Biologia Serbica, 2017, 39(1): 74-82

DOI 10.5281/zenodo.826908

Minireview

Annotation of the functional impact of coding genetic variants

Vladimir Perović*, Branislava Gemović, Veljko Veljković, Sanja Glišić and Nevena Veljković

Center for Multidisciplinary Research, Institute of Nuclear Sciences Vinča, University of Belgrade, Mihajla Petrovića Alasa 12-14, 11001 Belgrade, Serbia

Summary. Coding genetic variants can have profound effects on protein function. Computational tools for the prediction of these effects are used to complement and guide experimental biological studies. Phylogenetic analyses that determine the evolutionary relationship among related sequences are commonly used to distinguish between functionally significant and insignificant gene variations. Here, we have reviewed applications of the non-alignment sequence analyses method for phylogenetic analyses, ISTREE. Furthermore, we assessed how an unsupervised ISTREE-d3 method based on the universal d3 measure responds to this task compared to supervised and semi-supervised ISTREE methods that were previously used in two studies. The findings presented here suggest that ISTREE-d3 can efficiently substitute for the corresponding supervised models, given that it is more suitable for automatic applications. In conclusion, the ISTREE-d3 method has a broad biological relevance and represents a promising approach in functional assessment of coding gene variations.

Keywords: amino acid sequence, coding genetic variation, electron-ion interaction potential, H5N1 influenza virus, informational spectrum method, phylogenetic analysis, TET2 protein.

Introduction

Single nucleotide polymorphisms (SNPs) are the most common changes in the human genome and the coding genetic variants can affect protein function (1000 Genomes Project Consortium 2012). The effect of an SNP depends on its type (silent, missense, nonsense) and position. In the case of missense variants, this leads to amino acid substitution (AAS). The majority of SNPs, marked as neutral SNPs, can be found in the healthy population, while the subset of SNPs in the human genome is disease-related and capable of driving disease onset and progression. However, the functional effects of AAS can be ambiguous, especially in complex human diseases like cancer, because not all the variations detected in the malignant cells are "drivers" of the disease (Stratton et al. 2009). AAS can have various consequences – from no biological effect (neutral SNPs) to the abolition of protein function or acquisition of a new function, leading to disease start or progression (somatic mutations) (Studer et al. 2013). In addition to sequencing of the human genome and detection of mutations associated with various diseases, it is important to monitor and categorize variations in genomes of major human pathogens, such as the influenza viruses. Detecting functional AAS is clinically useful, but experimentally determining the biological effects of all AASs is costly and time-consuming (Thusberg and Vihinen 2009). Therefore, great efforts are invested in the development of computational methods for predicting the functional relevance of AAS. Most of these methods are based on phylogenetic analysis, such as PolyPhen-2 (Adzhubei et al. 2010) and SIFT (Ng and Henikoff 2001).

Phylogenetic analysis determines the evolutionary relationship inside the family of related sequences. Depending on the level of sequence similarity, methods for constructing phylogenetic trees can be divided into three groups (David 2001): (i) maximum parsimony methods (Fitch 1971; Sankoff and Cedergren 1983) used for similar sequences, (ii) distance methods (Feng and Doolittle 1996) used for sequences that share recognizable similarity, and (iii) a probabilistic ap-

*Corresponding author, e-mail: vladaper@vinca.rs

proach that uses maximum likelihood (Felsenstain 1973) or sampling methods (Mau et al. 1996). In our current research, we focused on the distance methods, which are based on distance matrices representing the dissimilarity between each pair of sequences. The distance scores are determined from the alignment score (Feng and Doolittle 1996) or various distance measure models (Jukes and Cantor 1969; Kimura 1980). The matrix is then transformed into a phylogenetic tree using clustering algorithms: the neighbor-joining method (NJ) (Saitou and Nei 1987), Fitch-Margoliash method (Fitch and Margoliash 1987) or the unweighted pair group method with arithmetic mean (UPGMA) (Sokal and Michener 1958).

In the majority of these approaches evolutionary models are based on the multiple sequence alignment (MSA), which has been the standard for sequence comparison for the past decades owing to its simple and manifest mechanism. However, generating optimal alignment is computationally demanding, especially for efficiently querying constantly expanding public genomic and proteomic databases. On the other hand, non-alignment approaches, in addition to being broadly applicable, are ultra-rapid and not computationally demanding (Borozan et al. 2015). They have the potential to overcome not only computational, but also fundamental limitations of sequence comparison by alignment, such as incapability of recognizing more divergent but functionally related sequences, overlooking of long-range interactions or recognizing the significance of single AAS (Vinga and Almeida 2003; Schwende and Pham 2014). The informational spectrum method (ISM) is a non-alignment method based on the assumption that the protein-protein interaction encompasses two major steps: (i) recognition and targeting between interacting proteins (long-range interactions at distances greater than 100 Å), and (ii) chemical binding (short-range interactions at distances less than 5 Å) (Veljkovic and Slavic 1972; Veljkovic 1980). ISM transforms a protein sequence into a virtual spectrum encompassing peaks (defined by corresponding frequency and amplitude) that correlate with its protein interactions and functions.

We have recently developed the informational spectrum-based phylogenetic analysis ISTREE, an ISM-based phylogenetic algorithm, and demonstrated that it efficiently recognizes the functional significance of AASs in the highly pathogenic influenza virus (Perovic 2013). The core of the ISTREE method consists of distance measures, which are defined as follows: d1 – single frequency distance: the absolute difference of the amplitudes on predefined frequency F; d2 – amplitude ratio distance: the absolute difference between the amplitude ratios on predefined frequencies F1 and F2; d3 – full spectrum distance: the Minkowski L_1 distance (Minkowski 1953) between corresponding spectra.

Spectral based phylogenetic analyses, by applying different sequence representations and distance measures on protein sequence data, are capable of responding to fundamental questions related to the functional significance of AASs (Veljkovic et al. 2009b, 2015, 2016). The intrinsic characteristic of models that rely on d1 and d2 is the requirement of analyses prior to application of clustering algorithms in which characteristic frequencies are predefined, and as such they are not suitable for fully automated analyses. Here, in two case studies, we will review the capacity of different d measures to recognize genetic variations that affect protein biological activity, and examine how efficiently the universal d3 measure extracts this information compared to d1 and d2.

Informational spectrum method (ISM)

The physicochemical descriptor electron-ion interaction potential (EIIP) represents the main energy term of valence electrons (Veljkovic and Slavic 1972) and determines the long-range properties of biological molecules. The EIIP for organic molecules is determined by the following equation (Veljkovic 1973; Veljkovic et al. 1985):

$$W = 0.25Z^*\sin(1.04\pi Z^*)/2\,\pi \quad (1)$$

where Z* is the average quasi-valence number (AQVN) calculated by

$$Z^* = \Sigma^{m} n_i Z_i / N \quad (2)$$

where n_i is the number of atoms of the i-th atomic component, Z_i is the valence number of the i-th component, m is the number of atomic components in the molecule and N is the total number of atoms.

The ISM technique assigns the corresponding value of EIIP to each amino acid in the protein sequence (Table 1). This sequence of numbers, corresponding to the protein sequence, is then transformed to a spectrum using a discrete Fourier transform defined as:

$$X(n) = \Sigma_m\, x(m) e^{-i(2/N)nm},\ n = 1, 2, ..., N/2 \quad (3)$$

where N is the total number of points in this series, $x(m)$ is the m-th member of a given series and $X(n)$ are the Fourier transformation coefficients that describe the amplitude, phase and frequency of the sinusoids of which the original EIIP signal is composed. The absolute values of a complex discrete Fourier transformation determine the amplitude spectrum, which is presented as an energy density spectrum (Veljkovic et al. 1985):

$$S(n) = X(n)X^*(n) = |X(n)|^2,\ n = 1, 2, ..., N/2 \quad (4)$$

In this way protein sequences are analyzed as discrete signals, i.e. number series. It is assumed that their points are equidistant with the distance d = 1, and therefore the maximal frequency in a spectrum is F = 1/2d = 0.5. The fre-

Table 1. The EIIP values used to encode the amino acids.

Amino acid	One letter code	EIIP value (Ry)
Leucine	L	0.0000
Isoleucine	I	0.0000
Asparagine	N	0.0036
Glycine	G	0.0050
Valine	V	0.0057
Glutamic acid	E	0.0058
Proline	P	0.0198
Histidine	H	0.0242
Lysine	K	0.0371
Alanine	A	0.0373
Tyrosine	Y	0.0516
Tryptophan	W	0.0548
Glutamine	Q	0.0761
Methionine	M	0.0823
Serine	S	0.0829
Cysteine	C	0.0829
Threonine	T	0.0941
Phenylalanine	F	0.0954
Arginine	R	0.0956
Aspartic acid	D	0.1263

quency range is independent of the total number of points in the sequence, and only the resolution of the spectrum is influenced by the total number of points in a sequence. The resolution of the N-point sequence is $1/n$. The n-th point in the spectral function corresponds to a frequency $f(n) = n/N$. Thus, the initial information of the protein defined by the sequence of amino acids can be represented in the form of an informational spectrum (IS) as a series of frequencies and their amplitudes.

The frequencies in an IS correspond to the distribution of structural motifs with defined physicochemical properties that determine a biological function of a protein. When comparing proteins that share the same biological or biochemical function, the ISM technique allows the detection of code/frequency pairs which are specific for their common biological properties, or which correlate with their specific interaction. These common informational characteristics of sequences are discovered using a consensus informational spectrum (CIS). A CIS of N spectra is obtained by the following equation:

$$C(j) = \Pi\, S(i,j) \quad (5)$$

where $S(i,j)$ is the j-th element of the i-th power spectrum and $C(j)$ is the j-th element of CIS. Thus, CIS is the Fourier transform of the correlation function for the spectrum, and any spectral component (frequency) not present in all compared informational spectra is eliminated. Peak frequencies in the CIS represent the common information encoded in the primary structure of analyzed sequences. This information corresponds to the mutual long-range interaction between analyzed proteins or their interaction with the common interactor.

Informational spectrum phylogenetic analysis: IS-TREE method

Single frequency distance (d_1)

To construct an informational tree related to a certain biological function represented by the previously determined characteristic frequency F, derived from the CIS of a family of protein sequences, the distance between sequences can be defined as the absolute difference of the amplitudes on the frequency F. Let X_1 and X_2 be two sequences, S_1 and S_2 their corresponding spectra. Let F be the characteristic frequency. Let $A_1(F)$ and $A_2(F)$ be amplitudes on frequency F of spectra S_1 and S_2, respectively. Then the distance between X_1 and X_2 is defined as:

$$d_1(X_1,X_2) = \left| A_1(F) - A_2(F) \right| \quad (6)$$

Let P be the set of values $A(F)$ for every sequence X. Distance d_1 is the Euclidean distance on the set of real numbers R and therefore it is a valid metric measure on P. It satisfies:

1. $d_1(x,y) = 0$, non-negativity
2. $d_1(x,y) = 0 \Leftrightarrow x=y$, identity of indiscernibles
3. $d_1(x,y) = d_1(y,x)$, symmetry
4. $d_1(x,z) = d_1(x,y) + d_1(y,z)$, triangle inequality

It is also a valid additive evolutionary measure on P. It satisfies additivity (four-point condition): two of three sums $d_1(x,y)+d_1(z,w)$, $d_1(x,z)+d_1(y,w)$, $d_1(x,w)+d_1(y,z)$, are equal and larger than a third sum.

Amplitude ratio distance (d_2)

To infer the information that corresponds to the transfer between two biological functions represented by previously determined characteristic frequencies F1 and F2, derived from the CIS of family of protein sequences, the distance between sequences can be defined as the absolute difference of the amplitude ratios. Let X_1 and X_2 be two sequences, S_1 and S_2 their corresponding spectra. Let F_1 and F_2 be two characteristic frequencies. Let $A_1(F_1)$ and $A_1(F_2)$ be amplitudes of spectrum S_1 on frequencies F_1 and F_2, respectively. Let $A_2(F_1)$ and $A_2(F_2)$ be amplitudes of spectrum S_2 on frequencies F_1 and F_2, respectively. Then the distance between X_1 and X_2 is defined as:

$$d_2(X_1,X_2) = \left| A_1(F_1)/A_1(F_2) - A_2(F_1)/A_2(F_2) \right| \quad (7)$$

Let P be the set of values $A(F_1)/A(F_2)$ for every sequence

X, where $A(F_1)$ and $A(F_2)$ are spectrum amplitudes of sequence S on frequencies F_1 and F_2, respectively. Set P is a subset of the set of real numbers R. Like d_1, distance d_2 is then the Euclidean distance on R, and therefore it is a valid metric measure on P. It is also a valid additive evolutionary measure on P, it satisfies additivity.

Full spectrum distance (d_3)

To compute the informational tree that considers all information from the informational spectra of the protein sequences, the distance between two sequences can be defined as the Minkowski L_p distance (Manhattan distance for p = 1) between corresponding spectra. Let X_1 and X_2 be two sequences, $S_1 = \{S_1(n)\}$ and $S_2 = \{S_1(n)\}$, n = 1, 2, ..., N/2, their corresponding energy density spectra, then the distance between X_1 and X_2 is defined as:

$$d_3(X_1,X_2) = (\Sigma_{n=1..N/2}|S_1(n) - S_2(n)|)/N \quad (8)$$

where N is the length of the longest sequence. Distance d_3, as a Minkowski L_1 distance, is a valid metric measure as a result of the Minkowski inequality (Minkowski 1953).

Differences and similarities between IS distances

Distances d1 and d2 are valid additive evolutionary measures and are suitable for the neighbor-joining clustering method. Distance d3 is not an additive measure, therefore it is more appropriate to use d3 with the UPGMA clustering algorithm.

In order to use d1 or d2 distance, it is first necessary to determine the characteristic frequencies by means of ISM analysis, unlike the d3 distance, which does not need any prior analysis.

ISTREE algorithm

1. For each sequence calculate its IS:
   a. Convert amino acid sequence into EIIP signal with zero mean
   b. Zero-padding to the length of the longest signal
   c. Apply Fourier transform to signal to generate IS
2. Calculate the CIS of all spectra
3. Chose the IS distance (d1, d2 or d3) and:
   a. Determine characteristic frequency F on CIS (for d1 distance) or
   b. Determine two characteristic frequencies F_1 and F_2 on CIS (for d2 distance) or
   c. Do nothing (d3 distance)
4. Depending of the IS distance, calculate the distance matrix with the following distance measure:
   $d_1(X_i,X_j) = |A_i(F) - A_j(F)|$ or
   $d_2(X_i,X_j) = |A_i(F_1)/A_i(F_2) - A_j(F_1)/A_j(F_2)|$ or
   $d_3(X_i,X_j) = (\Sigma_{n=1..N/2}|S_1(n) - S_2(n)|) / N$
5. Infer the phylogenetic tree using neighbor-joining (NJ) or UPGMA method

ISTREE properties

The properties of the ISM-based phylogenetic approach and main advantages over standard methods are: (i) ISTREE is not based on MSA and does not use any substitution model, (ii) it is sensitive to the position of mutation and the type of the substituted residue, (iii) it is sensitive to a single mutation (Perovic 2013).

The ISTREE has high performance in terms of computing time due to its low algorithm complexity and absence of the MSA calculation phase.

Software

For generating the ISM-based trees, we used our service 'ISTREE' that was developed in the JAVA programming language which is freely available on: http://www.vin.bg.ac.rs/180/istree/. For conventional phylogenetic trees, we used MEGA5 software package (Tamura et al. 2011), and for MSA calculation we used the MUSCLE algorithm (Edgar 2004).

ISTREE-based semi-supervised analysis of HA1 H5N1 coding genetic variants

In the last decade, the influenza virus has reemerged as one of the most severe threats to human health. The most variable segment of the genome of influenza viruses codes for hemagglutinin (HA) which mediates the viruses' entrance into the host cells (Wiley and Skehel 1987). HA dominantly binds to the receptor specific for the preferable host, but this can be affected by genetic variations in HA, leading to a switch to a different host. Therefore, to be able to localize the potential center of a new influenza pandemic, it is important to predict the functional effects of variations in HA.

In a previous publication (Perovic et al. 2013), ISM was used in the analysis of the highly pathogenic avian influenza virus (HPAIV) type A subtype H5N1. The study of functional aspects of the evolution of the HA1 in Egypt after 2006 identified and predicted mutations that enhance the virus' human tropism. This was accomplished through a supervised analysis of the evolutionary relationship among different strains of HPAIV-H5N1. The dataset contained 526 HA1 sequences from HPAIV-H5N1 isolated in Egypt in the period 2006-2011, which were available in the NCBI (http://ncbi.nlm.nih.gov/protein) and GISAID (http://platform.gisaid.org) databases.

CIS analysis of HA1 protein sequences identified two characteristic frequencies that represent two biological functions. The IS frequency component F1 = 0.076 corresponds to the tropism of the H5N1-HPAIV, and F2 = 0.236 cor-

responds to the tropism of the seasonal H1N1 virus. Examination of the generated phylogenetic tree by the ISTREE algorithm identified cluster of specific AASs and predicted mutations that increase human tropism of Egyptian H5N1-HPAIV (Perovic et al. 2013). The predicted mutations were further experimentally confirmed to increase HPAIV human-to-human transmission *in vitro* (Schmier et al. 2015).

This type of study is semi-supervised because it combines i) an unsupervised ISTREE hierarchical clustering algorithm, with no prior information (annotation) about the functional effects of virus gene variants, and ii) calculation of d2 distances, which depends on the previous determination of characteristic F1 and F2 frequencies. The comparison of the d2-based HA phylogenetic tree generated by the ISTREE tool with a d3-based tree, which does not require prior analysis, showed similar clustering (Fig. 1). 99.2% of sequences are grouped equally in two major clusters in the ISTREE-d2 tree, d2-G1 and d2-G2 (Fig. 1A), and in the ISTREE-d3 tree, d3-G1 and d3-G2 groups (Fig. 1B). Subsequent analyses of these two clusters revealed a similar result in mutation prediction, since both approaches identified the same four HA1 mutations that have an increased potential for human H5N1-HPAIV infection.

In Fig. 2, the proteins in standard phylogenetic trees of HA1 are colored according to ISTREE-d2 clusters G1 and G2. The conventional trees show a similar grouping of proteins to that in ISTREE-generated trees, but with a less clear separation into two major clusters as in ISM-based trees (Figs 1, 2).

Comparison of two ISM-based phylogenetic trees (Fig. 1) suggests that the d3 distance could be used instead of the d2 distance in unsupervised analysis and assessment of genetic variants. This accelerates and simplifies the ISM-based process, which does not need prior ISM analysis and identification of characteristic frequencies.

ISTREE-based supervised analysis of TET2 coding genetic variants

A recent example of clinically useful detection of functional mutations arose from the analysis of Tet methylcytosine dioxygenase 2 (TET2), a protein involved in DNA demethylation (Tahiliani et al. 2009). TET2 is mutated in various cancers, including all types of myeloid malignancies (Delhommeau et al. 2009), in which it represents an important marker of disease progression (Grossmann et al. 2011; Metzeler et al. 2011), minimal residual disease (Jan et al. 2012) and response to treatment (Itzykson et al. 2011). Previously, several computational methods were used and compared to predict the functional effects of AASs in TET2 (Gemovic et al. 2013a). This study showed that it is especially difficult to correctly predict the biological effects of a subset of AASs positioned outside TET2's conserved functional domains (CFDs). This motivated us to develop an algorithm

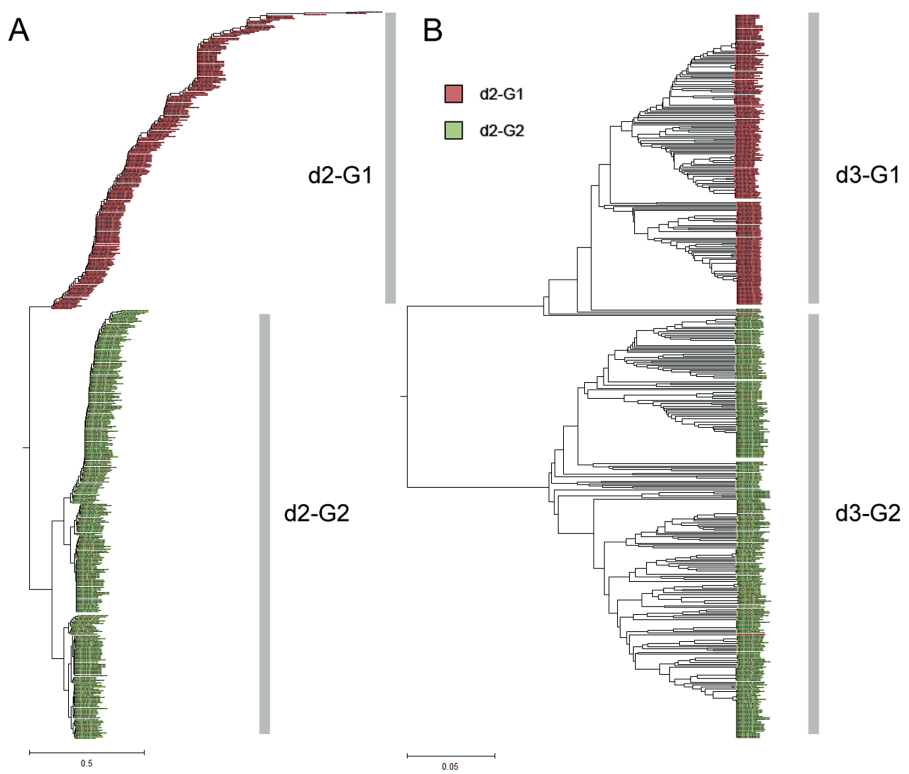

Fig. 1. ISM-based phylogenetic analysis of Egyptian HA1 H5N1. The trees were generated by ISTREE algorithm using **A**) d2 distance (A(0.236)/A(0.076)) and **B**) d3 distance.

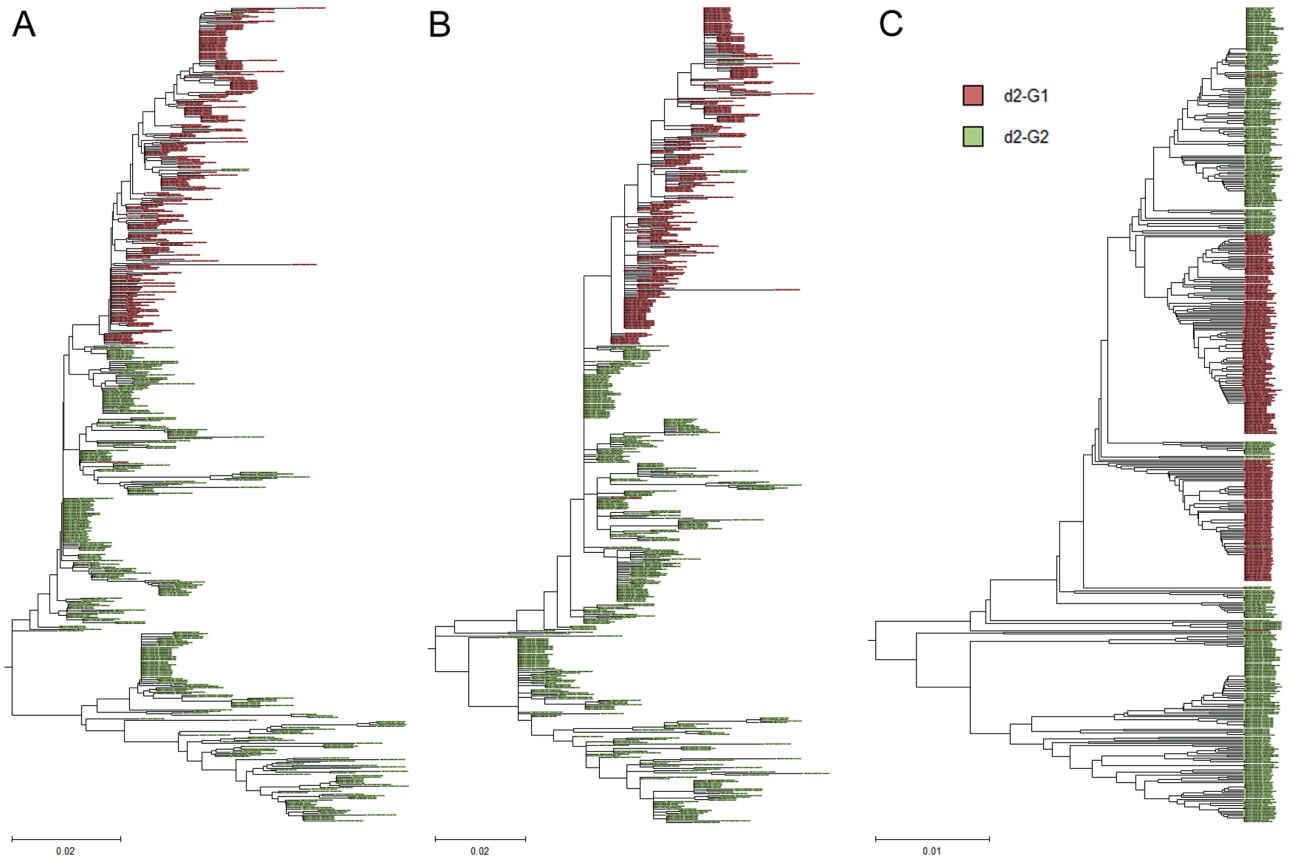

Fig. 2. MSA-based phylogenetic analysis of Egyptian HA1 H5N1. The trees were generated using standard methods: **A**) neighbor joining, **B**) maximum likelihood and **C**) UPGMA.

based on the ISM and statistical analysis for classification of the functional effects of AAS outside the CFDs in epigenetic regulators associated with myeloid malignancies. The ISM-based classification method outperformed the most commonly used phylogeny-based tools SIFT and PolyPhen-2 in terms of prediction efficacy.

From a total of 166 mutations in TET2, gathered from literature and the dbSNP database (Sherry et al. 2001), we analyzed all 69 coding variants outside CFDs. AAS was marked as neutral SNP if it had a predefined frequency in a healthy population and/or it was experimentally detected in the germline by the original study. On the other hand, AAS was marked as a mutation (MUT) if the original work experimentally confirmed its somatic status.

The IS frequency F = 0.491 was found to discriminate with statistical significance between neutral SNPs and pathogenic mutations and we based the ISM classification model for prediction of AASs in TET2 on this feature. When the ISTREE method with d1 distance for the frequency F = 0.491 is applied to the set of 69 mutated TET2 sequences, two distinct clusters emerge (Fig. 3A). Analysis of d1-G1 and d1-G2 clusters showed that 66.67% of sequences classified as MUT are grouped in the d1-G1 cluster, and 57.14% of sequences classified as SNP are grouped in the d1-G2 cluster. We used the following measures of predictive performance:

Accuracy = (TP+TN)/(TP+TN+FP+FN)

Recall = TP/(TP+FN)

Precision = TP/(TP+FP)

F1 = 2*Precision*Recall/(Precision+Recall),

where TP, TN, FP and FN indicate the number of true positives (correctly classified coding genetic variations), true negatives (correctly predicted coding genetic variations), false positives and false negatives, respectively.

If AAS is classified as MUT when positioned in the d1-G1 group and as SNP when positioned in d1-G2 group, the accuracy of this model is 0.61. Table 2 shows the performance statistics of this method compared to the ISM classification model. The prediction efficacy of the ISTREE-d1 model in terms of accuracy is a bit lower than that of the ISM model, but it still outperforms the SIFT and PolyPhen-2 tools (Table 2).

Using the ISTREE-d3 method, we generated the phylogenetic tree shown in Fig. 3B. Although there is no clear clustering into two distinct groups, the big branch in the middle could be used as a separation boundary and the tree can be parted into two distinct groups, the top cluster d3-G1 and

Table 2. Efficacy of predicting the effects of TET2 coding gene variants by ISM-based and the most commonly used methods SIFT and Polyphen-2.

	ISM	ISTREE-d1	ISTREE-d3	SIFT	Polyphen-2
Accuracy	0.623188	0.608696	0.623188	0.57971	0.565217
Recall	0.62963	0.666667	0.62963	0.592593	0.444444
Precision	0.515152	0.5	0.515152	0.470588	0.444444
F1 score	0.566667	0.571429	0.566667	0.52459	0.444444

the rest of the tree, d3-G2 group (Fig. 3B). The analysis of d3-G1 and d3-G2 clusters showed that 62.96% of sequences classified as MUT are grouped in the d3-G1 cluster, and 61.90% of sequences classified as SNP are grouped in the d3-G2 cluster. The predictive performance of the ISTREE-d3 method is equal to the ISM classification model (Table 2), but the advantage of the ISTREE-d3 model is that it does not require prior ISM and statistical analyses to define the characteristic IS frequency.

A detailed comparison of ISTREE clusters (Fig. 4) revealed that 70% of sequences belong to the same group, either G1 or G2, in both ISTREE-d1 and ISTREE-d3. This incongruity implies that these two models rely on complementary information for classification. In future, we will consider a classificatory relying on both features concurrently.

Finally, the conventional phylogenetic trees generated using a neighbor-joining algorithm with Jones-Taylor-Thornton (JTT) substitution model (Fig. 5A) and maximum likelihood method (Fig. 5B) failed to reveal distinct clusters and are obviously not sensitive to single AAS. This is in concordance with previous findings that MSA-based phylogenetic analyses are not suitable for the prediction of the functional effects of TET2 gene variations in coding regions (Gemovic et al. 2013b).

Conclusion

In analyzing the functional significance of coding gene variations, ISTREE has an advantage over conventional phylogenetic trees because it is significantly more sensitive to the effects of a single AAS. ISTREE is based on a non-alignment sequence analyses method, ISM, which has been successfully applied to this task for decades. However, thus far ISM analysis has never been used in a fully automated unsupervised procedure. Here we demonstrated that it is possible to overcome this drawback when the universal d3 measure and ISTREE-d3 phylogenetic analysis are applied. In future, we will perform a comprehensive comparison to conventional phylogenetic tree methods in order to establish the full potential of ISTREE-d3 in the functional assessment of coding genetic variants.

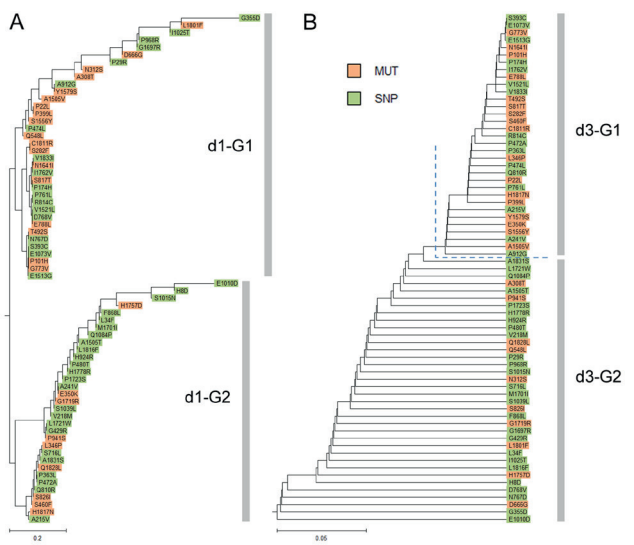

Fig. 3. Analysis of AASs outside CFDs in TET2 protein using the ISTREE tool. The phylogenetic trees were generated using **A**) d1 distance (A(0.491)) and **B**) d3 distance. Mutations are colored according to SNP and MUT classes.

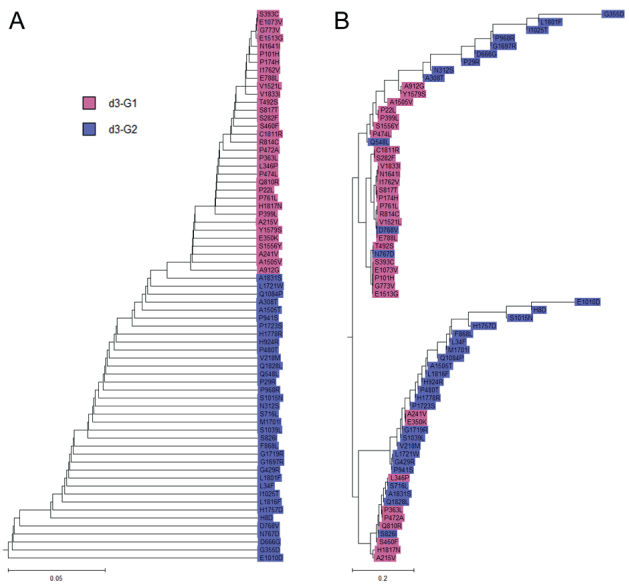

Fig. 4. Comparison of ISTREE phylogenetic trees of TET2 sequences. The trees were generated by **A**) d3 distance and **B**) d1 distance. Sequences are colored according to groups G1 and G2 in ISTREE-d3 tree.

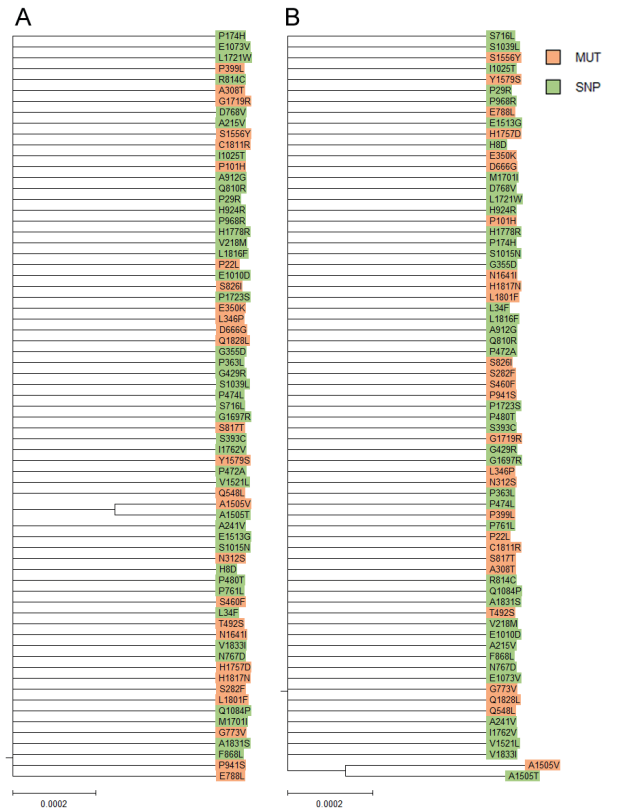

Fig. 5. Phylogenetic analysis of all 69 non-CFD mutated TET2 sequences generated by conventional MSA based tools. The trees are inferred using **A**) neighbor joining and **B**) maximum likelihood method. Sequences are colored according to their mutations, SNP or MUT class.

Acknowledgments

This research was funded by the Ministry of Education, Science and Technological Development of the Republic of Serbia (Grant No. 173001).

References

1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature. 491(7422):56-65.

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. Nature Methods. 7(4):248-9.

Borozan I, Watt S, Ferretti V. 2015. Integrating alignment-based and alignment-free sequence similarity measures for biological sequence classification. Bioinformatics. 31(9):1396-1404.

David WM. 2001. Bioinformatics: sequence and genome analysis. Cold Spring Harbor Laboratory Press. 6.

Delhommeau F, Dupont S, Valle VD, James C, Trannoy S, Massé A, Kosmider O, Le Couedic JP, Robert F, Alberdi A, et al. 2009. Mutation in TET2 in myeloid cancers. New England Journal of Medicine. 360(22):2289-301.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research. 32(5):1792-7.

Felsenstein J. 1973. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Systematic Zoology. 22:240-249.

Feng DF, Doolittle RF. 1996. Progressive alignment of amino acid sequences and construction of phylogenetic trees from them. Methods in Enzymology. 266:368-382.

Fitch WM, Margoliash E. 1987. Construction of phylogenetic trees. Science. 155:279-284.

Fitch WM. 1971. Toward defining the course of evolution: minimum change for a specific tree topology. Systematic Zoology. 35:406-416.

Gemovic B, Perovic V, Glisic S, Veljkovic N. 2013a. Feature-based classification of amino acid substitutions outside conserved functional protein domains. Scientific World Journal. 2013:948617.

Gemović B, Perović V, Glišić S, Veljković N. 2013b. Can We Use Standard Tools to Predict Functional Effects of Missense Gene Variations Outside Conserved Domains? TET2 Example. In: 2nd International Conference "Theoretical Approaches to BioInformation Systems" (TABIS. 2013). p. 65.

Global Initiative on Sharing All Influenza Data (GISAID) database Available: http://platform.gisaid.org.

Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C, et al. 2007. Patterns of somatic mutation in human cancer genomes. Nature. 446(7132):153-158.

Grossmann V, Kohlmann A, Eder C, Haferlach C, Kern W, Cross NC, Haferlach T, Schnittger S. 2011. Molecular profiling of chronic myelomonocytic leukemia reveals diverse mutations in >80% of patients with *TET2* and *EZH2* being of high prognostic relevance. Leukemia. 25(5):877-879.

Itzykson R, Kosmider O, Cluzeau T, Mansat-De Mas V, Dreyfus F, Beyne-Rauzy O, Quesnel B, Vey N, Gelsi-Boyer V, Raynaud S, et al. 2011. Impact of *TET2* mutations on response rate to azacitidine in myelodysplastic syndromes and low blast count acute myeloid leukemias. Leukemia. 25(7):1147-1152.

Jan M, Snyder TM, Corces-Zimmerman MR, Vyas P, Weissman IL, Quake SR, Majeti R. 2012. Clonal evolution of preleukemic hematopoietic stem cells precedes human acute myeloid leukemia. Science Translational Medicine. 4(149):149ra118.

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In: munro HN, editor. Mammalian protein metabolism. Vol. III. New York: Academic Press. p. 21-132.

Kimura M. 1980. A simple model for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolutiob. 16:111-120.

Mau B, Newton MA, Larget B. 1996. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Technical Report 961, Statistics Department, university of Wisconsin-Madison.

Metzeler KH, Maharry K, Radmacher MD, Mrózek K, Margeson D, Becker H, Curfman J, Holland KB, Schwind S, Whitman SP, et al. 2011. *TET2* mutations improve the new European LeukemiaNet risk classification of acute myeloid leukemia: a Cancer and Leukemia Group B study. Journal of Clinical Oncology. 29(10):1373-1381.

Minkowski H. 1953. Geometrie der Zahlen. London: Chelsea Pub. Co.

National Center for Biotechnology Information (NCBI) database. Available: http://ncbi.nlm.nih.gov/protein.

Ng PC, Henikoff S. 2001. Predicting deleterious amino acid substitutions. Genome Research. 11(5):863-874.

Perovic VR, Muller CP, Niman HL, Veljkovic N, Dietrich U, Tosic DD, Glisic S, Veljkovic V. 2013. Novel phylogenetic algorithm to monitor human tropism in Egyptian H5N1-HPAIV reveals evolution toward efficient human-to-human transmission. PLOS ONE. 8(4):e61572.

Perovic VR. 2013. Novel algorithm for phylogenetic analysis of proteins: application to analysis of the evolution of H5N1 influenza viruses. Journal of Mathematical Chemistry. 51(8):2238-2255.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution. 4(4):406-425.

Sankoff D, Cedergren RJ. 1983. Simultaneous comparison of three or more sequences related by a tree. In: Sankoff D, Kruskal JR, editors. Time warps, string edits, and macromolecules: The theory and

practice of sequence comparison. Ontario (CA): Addison-Wesley. p. 253-264.
Schmier S, Mostafa A, Haarmann T, Bannert N, Ziebuhr J, Veljkovic V, Dietrich U, Pleschka S. 2015. In Silico Prediction and Experimental Confirmation of HA Residues Conferring Enhanced Human Receptor Specificity of H5N1 Influenza A Viruses. Scientific Reports. 5:11434.
Schwende I, Pham TD. 2014. Pattern recognition and probabilistic measures in alignment-free sequence analysis. Briefings in Bioinformatics. 15(3):354-368.
Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. Nucleic Acids Research. 29(1):308-311.
Sokal RR, Michener CD. 1958. A Statistical Method of Evaluating Systematic Relationships. University of Kansas Scientific Bulletin. 28:1409-1438.
Stratton MR, Campbell PJ, Futreal PA. 2009. The cancer genome. Nature. 458(7239):719-724.
Studer RA, Dessailly BH, Orengo CA. 2013. Residue mutations and their impact on protein structure and function: detecting beneficial and pathogenic changes. Biochemical Journal. 449(3):581-594.
Tahiliani M, Koh KP, Shen Y, Pastor WA, Bandukwala H, Brudno Y, Agarwal S, Iyer LM, Liu DR, Aravind L, Rao A. 2009. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. Science. 324(5929):930-935.
Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Molecular biology and evolution. 28(10):2731-9.
Thusberg J, Vihinen M. 2009. Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods. Human Mutation. 30(5):703-714.
Veljkovic V. 1973. The dependence of the Fermi energy on the atomic number. Physics Letters A. 45(1):41-42.
Veljkovic V. 1980. Theoretical approach to preselection of cancerogens and chemical carcinogenesis. New York: Gordon & Breach.
Veljkovic V, Slavic I. 1972. Simple general-model pseudopotential. Physical Review Letters. 29:105-106.
Veljkovic V, Cosic I, Dimitrijevic B, Lalovic D. 1985. Is it possible to analyze DNA and protein sequence by the method of digital signal processing? IEEE Transactions on Biomedical Engineering. 32(5):337-341.
Veljkovic V, Veljkovic N, Muller CP, Müller S, Glisic S, Perovic V, Köhler H. 2009a. Characterization of conserved properties of hemagglutinin of H5N1 and human influenza viruses: possible consequences for therapy and infection control. BMC Structural Biology. 9:21.
Veljkovic V, Niman HL, Glisic S, Veljkovic N, Perovic V, Muller CP. 2009b. Identification of hemagglutinin structural domain and polymorphisms which may modulate swine H1N1 interactions with human receptor. BMC Structural Biology. 9:62
Veljkovic V, Paessler S, Glisic S, Prljic J, Perovic VR, Veljkovic N, Scotch M. 2015. Evolution of 2014/15 H3N2 Influenza Viruses Circulating in US: Consequences for Vaccine Effectiveness and Possible New Pandemic. Frontiers in Microbiology. 6:1456.
Veljkovic V, Veljkovic N, Paessler S, Goeijenbier M, Perovic V, Glisic S, Muller CP. 2016. Predicted enhanced human propensity of current avian-like H1N1 swine influenza virus from China. PLOS ONE. 11(11):e0165451.
Vinga S, Almeida J. 2003. Alignment-free sequence comparison-a review. Bioinformatics. 19(4):513-523.
Wiley DC, Skehel JJ. 1987. The structure and function of the hemagglutinin membrane glycoprotein of influenza virus. Annual Review of Biochemistry. 56(1):365-394.