Web-based tool for the annotation of pathological variants on proteins: PMut 2017 update.

Víctor López-Ferrando[1], Xavier de la Cruz[2], Modesto Orozco[1,3] , Josep Ll. Gelpí[1,3]

[1]*Barcelona Supercomputing Center (BSC). Joint Program BSC-CRG-IRB research Program for Computational Biology. Barcelona. Spain.*

[1]victor.lopez.ferrando@bsc.es

[2]*Vall d'Hebron Research Institute, Universitat Autònoma de Barcelona, Barcelona, Spain.*

[2]xavier.delacruz@vhir.org

[3]*Dept. of Biochemistry and Molecular Biomedicine, University of Barcelona. Spain.*

[3]modesto.orozco@irbbarcelona.org, gelpi@ubedu

***Keywords*— Pathology prediction of SAV, Machine learning, Single amino acid variants**

Extended Abstract

Assessing the impact of amino acid mutations in human health is an important challenge in biomedical research. As sequencing technologies are more available, and more individual genomes become accessible, the number of identified variants has dramatically increased. PMut, released back in 2005 [1], has been one of the popular predictors in this field. PMut was a neural-network-based classifier using sequence data to provide a pathology score for point mutations in proteins.

We now release a new, revised, and much more powerful version of PMut. It features PyMut prediction engine, a Python module that includes numerous machine learning capabilities aimed at the analysis of protein variant pathology annotation. We also release PMut2017 predictor, a full update of the PMut predictor based on the SwissVar [2] variation database. It achieves an accuracy of 82% and a Matthews Correlation Coefficient (MCC) of 0.62, and matches the most popular predictors' performance. The engine is implemented in Python using MongoDB engine for data management. It has been adapted to run at the HPC level to cover large scale annotation projects.

A. PMut prediction engine

PMut predictor engine (PyMut), is a Python 3 module, based on the popular scientific computing libraries NumPy (www.numpy.org), Scipy (www.scipy.org), Pandas (pandas.pydata.org), Scikit-learn (scikit-learn.org), Matplotlib (matplotlib.org), and Seaborn (seaborn.pydata.org).

PyMut performs all the operations involved in the machine learning process, such as: features computation and distribution analysis, most informative features selection, classifier training and evaluation using different metrics and Receiver Operating Characteristic (ROC) curves, the training and use of predictors.

The source code of PyMut is publicly available at https://gitbhub.com/vlopezferrando/pymut, and can be installed locally from the official Python package repository (https://pypi.python.org/pypi/pymut).

B. PMut2017 predictor

PMut2017 is a predictor trained using the SwissVar [2] variation database (October 2016 release), containing 27,203 disease and 38,078 neutral mutations on 12,141 proteins. 215 features were computed for all these mutations, describing protein interactome information, physical property differences between the wild type and mutated amino acids, and sequence conservation. Conservation features are derived from PSI-Blast [3] searches over UniRef100 and UniRef90 [4] cluster databases and Multiple Sequence Alignments performed by Kalign2 [5]. 12 of these features were selected by an iterative algorithm and used to train a Random Forest classifier. The Random Forest score (between 0 and 1) was analyzed and translated to a statistically meaningful reliability score.

PMut2017 was evaluated using two different approaches. First, we run a classic 10-fold cross-validation with 50% sequence identity exclusive folds (Fig. 1), getting an MCC of 0.62, which increases to 0.69 and 0.77 when keeping the most reliable predictions.

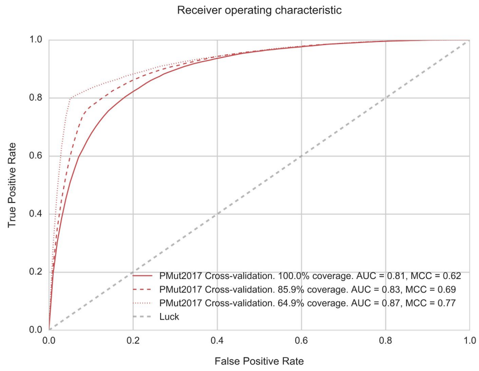


Fig. 1 Receiver Operating Characteristic (ROC) curve of a 10-fold cross-validation with 50% sequence identity exclusion between folds.

Secondly, we performed a blind validation using the SwissVar entries added during 2016. For this purpose, we trained a predictor using the SwissVar data of December 2015, and evaluated the predictions of the 1,656 pathological and 1,510 neutral variants on 762 proteins added during the year 2016. We compared the predictions to those of other widely used predictors (Table 1).

Table I

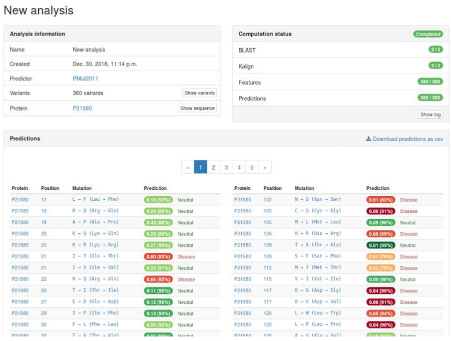
Comparative performance of PMut predictor

Method	% Cov.	Acc.	Spec.	Sens.	AUC	MCC
SIFT	89.6	0.61	0.33	0.88	0.60	0.25
Polyphen2	92.1	0.64	0.35	0.91	0.63	0.32
FATHMM	90.5	0.55	0.45	0.64	0.55	0.09
CADD	95.0	0.65	0.33	0.94	0.64	0.35
M-CAP	91.5	0.60	0.19	0.95	0.57	0.22
Condel	91.0	0.63	0.40	0.84	0.62	0.26
PON-P2	42.4	0.72	0.52	0.9	0.71	0.45
PROVEAN	91.5	0.64	0.41	0.87	0.64	0.31
LRT	95.1	0.73	0.58	0.87	0.73	0.47
MutationTas.	95.1	0.65	0.31	0.96	0.64	0.36
MutationAss .	95.1	0.63	0.46	0.78	0.62	0.26
MetaSVM	95.1	0.63	0.51	0.74	0.62	0.26
MetaLR	95.1	0.6	0.46	0.73	0.60	0.20
PMut	**100.0**	**0.71**	**0.65**	**0.76**	**0.71**	**0.42**
PMut (85%)	**81.0**	**0.76**	**0.76**	**0.77**	**0.76**	**0.53**
PMut (90%)	**51.2**	**0.81**	**0.78**	**0.84**	**0.81**	**0.62**

Coverage, Accuracy, Specificity, Sensitivity, Area Under the ROC Curve and Matthews Correlation Coefficient of the predictions of different methods of the new variants added to SwissVar during 2016 (3,166 mutations).

C. *PMut web portal*

The PMut web portal provides access to all the PMut functionalities. The portal is divided in 3 sections: 1) a data repository, with a set of precalculated features and predictions, 2) single-protein and batch analysis requests and 3) a frontend to the PyMut engine allowing the user to train their own custom predictors. Figures 2 and 3 show screenshots of the interface.



New analysis

Analysis information

Name	New analysis
Created	Dec. 30, 2016, 11:14 p.m.
Predictor	PMut2017
Variants	360 variants
Protein	P21580

Computation status Completed

BLAST	2 / 2
Kalign	2 / 2
Features	360 / 360
Predictions	360 / 360

Predictions

Protein	Position	Mutation	Prediction	
P21580	12	L → F (Leu → Phe)	0.18 (93%)	Neutral
P21580	16	R → Q (Arg → Gln)	0.29 (89%)	Neutral
P21580	18	A → P (Ala → Pro)	0.43 (85%)	Neutral
P21580	20	K → Q (Lys → Gln)	0.29 (89%)	Neutral
P21580	20	K → R (Lys → Arg)	0.27 (90%)	Neutral
P21580	21	I → T (Ile → Thr)	0.60 (83%)	Disease
P21580	21	I → V (Ile → Val)	0.25 (91%)	Neutral
P21580	22	R → Q (Arg → Gln)	0.66 (85%)	Disease
P21580	25	T → I (Thr → Ile)	0.11 (95%)	Neutral
P21580	27	E → D (Glu → Asp)	0.12 (95%)	Neutral
P21580	29	I → F (Ile → Phe)	0.13 (94%)	Neutral
P21580	30	F → L (Phe → Leu)	0.20 (92%)	Neutral
P21580	33	T → A (Thr → Ala)	0.07 (96%)	Neutral

Protein	Position	Mutation	Prediction	
P21580	102	N → S (Asn → Ser)	0.61 (83%)	Disease
P21580	103	C → G (Cys → Gly)	0.86 (91%)	Disease
P21580	105	M → L (Met → Leu)	0.09 (96%)	Neutral
P21580	106	H → R (His → Arg)	0.68 (85%)	Disease
P21580	108	T → A (Thr → Ala)	0.01 (99%)	Neutral
P21580	109	S → F (Ser → Phe)	0.51 (79%)	Disease
P21580	112	M → T (Met → Thr)	0.52 (79%)	Disease
P21580	115	V → I (Val → Ile)	0.09 (96%)	Neutral
P21580	117	D → G (Asp → Gly)	0.84 (90%)	Disease
P21580	117	D → V (Asp → Val)	0.86 (91%)	Disease
P21580	120	L → W (Leu → Trp)	0.65 (84%)	Disease
P21580	122	L → P (Leu → Pro)	0.84 (90%)	Disease
P21580	125	A → V (Ala → Val)	0.18 (93%)	Neutral

Fig. 2 Sample prediction analysis.

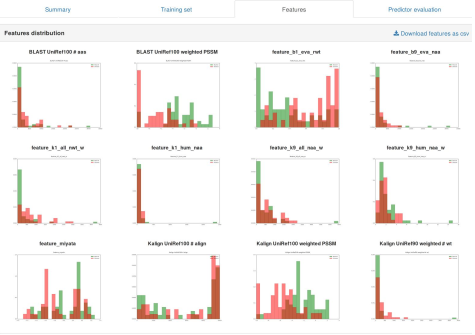


Fig. 3 Distributions of features computed to train a custom predictor.

D. *Conclusions*

The 2017 release of PMut includes an up to date predictor engine which matches the performance of state-of-the-art predictors, allows an easy access to all its capabilities via an intuitive web interface and offers a full set of tools bundled in the PyMut module to apply machine learning methods to the prediction of protein mutation pathology.

E. *ACKNOWLEDGEMENT*

We thank Dr. Patrick Aloy for providing protein interactome data. We thank Jose A. Alcantara, and Dr. Adam Hospital, for technical assistance.

References

[1] Ferrer-Costa, et al. «PMUT: A Web-Based Tool for the Annotation of Pathological Mutations on Proteins». Bioinformatics 21, num. 14 (15 July 2005): 3176-78.

[2] Mottaz, et al. «Easy Retrieval of Single Amino-Acid Polymorphisms and Phenotype Information Using SwissVar». Bioinformatics 26, num. 6 (15 March 2010): 851-52.

[4] Altschul, et al. «Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs». Nucleic Acids Research 25, num. 17 (9 January 1997): 3389-3402.

[3] Suzek, et al. «UniRef: Comprehensive and Non-Redundant UniProt Reference Clusters». Bioinformatics 23, num. 10 (15 May 2007): 1282-88.

[5] Timo Lassmann, Oliver Frings, and Erik L. L. Sonnhammer, "Kalign2: High-Performance Multiple Alignment of Protein and Nucleotide Sequences Allowing External Features," *Nucleic Acids Research* 37, no. 3 (January 2, 2009): 858–65, doi:10.1093/nar/gkn1006.

Author biography

Víctor López Ferrando was born in Castelló de la Plana in 1990. He studied Mathematics and Computer Engineering in Universitat Politècnica de Catalunya, having an special interest in algorithmics. His degree thesis was titled "Topology and Time Synchronization algorithms in wireless sensor networks". After finishing his studies, he worked as a software engineer for two years at Talaia Networks, a spin-off of the Department of Computer Architecture of UPC. In September 2014 he got a La Caixa Severo Ochoa Fellowship and started his PhD in Bioinformatics in the BSC.