ePat: extended PROVEAN annotation tool

Takumi Ito¹

Kazutoshi Yoshitake¹

Takeshi Iwata²

¹ The Univresity of Tokyo

² National Institute of Sensory Organs, National Hospital Organization Tokyo Medical Center

Abstract

The 'ePat' (extended PROVEAN annotation tool) is a software tool that extends the functionality of PROVEAN: a software tool for predicting whether amino acid substitutions and indels will affect the biological function of proteins. The 'ePat' extends the conventional PROVEAN to enable the following two things, which the conventional PROVEAN could not calculate the pathogenicity of these variants. First is to calculate and score the pathogenicity of indel mutations with frameshift and variants near splice junctions. Second is to use batch processing to calculate the pathogenicity of multiple variants into a variants list (VCF file) in a single step. ePat can help extract variants that affect biological functions by utilizing not only point mutations, and indel mutations that does not cause frameshift, but also frameshift, stop gain, and splice variants. These extended features will increase detection rate and improve diagnostic of inherited diseases or associate specific variant to phenotype.

1. Introduction

In recent years, improvements in sequencing technology have generated a large amount of information on mutations in patients with inherited diseases or organisms with specific phenotype compared to normal. Mutations such as amino acid substitutions, insertions, deletions, frameshifts, and acquisition of stop codons significantly affect protein function [1] [2] [3] [4] [5] [6].

Computational tools such as Polyphen [7], SIFT [8], PROVEAN [9], and other prediction tools [10] have been developed to search for variants that cause diseases or specific phenotypes, and to quantify the impact of variants on protein function (pathogenicity). PROVEAN made it enable to predict the pathogenicity of variants not limited to single amino acid substitutions but also in-frame insertions, deletions, and multiple amino acid substitutions [11]. However,

these tools have two problems: first, they are unable to calculate variants that are considered to have a significant impact on protein function, such as frameshift and stop gain, and second, they cannot calculate the pathogenicity of variants that spread across multiple genes in a single step.

We have developed 'ePat' (extended PROVEAN annotation tool), an extension of PROVEAN that solves the above two problems. Unlike existing tools, ePat is able to calculate the pathogenicity of variants near the splice junction, frameshift, stop gain, and start lost. In addition, batch processing is used to calculate the pathogenicity of all variants in a VCF file in a single step.

2. Method

With the given reference, a database for SnpEff is created and input vcf file is annotated with SnpEff. We then extract variants that have a HIGH or MODERATE pathogenicity level as a result of the SnpEff annotation. For each row of the VCF file, information of the variants annotated with SnpEff (INFO column) was extracted. With this information, the variants are classified into 5 categories. The first is a variant near the splice junction (splice variants), the second is frameshift, the third is stop gain, the fourth is start lost, and the last is in-frame variants (point mutation or indel mutations that do not cause frameshift).

Variants from category 1 to 4 are given pathogenicity as defined by ePat, and category 5 is given pathogenicity by PROVEAN. The pathogenicity defined by ePat is calculated with the following method. For each amino acid, calculate the pathogenicity when it is replaced by each of the 20 amino acids. The average of these pathogenicity is used as the pathogenicity for that position. The maximum pathogenicity for each position is the pathogenicity of this variants. The details of the algorithm are described below (Figure1).

2-1. Input file

The input data is a VCF file after variant call (Table 1), a FASTA file of the reference genome, and a GTF file with gene annotations. These input files are given as arguments to the ePat runtime.

2-2. SnpEff Annotation

With given reference, we create a database for SnpEff and annotate the VCF file with SnpEff. We then extract variants that have a HIGH or MODERATE pathogenicity level as a result of the SnpEff annotation.

2-3. Extract Variant Information

For each row of the VCF file, extract the information of the variants annotated with SnpEff ([gene ID, variant type, pathogenicity level, DNA mutation information, amino acid mutation information]) from the INFO column. With this information, the variants are classified into (1) variants near the splice junction (splice variants), (2) frameshift, (3) Stop Gain, (4) Start Lost, and (5) in-frame variants (point Mutation or indel mutations that do not cause frameshift).

2-4. Calculate pathogenicity

Variants from (1) to (4) are given pathogenicity as defined by ePat, and those (5) will be given pathogenicity by PROVEAN. The pathogenicity defined by ePat is calculated with the following method.

For each position, calculate the pathogenicity when it is replaced by each of the 20 amino acids. The average of these pathogenicity is used as the pathogenicity for that position. The maximum pathogenicity for each position is the pathogenicity of this mutation.

If the mutation is (1) mutation near splice junctions, then we calculate the pathogenicity defined by ePat in the range from the splice junction where the mutation occurs to the stop codon.

If the mutation is (2) frameshift, then we calculate the pathogenicity defined by ePat in the range from the amino acid where the frameshift starts to the stop codon.

If the mutation is (3) stop codon, then we calculate the pathogenicity defined by ePat in the range from the amino acid to be replaced by the stop codon to the original stop codon.

For Stop Lost, the pathogenicity is not calculated.

If the mutation is (4) start lost, then we calculate the pathogenicity defined by ePat in the range from the original start codon to the next methionine.

If the mutation is (5) in-frame variant, then we calculate the pathogenicity by PROVEAN.

3. Result and Discussion

After the calculation is completed, the PROVEAN_score and PROVEAN_pred columns are added to the Input file. In the PROVEAN_score column, the pathogenicity calculated by ePat or PROVEAN is described. In the PROVEAN_pred column, if the pathogenicity is less than -2.5, then D (Damaged) is described, meaning that the mutation is likely to affect amino acid function. If the pathogenicity is greater than -2.5, then N (Not Damaged) is described, meaning that the mutation is unlikely to affect amino acid function. The output is named as output_provean_{PREFIX_OF_YOUR_INPUTFILE}.txt and saved in the output directory (Table

2).

Lists of SNVs/INDELs with rough PROVEAN annotations for each family were created as before [6]. Of the number of SNVs/INDELs in this list that were not annotated by PROVEAN, we measured how many SNVs/INDELs were annotated with ePat. We created 1216 files and detected 491.5 SNVs/INDELs per family, of which 189.54 SNVs/INDELs per family were not annotated with conventional PROVEAN. Of the 189.54 SNVs/INDELs that were not annotated, we were able to annotate 151.63 SNVs/INDELs per family using ePat (Table 3).

In the past, PROVEAN was only able to annotate 61% of the total SNVs/INDELs, and the rest of the data was not usable. Therefore, we developed ePat, which can calculate the toxicity of indel mutations with frameshift and mutations near splice junctions. As a result, we were able to annotate 92% of the total SNVs/INDELs by using ePat. In conclusion, ePat can help extract variants that affect biological functions with not only point mutations, and indel mutations that does not cause frameshift, but also frameshift, stop gain, and splice variants.

4. Software and code availability

The source code of ePat and the documentation is available from <https://github.com/itotaku1225/ePat>.

The singularity image and test data is available from <https://zenodo.org/record/5482094#.YZSJtWDP2Uk>.

5. Reference

[1]Kawamura, Y.; Suga, A.; Fujimaki, T.; Yoshitake, K.; Tsunoda, K.; Murakami, A.; Iwata, T. LRRTM4-C538Y Novel Gene Mutation Is Associated with Hereditary Macular Degeneration with Novel Dysfunction of ON-Type Bipolar Cells. *J Hum Genet* 2018, 63 (8), 893–900. <https://doi.org/10.1038/s10038-018-0465-4>.

[2]Hayashi, T.; Mizobuchi, K.; Kameya, S.; Yoshitake, K.; Iwata, T.; Nakano, T. A New PDE6A Missense Variant p.Arg544Gln in Rod–Cone Dystrophy. *Doc Ophthalmol* 2021, 143 (1), 107–114. <https://doi.org/10.1007/s10633-021-09826-y>.

[3]Minegishi, Y.; Sheng, X.; Yoshitake, K.; Sergeev, Y.; Iejima, D.; Shibagaki, Y.; Monma, N.; Ikeo, K.; Furuno, M.; Zhuang, W.; Liu, Y.; Rong, W.; Hattori, S.; Iwata, T. CCT2 Mutations Evoke Leber

Congenital Amaurosis Due to Chaperone Complex Instability. *Sci Rep* 2016, 6, 33742. <https://doi.org/10.1038/srep33742>.

[4] Mizobuchi, K.; Hayashi, T.; Yoshitake, K.; Fujinami, K.; Tachibana, T.; Tsunoda, K.; Iwata, T.; Nakano, T. Novel Homozygous CLN3 Missense Variant in Isolated Retinal Dystrophy: A Case Report and Electron Microscopic Findings. *Mol Genet Genomic Med* 2020, 8 (8), e1308. <https://doi.org/10.1002/mgg3.1308>.

[5] Suga, A.; Mizota, A.; Kato, M.; Kuniyoshi, K.; Yoshitake, K.; Sultan, W.; Yamazaki, M.; Shimomura, Y.; Ikeo, K.; Tsunoda, K.; Iwata, T. Identification of Novel Mutations in the LRR-Cap Domain of C21orf2 in Japanese Patients With Retinitis Pigmentosa and Cone-Rod Dystrophy. *Invest Ophthalmol Vis Sci* 2016, 57 (10), 4255–4263. <https://doi.org/10.1167/iovs.16-19450>.

[6] Yang, L.; Fujinami, K.; Ueno, S.; Kuniyoshi, K.; Hayashi, T.; Kondo, M.; Mizota, A.; Naoi, N.; Shinoda, K.; Kameya, S.; Fujinami-Yokokawa, Y.; Liu, X.; Arno, G.; Pontikos, N.; Kominami, T.; Terasaki, H.; Sakuramoto, H.; Katagiri, S.; Mizobuchi, K.; Nakamura, N.; Mawatari, G.; Kurihara, T.; Tsubota, K.; Miyake, Y.; Yoshitake, K.; Iwata, T.; Tsunoda, K. Genetic Spectrum of EYS-Associated Retinal Disease in a Large Japanese Cohort: Identification of Disease-Associated Variants with Relatively High Allele Frequency. *Sci Rep* 2020, 10, 5497. <https://doi.org/10.1038/s41598-020-62119-3>.

[7] Adzhubei, I. A.; Schmidt, S.; Peshkin, L.; Ramensky, V. E.; Gerasimova, A.; Bork, P.; Kondrashov, A. S.; Sunyaev, S. R. A Method and Server for Predicting Damaging Missense Mutations. *Nat Methods* 2010, 7 (4), 248–249. <https://doi.org/10.1038/nmeth0410-248>.

[8] Kumar, P.; Henikoff, S.; Ng, P. C. Predicting the Effects of Coding Non-Synonymous Variants on Protein Function Using the SIFT Algorithm. *Nat Protoc* 2009, 4 (7), 1073–1081. <https://doi.org/10.1038/nprot.2009.86>.

[9] Choi, Y.; Chan, A. P. PROVEAN Web Server: A Tool to Predict the Functional Effect of Amino Acid Substitutions and Indels. *Bioinformatics* 2015, 31 (16), 2745–2747. <https://doi.org/10.1093/bioinformatics/btv195>.

[10] Suybeng, V.; Koeppel, F.; Harlé, A.; Rouleau, E. Comparison of Pathogenicity Prediction Tools on Somatic Variants. *The Journal of Molecular Diagnostics* 2020, 22 (12), 1383–1392. <https://doi.org/10.1016/j.jmoldx.2020.08.007>.

[11]Choi, Y.; Sims, G. E.; Murphy, S.; Miller, J. R.; Chan, A. P. Predicting the Functional Effect of Amino Acid Substitutions and Indels. PLOS ONE 2012, 7 (10), e46688. <https://doi.org/10.1371/journal.pone.0046688>.

6. Figure legends

Figure 1.

Overview of ePat

With the given reference, we create a database for SnpEff and annotate with SnpEff. We then extract variants that have a HIGH or MODERATE pathogenicity level as a result of the SnpEff annotation. For each row of the VCF file, information of the variants annotated with SnpEff was extracted. With this information, the variants are classified into 5 categories. The first is a variant near the splice junction (splice variants), the second is frameshift, the third is stop gain, the fourth is start lost, and the last is in-frame variants (point mutation or indel mutations that do not cause frameshift).

Variants from category 1 to 4 are given pathogenicity as defined by ePat, and category 5 is given pathogenicity by PROVEAN. The pathogenicity defined by ePat is calculated with the following method. For each amino acid, calculate the pathogenicity when it is replaced by each of the 20 amino acids. The average of these pathogenicity is used as the pathogenicity for that position. The maximum pathogenicity for each position is the pathogenicity of this variants. The details of the algorithm are described below.

Table 1 Input overview of ePat

#CHRO	POS	ID	RE	AL	QUA	FILTE	INFO	FORM	HG0009
M			F	T	L	R		AT	6
22	2469829	rs2021421	C	G	100.	PASS	AC=1;AF=0.000199681;AN=5008	GT	0 0
	4	65			0				
22	2470931	rs5480176	G	A	100.	PASS	AC=1;AF=0.000199681;AN=5008	GT	0 0
	7	12			0				
22	2470932	rs5695942	G	C	100.	PASS	AC=1;AF=0.000199681;AN=5008	GT	0 0
	1	48			0				
22	2470942	rs3578391	C	T	100.	PASS	AC=4;AF=0.000798722;AN=5008	GT	0 0
	0	4			0				
22	2470942	rs3717804	A	G	100.	PASS	AC=1;AF=0.000199681;AN=5008	GT	0 0
	3	53			0				

Table 2 Output overview of ePat

#	PO	ID	R	A	Q	FI	INFO	F	H	PR	PRO
C	S		E	L	U	L		O	G	OVE	VEA
H			F	T	A	T		R	00	AN_	N_s
R					L	E		M	09	pre	core
O						R		A	6	d	
M								T			
2	24	rs2	C	G	1	P	AC=1;AF=0.000199681;AN=5008;ANN=G missense_va	G	0	N	-
2	69	021			0	A	riant MODERATE SPECC1L SPECC1L transcript NM_	T	0		1.0
	82	421			0	S	015330 protein_coding 3/17 c.95C>G p.Ser32Cys				05
	94	65			.	S	389/6763 95/3354 32/1117				
					0						
2	24	rs5	G	A	1	P	AC=1;AF=0.000199681;AN=5008;ANN=A missense_va	G	0	N	-
2	70	480			0	A	riant MODERATE SPECC1L SPECC1L transcript NM_	T	0		2.1
	93	176			0	S	015330 protein_coding 4/17 c.190G>A p.Gly64Ar				65
	17	12			.	S	g 484/6763 190/3354 64/1117				
					0						
2	24	rs5	G	C	1	P	AC=1;AF=0.000199681;AN=5008;ANN=C missense_va	G	0	N	-
2	70	695			0	A	riant MODERATE SPECC1L SPECC1L transcript NM_	T	0		0.8
	93	942			0	S	015330 protein_coding 4/17 c.194G>C p.Gly65Al				81
	21	48			.	S	a 488/6763 194/3354 65/1117				
					0						
2	24	rs3	C	T	1	P	AC=4;AF=0.000798722;AN=5008;ANN=T missense_va	G	0	N	-
2	70	578			0	A	riant MODERATE SPECC1L SPECC1L transcript NM_	T	0		1.0
	94	391			0	S	015330 protein_coding 4/17 c.293C>T p.Ser98Ph				51
	20	4			.	S	e 587/6763 293/3354 98/1117				
					0						
2	24	rs3	A	G	1	P	AC=1;AF=0.000199681;AN=5008;ANN=G missense_va	G	0	N	-
2	70	717			0	A	riant MODERATE SPECC1L SPECC1L transcript NM_	T	0		0.3
	94	804			0	S	015330 protein_coding 4/17 c.296A>G p.Lys99Ar				88
	23	53			.	S	g 590/6763 296/3354 99/1117				
					0						

Table 3 Statistical information of 1216 families used in the analysis

	average	rate[%]
The Number of SNVs/INDELs	491.5	100
The Number of SNVs/INDELs annotated with PROVEAN	301.95	61
The Number of SNVs/INDELs not annotated with PROVEAN	189.54	39
The Number of SNVs/INDELs newly annotated with ePat	151.63	31
The Number of SNVs/INDELs annotated with ePat	453.58	92