Iran. J. Biotechnol. April 2024;22(2): e3787 DOI: 10.30498/ijb.2024.413800.3787

Brief Report

In-Silico Method for Predicting Pathogenic Missense Variants Using Online Tools: *AURKA* Gene as a Model

Eric Jonathan Maciel-Cruz[1,2], Luis Eduardo Figuera-Villanueva[1,2], Liliana Gómez-Flores-Ramos[3], Rubiceli Hernández-Peña [1,2], Martha Patricia Gallegos-Arreola[2*]

[1]Doctorado en Genética Humana, Instituto de Genética Humana "Dr. Enrique Corona Rivera", Centro Universitario de Ciencias de la Salud (CUCS), Universidad de Guadalajara (UdG), Guadalajara, Jalisco, México
[2]División de Genética, Centro de Investigación Biomédica de Occidente (CIBO), Instituto Mexicano del Seguro Social (IMSS), Guadalajara, Jalisco, México
[3]CONAHCYT- Centro de Investigación en Salud Poblacional, Instituto Nacional de Salud Pública, Cuernavaca, Morelos, Mexico.

[*]*Corresponding author*: Martha Patricia Gallegos-Arreola, División de Genética, Centro de Investigación Biomédica de Occidente (CIBO), Instituto Mexicano del Seguro Social (IMSS), Guadalajara, Jalisco, México. Tel/Fax: +52- 13331158793, E-mail: marthapatriciagallegos08@gmail.com

Received: 2023/08/30 ; **Accepted:** 2024/03/09

Background: *In-silico* analysis provides a fast, simple, and cost-free method for identifying potentially pathogenic single nucleotide variants.

Objective: To propose a simple and relatively fast method for the prediction of variant pathogenicity using free online *in-silico* (IS) tools with *AURKA* gene as a model.

Materials and Methods: We aim to propose a methodology to predict variants with high pathogenic potential using computational analysis, using *AURKA* gene as model. We predicted a protein model and analyzed 209 out of 64,369 *AURKA* variants obtained from Ensembl database. We used bioinformatic tools to predict pathogenicity. The results were compared through the VarSome website, which includes its own pathogenicity score and the American College of Medical Genetics (ACMG) classification.

Results: Out of the 209 analyzed variants, 16 were considered pathogenic, and 13 were located in the catalytic domain. The most frequent protein changes were size and hydrophobicity modifications of amino acids. Proline and Glycine amino acid substitutions were the most frequent changes predicted as pathogenic. These bioinformatic tools predicted functional changes, such as protein up or down-regulation, gain or loss of molecule interactions, and structural protein modifications. When compared to the ACMG classification, 10 out of 16 variants were considered likely pathogenic, with 7 out of 10 changes at Proline/Glycine substitutions.

Conclusion: This method allows quick and cost-free bulk variant screening to identify variants with pathogenic potential for further association and/or functional studies.

*Keywords***:** Computational Biology, Genomic Structural Variation, Missense Mutation, Single Nucleotide Polymorphism

1. Background

Among the wide range of human genetic variation (insertion, deletion, substitutions, etc.), the non-synonymous single nucleotide variants (SNV) can lead to protein malfunction and alter cellular processes. Within these, missense variants (MSV) are defined as SNVs that convert a single codon into a different amino acid (AA); such MSVs have the potential to cause a deleterious effect depending on the mutated AA based on the hydrophobicity, charge, size, and physical contacts (1, 2).

The pathogenicity of specific variants of several genes is still under discussion (3). According to the American College of Medical Genetics standards and guidelines (4), many SNVs are considered variants of unknown significance (VUS). As an example, Aurora Kinases (AK) is a family of serin-threonine kinases encompassing three proteins: Aurora A (AURKA), Aurora B, and Aurora C. AK possess similar structures in their catalytic domain but vary greatly in their N- and C- terminal domains (5). *AURKA* MSVs have been previously studied in breast cancer; however, these studies are controversial (3).

2. Objectives

Our aim is to propose a simple and relatively fast method for the prediction of variant pathogenicity using free online *in-silico* (IS) tools with *AURKA* as a model.

3. Material and Methods

3.1. Data Mining

AURKA data were obtained from Ensembl (ENSG00000087586); the datasheet was downloaded and filtered by specific criteria described in **Figure 1**. We obtained 209 variants for computational analysis after the elimination of repeated data. Variant information was used for the analysis (**Fig. 1**). The protein sequence was obtained in FASTA format from UniProt (O14965). The database was created and modified with Microsoft Excel 2019 and VSCode v.1.81.1 using Python language v.3.11.4.

3.2. In Silico Data Analysis

Pathogenicity prediction was carried out using a filtering stage methodology. As each platform reports its results differently (score, accuracy, pathogenicity, or effect prediction), in order to simplify our results, a "benign" or "pathogenic", we used prediction outcome. If more than one tool of the filter stage considered the variant as pathogenic, such variant was analyzed in the next stage; if no tool considered it pathogenic, no further analysis was done, and the variant was considered benign. This process was done for the first three PP filters. The filters were grouped according to similar algorithms and prediction properties. For the final step, only when Missense3D detected structural damage, the variant was considered pathogenic. The HOPE tool was employed to provide physicochemical properties of the AA, conservation, and protein domain information.

3.2.1. 1st Filter: Sequence-Based Homology Prediction

Four tools were applied to analyze 209 variants on the first filter: PROVEAN, SIFT, Mutation Taster, and PredictSNP2. PROVEAN (http://provean.jcvi.org/index.php) analyzes the nucleotide sequence and predicts the variant effect on the protein; SIFT (https://sift.bii.a-star.edu.sg/www/SIFT_dbSNP.html) analyzes the physical properties of the AA and predicts if it can disturb the protein function. If a score of ≤-2.5 is observed, the variant is considered deleterious for both PROVEAN and SIFT tools.

Mutation Taster (https://www.mutationtaster.org/) predicts the potential disease of the SNV alteration by employing Bayesian classifiers according to the most probable prediction for the mutation. A ≥0.51 value predicts a deleterious variant.

PredictSNP2 (https://loschmidt.chemi.muni.cz/predictsnp2/) evaluates the SNV effects using a consensus of five prediction tools (CADD, DANN, FATHMM, FunSeq2 and GWAVA). The results are color-coded (green for neutral, red for deleterious, and gray for unknown).

3.2.2. 2nd Filter: Structure-Based Homology Prediction

Three tools were used to analyze 124 variants on the second filter: PolyPhen, PANTHER, and SNAP2.

PolyPhen (http://genetics.bwh.harvard.edu/pph2/) analyzes the substitution impact on the structure through physical and phylogenetic comparative considerations, trying to identify the specific alteration site and compare it with available databases to evaluate the effects.

PANTHER (http://www.pantherdb.org/tools/csnp-

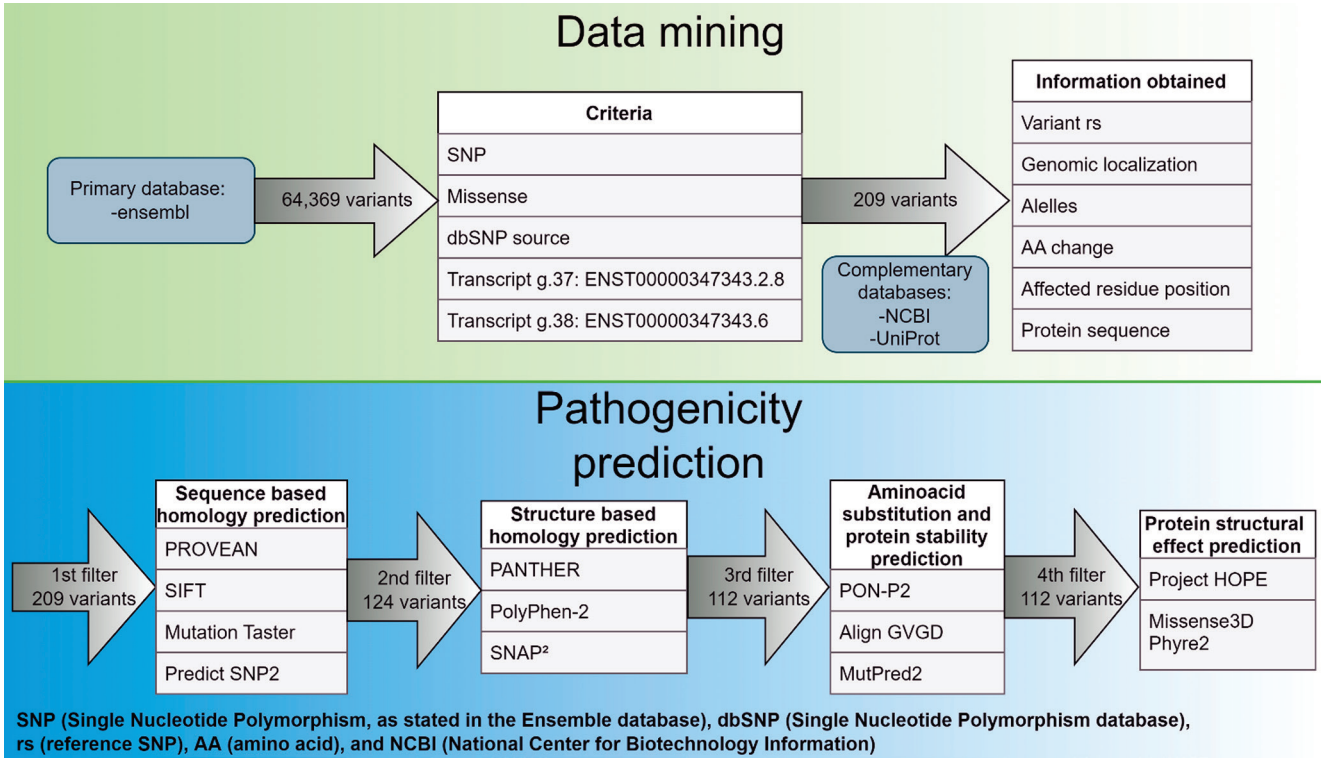

Figure 1. Data mining and PP proposed methodology. After the final prediction filter, "Protein structural effect prediction," only 16 variants were considered as possibly likely-pathogenic.

ScoreForm.jsp) analyzes the functional impact of the protein-coding SNV through evolutionary conservation. It categorizes the variant as "possibly damaging" or "possibly benign".

SNAP2 (https://rostlab.org/services/snap2web/) categorizes the variants as "neutral" or "effect" based on evolutionary information through multiple sequence alignment, secondary structure, and solvent accessibility analysis. The results vary from -100 (strong neutral prediction) to +100 (strong effect prediction). A higher effect value has a correlation degree to the effect severity.

3.2.3. 3rd Filter: Aminoacid Substitution and Protein Stability Prediction

Three tools were used to analyze 112 variants on the third filter: PON-P2, Align GVGD, and MutPred2.

PON-P2 (http://structure.bmc.lu.se/PON-P2/) predicts the pathogenicity substitutions based on AA features, gene ontology, evolutionary conservation, and functional protein site. It classifies the variants into pathogenic, unknown, or neutral.

Align GVGD (http://agvgd.hci.utah.edu/index.php) analyzes the biophysical characteristics of AA and protein sequence alignments to predict if a mutation falls in an enriched deleterious or neutral spectrum. The output categorizes the mutations into C0 to C65, resulting in tolerated/neutral variants (C0) to "untolerated"/pathogenic variants (C15-C65).

MutPred2 (http://mutpred2.mutdb.org/index.html) allows AA substitutions to score into benign (≤0.50) or pathogenic (0.51-1.0) based on known pathogenic and neutral variants and inter-species pairwise alignment.

3.2.4. 4th Filter: Protein Structural Effect Prediction

For the final step, we used HOPE and Missense3D were used to analyze 112 variants. A protein model was made through Phyre2.

Phyre2 (http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index) provides three-dimensional structure prediction from a given sequence. The modeling can be done (according to the website limitations) to

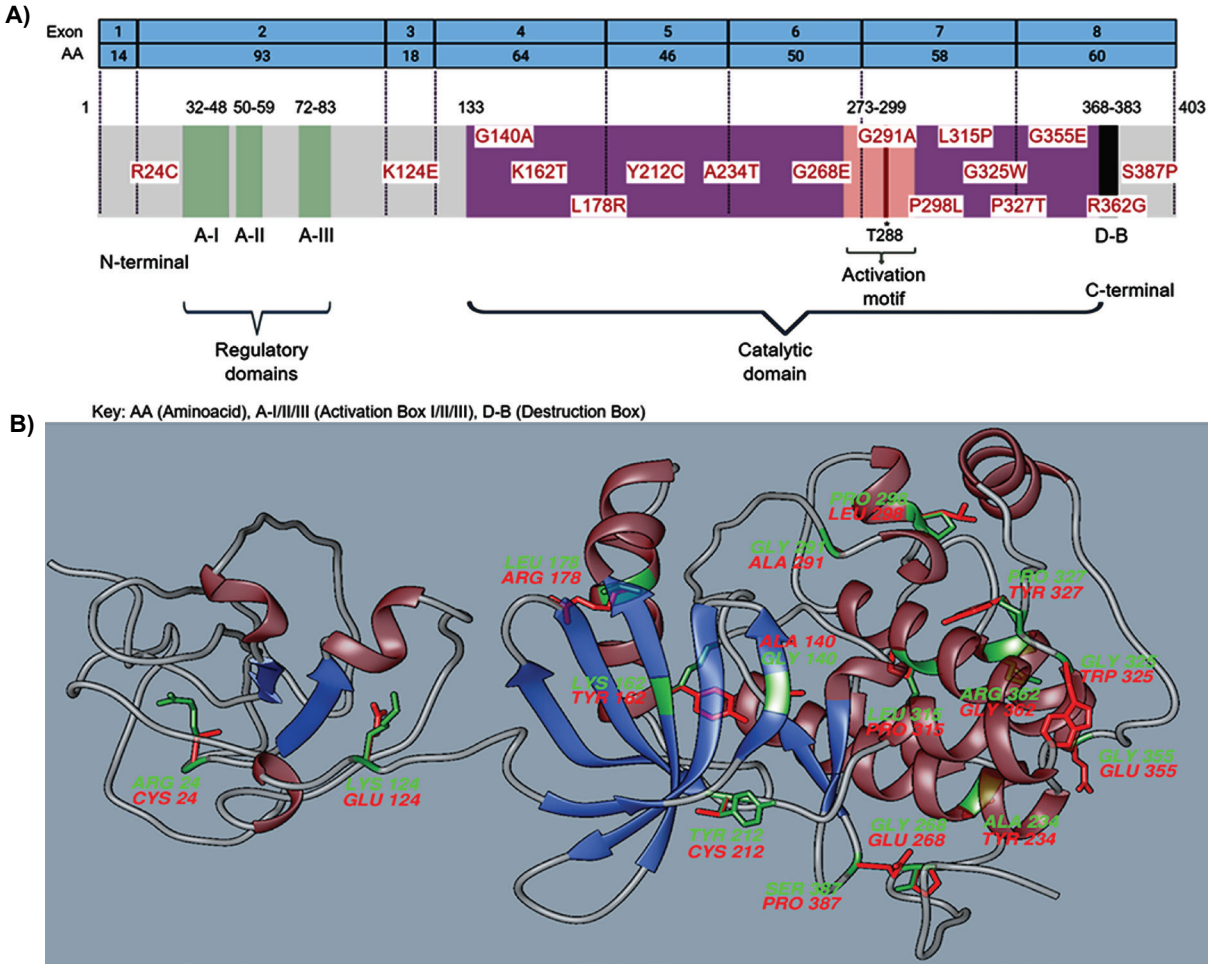

Figure 2. Schematic representation of *AURKA* exons, protein functional domains, and main predicted mutations. A) Purple lines correspond to exon translation regions; green boxes correspond to activation/regulation domains; catalytic domain (purple box) harbors the activation motif (pink box), containing a phosphorylated Tyrosine residue (red line); black box corresponds to the destruction box (D-B): predicted pathogenic mutations are shown in red text. **B)** Tertiary structure prediction of the wild-type (green) and mutant-type (red) residues.

carry out the prediction through an intensive method that allows complete modeling using multiple templates and *ab initio* techniques. Model refining was done using ModRefiner (https://zhanggroup.org/ModRefiner/), quality was verified with a Z-score and Ramachadran plot (6) using ProSa (https://prosa.services.came.sbg.ac.at/prosa.php) and PROCHECK (https://saves.mbi.ucla.edu/) webtools, respectively.

Missense3D (http://missense3d.bc.ic.ac.uk/~missense3d/) is an AA structural substitution predictor based on physicochemical properties. It requires a .pdb file containing the three-dimensional structure of a molecule; the results include disulfide breakage, buried Proline introduction, or clash introduction, among others. Depending on each parameter value predicting structural damage, a neutral or altered result is provided.

Project HOPE (https://www3.cmbi.umcn.nl/hope/) analyzes point protein mutation structural effects. The output provides results regarding AA properties, structure analysis, physical contacts, evolutionary conservation, and domain affectation.

4. Results

Of the 64,369 *AURKA* SNVs, only 209 met the criteria for *in-silico* analysis, and only 16 variants were considered pathogenic (**Fig. 2**).

Table 1. Project HOPE and Missense3D predictions for the 16 final PMSV.

rs	AA change	Project HOPE				Missense 3D altered structure	ACMG	VarSome[b]
		Different size	Different AA hydrophobicity	Different AA charge	Evolutionarily conserved			
rs536637669	Ser387Pro	+	+	-	+	Disallowed phi/psi	VUS	9/5
rs1284841822	Arg362Gly	+	+	+	-	Cavity altered	VUS	8/6
rs747506381	Gly355Glu	+	+	+	-	Disallowed phi/psi	B	8/6
rs928987283	Pro327Thr	-	+	-	-	Secondary structure altered	PP	12/3
rs11539196	Gly325Trp	+	+	-	+	Disallowed phi/psi	PP	12/3
rs1377907944	Leu315Pro	+	-	-	+	Cavity altered; Buried Pro introduced; Buried/exposed switch	PP	12/3
rs1255490947	Pro298Leu	+	-	-	+	Buried/exposed switch	PP	13/2
rs560948705	Gly291Ala	+	+	-	+	Disallowed phi/psi; Gly in a bend	PP	13/2
rs747008066	Gly268Glu	+	+	+	-	Disallowed phi/psi; Cavity altered; Gly in a bend	PP	12/4
rs1015771390	Ala234Thr	+	+	-	+	Cavity altered	PP	11/5
rs1304208982	Tyr212Cys	+	+	-	+	Cavity altered	VUS	11/5
rs948288770	Leu178Arg	+	+	+	-	Cavity altered	PP	11/5
rs1197614826	Lys162Thr	+	+	+	+	Cavity altered	PP	13/3
rs879169420	Gly140Ala	+	+	-	-	Cavity altered	PP	14/2
rs751452141	Lys124Glu	+	-	+	+	Buried charge switch	VUS	6/10
rs188825988	Arg24Cys	+	+	+	+	Buried charge replaced	B	8/5

AA (Amino acid), MR (Mutant-type residue), WR (Wild-type residue), RSA (Relative Solvent Accessibility), ACMG (American College of Medical Genetics), PP (Probably Pathogenic), VUS (Variant of Unknown Significance), B (Benign).

A) Considered positive when the MR is not observed in the AA position.

B) Score based on individual predictions (pathogenic/benign). Not all predictions are available for each mutation.

4.1. Sequence-Based Homology Prediction

PROVEAN predicted 132 (64.1%) benign and 74 (35.9%) pathogenic variants; SIFT reported 108 (52.4%) benign and 98 (47.6%) pathogenic; MutTaster reported 126 (61.2%) benign and 80 (38.8%) pathogenic; PredictSNP2 reported 140 (68.0%) benign and 66 (32.0%) pathogenic; 85 variants were considered as benign by all platforms. Cumulative analysis reported that 33 variants were considered pathogenic by one tool, 28 by two tools, 11 by three tools, and 52 by all tools. 124 variants were selected for further analysis.

4.2. Structure-Based Homology Prediction

From 124 that surpassed the first filters, PANTHER reported 19 (15.3%) benign and 105 (84.7%) as pathogenic variants; PolyPhen2 reported 37 (29.8%) benign and 87 (70.2%) pathogenic; SNAP2 reported 52 (41.9%) benign and 72 (58.1%) as pathogenic.

Only 12 variants were considered benign by all tools. Furthermore, 21 variants were considered pathogenic by one platform, 30 variants by two tools, and 61 variants by all tools. Of the analyzed variants, 112 were selected for further analysis.

4.3. Aminoacid Substitution and Protein Stability Prediction

Both PONP2 and Align GVGD reported only one (0.9%) benign and 111 (99.1%) pathogenic variants. MutPred analysis considered 47 (42.0%) benign and 65 (58.0%) as pathogenic.

No variants were considered neutral. Furthermore, only two variants were considered benign by one platform, 45 were considered pathogenic by two platforms, and

63 by all platforms. All variants were considered for further analysis.

4.4. Protein Structural Effect Prediction

The Phyre2 model reported that 100% of residues were modeled with confidence >90%. Protein ends were modeled with low confidence (1-16 and 389-403 residues), while the rest of the residues were modeled with high confidence. After using ModRefiner, the Z-score quality check reported that the model is within the range of scores found for native proteins of similar size, and ProCheck reported that 99.1% of residues were within the favored in allowed regions. Only 16/112 variants were considered pathogenic by HOPE and Missense3D.

Missense3D reported 16 (15.8%) pathogenic and 85 (84.2%) benign variants; the reports include the specific structural damage caused by the specific mutant-type residue (MR). Missense3D alteration criteria are described on the webpage (7). The most frequent structural damage detected were cavity alteration (8/16 variants) and disallowed phi/psi (5/16) variants.

According to HOPE, most variants were related to the kinase domain (133 to 383 residue position). Contact gains/losses were predicted in four variants, described as spatially closed residues that form bridges or bonds in specific atoms (6): R362G wild-type residue (WR) forms a hydrogen bond with 358D, while the MR forms a salt bridge with 358D and 376E; Y212C and K162T WR form a chemical bond with Protein Data Bank (PDB) Chemical Component 626, WR size can abolish this bond, Y212C MR can cause a loss of interaction with ATP nucleotide which can cause a protein function abolition; K162T WR forms a hydrogen bond with 276G and 277W while WR forms a salt bridge with 181E; G140A WR is not in direct contact with a ligand, however, MR might affect ligand-contact made by one of the neighboring residues. Fourteen variants were found to affect both the protein kinase and aurora kinase domains (R362G, G355E, P327T, G325W, L315P, P298L, G291A, G268E, A234T, Y212C, L178R, K162T, G140A, K124E), while two are located at the protein ends and no specific domains are associated (S387P, R24C) (**Fig. 2**). HOPE results are resumed in **Table 1**.

5. Discussion

Our analysis predicted 16/209 (7.65%) pathogenic variants considering our methodology in order to increase sensibility. However, no association studies were found for any of the final PMSV in public databases. This may be due to the global allele frequencies, as all variants had frequencies <0.01 or only reported as sporadic. However, as mentioned earlier, these databases focus on global frequencies, which means that specific populations may have different allele frequencies. According to the ACMG, 10/16 of the variants were considered likely pathogenic, with seven of them involving Proline/Glycine changes.

The frequent alterations observed in the final pathogenic variants are changes in AA size and hydrophobicity. These physicochemical properties may alter the structure and/or function of the protein (8). The most common mutations in proteins are the replacement of Proline and Glycine, either as a loss (such as G355E or P327) or a gain (such as R362G or L315P). Glycine is the smallest and most flexible of the AA, while Proline's side chain has a rigid cyclic structure (9, 10). Therefore, the specific losses or gains of Glycine can affect the bending of the protein structure, while Proline's mutations might affect the backbone rigidity of the normal tertiary structure, which can result in the alteration of the local binding sites or protein stability (11).

Depending on the location of the AA in the protein (main activity or regulatory domain, surface or core), interaction zone residues, or relevance to protein-ligand activities, the MR may cause a loss or gain of function or regulation, gain or loss of other molecule interactions, or production of core bumps, respectively (12). The catalytic domain was affected in 13/16 of the pathogenic variants. This domain allows for a reversible phosphorylation process, producing a conformational change affecting the protein function, such as enzyme activity, protein-protein associations, or even cellular location (13, 14).

One limitation is the lack of supporting studies (functional assays or associations), to confirm our results. The intrinsic prediction limits of the IS tools must be considered, as IS results should be verified by *in vitro* or *in vivo* studies to be conclusive (15).

Regarding future directions, further IS analysis, such as hydrogen bonds, clashes or contacts, molecular docking, or molecular dynamics, surely will improve our findings.

6. Conclusions

IS analysis allows a first-hand, fast, and free approach

to studying SNV. Our methodology can be applied to a single or multiple SNV (hundreds to thousands) with relatively small-time differences. In the *AURKA* model, a filter was not useful to discriminate variants; however, for other genes, each filter might discard none, few, or several variants. This is due to the nature of each variant and highlights the difference between each prediction tool.

To the best of our knowledge, no single tool or IS methodology can fully effectively predict variant pathogenicity. However, a feasible approach is to select those SNVs with higher pathogenic potential and carry out further studies (experimental or computational), which surely help clarify a variant biological impact.

Abbreviations

AURKA: Aurora Kinase A; AK: Aurora Kinases; MSV: Missense Variants; AA: Amino acid; VUS: Variant of Unknown Significance; IS: In Silico; SNV: Single nucleotide variant; PDB: Protein Data Bank; RSA: Relative Solvent Accessibility.

Acknowledgments

The authors thank CONAHCYT, CUCS-UDG, and CIBO-IMSS for providing the facilities to carry out this work.

Declarations

Ethics approval and consent to participate

Used data is publicly available. No ethics approval was required for this research; thus, neither consent to participate was required.

Consent for publication

Due to the required data being publicly available, no consent for publication was required for this research

Availability of data and material

All required data can be sent to replicate results or verify the steps upon reasonable request.

Competing interests

The authors declare no competing interests.

Funding

No funding was required for this research.

Authors' contributions

Study design EJMC, MPGA; Data mining EJMC, RHP, LGFR, LEFV; *In silico* tools management EJMC; Data analysis EJMC, MPGA, RHP, LGFR, LEFV; Manuscript preparation EJMC, MPGA, RHP, LGFR, LEFV.

References

1. Arshad S, Ishaque I, Mumtaz S, Rashid MU, Malkani N. In-Silico Analyses of Nonsynonymous Variants in the BRCA1 Gene. *Biochem Genet*. 2021;**59**(6):1506-1526. doi: 10.1007/s10528-021-10074-7
2. Venselaar H, Te Beek TA, Kuipers RK, Hekkelman ML, Vriend G. Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. *BMC Bioinformatics*. 2010;**11**:548. doi: 10.1186/1471-2105-11-548
3. Golmohammadi R, Namazi MJ, Going JJ, Derakhshan MH. A single nucleotide polymorphism in codon F31I and V57I of the AURKA gene in invasive ductal breast carcinoma in Middle East. *Medicine (Baltimore)*. 2017;**96**(37):e7933. doi: 10.1097/md.0000000000007933
4. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, *et al*. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*. 2015;**17**(5):405-424. doi: 10.1038/gim.2015.30
5. Crane R, Gadea B, Littlepage L, Wu H, Ruderman JV. Aurora A, Meiosis and Mitosis. *Biology of the Cell*. 2004;**96**(3):215-229. doi: 10.1016/j.biolcel.2003.09.008
6. Mahmoodi S, Amirzakaria JZ, Ghasemian A. In silico design and validation of a novel multi-epitope vaccine candidate against structural proteins of Chikungunya virus using comprehensive immunoinformatics analyses. *PLOS ONE*. 2023;**18**(5):e0285177. doi: 10.1371/journal.pone.0285177
7. Ittisoponpisan S, Islam SA, Khanna T, Alhuzimi E, David A, Sternberg MJE. Can Predicted Protein 3D Structures Provide Reliable Insights into whether Missense Variants Are Disease Associated? *J Mol Biol*. 2019;**431**(11):2197-2212. doi: 10.1016/j.jmb.2019.04.009
8. Ayariga JA, Villafane R. Single Amino Acid Change Mutation in the Hydrophobic Core of the N-terminal Domain of P22 TSP affects the Proteins Stability. 2021.
9. Melnikov S, Mailliot J, Rigger L, Neuner S, Shin BS, Yusupova G, *et al*. Molecular insights into protein synthesis with proline residues. *EMBO Rep*. 2016;**17**(12):1776-1784. doi: 10.15252/embr.201642943
10. Högel P, Götz A, Kuhne F, Ebert M, Stelzer W, Rand KD, *et al*. Glycine Perturbs Local and Global Conformational Flexibility of a Transmembrane Helix. *Biochemistry*. 2018;**57**(8):1326-1337. doi: 10.1021/acs.biochem.7b01197
11. Bauer F, Sticht H. A proline to glycine mutation in the Lck SH3-domain affects conformational sampling and increases ligand binding affinity. *FEBS Lett*. 2007;**581**(8):1555-1560. doi: 10.1016/j.febslet.2007.03.012
12. Gerasimavicius L, Livesey BJ, Marsh JA. Loss-of-function, gain-of-function and dominant-negative mutations have profoundly different effects on protein structure. *Nature Communications*. 2022;**13**(1). doi: 10.1038/s41467-022-31686 -6

13. Hanks SK, Quinn AM, Hunter T. The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. *Science*. 1988;**241**(4861):42-52. doi: 10.1126/science.3291115
14. Khan FA. Biotechnology fundamentals2020.
15. Ernst C, Hahnen E, Engel C, Nothnagel M, Weber J, Schmutzler RK, *et al*. Performance of in silico prediction tools for the classification of rare BRCA1/2 missense variants in clinical diagnostics. *BMC Medical Genomics*. 2018;**11**(1). doi: