IEEE EMBS Int Conf Biomed Health Inform, 2019, Accepted

Structure-based Method for Predicting Deleterious Missense SNPs

Boshen Wang, Wei Tian, Xue Lei, Alan Perez-Rathke, Yan Yuan Tseng and Jie Liang

***Abstract*— Missense SNPs are key factors contributing towards many Mendelian disorders and complex diseases. Identifying whether a single amino acid substitution will lead to pathological effects is important for interpreting personal genome and for precision medicine. In this study, we describe a novel method for predicting whether a missense SNP likely brings about pathological effects. Our approach integrates sequence information, biophysical properties, and topological properties of protein structures. In our test dataset consisting of 500 deleterious variants and 500 neutral, our method achieves an accuracy of 0.823. The ROC curve of model has an AUC of 0.910. Our methods outperforms two well known methods, and is comparable with the widely used Polyphen-2 method, while requiring a much smaller amount (approximately 25%) of training data. Our method can be used to aid in distinguishing driver and passenger mutations in cancer and in assessing missense mutations assocaited with rare diseases. It can also be used to identifying mutations in rare disease where only limited patient exome data exsit.**

I. INTRODUCTION

In cells, proteins carry out essential functions such as DNA replication, signal transduction, metabolic catalysis, and molecular transport. Missense mutations may affect protein function and lead to a pathological phenotype [1]. A well-known missense variant is the V600E/K substitution in the BRAF gene. This variant has been confirmed as the driver mutation in several cancer types [2]. The mutated B-Raf protein deregulates activation of the downstream MEK/ERK effectors, contributing to uncontrolled cellular growth [3]. Several FDA-approved drugs, including Vemurafenib [4] and Dabrafenib [5], have already been developed that target the effects of this mutation.

In general, an effective prediction tool that can identify deleterious missense SNPs can help to discriminate driver mutation from passenger mutations in heterogeneous cancer genome data. It can also help to identify residues essential for maintaining enzyme function. Several pathological SNP prediction tools based on sequence analysis have already been developed. The Polyphen-2 method is based on multiple sequence alignment and limited protein structural information [6]. The FATHMM method builds a hidden Markov model based on profiles of protein families [7]. The CHASM method combines sequence information, clinical data, and predictive values to identify cancer driver mutations [8]. The PMUT method is a neural network classifier based on sequence conservation information and predicted physicochemical properties [9]. A more recent study named RAPSODY utilizes sequence features and elastic network models from the corresponding protein's 3-D stuctural coordinates to determine functional significant missense variants [10].

However, these methods requires large training dataset, which may not be applicable for ceratin rare diseases. In this study, we describe a method for predicting whether a missense SNP likely brings about pathological effects. Our approach integrates sequence information, biophysical properties, and topological descriptors of protein structures. Our method outperformes several methods, including FATHMM and PMUT. With a much smallar training dataset (~25%) the performance of our method is comparable with that of Polyphen-2, indicating that our method can be useful in identifying mutations in rare disease where only limited patient exome data exsit.

II. METHODS

To avoid overfitting, we choose the random forest classifier [11] to estimate the effect by a missense variant, as many other machine learning methods tend to overfit easily. This part provides description of our method, including the features implemented and model setting, Figure 1 provides an overview of our method.

A. Model Design

The random forest classifier is used to predict whether a missense variant will lead to pathological effect, when presented with the set of input features of the residue of interest. As an ensemble learning method, random forest will generate plenty of decision trees, and the predicted class is labelled by the majority votes.

We use the "randomForest" R package [12]. We choose stratified resampling to overcome potential biases brought by slightly imbalanced nature of the dataset. As the result, the training dataset contains an equal number of positive (deleterious) and negative (neutral) variants to each tree in the random forest. We set the random forest consisting 2,000 individual trees.

This work was supported by NIH grants

B. Wang is with the Bioinformatics Program, Department of Bioengineering, University of Illinois at Chicago, Chicago, IL 60607, USA. bwang54@uic.edu

W. Tian is with the Bioinformatics Program, Department of Bioengineering, University of Illinois at Chicago, Chicago, IL 60607, USA. wtain7@uic.edu

X. Lei is with the Bioinformatics Program, Department of Bioengineering, University of Illinois at Chicago, Chicago, IL 60607, USA. xlei4@uic.edu

A. Perez-Rathke is with the Bioinformatics Program, Department of Bioengineering, University of Illinois at Chicago, Chicago, IL 60607, USA. perezrat@uic.edu

Y. Tseng is with the Molecular Medicine and Genetics, Wayne State University, Detroit, MI 48201, USA. ytseng@wayne.edu

J. Liang is with the Bioinformatics Program, Department of Bioengineering, University of Illinois at Chicago, Chicago, IL 60607, USA. jliang@uic.edu

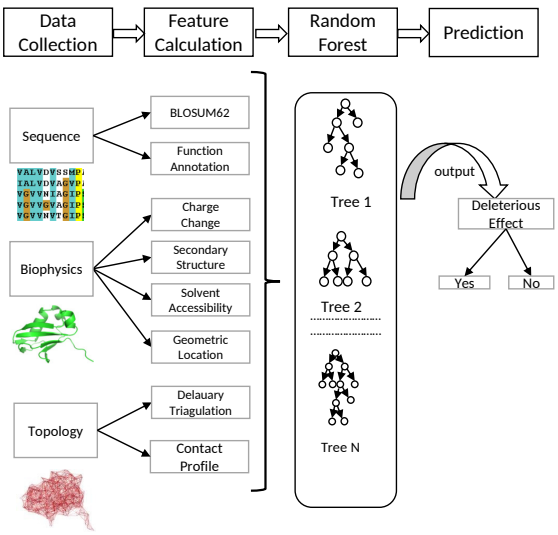

Fig. 1. Prediction of deleterious missense SNP using random forest.

B. Data Collection

We use the HumDiv dataset from Polyphen-2 as the gold-standard. HumDiv contains a collection of deleterious missense mutations causing Mendelian diseases from the UniProtKB database. HumDiv also contains neutral variants obtained from multiple sequence alignment of human proteins with related mammalian homologs, where the human-specific mutated sites not in UniProtKB are assumed to be neutral [6].

C. Structure Retrieval and Residue Mapping

While there have been drastic improvement in techniques such as X-ray, NMR and cryo-EM for determining protein structures, many human proteins still have no known structures. Furthermore, sequence isoforms and experimental artifacts complicate the mapping between sequence and structure.

In this study, we use two sequence alignment steps to map protein sequence to the corresponding structures at the residue-level. We first use BLASTP [13] for local alignment to query a sequence against the whole PDB database [14] and obtain candidate structures. We then use CLUSTALW [15] to generate pairwise alignments between the initial query sequence and the set of potential candidates. If the two-way sequence identity is above 80% after CLUSTAL alignment, the candidate structure is selected, and residue-level mapping information derived. If there are multiple potential candidates, we select the candidate with highest sequence identity, with random selection in the case of ties.

After this mapping process, our dataset consists of 2,106 deleterious mutations and 2,345 neutral mutations from 240 proteins. This dataset is only approximately 25% of the original HumDiv training dataset used in Polyphen-2 [6]. This more stringent selection procedure will help mitigate bias that would otherwise be present when using structures with lower sequence identity.

D. Substitution Score

In previous studies, amino acid substitution matrices, site-specific conversation scores, and hidden Markov probability models are the most widely-used features [6]–[8], [10], [16]. In our study, we simply use the substitution score from the BLOSUM62 matrix [17].

E. Function Annotation

Many proteins interacts with other proteins or ligand molecules to catalyze biochemical reactions. We assume that the functional sites or regions where such interactions occur are more likely to be intolerant to substitution. For functional features, we extract wild-type residue functional annotations from UniProt [18].

Specifically, we incorporate two classes of functional features: functional sites and functional regions. Functional sites are categorical features based on the annotations of active site, binding site, glycosilation site, metal ion-binding site, initiator methionine, and splice variant. Functional regions are also categorical features based on the annotations of short sequence motif, topological domain, signal peptide, nucleotide phosphate-binding region, lipid moiety-binding region, calcium-binding region, transmembrane region, intramembrane region, and zinc finger region.

F. Change of Side-chain Charge

Side-chain charge may contribute to the formation or disruption of disulfide bonds, salt bridgse, metal-ion bonding, and hydrogen bonding. We assign basic residues His, Arg, and Lys a positive charge (+1). We assign acid residues Asp and Glu a negative charge (–1). All other residues are neutral (0).

Changes in side-chain charge may disrupt electrostatic equilibrium and influence enzyme behavior. At each residue substitution, we tally the change of side-chain charge between mutant and wild types, with possible values in the range of { -2, -1, 0, 1, 2.}.

G. Protein Secondary Structure

We also incorporate information on the secondary structure associated with the wild-type residue. Protein secondary structure such as α-helices and β-sheets have characteristic patterns of hydrogen bonds among the backbone atoms. We use the DSSP [19] method to assign a residue to one of the following secondary structure categories: α-helix, 3_{10}-helix, π-helix, β-sheet, hydrogen-bonded turn, or loop.

H. Solvent Accessibility and Geometric Location

Solvent accessible (SA) surface area quantifies the amount of atomic surface exposed to solvent. This important property can convey information regarding hydrophobic versus hydrophilic trends. We use the CASTp [20] method for calculating the SA of the wild-type residue.

In addition to SA, CASTp also provides topological information and regional location of each wild-type residue. Specifically, CASTp classifies location of residues into three types of locations: buried, pocket, and surface. Buried

residues are at the hydrophobic core of the protein with no SA. Pocket residues are located on solvent accessible concavities within the protein. All other residues are considered to be located on the surface. Our model uses SA as a numeric feature and regional location as a categorical feature.

I. Contact Profile

Quantifying the surrounding environment of a residue based on Euclidean distance is a well-known approach in structural bioinformatics research, for instance, in prediction of free energy changes from point mutation [21]. The Mutation3D method detects hotspot region for cancer mutation based on bootstrap sampling of Euclidean distance [22].

In this study, we calculate a three-layer nearby-residue composition by the star-operator derived from the simplicies obtained from the weighted Delanuay triangulation of the protein structure. An alpha shape is also used to trim the original Delanuay triangulation. In our model, each layer is represented as a 20 element integer vector consisting of the number of residues of each each type present within that layer.

We adopt a breadth-first search (BFS) algorithm to ientify residues within three layers by Delaunay edges. For the wild-type residue of interest, we iteratively find the surrounding residues which share edges with it, which is classified as the first layer. Then for each residue in the first layer, we detect the nearby residues connected to it to obtain residue information on second layer, excluding residue of interest and other residues in the same layer. The same procedure is applied for identifying residues in the third layer.

For comparison, we also built a conventional three-layer feature set based on Euclidean distance from residues within 3, 4, and 5 Å respectively. We found the star-shape model provides an 2% improvement in accuracy.

III. RESULTS AND DISCUSSION

A. Performance

We build a test dataset consisting of 500 deleterious variants and 500 neutral mutations from HumDiv through random selection without replacement from mutations not used in the training set. We use the default threshold probability of 0.5 as the threshold to distinguish deleterious and neutral mutation.

Overall, our methods perform well in distinguishing deleterious variants from neutral variants, with an accuracy of 0.822, a recall value of 0.820, and a precision value of 0.823 (Table I). The AUC of our method is 0.910, as shown in the ROC curve 2. High AUC value indicates that our model give good performance with different choices of threshold.

We also compare our performance with several leading methods in predicting pathological missense mutations, including FATHMM [7], Polyphen-2 [6], and PMUT [9]. Table I provides detailed performance information.

FATHMM [7] has the highest recall, but gives many false positives, with a very poor specificity of 0.346 compared to 0.824 of our method. That is, FATHMM cannot recoginze

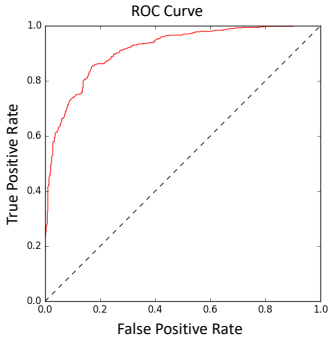

Fig. 2. Receiver operating characteristic (ROC) curve.

neutral variant reliably. Given the heterogeneous nature of cancer exome data, neutral mutations will constitute most proportion of observed variants. Therefore, FATHMM is unlikely to be able to predict driver mutations reliably.

Although PMUT [9] exhibits balanced performance on identifying deleterious and neutral mutations, our method outperforms PMUT signifantly in all measures (Table I).

Polyphen-2 is a well-established method [6] and has the best overall performance, although the difference with our method is small (0.910 in AUC for our method and 0.941 for Polyphen-2). However, our training data is only about 25% of Polyphen-2 (240 vs. 978 proteins).

Overall, our method out-performas several existing methods, and have comparable performance with the leading method of Polyphen-2, while requiring a much smaller training data set. These results suggest that our method can be useful in identifying mutations in rare disease carried by small poplution, where the exome data is more limited.

TABLE I

	Our model	FATHMM	PMUT	Polyphen-2
Accuracy	0.822	0.624	0.727	0.877
Recall	0.820	0.901	0.752	0.893
Precision	0.823	0.579	0.716	0.866
Specificity	0.824	0.346	0.702	0.862
F1 Score	0.821	0.705	0.734	0.879

B. Discussion

While our model can discriminates pathological missense mutations from neutral mutations as shown in the ROC curve (figure 2), and compares well with other methods, there are several aspects our model can be further improved. For example, it is likely that once more human protein structures are available and our stringently derived training data is significantly enlarged, our method will likely perform better. Below we discuss two additional aspects where our method can be imrpoved.

Recent studies have found that epigenetic modifications also impact phenotype [23], [24]. A variant of FATHMM, called FATHMM-XF, shows improved performance over the base model by incorporating epigenetic data such as histone modification, open chromatin, methylation, and transcription factor binding sites [24]. Incorporating such epigenetic data

will likely further improve our prediction of pathological effects of missense mutations.

Currently, our model predicts pathological effect from missense SNPs only. However, other non-synonymous mutation in exon regions may also be pathological. For example, excessive trinucleotide (CAG) repeats in the HTT gene can lead to Huntingtion's disease [25]. Future extension in considering mutated protein fragments may help to predict effects from insertion and deletion.

IV. CONCLUSION

By integrating features based on sequence, structural, and topological properties of proteins, we have developed a method that can be applied to any missense SNP with known structure. Our method achieves an accuracy of 0.822 in predicting deleterious effects, at a precision value of 0.823. The AUC of the ROC curve of our model is 0.910, suggesting that our model has stong discriminative performance under many different classification thresholds. Furthermore, our method is applicable with reduced training data, and may be useful for predicting effects of missense mutations related to rare diseases.

V. ACKNOWLEDGMENTS

This work is supported by NIH grants R01CA204962-01A1, R35GM127084, and R21 AI126308.

REFERENCES

[1] Z. Wang and J. Moult, "Snps, protein structure, and disease," *Human mutation*, vol. 17, no. 4, pp. 263–270, 2001.

[2] P. A. Ascierto, J. M. Kirkwood, J.-J. Grob, E. Simeone, A. M. Grimaldi, M. Maio, G. Palmieri, A. Testori, F. M. Marincola, and N. Mozzillo, "The role of braf v600 mutation in melanoma," *Journal of translational medicine*, vol. 10, no. 1, p. 1, 2012.

[3] E. W. Joseph, C. A. Pratilas, P. I. Poulikakos, M. Tadi, W. Wang, B. S. Taylor, E. Halilovic, Y. Persaud, F. Xing, A. Viale, *et al.*, "The raf inhibitor plx4032 inhibits erk signaling and tumor cell proliferation in a v600e braf-selective manner," *Proceedings of the National Academy of Sciences*, vol. 107, no. 33, pp. 14903–14908, 2010.

[4] P. B. Chapman, A. Hauschild, C. Robert, J. B. Haanen, P. Ascierto, J. Larkin, R. Dummer, C. Garbe, A. Testori, M. Maio, *et al.*, "Improved survival with vemurafenib in melanoma with braf v600e mutation," *New England Journal of Medicine*, vol. 364, no. 26, pp. 2507–2516, 2011.

[5] A. Hauschild, J.-J. Grob, L. V. Demidov, T. Jouary, R. Gutzmer, M. Millward, P. Rutkowski, C. U. Blank, W. H. Miller Jr, E. Kaempgen, *et al.*, "Dabrafenib in braf-mutated metastatic melanoma: a multicentre, open-label, phase 3 randomised controlled trial," *The Lancet*, vol. 380, no. 9839, pp. 358–365, 2012.

[6] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev, "A method and server for predicting damaging missense mutations," *Nature methods*, vol. 7, no. 4, p. 248, 2010.

[7] H. A. Shihab, J. Gough, D. N. Cooper, P. D. Stenson, G. L. Barker, K. J. Edwards, I. N. Day, and T. R. Gaunt, "Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden markov models," *Human mutation*, vol. 34, no. 1, pp. 57–65, 2013.

[8] W. C. Wong, D. Kim, H. Carter, M. Diekhans, M. C. Ryan, and R. Karchin, "Chasm and snvbox: toolkit for detecting biologically important single nucleotide mutations in cancer," *Bioinformatics*, vol. 27, no. 15, pp. 2147–2148, 2011.

[9] C. Ferrer-Costa, J. L. Gelpí, L. Zamakola, I. Parraga, X. De La Cruz, and M. Orozco, "Pmut: a web-based tool for the annotation of pathological mutations on proteins," *Bioinformatics*, vol. 21, no. 14, pp. 3176–3178, 2005.

[10] L. Ponzoni and I. Bahar, "Structural dynamics is a determinant of the functional significance of missense variants," *Proceedings of the National Academy of Sciences*, p. 201715896, 2018.

[11] A. Liaw, M. Wiener, *et al.*, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[12] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[13] J. Ye, S. McGinnis, and T. L. Madden, "Blast: improvements for better sequence analysis," *Nucleic acids research*, vol. 34, no. suppl_2, pp. W6–W9, 2006.

[14] H. M. Berman, T. Battistuz, T. Bhat, W. F. Bluhm, P. E. Bourne, K. Burkhardt, Z. Feng, G. L. Gilliland, L. Iype, S. Jain, *et al.*, "The protein data bank," *Acta Crystallographica Section D: Biological Crystallography*, vol. 58, no. 6, pp. 899–907, 2002.

[15] J. Thompson, D. Higgins, and T. Gibson, "Clustalw: improving the sensitivity of progressive weighting, position-specific gap penalties and weight matrix choice," *Nucleic Acids Res*, vol. 22, pp. 4673–4680, 1994.

[16] A. González-Pérez and N. López-Bigas, "Improving the assessment of the outcome of nonsynonymous snvs with a consensus deleteriousness score, condel," *The American Journal of Human Genetics*, vol. 88, no. 4, pp. 440–449, 2011.

[17] S. Henikoff and J. G. Henikoff, "Amino acid substitution matrices from protein blocks," *Proceedings of the National Academy of Sciences*, vol. 89, no. 22, pp. 10915–10919, 1992.

[18] E. Boutet, D. Lieberherr, M. Tognolli, M. Schneider, and A. Bairoch, "Uniprotkb/swiss-prot," in *Plant bioinformatics*, pp. 89–112, Springer, 2007.

[19] W. Kabsch and C. Sander, "Dssp: definition of secondary structure of proteins given a set of 3d coordinates," *Biopolymers*, vol. 22, pp. 2577–2637, 1983.

[20] W. Tian, C. Chen, X. Lei, J. Zhao, and J. Liang, "Castp 3.0: computed atlas of surface topography of proteins," *Nucleic acids research*, 2018.

[21] D. E. Pires, D. B. Ascher, and T. L. Blundell, "mcsm: predicting the effects of mutations in proteins using graph-based signatures," *Bioinformatics*, vol. 30, no. 3, pp. 335–342, 2013.

[22] M. J. Meyer, R. Lapcevic, A. E. Romero, M. Yoon, J. Das, J. F. Beltrán, M. Mort, P. D. Stenson, D. N. Cooper, A. Paccanaro, *et al.*, "mutation3d: cancer gene prediction through atomic clustering of coding variants in the structural proteome," *Human mutation*, vol. 37, no. 5, pp. 447–456, 2016.

[23] A. Portela and M. Esteller, "Epigenetic modifications and human disease," *Nature biotechnology*, vol. 28, no. 10, p. 1057, 2010.

[24] M. F. Rogers, H. A. Shihab, M. Mort, D. N. Cooper, T. R. Gaunt, and C. Campbell, "Fathmm-xf: accurate prediction of pathogenic point mutations via extended features," *Bioinformatics*, vol. 34, no. 3, pp. 511–513, 2017.

[25] H. Telenius, B. Kremer, Y. P. Goldberg, J. Theilmann, S. E. Andrew, J. Zeisler, S. Adam, C. Greenberg, E. J. Ives, L. A. Clarke, *et al.*, "Somatic and gonadal mosaicism of the huntington disease gene cag repeat in brain and sperm," *Nature genetics*, vol. 6, no. 4, p. 409, 1994.