictor. We also used features developed for it here. These include the length of the chain divided by 1000, the residue type density of the whole chain (25 values), and the directional residue-pair density (625 values). We have 25 residue types because we account for atypical residues ('B','Z'), unknowns ('X'), and chain gaps ('!','-'). The output and inputs are stored in separate files in a directory named for that specific mutant.

For a given residue mutation, we also used a window of neighboring residues as input. Based on previous experience we chose a 10-residue window on both sides of the mutated residue (21 residues total), capturing some of the local sequential environment of the residue. Each residue in this window is represented by its ASAquick average RASA prediction and the associated standard deviation of the prediction, and seven parameters characterizing their physical and chemical properties. These include a

steric parameter (graph shape index), hydrophobicity, volume, polarizability, isoelectric point, helix probability, and sheet probability [46, 47]. We also include the prediction of ASAquick for this window size on the original (unmutated) sequence. Additionally, we take the difference between the RASA predicted for the original and mutated sequence over a window of 15 residues (31 residues total) and the differences between their predicted errors. We include in this input the average difference in RASA and error predictions over the entire window. The nomenclature, size, and short explenation of the input features is given in Table 1.

We use a two-vector for the output. We code the phenotype, neutral or deleterious, as, $(-1, 1)$ or $(1, -1)$, respectively. This scale is used since we are using a bi-level hyperbolic tangent for our activation function, meaning that our networks predict in the $[-1, 1]$ interval. When making a phenotype assignment we reverse the sign of the second coordinate and take the average over the two-vector. We tested several architectures for our networks. We used a momentum value of 0.035 and 0.05, a hyperbolic tangent activation function with an activation parameter of 0.05, a learning rate of 0.00021 and trained the networks until there was no improvement in the over-fit protection set accuracy for 400 epochs (training iterations). The networks achieved this point after a few hundred to a few thousand epochs. For this study we used a general two hidden layer neural network (GENN) [48]. The parameters were selected based on previous experience and by grid searches on random sets with a few thousand instances.

Data for human MMs was collected on May 27, 2021 from the ClinVar database [49–52] (ncbi.nlm.nih.gov/pub/ClinVar/) which provides information about the association between disease and mutation for human genes. This data is a list of mutations and their clinical associations. Sequence changes are described by a gene identifier, mutation positions, and the sequence change. To limit the complexity of the prediction problem, we collected only MMs. We found approximately 2,300,000 entries in ClinVar. Out of these, approximately 700,000 were of the 'single nucleotide variant' type, with approximately 80,000 MM entries with conclusive or likely pathogenic label.

We have found instances of MMs with multiple entries in ClinVar. For 610 of these MMs, their pathogenicity labels were identical and we accepted them as labeled to the dataset. For four MMs the pathogenicity labels were in conflict. Out of these four, for three the deleterious labels involved a specific disease association. We accepted these three to the dataset as deleterious instances. One entry without specific disease association and conflicting pathogenicity was removed from the dataset.

Non-human data was collected from the Swiss-Prot dataset [53, 54] (www.uniprot.org). There are two types of sequence variation described in Swiss-Prot: naturally occurring sequence variants and mutagenesis variants. For some of the mutagenesis entries, clear phenotype information is given. Unfortunately, there is little uniformity in the description of phenotypes. We downloaded the Swiss-Prot database and then developed manually curated key-word searches to assign a given phenotype descriptor in Swiss-Prot as benign or deleterious. We consider the natural variant entries to be benign mutations. For mutagenesis variants, we use terms such as "abolish" and "inhibit" to find MMs that cause a change in phenotype. In May 2021 we collected 71,460 non-human MMs. Out of these, 18,873 MMs were categorized as benign and 52,587 as deleterious.

On December 6, 2021 we again collected data. In this case we gathered 81,902 human MMs and 72,094 non-human MMs. In the December dataset, 3325 human

Table 1 Input features description

File	Size[a]	Description
physpar.zs	147	z-scored physical parameters in 21-residue window
asawinprf.zs	42	z-scored RASA prediction and standard deviation in 21-residue window
asawinprfmut	42	RASA prediction and standard deviation in 21-residue window for mutated sequence
momintdj.asa	400	Disjoint moment integral for predicted RASA for $n = 1, ..., 400$
genn.gin.orig	651	Single and two residue composition and sequence size for unmutated sequence
genn.gin	651	Single and two residue composition and sequence size for mutated sequence
asamutdif.zs	70	Predicted RASA change upon mutation
momintdj.asa.err	400	Disjoint moment integral for RASA estimated error
momintj.asa	400	Joint moment integral for predicted RASA
momintj.asa.err.zs	400	z-scored joint moment integral for RASA estimated error

Description of individual-input-file-predictors.

[a] The number of values in the input file

Table 2 Datasets

	May 27, 2021		December 6, 2021	
	Human	Non-Human	Human	Non-Human
Benign	41,123	18,873	44,342	18,931
Pathogenic	36,339	52,587	37,560	53,163

Summary of data used for generating and testing server

MMs and 634 non-human MMs were new and not present in the May dataset. Out of the new human MMs, 2638 were categorized as benign and 687 as pathogenic. Out of the new non-human MMs, 31 MMs were categorized as benign and 603 MMs as deleterious. The prediction server used in CAGI6 was trained with the datasets collected in Dec. 2021. Estimated prediction accuracy was obtained from the prediction of exclusive December data by networks trained on May 2021 data. The size of the datasets is summarized in Table 2.

At the beggining of this project we applied a naive and quick approach of generating training and over-fit sets. We randomly selected those sets from the collected MM data. An outcome of this is that the server that participated in the CAGI 6 experiment was trained on unbalanced sets in terms of pathogenicity assignment. Additionally, training and over-fit sets contained different variations of identical genes. The underpinning assumption is that pathogenicity is determined by local biochemistry. This ignores complex and non-local processes that premeate biology. These effects resulted in skewed input feature importance ranking and poor prediction quality in CAGI 6.

We report accuracy in several ways. We used the RMSE between predicted and labeled states as the main accuracy measure used by the neural networks during training. We also estimated the classification ability of the predictor using the Area Under the the receiver operating Curve (AUC), and the Matthews Correlation Coefficient (MCC).

For the CAGI6 experiment we generated six models with different combinations of trained weights. For each model we selected 12 networks that performed best on their respective over-fit set. We averaged the predicted pathogenicity from the 12 networks and used the standard deviation as error estimates. For models 1–3 we used networks with 42 nodes per hidden layer. Weights for model 1 were trained exclusively on human MMs, weights for model 2 were trained exclusively on non-human MMs, and weights for model 3 were trained on both human and non-human MMs. For models 4–6 we used networks with 22 and 32 nodes per hidden layer. Weights for model 4 were trained exclusively on human MMs, weights for model 5 were trained exclusively on non-human MMs, and weights for model 6 were trained on both human and non-human MMs.

Results

The choice which input features to use in this prediction server was critical and involved several considerations. The first was the speed of generating the input feature. For the work here we use only single sequence features, i.e., features that don't need multiple sequence alignment to be calculated. We then considered the effectiveness of these features. The server that participated in the CAGI 6 experiment was trained on a unbalanced sets and somewhat noisy data. As we shall see below that resulted in skewed input feature importance ranking and poor prediction quality. In Table 3 we show the average over-fit protection set (30% of data excluded from training) prediction accuracy for each individual input feature. For some features a clear advantage is evident. To obtain the list of features for our predictor, we added individual input features, ranked by lowest error, until adding them does not improve the prediction, the line in the table represents this cutoff. Note that we have also tested adding an input feature to the top performing feature (genn.gin.zs) and arrived at a list similar to the one in Table 3. In this case we found input features genn.gin.zs and genn.gin.orig.zs are most effective. This is an artifact from the unbalanced nature of our data. Since for many genes the distribution of phenotypes in the database is heavily skewed in one direction (benign or deleterious), the network found that it would be most effective to recognize the gene as a way of determining the assignment. Work

Table 3 Test set prediction error for individual input features

Feature	Error	STDEV
genn.gin.orig.zs	0.58	0.03
genn.gin.zs	0.59	0.01
asamutdif.zs	0.61	0.03
momintdj.asa.zs	0.65	0.01
physpar.zs	0.65	0.02
genn.gin	0.66	0.03
momintj.asa.err	0.66	0.03
momintj.asa.err.zs	0.66	0.03
momintdj.asa.err.zs	0.66	0.02
genn.gin.orig	0.67	0.02
momintj.asa	0.67	0.03
physpar	0.68	0.01
momintdj.asa	0.68	0.02
momintj.asa.zs	0.69	0.02
asamutdif	0.99	0.00

Individual test set error for each of the input features used (above line) and some input feature not used (below line) in our server

done following the completion of the CAGI 6 experiment demonstrates that *physpar.zs* is the most informative feature from our set of predictors.

The following describes the nomenclature used in Table 3. *genn.gin* contains the length of the mutated protein divided by 1000 (1 value), the residue composition with 25 residue types accounting for atypical residues, unknowns and chain gaps (25 values), and the directional two-residue composition (625 values); *genn.gin.orig* is identical to *genn.gin* except that it is calculated for the original (unmutated) sequence; *asamutdif* contains the average and standard deviation, taken along the sequence, for the difference between the predicted RASA for the mutated and original protein, and similarly for predicted RASA error (4 values), the same procedure is done for the absolute value of the difference in predicted RASA and its error between mutated and original protein (4 values), and, it contains the difference in RASA and error prediction, between mutated and original protein, for a window of 15 residues to each side of the mutated residue (62 values); *momint* refers to the moment integral with the suffices '*j*' for *joint* and '*dj*' for *disjoint*, and '*asa*' and '*asa.err* refer to the predicted RASA and its error respectively (400 values). An ending of '*zs*' indicated that the input features were z-scored, its absence indicates that features were not z-scored.

We also analyzed the benefit of averaging over different realizations of the networks. In Fig. 1 we present the AUC and MCC, for prediction of phenotype on the validation set, as a function of the number of networks used in the average. From the plot it appears that in this case there is no significant improvement in accuracy for averaging over more than about a dozen networks. Therefore, we trained six randomly initialized networks with a momentum value of 0.035 and six randomly initialized networks with a momentum value of 0.05. We then average the 12 networks for each case to obtain both an average prediction and a prediction error estimate for the server. We assign a prediction by averaging the raw score of the networks and use the standard deviation for error estimate. Estimated over-fit prediction error and standard deviation for the trained networks are 0.411 and 0.004 respectively.

Evaluation of the prediction was done by recollecting new MMs from the ClinVar dataset, and testing our approach on this new data. To evaluate the accuracy of phenotype prediction we calculated the MCC and AUC. Student's t-test analysis revels for this case a t-value of 2.6 with more than 3000 degrees of freedom indicating more than 99% confidence in rejecting the assumption of no connection between phenotype prediction and label. We have compared our prediction server to the predictions of Provean [28, 55] and PolyPhen-2 [24, 56] on our validation test set. For Provean we find an AUC of 0.835 and an MCC of 0.481, while for PPH2 we find an AUC of 0.811 and an MCC of 0.471. Our approach gives an AUC of 0.953 and an MCC of 0.780 for this set. One should stress that our high predictive power here results from the unbalanced nature of ClinVar data, and reflects the neural networks ability to learn this artifact; and not to discern between deleterious and benign MMs. This is further displayed in the low ranking our method received in CAGI6.

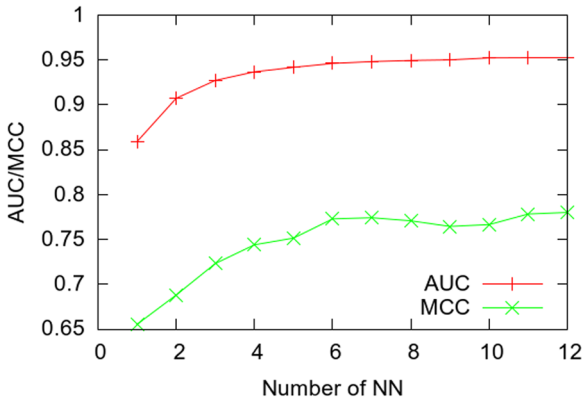

Fig. 1 Values for the area under the receiver operating curve (AUC) and the Matthews correlation coefficient (MCC) for the prediction of missense mutation, versus the number of neural networks averaged over, in a simulated read-world test. Networks trained on human ClinVar data up to May 27, 2021 and tested with MMs collected from ClinVar on Dec. 6, 2021 that do no appear in the May 27, 2021 dataset

Discussion

For our limited training dataset with an unbalanced distribution of phenotypes, global features became more dominant, as shown in Table 3. In our data, 4829 human genes had only benign MMs and 1131 had only deleterious MMs. 3243 had both phenotypes, however, 1210 of these had a single MMs example for one of the phenotypes. Our testing allows us some confidence in the usefulness of our predictor for genes for which balanced training data exist. However, properly balanced data is a critical issue, and gathering and processing it for this work was a major difficulty.

Conclusions

We have designed and built a fast machine learning server for the prediction of the phenotype of MMs in a single sequence-based approach. Its speed results from not using sequence alignments. To compensate for some of the information loss due to the lack of sequence alignments, we designed a new type of input feature that captures some of the sequence information at the global level

by integrating the sequence with periodic weights. The method produced fast and relatively accurate predictions for human genes as estimated from testing on post-training data from ClinVar. However, further examination of the prediction reveals a strong bias in the predictor resulting from an unbalanced training set and noisy data. For genes for which the training data is either sparse or unbalanced the prediction accuracy is poor. Although global features were found useful, more work, especially on capturing the phase information in the sequence, can potentially significantly improve their usefulness in protein properties prediction. The resulting prediction server is available on-line at http://www.mamiris.com/Shoni.cagi6.

Acknowledgements
This research was supported by the National Science Foundation grant DBI-1661391, and the National Institute of Health grants R01HG012117 and R01GM127701. Computational infrastructure was supported in part by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute, by the National Science Foundation under Grant No. CNS-0521433, and by Shared University Research grants from IBM, Inc., to Indiana University. The CAGI experiment is supported by National Institute of Health grant U24 HG007346. Any opinions, findings and conclusions, or recommendations expressed in this material are those of the authors, and do not necessarily reflect the views of any of these organizations.

Author Contributions
E.F. R.J., and A.K. planned the research, E.F. conducted the research and wrote the main manuscript text and E.F. prepared figure 1. All authors reviewed the manuscript.

Declarations

Competing Interests
The authors declare no competing interests.

Received: 15 June 2023 Accepted: 8 August 2024
Published online: 27 August 2024

References

1. Chin IS, Khan A, Olsson-Brown A, Papa S, Middleton G, Palles C. Germline genetic variation and predicting immune checkpoint inhibitor induced toxicity. npj Genomic Med. 2022;7(1):73.
2. Keller J, Gomez R, Williams G, Lembke A, Lazzeroni L, Murphy GM, Schatzberg AF. HPA axis in major depression: cortisol, clinical symptomatology and genetic variation predict cognition. Mol Psychiatry. 2017;22(4):527–36.
3. Battey CJ, Ralph PL, Kern AD. Predicting geographic location from genetic variation with deep neural networks. Elife. 2020;9: e54507.
4. Okbay A, Beauchamp JP, Fontana MA, Lee JJ, Pers TH, Rietveld CA, Turley P, Chen G-B, Valur Emilsson S, Meddens FW, et al. Genome-wide association study identifies 74 loci associated with educational attainment. Nature. 2016;533(7604):539–42.
5. Marioni RE, Ritchie SJ, Joshi PK, Hagenaars SP, Okbay A, Fischer K, Adams MJ, Hill WD, Davies G, Social Science Genetic Association Consortium, et al. Genetic variants linked to education predict longevity. Proc Natl Acad Sci. 2016;113(47):13366–71.
6. Cheng J, Nguyen TYD, Cygan KJ, Çelik MH, Fairbrother WG, Gagneur J, et al. MMSplice: modular modeling improves the predictions of genetic variant effects on splicing. Genome Biol. 2019;20(1):1–15.
7. Davies RW, Fiksinski AM, Breetvelt EJ, Williams NM, Hooper SR, Monfeuga T, Bassett AS, Owen MJ, Gur RE, Morrow BE, et al. Using common genetic variation to examine phenotypic expression and risk prediction in 22q11. 2 deletion syndrome. Nat Med. 2020;26(12):1912–8.
8. Trépo E, Valenti L. Update on NAFLD genetics: from new variants to the clinic. J Hepatol. 2020;72(6):1196–209.
9. Bouafi H, Bencheikh S, Mehdi Krami AL, Morjane I, Charoute H, Rouba H, Saile R, Benhnini F, Barakat A. Prediction and structural comparison of deleterious coding nonsynonymous single nucleotide polymorphisms (nsSNPs) in human LEP gene associated with obesity. BioMed Res Int. 2019;2019:1832084.
10. Genome Interpretation Consortium et al. Cagi, the critical assessment of genome interpretation, establishes progress and prospects for computational genetic variant interpretation methods. arXiv e-prints, pages arXiv:2205, 2022.
11. Cagi. The critical assessment of genome interpretation, establishes progress and prospects for computational genetic variant interpretation methods. Genome Biol. 2024;25(1):53.
12. Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J. Critical assessment of methods of protein structure prediction(CASP)-Round XIV. Proteins Struct Funct Bioinform. 2021;89(12):1607–17.
13. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–9.
14. Baek M, Baker D. Deep learning and protein structure modeling. Nat Methods. 2022;19(1):13–4.
15. Karchin R, Diekhans M, Kelly L, Thomas DJ, Pieper U, Eswar N, Haussler D, Sali A. LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. Bioinformatics. 2005;21(12):2814–20.
16. Bao L, Cui Y. Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information. Bioinformatics. 2005;21(10):2185–90.
17. Dobson RJ, Munroe PB, Caulfield MJ, Saqi MAS. Predicting deleterious nsSNPs: an analysis of sequence and structural attributes. BMC Bioinform. 2006;7(1):217.
18. Ng PC, Henikoff S. Predicting the effects of amino acid substitutions on protein function. Annu Rev Genomics Hum Genet. 2006;7:61–80.
19. Care MA, Needham CJ, Bulpitt AJ, Westhead DR. Deleterious SNP prediction: be mindful of your training data! Bioinformatics. 2007;23(6):664–72.
20. Cooper GM, Shendure J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat Rev Genet. 2011;12(9):628–40.
21. Tian J, Ningfeng W, Guo X, Guo J, Zhang J, Fan Y. Predicting the phenotypic effects of non-synonymous single nucleotide polymorphisms based on support vector machines. BMC Bioinform. 2007;8(1):450.
22. Teng S, Michonova-Alexova E, Alexov E. Approaches and resources for prediction of the effects of non-synonymous single nucleotide polymorphism on protein function and interactions. Curr Pharm Biotechnol. 2008;9(2):123–33.
23. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the sift algorithm. Nat Protoc. 2009;4(7):1073.
24. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248–9.
25. Huang T, Wang P, Ye Z-Q, Heng X, He Z, Feng K-Y, LeLe H, Cui WR, Wang K, Dong X, et al. Prediction of deleterious non-synonymous SNPs based on protein interaction network and hybrid properties. PLoS ONE. 2010;5(7): e11900.
26. Capriotti E, Altman RB. Improving the prediction of disease-related variants using protein three-dimensional structure. BMC Bioinform. 2011;12(S4):S3.
27. Capriotti E, Altman RB. A new disease-specific machine learning approach for the prediction of cancer-causing missense variants. Genomics. 2011;98(4):310–7.
28. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. PLoS ONE. 2012;7(10): e46688.

29. Lopes MC, Joyce C, Ritchie GRS, John SL, Cunningham F, Asimit J, Zeggini E. A combined functional annotation score for non-synonymous variants. Hum Hered. 2012;73(1):47–51.
30. Wu J, Jiang R. Prediction of deleterious nonsynonymous single-nucleotide polymorphism for human diseases. Sci World J. 2013;2013: 675851.
31. Dakal TC, Kala D, Dhiman G, Yadav V, Krokhotin A, Dokholyan NV. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms in il8 gene. Sci Rep. 2017;7(1):1–18.
32. Desai M, Chauhan JB. Computational analysis for the determination of deleterious nsSNPs in human MTHFR gene. Comput Biol Chem. 2018;74:20–30.
33. Desai M, Chauhan JB. Predicting the functional and structural consequences of nsSNPs in human methionine synthase gene using computational tools. Syst Biol Reprod Med. 2019;65(4):288–300.
34. Ponzoni L, Peñaherrera DA, Oltvai ZN, Bahar I. Rhapsody: predicting the pathogenicity of human missense variants. Bioinformatics. 2020;36(10):3084–92.
35. Peng Y, Alexov E. Investigating the linkage between disease-causing amino acid variants and their effect on protein stability and binding. Proteins Struct Funct Bioinform. 2016;84(2):232–9.
36. Tang H, Thomas PD. Tools for predicting the functional impact of nonsynonymous genetic variation. Genetics. 2016;203(2):635–47.
37. Van Rappard DF, Boelens JJ, Wolf NI. Metachromatic leukodystrophy: disease spectrum and approaches for treatment. Best Pract Res Clin Endocrinol Metab. 2015;29(2):261–73.
38. Faraggi E, Zhou Y, Kloczkowski A. Accurate single-sequence prediction of solvent accessible surface area using local and global features. Proteins Struct Funct Bioinform. 2014;82(11):3170–6.
39. Eisenberg D, Weiss RM, Terwilliger TC. The hydrophobic moment detects periodicity in protein hydrophobicity. Proc Natl Acad Sci. 1984;81(1):140–4.
40. Orlin Ch Ivanov and Berthold Förtsch. Universal regularities in protein primary structure: preference in bonding and periodicity. Orig Life Evol Biosph. 1986;17(1):35–49.
41. Rackovsky S. "hidden" sequence periodicities and protein architecture. Proc Natl Acad Sci. 1998;95(15):8580–4.
42. Marsella L, Sirocco F, Trovato A, Seno F, Tosatto SCE. Repetita: detection and discrimination of the periodicity of protein solenoid repeats by discrete Fourier transform. Bioinformatics. 2009;25(12):i289–95.
43. Rackovsky S. Global characteristics of protein sequences and their implications. Proc Natl Acad Sci. 2010;107(19):8623–6.
44. Rackovsky S. Sequence determinants of protein architecture. Proteins Struct Funct Bioinform. 2013;81(10):1681–5.
45. Scheraga HA, Rackovsky S. Homolog detection using global sequence properties suggests an alternate view of structural encoding in protein sequences. Proc Natl Acad Sci. 2014;111(14):5225–9.
46. Meiler J, Müller M, Zeidler A, Schmäschke F. Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks. Mol Model Annu. 2001;7(9):360–9.
47. Zhou Y, Faraggi E. Prediction of one-dimensional structural properties of proteins by integrated neural networks. In: Rangwala H, Karypis G, editors. Introduction to protein structure prediction: methods and algorithms. Hoboken: Wiley; 2010. p. 45–74.
48. Faraggi E, Kloczkowski A. Genn: a general neural network for learning tabulated data with examples from protein structure prediction. In: Artificial Neural Networks. Berlin: Springer; 2015. p. 165–78.
49. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. Clinvar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014;42(D1):D980–5.
50. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, Baoshan G, Hart J, Hoffman D, Hoover J, et al. Clinvar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 2016;44(D1):D862–8.
51. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, Baoshan G, Hart J, Hoffman D, Jang W, et al. Clinvar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018;46(D1):D1062–7.
52. Landrum MJ, Chitipiralla S, Brown GR, Chen C, Baoshan G, Hart J, Hoffman D, Jang W, Kaur K, Liu C, et al. Clinvar: improvements to accessing data. Nucleic Acids Res. 2020;48(D1):D835–44.
53. Boeckmann B, Bairoch A, Apweiler R, Blatter M-C, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, et al. The SWISS-PROT protein knowledgebase and its supplement trEMBL in 2003. Nucleic Acids Res. 2003;31(1):365–70.
54. UniProt Consortium. Uniprot: the universal protein knowledgebase in 2021. Nucleic acids research. 2021;49(D1):D480–9.
55. Choi Y. A fast computation of pairwise sequence alignment scores between a protein and a set of single-locus variants of another protein. In: Proceedings of the ACM conference on bioinformatics, computational biology and biomedicine. 2012. pp. 414–417.
56. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using polyphen-2. Curr Protoc Hum Genet. 2013;76(1):7–20.

Publisher's Note