P r o c e e d i n g s

of the

2nd International Conference "Theoretical Approaches to BioInformation Systems" (TABIS.2013)

September 17 – 22, 2013, Belgrade, Serbia

Editors

B. Dragovich, R. Panajotović, D. Timotijević

Institute of Physics

Belgrade, 2014, SERBIA

Autor: Grupa autora

Naslov: 2nd INTERNATIONAL CONFERENCE
"THEORETICAL APPROACHES TO BIOINFORMATION
SYSTEMS" (TABIS.2013)
(Druga medjunarodna konferencija "Teorijski pristupi bioinformacionim
sistemima" - TABIS.2013)

Izdavač: Institut za fiziku, Beograd, Srbija

Izdanje: Prvo izdanje

Štampar: Ton Plus, Beograd

Tiraž: 100

ISBN: 978-86-82441-40-3

1. Dragović Branko

CIP – Katalogizacija u publikaciji
Narodna biblioteka Srbije, Beograd

Can We Use Standard Tools to Predict Functional Effects of Missense Gene Variations Outside Conserved Domains? TET2 Example

Branislava Gemović [a]
Vladimir Perović [b]
Sanja Glišić [c]
Nevena Veljković [d]
Centre for Multidisciplinary Research and Engineering,
Vinča Institute of Nuclear Sciences, University of Belgrade, Belgrade, Serbia

Abstract

The most common genetic variations in humans are Single Nucleotide Polymorphisms (SNPs), so predicting their associations with cancers is a significant issue. Here, we were particularly interested in SNPs occurring outside protein Conserved Domains

[a] e-mail address: gemovic@vinca.rs
[b] e-mail address: vladaper@vinca.rs
[c] e-mail address: sanja@vinca.rs
[d] e-mail address: nevenav@vinca.rs

(CDs) of TET2, a recently discovered epigenetic regulator involved in leukemogenesis. Functional effects of TET2 gene variations were assessed with four publicly available tools: PhD-SNP, MutPred, PolyPhen-2 and SIFT. The methods were tested on the dataset of 166 SNPs and somatic TET2 mutations, and separately on the subset of 69 variations outside TET2 CDs. Abilities of tested tools to separate neutral SNPs from pathogenic mutations were similar to previously reported on complete TET2 dataset. However, we observed significantly lower accuracy of predictions outside CDs, ranging from 0.54 to 0.62. Also, areas under the ROC curves were low, 0.51-0.55. Correlations between predictions and positions of variations inside/outside CDs were significant and high, 0.46-0.78. Low efficiency of commonly used tools in predicting functional effects of variations outside CDs emphasize the need for new or modified algorithms.

1 Introduction

The most frequent human genetic variations are SNPs, of which an important subset contains SNPs resulting in the amino acid substitutions (AAS). These mutations play one of the most important roles in cancer transformation [1, 2]. A number of tools have been developed to computationally predict which AAS have relevant phenotypic effect [for review see 3]. In this study we evaluated four widely used tools PhD-SNP [4], MutPred [5], PolyPhen-2 [6] and SIFT [7]. The stated tools use different protein features for predicting pathogenic effects of AAS. SIFT uses only evolutionary information, PhD-SNP combines it with sequence properties, while PolyPhen-2 and MutPred use a number of structural and functional data, in addition.

Several previous studies showed that more than 50% of cancer-associated mutations are positioned outside CDs [8, 9]. Also, extensive analysis of mutations in the important cancer-associated protein family, protein kinases, showed that numerous driver mutations are not in the kinase domains [10]. Nonetheless, performance evaluation of prediction tools has never been specifically focused on the effects of variations outside protein CDs.

TET2 is epigenetic regulator acting as an enzyme, normally converting 5-methylcytosine to 5-hydroxymethylcytosine in DNA [11]. It has been frequently mutated in all types of myeloid malignancies [12]. TET2 mutations predispose hematopoietic stem cells towards uncontrolled self-renewal and consequently development of myeloid malignancies [13, 14]. Even more,

mutations in TET2 are prognostic markers in acute myeloid leukemia [15] and play a role in leukemia transformation [16]. Having two well defined CDs and numerous AAS identified along entire sequence, TET2 represents a good candidate gene for pilot testing on the ability of published computational tools to discriminate between neutral SNPs and pathogenic mutations outside CDs.

2 Materials and Methods

Missense variations in TET2 gene were collected from literature, COSMIC [17] and dbSNP database [18].To label an AAS as a mutation, besides its association with a myeloid malignancy, we looked in original papers for evidence of its somatic nature. There were two criteria to label an AAS as a SNP: first included evidence in original papers of its presence in germline and the second implied described frequency of the polymorphism in healthy population. All-TET2 dataset contained 166 TET2 variations, of which 121 were mutations associated with myeloid malignancies. Also, we constructed a sub-dataset nCD-TET2 from all-TET2 that contained 69 variations outside TET2 CDs, 42 neutral SNPs and 27 mutations. TET2 CDs and non CD regions were determined from the relevant literature [19].

The pathogenicity of TET2 variations were predicted by the tools PhD-SNP [4], MutPred [5], PolyPhen-2 [6] and SIFT [7]. For all tools, we applied default parameters. Contrary to other three tools, PhD-SNP does not give probability scores as a result, so all statistical analyses for this method was done solely on the predictions. PolyPhen-2 and MutPred provide probability scores for a hypothesis that a variation is a damaging mutation and score of 0.5 was used as a predictions threshold. In the case of SIFT, variation is predicted to be a damaging mutation if the probability score is less than 0.05.

The performance of the four tools was assessed by three parameters: accuracy, sensitivity and specificity. For the additional evaluation of prediction tools, we constructed receiver operating characteristic (ROC) curves for both probability scores, where applicable, and predictions. The parameter used was area under the curve (AUC). Correlations between the predictions of tools and position of the variations inside/outside TET2 CDs were calculated using Spearman's rank correlation coefficients. For the determination of the significance of the results, we used chi-square test. The p-values were estimated in a two-tailed fashion. The significance threshold was p-value ≤ 0.01.

3 Results and Discussion

First, we evaluated the performance of PhD-SNP, MutPred, PolyPhen-2 and SIFT in predicting the pathogenicity of missense variants positioned outside TET2 CDs (Table 1). Although accuracies of PhD-SNP and MutPred were somewhat higher than accuracies of PolyPhen-2 and SIFT, the sensitivity and specificity of these tools were quite

	nCD-TET2 dataset			all-TET2 dataset		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
PhD-SNP	0.61	0.04	0.98	0.75	0.67	0.98
MutPred	0.62	0.04	1.00	0.52	0.34	1.00
PolyPhen-2	0.54	0.37	0.64	0.78	0.83	0.62
SIFT	0.55	0.52	0.57	0.78	0.85	0.58

unbalanced. So, we used AUC as additional measure of the performance of these four tools (Fig.1A). PhD-SNP, PolyPhen-2 and SIFT had extremely low AUC values ranging from 0.55 to 0.59 for probability scores and 0.51-0.55 for predictions. MutPred showed high discrepancy between AUC values of its probability scores (AUC=0.68) and predictions (AUC=0.52). This implies that predictions threshold of 0.5, suggested by authors, doesn't represent the optimal value for this particular dataset. But, although higher than for other three tools, performance of MutPred, still, cannot be considered satisfactory.

The accuracy of tested tools predicting pathogenicity of myeloid malignancies-associated variations positioned outside TET2 CDs was shown to be much lower than in the case of more comprehensive datasets, containing mutations not restricted to nCD-regions and originating from various diseases [20, 21]. So, we tested if our findings are specific for the TET2 variations, by evaluating the same tools on the complete all-TET2 dataset. As can be observed from Table 1, PhD-SNP, PolyPhen-2 and SIFT performances were in accordance with previously mentioned studies. Of note, MutPred prediction capacity on the all-TET2 dataset was significantly lower than reported by Thusberg et al. [20] and Li et al. [5]. We are speculating that this is, again, on the account of the predefined prediction threshold which is not appropriate, similarly to the nCD-TET2 dataset. Nevertheless, differences in AUC values for all tested tools between all-TET2 and nCD-TET2 datasets (Fig.1B), also, reflect decrease of their performance when dataset contains only variations outside CDs.

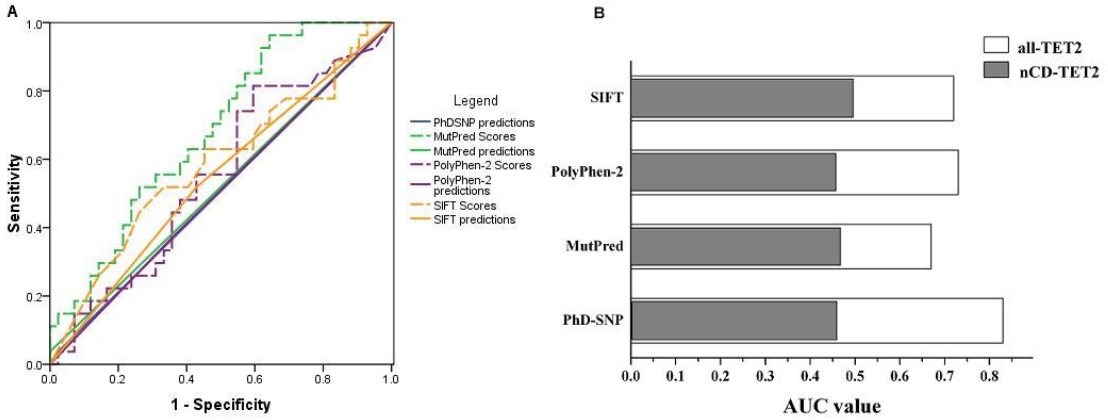

Figure 1: ROC analysis of PhD-SNP, MutPred, PolyPhen-2 and SIFT predictions of pathogenicity of nCD-TET2 variations. **A** ROC curves for probability scores and predictions (only predictions were available for PhD-SNP); **B** Difference between AUC values of predictions on all-TET2 and nCD-TET2 datasets

All tested tools base their predictions on the conservation of the amino acid position in a sequence, so we assumed that their predictions correlate significantly with the position of AAS in TET2 sequence, i.e. whether it is placed in the CD or not. To test this, we compared, pairwise, predictions of each tool and positions of variations in the CD/nCD (Table 2) and observed significant correlations ($p < 0.001$).

	PhD-SNP	MutPred	PolyPhen-2	SIFT
CD/nCD	0.78	0.46	0.65	0.52

Together, our results suggest that tested tools tend to use information about the position of variation in the protein CDs to annotate this variation as a mutation. On TET2 example, this is reflected by the accuracy of 0.95 of PolyPhen-2 and SIFT when we tested variations placed inside CDs (data not shown). But, tendency of these tools to annotate variations outside CDs as neutral SNPs can result in high number of false negatives and this can be the reason for the poor performance on our nCD-TET2 dataset.

In this pilot study, we intended to emphasize the importance of considering the information other than evolutionary in computational tools that predict disease related mutations in complex diseases.

Acknowledgements

This work is supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia (Grant No. 173001).

References

[1] T.J. Ley, E.R. Mardis, L. Ding, B. Fulton, M.D. McLellan, K. Chen, D. Dooling, B.H. Dunford-Shore et al., *Nature* **456** (2008) 66.

[2] E.D. Pleasance, R.K. Cheetham, P.J. Stephens, D.J. McBride, S.J. Humphray, C.D. Greenman, I. Varela, M.L. Lin et al., *Nature* **463** (2010) 191.

[3] D.M. Jordan, V.E. Ramensky and S.R. Sunyaev, *Curr. Opin. Struct. Biol.* **20** (2010) 342.

[4] E. Capriotti, R. Calabrese and R. Casadio, *Bioinformatics* **22** (2006) 2729.

[5] B. Li, V.G. Krishnan, M.E. Mort, F. Xin, K.K. Kamati, D.N. Cooper, S.D. Mooney and P. Radivojac, *Bioinformatics* **25** (2009) 2744.

[6] I.A. Adzhubei, S. Schmidt, L. Peshkin, V.E. Ramensky, A. Gerasimova, P. Bork, A.S. Kondrashov and S.R. Sunyaev, *Nat. Methods.* **7** (2010) 248.

[7] P.C. Ng and S. Henikoff, *Nucleic Acids Res.* **31** (2003) 3812.

[8] P. Yue, W.F. Forrest, J.S. Kaminker, S. Lohr, Z. Zhang and G. Cavet, *Hum. Mutat.* **31** (2010) 264.

[9] T.A. Peterson, N.L. Nehrt, D. Park and M.G. Kann, *J. Am. Med. Inform. Assoc.* **19** (2012) 275.

[10] C. Greenman, P. Stephens, R. Smith, G.L. Dalgliesh, C. Hunter, G. Bignell, H. Davies, J. Teague et al., *Nature* **446** (2007) 153.

[11] M. Tahiliani, K.P. Koh, Y. Shen, W.A. Pastor, H. Bandukwala, Y. Brudno, S. Agarwal, L.M. Iyer et al., *Science* **324** (2009) 930.

[12] F. Delhommeau, S. Dupont, V. Della Valle, C. James, S. Trannoy, A. Massé, O. Kosmider, J.P. Le Couedic et al., *N. Engl. J. Med.* **360** (2009) 2289.

[13] K. Moran-Crusio, L. Reavie, A. Shih, O. Abdel-Wahab, D. Ndiaye-Lobry, C. Lobry, M.E. Figueroa, A. Vasanthakumar et al., *Cancer Cell* **20** (2011) 11.

[14] C. Quivoron, L. Couronné, V. Della Valle, C.K. Lopez, I. Plo, O. Wagner-Ballon, M. Do Cruzeiro, F. Delhommeau et al., *Cancer Cell* **20** (2011) 25.

[15] K.H. Metzeler, K. Maharry, M.D. Radmacher, K. Mrózek, D. Margeson, H. Becker et al., *J. Clin. Oncol.* **29** (2011) 1373.

[16] O. Abdel-Wahab, T. Manshouri, J. Patel, K. Harris, J. Yao, C. Hedvat, A. Heguy, C. Bueso-Ramos et al., *Cancer Res.* **70** (2010) 447.

[17] S.A. Forbes, N. Bindal, S. Bamford, C. Cole, C.Y. Kok, D. Beare, M. Jia, R. Shepherd et al., *Nucleic Acids Res.* **39** (2011) D945.

[18] S.T. Sherry, M.H. Ward, M. Kholodov, J. Baker, L. Phan, E.M. Smigielski and K. Sirotkin, *Nucleic Acids Res.* **29** (2001) 308.

[19] S.M. Langemeijer, R.P. Kuiper, M. Berends, R. Knops, M.G. Aslanyan, M. Massop, E. Stevens-Linders, P. van Hoogen et al., *Nat. Genet.* **41** (2009) 838.

[20] J. Thusberg, A. Olatubosun and M. Vihinen, *Hum. Mutat.* **32** (2011) 358.

[21] S. Hicks, D.A. Wheeler, S.E. Plon and M. Kimmel, *Hum. Mutat.* **32** (2011) 661.