Bioinformatics, 2024, **40(4)**, btae197
https://doi.org/10.1093/bioinformatics/btae197
Advance Access Publication Date: 11 April 2024
Applications Note

OXFORD

Data and text mining

DrivR-Base: a feature extraction toolkit for variant effect prediction model construction

Amy Francis [1,*], **Colin Campbell**[2], **Tom R. Gaunt** [1]

[1]MRC Integrative Epidemiology Unit, Bristol Medical School (PHS), University of Bristol, Bristol BS8 2BN, United Kingdom
[2]Intelligent Systems Laboratory, University of Bristol, Bristol BS1 5DD, United Kingdom

*Corresponding author. MRC Integrative Epidemiology Unit, Bristol Medical School (PHS), University of Bristol, Oakfield House, Bristol BS8 2BN, United Kingdom E-mail: amy.francis@bristol.ac.uk (A.F.)

Associate Editor: Jonathan Wren

Abstract

Motivation: Recent advancements in sequencing technologies have led to the discovery of numerous variants in the human genome. However, understanding their precise roles in diseases remains challenging due to their complex functional mechanisms. Various methodologies have emerged to predict the pathogenic significance of these genetic variants. Typically, these methods employ an integrative approach, leveraging diverse data sources that provide important insights into genomic function. Despite the abundance of publicly available data sources and databases, the process of navigating, extracting, and pre-processing features for machine learning models can be highly challenging and time-consuming. Furthermore, researchers often invest substantial effort in feature extraction, only to later discover that these features lack informativeness.

Results: In this article, we introduce *DrivR-Base*, an innovative resource that efficiently extracts and integrates molecular information (features) related to single nucleotide variants. These features encompass information about the genomic positions and the associated protein positions of a variant. They are derived from a wide array of databases and tools, including structural properties obtained from *AlphaFold*, regulatory information sourced from ENCODE, and predicted variant consequences from *Variant Effect Predictor*. *DrivR-Base* is easily deployable via a Docker container to ensure reproducibility and ease of access across diverse computational environments. The resulting features can be used as input for machine learning models designed to predict the pathogenic impact of human genome variants in disease. Moreover, these feature sets have applications beyond this, including haploinsufficiency prediction and the development of drug repurposing tools. We describe the resource's development, practical applications, and potential for future expansion and enhancement.

Availability and implementation: *DrivR-Base* source code is available at https://github.com/amyfrancis97/DrivR-Base.

Introduction

The rapid advancement in Next Generation Sequencing technologies has facilitated the extensive identification of variants within the human genome. A significant number of these variants have an unknown functional impact. Among these, many could potentially contribute to disease phenotypes as driver variants, while others are likely to be passively involved and causatively neutral in nature.

In response, a diverse range of machine learning methodologies have been proposed, with the primary objective of integrating genome-level information (features) to identify driver variants. Notable tools in this context include DeepMinds' most recent piece of work, *AlphaMissense*, (Cheng *et al.* 2023), our *FATHMM-MKL* (Shihab *et al.* 2015) and *CScape* (Rogers *et al.* 2017) predictors, as well as *CADD* (Rentzsch *et al.* 2019), *DANN* (Quang *et al.* 2015), *PolyPhen-2* (Adzhubei *et al.* 2013), and *EVE* (Frazer *et al.* 2021). While these tools employ diverse methodologies to tackle genomic prediction problems, the datasets, or features, integrated into the models prove equally crucial, and the utility of these classifiers heavily relies on the availability of feature data.

To our knowledge, *DrivR-Base* represents the first tool available to the research community that offers such a comprehensive and extensive compilation of annotations across the entire genome (Wang *et al.* 2010, McLaren *et al.* 2016, Liu *et al.* 2020). With its unique capability to integrate a wide array of detailed features and annotations from numerous databases, *DrivR-Base* stands out for its unparalleled breadth and depth of genomic and protein-level information accessible for extraction. Moreover, most modern tools focus on aggregating scores from machine learning models associated with a variant, rather than providing access to the raw annotations themselves (Liu *et al.* 2020). *DrivR-Base*, therefore, provides an unprecedented resource for the direct application in machine learning models to accelerate the development of variant prediction tools.

To date, numerous features have demonstrated their effectiveness in assessing the likelihood of a variant driving disease. Conservation-based features, such as PhyloP and PhastCons scores (Siepel *et al.* 2005, Pollard *et al.* 2009), quantify sequence conservation across species. Studies have suggested that regions with lower conservation tend to be less functionally significant (Woodruff 2001). These features have proven informative in several predictors (Shihab *et al.* 2015, Rentzsch *et al.* 2019, Sun and Yu 2019, Cabrera-Alarcon *et al.* 2022).

Received: 5 October 2023; **Revised:** 1 March 2024; **Editorial Decision:** 8 April 2024; **Accepted:** 9 April 2024

Additionally, various other features have played vital roles in driver-variant prediction. For instance, the *Variant Effect Predictor (VEP)* (McLaren *et al.* 2016) has been instrumental in developing widely-used prediction tools (Shihab *et al.* 2015, Rentzsch *et al.* 2019). *VEP* provides valuable insights into variant effects on transcripts within protein-coding regions, introns, and regulatory elements. Moreover, this context has seen the utilization of features such as sequence-based similarity measures, enabling mathematical comparisons of wild-type and mutant string patterns (e.g. spectrum kernels), as well as regulatory features from ENCODE (Dunham *et al.* 2012, Quang *et al.* 2015, Shihab *et al.* 2015, Rogers *et al.* 2017, Rentzsch *et al.* 2019). Additionally, information on GC content and CpG islands has proven valuable in these prediction tasks (Shihab *et al.* 2015, Rogers *et al.* 2017). Elevated GC content has been associated with increased bendability and the ability to undergo B-Z transitions, which are spatial features linked to open chromatin and active transcription (Vinogradov 2003).

While various feature groups are currently in use, additional molecular datasets could likely offer valuable insights in predicting driver variants. For instance, exploring the influence of single nucleotide variants (SNVs) on DNA shape properties is one such illustration. Multiple DNA shape properties have been implicated in DNA–protein interactions (Jones *et al.* 2003, Rohs *et al.* 2009, Chiu *et al.* 2017). Specifically, high electrostatic potentials have been linked to DNA binding sites (Jones *et al.* 2003, Chiu *et al.* 2017), and the narrowing of minor grooves has been associated with A-tracts, resulting in bending toward the minor groove (Rohs *et al.* 2009). As a result, SNVs occurring at these sites may disrupt these interactions and could lead to functional consequences.

Furthermore, other features that have not been extensively explored in this context include structural information sourced from the AlphaFold (Jumper *et al.* 2021) and PDB (Berman *et al.* 2000) databases. These databases contain a wealth of information that could prove valuable when assessing whether a genomic variant is likely to lead to disease. Other examples of feature groups that have not been widely employed thus far and are presented in this work include dinucleotide and amino acid properties.

In this paper, we introduce the creation of a novel repository, named *DrivR-Base*, designed to streamline the data acquisition process for constructing robust predictors of variant driver status. These datasets have broader applications, including the development of haploinsufficiency prediction models (Shihab *et al.* 2017) and potential adaptation for advancing drug repurposing tools (Irham *et al.* 2022). We focus on the human genome, providing users with a comprehensive toolkit of scripts, documentation, and links to original sources to build the required feature set. The deployment of bioinformatics tools across varied computational environments often presents a significant challenge due to dependency management and configuration issues. To address this, we have containerized *DrivR-Base* using Docker, ensuring that researchers can deploy our toolkit effortlessly, without the need to manage individual software dependencies. Further details can be found in the Supplementary section.

Description and implementation

DrivR-Base is a feature extraction toolkit that enables efficient integration of genomic and protein-level annotations for all possible combinations of single nucleotide variants in the GRCh38 build of the human genome (including all four possible nucleotides at a given position). The resulting features have a wide range of applications, including direct integration into machine learning models for variant effect prediction. The output of *DrivR-Base* is a single file where the variants are represented as rows, with a column dedicated to feature values for each of the attributes described below. The tool is fully containerised for Docker, facilitating straightforward installation and execution. The tool extracts information for ten different feature groups (FG) from human single nucleotide variants, which are mainly extracted from public databases:

i) *Conservation-based features*: Conservation-based features encompass several crucial metrics. These include PhyloP and PhastCons (Siepel *et al.* 2005, Pollard *et al.* 2009) scores, which assess whether nucleotide substitution rates deviate from the expectations under neutral drift. Each of these scores is obtained using seven different alignment methods. Additionally, our analysis incorporates Umap and Bismap mappability data (Karimzadeh *et al.* 2018), measured using four different types of species alignment methods. These metrics assess the extent to which a genomic region can be accurately mapped during sequencing, providing insights into the reliability of genomic or epigenomic characteristics. Regions exhibiting lower mappability readings may be more prone to error. To obtain these datasets for the entire genome, we retrieve data from the UCSC genome browser (Kent *et al.* 2002) and tailor our queries to specific input variants.

ii) *Variant Effect Predictor:* The *VEP* (McLaren *et al.* 2016) is organized into three main groups of features. Firstly, we extract all predicted transcript consequences for each variant and encode them using one-hot encoding. The outcome is a file that displays a "1" in the corresponding row for each variant if the transcript consequence is predicted. Next, we retrieve the predicted wild-type and mutant amino acids, presenting the results in two files. The first file follows a BED+2 format, with the final two rows representing the wild-type and mutant amino acids, respectively. For synonymous variants, the amino acids will be the same. Additionally, we generate another file that is one-hot encoded, making it suitable for direct integration into the user's models. Finally, we extract distances to transcripts. When variants are predicted to affect multiple transcripts, we calculate their mean, maximum, and minimum distances.

iii) *Dinucleotide properties:* This feature dataset is sourced from DiProDB, an extensive database encompassing 125 conformational and thermodynamic dinucleotide properties (Friedel *et al.* 2009), which provides values for four dinucleotide configurations: (a) The wild-type allele paired with the adjacent allele on the left, (b) The wild-type allele paired with the adjacent allele on the right, (c) The mutant allele paired with the adjacent allele on the left, and (d) The mutant allele paired with the adjacent allele on the right. The resulting table contains columns, each representing one of the 125 different properties. Column names include a prefix specifying which of the four configurations it pertains to. For example,

"1_Propeller_Twist" denotes the value for the propeller twist property in the first configuration.

iv) *DNA shape properties:* Here, we incorporate five DNA shape properties from DNAShapeR (Chiu *et al.* 2016). DNAShapeR employs a sliding-window approach to calculate minor groove width (MGW), helix twist (HelT), propeller twist (ProT), roll (Roll), and electrostatic potential (EP). In our scripts, we extract DNA shape features within a window of +10 and −10 on either side of the variant, but this can be easily adjusted by the user. The output is presented in a table, displaying the value for each DNA shape feature for every calculated base pair, where position 11 corresponds to the variant of interest.

v) *GC content and CpG sites: DrivR-Base* also calculates GC content, CpG counts and observed CpG versus expected CpG ratios for nine different window sizes.

vi) *Kernel-based sequence similarity:* Our approach also employs sequence-based p-spectrum kernels to capture potential disruptions in sequences flanking a single nucleotide variant (Campbell and Ying 2011). Spectrum kernels allow us to assess the composition of k-mers within the genomic regions surrounding a mutation. We explore various window sizes ranging from 2 to 20 and k-mer sizes ranging from 1 to 20. For each chosen window size (w), we systematically generate all possible combinations of specified k-mer sizes for both wild-type and mutant sequences. We then determine the frequency of occurrence for each k-mer in the respective sequences using the following mapping function:

$$\Phi_u^p(s) = |\{(v_1, v_2) : s = v_1 u v_2\}|$$

Here, u represents the sub-string k-mer of length p, v_1 denotes the wild-type sequence, v_2 refers to the mutant sequence, and s represents the sequence of interest. We subsequently derive a p-spectrum kernel by summing the products of corresponding row entries for the two sequences:

$$K_p(s,t) = \sum_{u \in \Sigma_p} \Phi_u^p(s) \Phi_u^p(t)$$

In this equation, s corresponds to the wild-type sequence, and t corresponds to the mutant sequence. We calculate the diagonals of the p-spectra by summing the squares of corresponding row entries within the mapping function matrix. For a more comprehensive explanation and detailed Python implementation, please refer to our Supplementary material and GitHub Repository.

vii) *Amino acid substitution matrices:* In this study, we extract amino acid substitution rates from a variety of matrices for non-synonymous variants sourced from the *Bio2mds* package in R (Pelé *et al.* 2012). The matrices used and their sources are shown in Table 1.

viii) *Amino acid properties: DrivR-Base* retrieves 532 amino acid properties for both wild-type and mutant amino acid sequences. These properties were sourced from the *AAindex* data within the *AAsea* package in R (Reddy 2019). They encompass information related to factors such as polarity, hydrophobicity, local flexibility, and helix-bend preferences.

Table 1. Amino acid substitution matrices and their sources.

Matrix type	Source
PAM40	Pelé *et al.* 2012
PAM160	Pelé *et al.* 2012
PAM250	Pelé *et al.* 2012
BLOSUM30	Henikoff and Henikoff 1992
BLOSUM45	Henikoff and Henikoff 1992
BLOSUM62	Henikoff and Henikoff 1992
GONNET	Gonnet *et al.* 1992
JTT	Jones *et al.* 1992
JTT_TM	Jones *et al.* 1994
PHAT	Ng *et al.* 2000

ix) *ENCODE database features:* ENCODE offers a wealth of functional information about the human genome (Dunham *et al.* 2012). In this work, we extract eight features potentially informative for variant pathogenicity:
   a) Transcription Factor ChIP-seq
   b) Histone ChIP-seq
   c) DNase-seq
   d) Mint-ChIP-seq
   e) ATAC-seq
   f) eCLIP
   g) ChIA-PET
   h) GM DNase-seq

To achieve this, we retrieve all available files for each feature group from ENCODE via the ENCODE API. Subsequently, we download, convert, and consolidate ENCODE peak files into comprehensive data frames for each feature group. These data frames include metadata like accession, target (e.g. transcription factor), biosample (e.g. cell/tissue type), and output type (e.g. narrow peak). Note that this script downloads all ENCODE data locally, requiring approximately 160GB of space.

Next, we cross-reference feature-specific databases with target SNVs, extracting relevant information overlapping with SNV locations. We then extract crucial data such as signal values, P-values, q-values, and peaks for each variant. For cases with multiple peaks, such as when replicate assays are involved, we also record minimum, maximum, mean, and range values.

x) *AlphaFold structural features: DrivR-Base* incorporates structural data from the AlphaFold database (Jumper *et al.* 2021) and PDB (Berman *et al.* 2000). Using the VEP query output, we identify genes and protein positions affected by coding variants. Gene names are converted to UniProtKB IDs, and an API retrieves corresponding crystallographic information files (CIF; *.cif*) from AlphaFold based on the UniProtKB ID. We extract structural information, including X, Y, and Z atom coordinates, isotropic atomic displacement parameters (IADP), and structural conformation types. The output includes two data frames: one containing the first four features (X, Y, Z coordinates, and IADP) for each variant, and another data frame with one-hot-encoded structural conformation types indicating potential effects on amino acids, such as bends or helical structures.

A detailed list of feature groups, their sources, and their implementation can be found in our Supplementary material.

Conclusions and future efforts

In summary, *DrivR-Base* is a versatile cross-database toolkit that consolidates diverse features for human SNVs. These

features have various applications, including constructing high-dimensional machine-learning models for predicting variant driver status. As previously commented, *DrivR-Base* can also be applied to predict haploinsufficient genes and to identify functional similarities to known drug targets, potentially aiding drug repurposing efforts. This tool streamlines feature extraction, saving researchers time and advancing their work. Our future goals include expanding the tool's capabilities to encompass a broader range of mutations, such as indels, deletions, and structural rearrangements, and diversifying the available feature groups for extraction. *DrivR-Base* is fully containerised for easy deployment using Docker, ensuring a reproducible and streamlined setup process. Detailed instructions for Docker deployment, including pulling the image, running the container, and executing the toolkit, are available in our comprehensive GitHub documentation at https://github.com/amyfrancis97/DrivR-Base. Researchers are encouraged to contact the authors to discuss the inclusion of additional feature groups in *DrivR-Base* or the enhancement of existing feature groups.

Acknowledgements

This work was carried out in the UK Medical Research Council Integrative Epidemiology Unit (MC_UU_00032/03) and using the computational facilities of the Advanced Computing Research Centre, University of Bristol. For the purpose of open access, the author(s) has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

Supplementary data

Supplementary data are available at *Bioinformatics* online.

Conflict of interest

None declared.

Funding

This work was funded by Cancer Research UK [C18281/A30905].

Data availability

DrivR-Base is open sourced and all code is available on GitHub https://github.com/amyfrancis97/DrivR-Base.

References

Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using polyphen-2. *Curr Protoc Hum Genet* 2013;Chapter 7:Unit7.20.

Berman HM, Westbrook J, Feng Z *et al.* The protein data bank. *Nucleic Acids Res* 2000;**28**:235–42.

Cabrera-Alarcon JL, Martinez JG, Enríquez JA *et al.* Variant pathogenic prediction by locus variability: the importance of the current picture of evolution. *Eur J Hum Genet* 2022;**30**:555–9.

Campbell C, Ying Y. *Learning with Support Vector Machines*. Kentfield, CA 94914, US: Morgan & Claypool Publishers, 2011.

Cheng J, Novati G, Pan J *et al.* Accurate proteome-wide missense variant effect prediction with alphamissense. *Science* 2023; **381**:eadg7492.

Chiu TP, Comoglio F, Zhou T *et al.* Dnashaper: an r/bioconductor package for dna shape prediction and feature encoding. *Bioinformatics* 2016;**32**:1211–3.

Chiu TP, Rao S, Mann RS *et al.* Genome-wide prediction of minor-groove electrostatic potential enables biophysical modeling of protein–DNA binding. *Nucleic Acids Res* 2017;**45**:12565–76.

Dunham I, Kundaje A, Aldred SF *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;**489**:57–74.

Frazer J, Notin P, Dias M *et al.* Disease variant prediction with deep generative models of evolutionary data. *Nature* 2021;**599**:91–5.

Friedel M, Nikolajewa S, Sühnel J *et al.* Diprodb: a database for dinucleotide properties. *Nucleic Acids Res* 2009;**37**:D37–D40.

Gonnet GH, Cohen MA, Benner SA. Exhaustive matching of the entire protein sequence database. *Science*1992;**256**:1443–5.

Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* 1992;**89**:10915–9.

Irham LM, Adikusuma W, Perwitasari DA *et al.* The use of genomic variants to drive drug repurposing for chronic hepatitis b. *Biochem Biophys Rep* 2022;**31**:101307.

Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* 1992; **8**:275–82.

Jones DT, Taylor WR, Thornton JM. A mutation data matrix for transmembrane proteins. *FEBS Lett* 1994;**339**:269–75.

Jones S, Shanahan HP, Berman HM *et al.* Using electrostatic potentials to predict dna-binding sites on dna-binding proteins. *Nucleic Acids Res* 2003;**31**:7189–98.

Jumper J, Evans R, Pritzel A. Highly accurate protein structure prediction with alphafold. *Nature* 2021;**596**:583–9.

Karimzadeh M, Ernst C, Kundaje A *et al.* Umap and bismap: quantifying genome and methylome mappability. *Nucleic Acids Res* 2018; **46**:e120.

Kent WJ, Sugnet CW, Furey TS *et al.* The human genome browser at ucsc. *Genome Res* 2002;**12**:996–1006.

Liu X, Li C, Mou C *et al.* Dbnsfp v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site snvs. *Genome Med* 2020; **12**:103–8.

McLaren W, Gil L, Hunt SE *et al.* The ensembl variant effect predictor. *Genome Biol* 2016;**17**:122–14.

Ng PC, Henikoff JG, Henikoff S. Phat: a transmembrane-specific substitution matrix. predicted hydrophobic and transmembrane. *Bioinformatics*2000;**16**:760–6.

Pelé J, Bécu JM, Abdi H *et al.* Bios2mds: an r package for comparing orthologous protein families by metric multidimensional scaling. *BMC Bioinformatics* 2012;**13**:133–7.

Pollard KS, Hubisz MJ, Rosenbloom KR *et al.* Detection of non-neutral substitution rates on mammalian phylogenies. *Genome Res* 2009; **20**:110–21.

Quang D, Chen Y, Xie X. Dann: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics*2015; **31**:761–3.

Reddy R. aasea: amino acid substitution effect analyser version 1.1.0 from cran, 2019. https://rdrr.io/cran/aaSEA/ (24 April 2024, date last accessed).

Rentzsch P, Witten D, Cooper GM *et al.* Cadd: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res* 2019;**47**:D886–D894.

Rogers MF, Shihab HA, Gaunt TR *et al.* Cscape: a tool for predicting oncogenic single-point mutations in the cancer genome. *Sci Rep* 2017;**7**:11597–10.

Rohs R, West SM, Sosinsky A *et al.* The role of dna shape in protein-dna recognition. *Nature* 2009;**461**:1248–53.

Shihab HA, Rogers MF, Campbell C *et al.* Hipred: an integrative approach to predicting haploinsufficient genes. *Bioinformatics* 2017; **33**:1751–7.

Shihab HA, Rogers MF, Gough J *et al.* An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics*2015;**31**:1536–43.
Siepel A, Bejerano G, Pedersen JS *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 2005;**15**:1034–50.
Sun H, Yu G. New insights into the pathogenicity of non-synonymous variants through multi-level analysis. *Sci Rep* 2019;**9**:1667.
Vinogradov AE. Dna helix: the importance of being gc-rich. *Nucleic Acids Res* 2003;**31**:1838–44.
Wang K, Li M, Hakonarson H. Annovar: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010;**38**:e164.
Woodruff DS. Populations, species, and conservation genetics. In: *Encyclopedia of Biodiversity* Princeton, New Jersey, USA: Simon Asher Levin 2001:811.

Bioinformatics, 2024, 40, 1–5
https://doi.org/10.1093/bioinformatics/btae197
Applications Note