Short communication

Unleashing the Power of MutationTaster2 and MutationTaster2021: The Machine Learning Approach to Genetic Variant Analysis

Desencadenando el poder de MutationTaster2 y MutationTaster2021: el enfoque de aprendizaje automático para el análisis de variantes genéticas

Neelabh Datta 0000-0002-1577-5461

Department of Biochemistry. Asutosh College (Affiliated to University of Calcutta). India.

Corresponding author: neelabhdatta@gmail.com

ABSTRACT

MutationTaster is a widely used web-based tool that predicts the functional impact of genetic variants. In recent years, the software has undergone significant improvements, leading to the development of MutationTaster2 and MutationTaster2021. The main difference between these two versions is the use of updated reference datasets and an improved algorithm for variant classification. MutationTaster2 utilizes the dbNSFP database, while MutationTaster2021 incorporates gnomAD and ClinVar data. Both versions employ a machine learning approach that combines multiple features to predict variant pathogenicity, including evolutionary conservation, physical properties of amino acid changes, and the potential effect on protein function. The output of MutationTaster is a score indicating the likelihood of a variant being disease causing, with a high score indicating a high likelihood of pathogenicity. Overall, MutationTaster2 and MutationTaster2021 represent valuable tools for researchers and clinicians in the field of genetic variant analysis, providing accurate and efficient predictions of variant pathogenicity.

Keywords: MutationTaster2; MutationTaster2021; ExAC.

RESUMEN

MutationTaster es una herramienta web ampliamente utilizada que predice el impacto funcional de las variantes genéticas. En los últimos años, el software ha experimentado mejoras significativas, lo que ha llevado al desarrollo de MutationTaster2 y MutationTaster2021. La principal diferencia entre estas dos versiones es el uso de conjuntos de datos de referencia actualizados y un algoritmo mejorado para la clasificación de variantes. MutationTaster2 utiliza la base de datos dbNSFP, mientras que MutationTaster2021 incorpora datos de gnomAD y ClinVar. Ambas versiones emplean un enfoque de aprendizaje automático que combina múltiples características para predecir la patogenicidad variante, incluida la conservación evolutiva, las propiedades físicas de los cambios de aminoácidos y el

efecto potencial en la función de la proteína. El resultado de MutationTaster es una puntuación que indica la probabilidad de que una variante cause una enfermedad; una puntuación alta indica una alta probabilidad de patogenicidad. En general, MutationTaster2 y MutationTaster2021 representan herramientas valiosas para investigadores y médicos en el campo del análisis de variantes genéticas, ya que proporcionan predicciones precisas y eficientes de la patogenicidad de variantes.
Palabras clave: MutationTaster2; MutationTaster2021; ExAC.

Received: 26/12/2022
Approved: 27/03/2023

MutationTaster2 and MutationTaster2021

“Here we present MutationTaster2 (http://www.mutationtaster.org/), the latest version of our web-based software MutationTaster1, which evaluates the pathogenic potential of DNA sequence alterations”. Jana Marie Schwarz and D. Cooper's paper titled "MutationTaster2: mutation prediction for the deep-sequencing age" presents a novel tool for predicting the functional consequences of genetic mutations. [(1)] The tool, called MutationTaster2, is based on machine learning algorithms and is designed to analyse whole genome and exome sequencing data to identify potentially deleterious mutations. MutationTaster2 is a powerful bioinformatics tool designed to predict the functional effects of DNA sequence variations, specifically in the context of human genetics. This tool is a sequel to the original MutationTaster tool, which was released in 2008, optimized in 2014 to handle the enormous data output generated by modern high-throughput sequencing technologies, including whole-genome and whole-exome sequencing. MutationTaster2 is an important contribution to the field of genetics and bioinformatics, as it provides valuable information for predicting the impact of genetic variants on protein function, and therefore disease causation.

The article "MutationTaster2: mutation prediction for the deep-sequencing age" by Schwarz et al., published in the journal Nature Methods in 2014, describes the development and validation of the MutationTaster2 tool. The article provides a detailed description of the various features and functions of the tool, as well as a comprehensive evaluation of its performance and accuracy. Another article “MutationTaster2021” highlights the advantages of the latest version of MutationTaster2, named MutationTaster2021, compared to the previous version. [(2)] In this summary, I will delve into the details of MutationTaster2, including its strengths and limitations, its relevance and utility in modern genetic research, its restrictions and the advantages of its latest version MutationTaster2021.

MutationTaster2 is based on a set of bioinformatics algorithms that use various features and characteristics of a DNA sequence variant to predict its functional consequences. These algorithms take into account factors such as conservation, physical properties of the amino

acid changes, and the location of the variant in relation to known functional elements such as splice sites, regulatory regions, and protein domains. MutationTaster2 provides two types of prediction scores for each variant: a binary classification into either “disease-causing” or “benign” categories or a probability score indicating the likelihood of the prediction being correct. The output of MutationTaster2 is also supplemented with various annotations and functional predictions, including splice site analysis, predictions of altered transcription factor binding, and protein domain predictions.

One of the major findings of the paper is that MutationTaster2 is able to accurately predict the functional consequences of mutations with high sensitivity and specificity. The authors demonstrated the performance of MutationTaster2 using a variety of datasets, including whole genome sequencing data from individuals with Mendelian disorders and large scale exome sequencing data from healthy individuals. [1] They found that MutationTaster2 was able to accurately classify mutations as either pathogenic or benign with high accuracy, outperforming other available tools. [1] This is a significant finding, as it demonstrates the ability of MutationTaster2 to accurately identify potentially pathogenic mutations among the many neutral or benign variations that are present in the genome. This is particularly important for the identification of mutations associated with disease, as it allows for the accurate identification of mutations that may be contributing to the development or progression of a particular disorder. Another important aspect of the paper is the efficiency of MutationTaster2. The authors demonstrated that MutationTaster2 was able to analyse large amounts of data in a relatively short amount of time, making it a valuable tool for the analysis of whole genome and exome sequencing data. [1] This is particularly important in the era of deep sequencing, where large amounts of data are generated and the analysis of this data can be challenging.

MutationTaster2 has several advantages over other mutation prediction tools. One of the key strengths of MutationTaster2 is its high accuracy, which has been demonstrated in several benchmarking studies. For example, the authors of the article conducted an analysis of 18,832 genetic variants from the Human Gene Mutation Database (HGMD) and found that MutationTaster2 achieved a sensitivity of 88.8% and a specificity of 79.7%. This performance was better than that of several other widely used prediction tools, including PolyPhen-2, SIFT, and MutationAssessor. [2-5] Another advantage of MutationTaster2 is its ability to handle large-scale datasets, making it well-suited for the analysis of high-throughput sequencing data. One of the major limitations of MutationTaster2 is its reliance on a set of pre-defined rules and algorithms, which may not always capture the complex and context-dependent effects of genetic variation. For example, MutationTaster2 may not be able to accurately predict the functional consequences of variants that affect protein-protein interactions, or variants that act in a tissue-specific manner. Furthermore, MutationTaster2 may be less effective for variants that are rare or that have not been previously reported in the literature or in variant databases. Nevertheless, the authors of the article acknowledge these limitations and suggest that on-going improvements and enhancements to the tool will help to address these issues.

A new version of MutationTaster, named MutationTaster2021, which employs a different prediction model than its predecessor, achieves more accuracy, particularly for uncommon benign varieties. MutationTaster now offers details on the illnesses they cause, making it easier to evaluate the relation of discovered recognised disease mutations to the clinical phenotype of the patient. To prioritise variations from VCF files based on the patient's clinical phenotype, MutationTaster2021 incorporates a disease mutation search engine, MutationDistiller. [2), (6] High-throughput sequencing has totally altered the picture, because in the past the inheritance of disease-linked areas was explored by linkage analysis, and positional and functional candidate genes were subsequently sequenced for potential mutations. [2] Due to the lack of need for linkage data, biomedical researchers may now concentrate on variations that are expected to have a negative impact on genes involved in disease aetiology. Hence, thousands of potentially harmful variations still need to be evaluated considering that, tens of thousands of DNA variants are discovered in each WES run. ExAC, gnomAD, or other large-scale sequencing initiatives like the 1000 Genomes Project, as well as other known polymorphisms, can be used to further minimise the number of potentially disease-causing variations. [7-9] Nevertheless, this method can only exclude a percentage of the benign variations since many polymorphisms are population-specific. A Random Forest model has been used in place of the Bayes classifier to improve prediction accuracy in both benign and harmful variants. Even though the false positive rate was significantly lowered by filtering out common polymorphisms, numerous uncommon or population-specific variations continued to be false positives. This problem was overcome by using all intragenic gnomAD mutations for which there was at least one homozygous carrier as benign training instances. [2] The prediction method was changed from Naive Bayes to Random Forest models in order to enhance the results. Grid searching revealed that using Random Forests with only one-third the size of the "ideal forest" may be employed in two prediction models without sacrificing more than 0.12% of balanced accuracy. [2),(9] The fact that these predictors were specifically trained for balanced accuracy—that is, the same predictive performance for benign and harmful variants—should be underlined. In contrast to predictors trained for specificity, this minimises the danger of missing an actual disease mutation even while it increases the frequency of false positive predictions.

MaxEntScan does have the limitation that it can only detect variations in canonical splice sites. It should be noted that MutationTaster2 and MutationTaster2021 do not look for cryptic splice sites that are activated by DNA variations since it was discovered that doing so would result in an excessive number of false positive predictions. [2] Especially notable are ExAC pLI scores to determine if a gene is tolerant of loss-of-function mutations and genotype counts from ExAC and gnomAD for the elimination of variations prevalent in healthy persons. [2),(7),(8] ExAC, gnomAD, and homozygous individuals from the 1000 Genomes Project are utilised to automatically identify variations as benign, but the pLI scores are not. [7-9] A query of several variations may be made automatically from within other programmes using an API provided by MutationTaster2021. Considering the VCF analysis pipeline generates predictions, the API has been limited to 50 variations per call instead of allowing users to upload VCF files for larger variant sets. [2] These predictions are then stored in the database. The updated version's

modifications enable a much quicker and more precise forecast of the impact of DNA variations. [2] The overall accuracy is improved by the Random Forest classifiers for non-coding variations from 92.2% to 97.0%, for variants producing single amino acid substitutions from 88.6% to 95.8%, and for variants causing changes that are more substantial in the amino acid sequence from 90.7% to 93.3%. [2]

Overall, the development of MutationTaster2 represents a significant advance in the field of mutation prediction. It has the potential to improve our understanding of the functional consequences of mutations and their role in the development of diseases and disorders. It is a valuable tool for the analysis of genomic data and will be useful for researchers studying the genetic basis of a variety of diseases and conditions. MutationTaster2021 is explicitly aimed at biomedical researchers who want to identify the pathological mutation in a patient suffering from a suspected monogenic disease. The information associated with a variant is presented in a user-friendly interface unlike other tools such as CADD. As with any classifier, a number of variants will be misclassified. [12] This becomes especially apparent for benign variants. In conclusion, MutationTaster2 and MutationTaster2021 are powerful tools for the prediction of the functional impact of genetic variants. Both methods are able to analyse both known and novel variants, and use a range of bioinformatics features to predict the potential impact of the variant on protein function. The original MutationTaster2 algorithm has been widely used and has been shown to be highly accurate, with a good balance between sensitivity and specificity. The more recent MutationTaster2021 algorithm builds upon the success of MutationTaster2 and adds new features to improve the accuracy of the predictions, particularly for non-coding variants. The incorporation of regulatory regions, splicing features, and machine learning models has resulted in a higher accuracy and a more nuanced prediction of variant impact. Both MutationTaster2 and MutationTaster2021 are useful tools for researchers and clinicians working in the field of genetics. They can be used to prioritize variants for further functional analysis or to guide clinical decision-making. As the field of genomics continues to expand, tools like these will become increasingly important for accurately interpreting genetic data and improving our understanding of genetic disease.

References

1. Schwarz JM, Cooper DN, Schuelke M, Seelow D. (2014). MutationTaster2: mutation prediction for the deep-sequencing age. Nature Methods, 11, 361-36 [Cited Feb 2023] Available: https://www.researchgate.net/publication/261220703_MutationTaster2_Mutation_prediction_for_the_deep-sequencing_age

2. Steinhaus R, Proft S, Schuelke M, Cooper DN, Schwarz JM, Seelow D, (2021), MutationTaster2021, Nucleic Acids Research, Volume 49, Issue W1, Pages W446–W451. [Internet]. [cited Feb 2023] Available in: https://doi.org/10.1093/nar/gkab266

3. 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), 248–249. [Internet]. [cited Feb 2023] Available in: https://doi.org/10.1038/nmeth0410-248
4. Ng PC. and Henikoff S. (2001) Predicting deleterious amino acid substitutions. Genome Res, 11, 863–874. [Cited Feb 2023] Available: https://europepmc.org/backend/ptpmcrender.fcgi?accid=PMC311071&blobtype=pdf
5. Reva B, Antipin Y, Sander C. (2011). Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic acids research, 39(17), e118. [Internet]. [cited Feb 2023]Available in: https://europepmc.org/article/MED/21727090
6. Hombach D, Schuelke M, Knierim E, Ehmke N, Schwarz JM, Fischer-Zirnsak B, et Al. MutationDistiller: user-driven identification of pathogenic DNA variants, *Nucleic Acids Research*, Volume 47, Issue W1, 02 July 2019, Pages W114–W120, [Cited Mar. 2023] Available: https://doi.org/10.1093/nar/gkz330
7. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, et Al. Exome Aggregation Consortium (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature, 536(7616), 285–291. [Internet]. [cited Feb 2023] Available in: https://doi.org/10.1038/nature19057
8. Karczewski KJ, Francioli LC, Tiao G. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 581, 434–443 (2020). [Internet]. [cited Feb 2023] Available in: https://doi.org/10.1038/s41586-020-2308-7
9. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini, JL, McCarthy S, McVean GA, Abecasis GR. (2015). A global reference for human genetic variation. Nature, 526(7571), 68–74. [Internet]. [cited Feb 2023] Available in: https://doi.org/10.1038/nature15393
10. Ho TK. (1995). Random decision forests. In Proceedings of 3rd international conference on document analysis and recognition (Vol. 1, pp. 278–282).
11. Yeo G, Burge CB.(2003). Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals, RECOMB. [cited Mar 2023 2023] Available in: https://www.researchgate.net/publication/8423973_Maximum_Entropy_Modeling_of_Short_Sequence_Motifs_with_Applications_to_RNA_Splicing_Signals
12. Kircher M, Witten DM JP, O'Roak BJ, Cooper GM, Shendure J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nature genetics, 46(3), 310–315. [Internet]. [cited Mar 2023] Available: https://www.researchgate.net/publication/260039664_A_General_Framework_for_Estimating_the_Relative_Pathogenicity_of_Human_Genetic_Variants

Conflicts of interests

There are no conflicts of interests to declare.