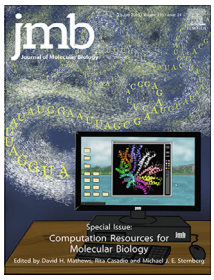

EvoRator2: Predicting Site-specific Amino Acid Substitutions Based on Protein Structural Information Using Deep Learning

Natan Nagar [1], Jérôme Tubiana [2], Gil Loewenthal [1], Haim J. Wolfson [2], Nir Ben Tal [3] and Tal Pupko [1]*

*1 - **The Shmunis School of Biomedicine and Cancer Research,** George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 69978, Israel*
*2 - **Blavatnik School of Computer Science,** Raymond & Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel*
*3 - **School of Neurobiology,** Biochemistry & Biophysics, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 69978, Israel*

***Correspondence to Tal Pupko:** talp@tauex.tau.ac.il (T. Pupko)*
https://doi.org/10.1016/j.jmb.2023.168155
Edited by Rita Casadio

Abstract

Multiple sequence alignments (MSAs) are the workhorse of molecular evolution and structural biology research. From MSAs, the amino acids that are tolerated at each site during protein evolution can be inferred. However, little is known regarding the repertoire of tolerated amino acids in proteins when only a few or no sequence homologs are available, such as orphan and *de novo* designed proteins. Here we present EvoRator2, a deep-learning algorithm trained on over 15,000 protein structures that can predict which amino acids are tolerated at any given site, based exclusively on protein structural information mined from atomic coordinate files. We show that EvoRator2 obtained satisfying results for the prediction of position-weighted scoring matrices (PSSM). We further show that EvoRator2 obtained near state-of-the-art performance on proteins with high quality structures in predicting the effect of mutations in deep mutation scanning (DMS) experiments and that for certain DMS targets, EvoRator2 outperformed state-of-the-art methods. We also show that by combining EvoRator2′s predictions with those obtained by a state-of-the-art deep-learning method that accounts for the information in the MSA, the prediction of the effect of mutation in DMS experiments was improved in terms of both accuracy and stability. EvoRator2 is designed to predict which amino-acid substitutions are tolerated in such proteins without many homologous sequences, including orphan or *de novo* designed proteins. We implemented our approach in the *EvoRator* web server (https://evorator.tau.ac.il).

Introduction

Sequence variation in proteins stems from the filtering of random mutations by evolutionary pressures that act to maintain protein structural and functional integrity.[1–2] Characterizing the residue-level distribution of substitutions is an important task in evolutionary studies, with implications in variant prioritization for clinical diagnostics, *de novo* protein design, and identification of functional sites.[3–4] Experimental methods such as deep mutational scan (DMS) are used to quantify the effect of non-synonymous mutations on a specific phenotypic outcome.[5–7] Such approaches are costly and time consuming and cannot be applied to analyze the large number of protein records accumulating in public databases.

The most common approaches for predicting residue-level tolerated sets of amino acids heavily rely on multiple sequence alignments (MSA):

given a query sequence, homologous sequences are identified and aligned. The various existing methods mainly differ in how they process the resulting MSA to make predictions. Methods such as SIFT[8] and EVCouplings[9] totally depend on MSAs, while methods such as PolyPhen2[10] combine both conservation and structure to make predictions. The highly accurate predictions obtained by methods such as DeepSequence[11] and EVE,[12] depend on the extraction of MSA-based latent variables, which are informative for estimating the probability of observing each substitution. These methods implement deep-learning algorithms that require separate training for each protein family of interest, which can be computationally demanding in large scale analyses. The requirement for an alignment means that residue-level distribution of allowable substitutions cannot be accurately predicted for proteins with a few or no homologs, including orphan proteins, *de novo* designed proteins, and newly emerged viral proteins. In other cases, the information content of the MSA in the presence of indels is the culprit, e.g., it is difficult to predict the effect of substitutions in an alignment site that is mostly gap characters.[1] Novel sequence-based, MSA-independent approaches rely on protein language models that are trained on raw sequence data. As noted by Laine et al., such models are huge in size, require delicate fine tuning, depend on multiple, slow training steps, and require suitable hardware.[13] The most accurate predictions are obtained by an ensemble approach, which combines the large langugae model Tranception and the MSA based model EVE.[14] However, prediction accuracy still highly depends on the number of similar sequences a given protein has in public sequence databases.

The primary determinant of site-specific distribution of substitutions is considered to be the structural context of the site in question.[1–2,15–16] Purifying selection is expected to act on sites with many intramolecular interactions. Such sites are typically located at the tightly-packed core of the proteins.[17–18] Thus, only substitutions to amino acids with similar spatial physicochemical properties as the wild-type are expected in such sites. In contrast, sites that face the solvent are generally expected to tolerate a much larger number of substitutions. In case of sites that are also involved in intermolecular interactions, the type of selection regime and its effect on the distribution of substitutions is context dependent.[1–2,19] For example, strong purifying selection is expected to act on catalytic sites,[20] allosteric sites,[21] post-translationally modified sites,[22] and on sites essential for complex formation.[23] Positive selection due to a changing environment is expected to affect sites that are directly involved in the selected function of a given protein-coding gene.[22] Examples include sites in viral proteins under drug pressure,[24] B-cell epitopes,[25] and toxins.[26] Thus, the residue-level distribution of substitutions observed across protein families evolved under constraints either on intramolecular or intermolecular interactions, or a combination of both. A complete description of these constraints requires knowing all interaction partners and conformational states of a given protein. In most practical cases such data are impossible to obtain. By identifying disagreements between empirical and structure-based expectations on the distribution of substitutions at residue level, one can potentially identify sites that are involved in biologically significant interactions.[1]

Predicting the set of tolerated amino acids at each position based on structure only can be obtained using methods such as FoldX[27] and Rosetta.[28] These methods use force-field models to estimate variants' impact on structural stability. Machine-learning (ML) based models can serve as an alternative to force-field calculations, as they are capable of generalizing to unseen protein families. We recently introduced EvoRator, a web server that implements an ML-regression algorithm to predict residue-level evolutionary rates based on protein structures.[29] Here we present EvoRator2, a user-friendly web server that exploits deep learning to predict the per-site distribution of substitutions based on protein structure. EvoRator2 utilizes a unique structure-based representation that is created by combining a set of physicochemical and structural characteristic (e.g., amino-acid composition, relative solvent accessibility, secondary structure) with features of atoms and amino acids that are based on their network topology and on the spatial-chemical patterns of their neighbors. EvoRator2 is designed to predict per-site distribution of substitutions without using MSAs. Discrepancies between the MSA-based and structure-based estimates are inferred to reflect functional constraints beyond those imposed by the structure. Using a previously published standardized experimental DMS data, we demonstrate that EvoRator2 can accurately predict substitutions. We study the performance as a function of the three-dimensional (3D) structure accuracy. We also show that when EvoRator2 is integrated with a method that relies on MSAs (generating the EvoRator2-MSA model), the combined model outperforms existing approaches, especially for proteins characterized by inexistent or non-informative MSAs.

Methods

Data preparation

We extracted features for a set of 20,691 unique chains (obtained from 19,683 randomly chosen distinct PDB files) with matching position-weighted scoring matrices (PSSMs) mined from the ConSurf-DB,[30–37] which stores over 100,000 unique chains, their MSAs and conservation scores at the single residue level. The PSSM of each

record in ConSurf-DB is based on an MSA of a non-redundant set of homologues obtained by clustering candidate homologues sequences at 95% using CD-HIT.[36–38]

Features

EvoRator2 exploits features extracted by ScanNet[39–40] and EvoRator[29] to predict the site-specific distribution of allowable substitutions. Here we briefly describe these features. Both ScanNet and EvoRator extract features from PDB files. These features are extracted from the biological assembly file if it exists. ScanNet implements a geometric deep learning algorithm that builds representations of atoms and amino acids based on the spatio-chemical arrangement of their neighbors. These representations implicitly capture structural parameters such as solvent accessibility, secondary structure, and surface convexity. EvoRator extracts these and other features directly, including glycosylation sites, binding sites, and protein–protein interaction sites, as well as features extracted from graph-based representations of proteins, in which the nodes represent C_α atoms, and the edges represent interactions between C_α atoms that are within less than 7 Å from each other. We used Scikit-learn[41] for processing the above features as follows: duplicate and constant features are removed; features with missing values are filled with the median of their existing values; categorical features are one-hot encoded, and numerical features are scaled by subtracting the mean of the feature from each of its values and dividing by the standard deviation of the feature. In total, we used 515 features (Supplementary Table S1).

Evaluation criterion

The goal of our algorithm is to predict the spectrum of allowable substitutions at each site. This spectrum is mathematically described as a probability distribution over the 20 amino acids, at each site. The probability of each amino acid at a specific site is henceforth referred to as the score of that amino acid. To train and estimate model performance, true probabilities of amino-acids at each site should be known. In the following we assume that amino-acid frequencies obtained by analyzing large MSAs reflect close enough estimates to the true probabilities, and we term them "true" scores. Accuracy is then estimated in terms of the Spearman's rank correlation coefficient ρ between the model scores and the true scores. For benchmarking, model scores are also compared to scores obtained from DMS experiments (see below).

Deep learning

For predicting the site-specific distribution of substitutions using the above data, we trained a feed-forward multi-layer perceptron architecture with back-propagation[42] to minimize the Kullback-Leibler divergence[43] between the predicted and true scores of each site, using the Keras[44] implementation in the deep learning library Tensorflow.[45] The model consists of an input layer that has 515 nodes—one node per feature, followed by two hidden layers of 515 nodes with a rectified linear unit activation function, with l2-regularization on each layer's weights ($\lambda = 5\times\ 10^{-4}$, selected based on previous experience with similar datasets), and batch normalization following each hidden layer. Such a design is sufficient to approximate most discrimination tasks using less computational resources compared to a network with more hidden layers.[46] Finally, there is an output layer that has 20 nodes with softmax activation function that predicts a vector of residue probabilities. To avoid overfitting and in order to reliably estimates of model performance, we partitioned our data to training, validation and test sets, comprising 16,135 proteins, 360 proteins, and 711 proteins, respectively. This partition is based on the CATH[47–49] category of each record (proteins with unknown CATH category were excluded from this analysis), such that similarly structured proteins are included in either the training, validation or test set. The model was trained for a maximum of 50 epochs. An early stopping condition of 10 epochs interrupted training early if no improvement was observed in the validation set in terms of the Kullbeck-Leibler divergence after 10 training epochs. In practice, the performance on the validation set did not improve after 10 to 15 epochs, so we retrained the final model over the complete dataset for 10 epochs.

Benchmarking

For benchmarking EvoRator2, we used the ProteinGym substitution benchmark,[14] a standardized dataset of 72 proteins targeted in 88 different DMS assays. Note that DMS data are only used for testing the performance of the model, not for training. The ProteinGym dataset summarizes the performance of several MSA-based and large language models over a wide range of protein functions, taxonomic groups, and fitness measures. In ProteinGym, the DMS score is positively correlated to fitness, and the performance is quantified in terms of Spearman's rank correlation coefficient ρ and the area under the ROC curve (AUC) between model scores and the experimental measurements as the standard measure of model performance. We evaluated EvoRator2′s performance using AlphaFold[50] predicted structures, because they are obtained for full-length proteins and therefore can be more readily mapped to the sequences that are targeted in DMSs. We managed to obtain the predicted structures of 46 proteins targeted in 59 DMSs from AlphaFold DB.[51] This set of DMSs was used to evaluate the performance of EvoRator2. For model comparisons, we considered only

those AlphaFold DB records whose sequences perfectly match the ones that were targeted in the DMS experiments. Based on this criterion, we used a subset of 48 DMSs of 38 proteins for model comparisons. The frequencies predicted by EvoRator2 were transformed to standard DMS score[11] using the following formula:

$$\text{Predicted DMS score} = \log\frac{P(x_{mut})+\varepsilon}{P(x_{wt})+\varepsilon}, \varepsilon = 0.00001$$

Where $P(x_{mut})$ and $P(x_{wt})$ represent the predicted frequencies for a mutated and wild-type protein sequences, respectively.

Overview of EvoRator2 Web Server

EvoRator2′s approach is implemented as a public web server accessible from: https://evorator.tau.ac.il/. The web server provides an estimate of the site conservation using the algorithm described in EvoRator's paper[29] and the allowable substitutions using the algorithm described in this study (EvoRator2). EvoRator2 is implemented in Python 3.7. The web server jobs are processed on ProLiant XL170r Gen9 servers, equipped with 128 GB RAM and 28 CPU cores per node.

Results

EvoRator2 is tailored to predict the per-residue distribution of substitutions based on protein structure and to map substitution profiles that cannot be well explained using structural information alone. The features for our machine-learning approach are extracted from ScanNet and EvoRator. For training and evaluating the predictive performance of EvoRator2, we used a dataset of 20,691 unique chains with their corresponding residue scores that we mined from ConSurf-DB. We evaluated the predictive performance of EvoRator2 in terms of the Spearman's rank correlation coefficient ρ between true (MSA-based) and predicted (structure-based) substitution scores. EvoRator2 showed satisfying performance over the test dataset (Spearman's ρ = 0.610). We observed a minor difference between the performance of the different methods on the training (Spearman's ρ = 0.617), validation (Spearman's ρ = 0.621) and test sets, indicating minimal overfitting. The predictions and features for the test data are provided in Supplementary Data S1 (https://doi.org/10.5281/zenodo.7709583).

For benchmarking EvoRator2, we examined the relationship between its predictions and scores obtained in DMSs, which are considered as a gold standard for assessing the performance of protein models.[52] The DMS data were taken from the ProteinGym benchmark for substitutions.[14] We used EvoRator2 to predict the residue-level scores for 46 predicted AlphaFold structures of proteins that were targeted in 59 DMS experiments (Supplementary Data S2, https://doi.org/10.5281/zenodo.7709583). EvoRator2 performance varied widely across datasets, ranging from no correlation to high correlation (Figure 1, Figure S1 and Supplementary Data S3). Interestingly, in a few cases, EvoRator2 predictions were better correlated with the experimental DMS data compared to state-of-the-art deep learning methods (Supplementary Data S4). We suspected that this variation reflects differences in the quality of structures predicted by AlphaFold. The predicted aligned error (PAE) is a primary quality measure of AlphaFold structures.[50] Briefly, PAE, which is calculated for each pair of residues in the predicted structure, estimates the confidence in domain packing and large-scale topology. The lower the PAE score is, the higher the confidence in the relative position and orientation of different parts of the model. We therefore examined the relationship of the mean PAE (i.e. PAE averaged across all residue pairs) to EvoRator2′s performance (Figure S2A). We found that EvoRator2′s predictive performance increases in terms of accuracy and robustness as the mean PAE decreases, reaching optimal performance at mean PAE values that characterize well-predicted structures (mean PAE < 5). We acknowledge that fact that some sequence and structure similarity may exist between our training set and the ProteinGym test set. However, we found no significant correlation between EvoRator2′s performance and the sequence identity to the most similar protein in the training set (Figure S2B).

We further hypothesized that the integration of structural information and MSA can improve the accuracy of current methods, presumably by compensating for potential biases introduced by poorly aligned regions or insufficient or excessive divergence in the MSA.[1] To test this hypothesis, we compared the predictive performance of an MSA based model to that of an integrated structure and MSA based model, across the ProteinGym substitution benchmark.[14] The MSA based model includes predictions supplied by an ensemble of Tranception and EVE models (ETEVE).[14] This model was chosen as baseline model for the analysis, because it is based only on the predictions made by the most accurate prediction method reported in ProteinGym substitution benchmark, and it requires MSA data to make a prediction.[14] The second model integrates predictions supplied by EvoRator2 and ETEVE (EvoRator2 + ETEVE), which require structural and MSA data, respectively. The input of the EvoRator2 + ETEVE model is the pair of vectors for a specific position for all possible substitutions (EvoRator2 scores, ETEVE scores). The output is a single score-vector for each substitution. The relationship between this pair and the final score is modeled separately for each protein. Specifically, a linear regression model is assumed, in which the EvoRator2 and ETEVE vectors are transformed by restricted cubic splines with

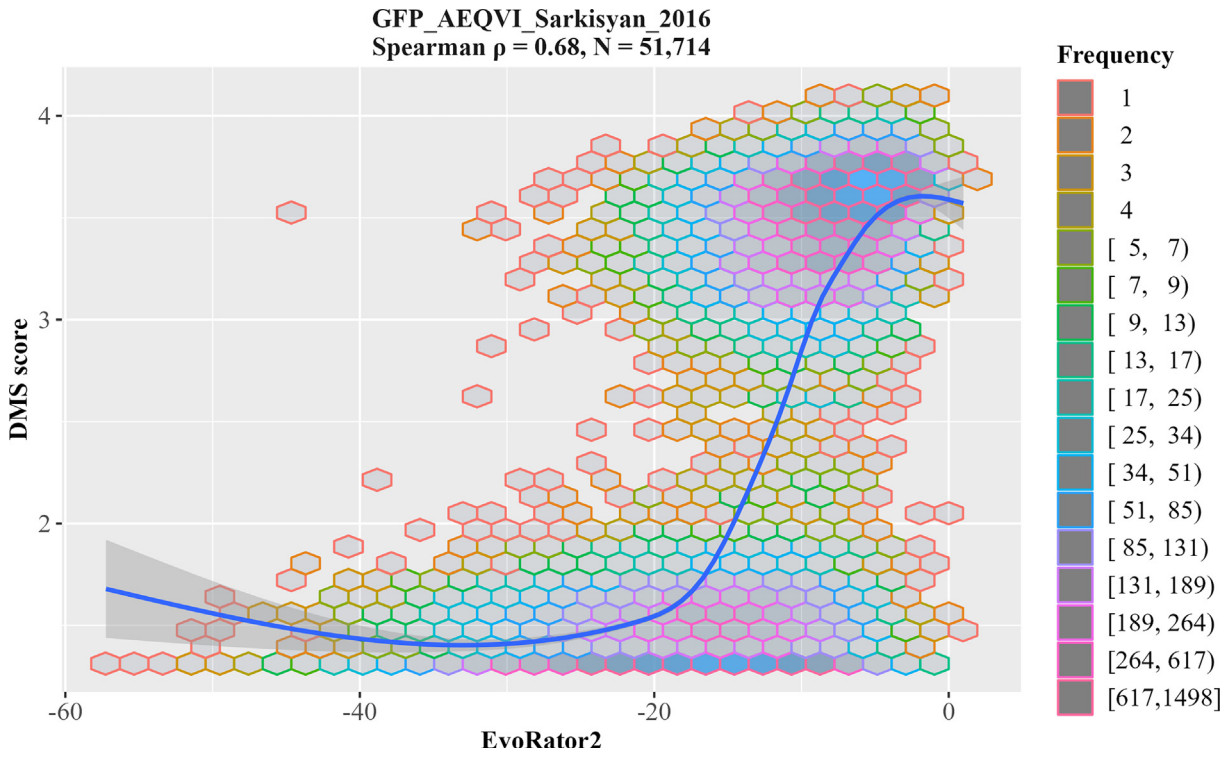

Figure 1. The relationship between EvoRator2′s predictions and experimental DMS data. DMS scores for the GFP protein (reflecting fluorescence levels) were taken from ProteinGym. Plotted are the DMS scores versus EvoRator2 scores (see Figure S1 for more proteins). A generalized additive model (GAM) with smooth functions (solid curve) with 95% confidence bands (in gray shade) is used to plot the relationship between EvoRator2′s predictions and experimental DMS scores. The scattered data points were binned using hexagonal binning because of the large sample size.

five knots.[54] This allows for accounting for the non-linear association between the input and the output (Figure S1).[53] In some cases, the output of DMS experiments is binary, i.e., for each substitution whether or not it is pathogenic.[14] In this case, the linear regression model is replaced by a logistic regression classification model using the same input. To ensure comparability with respect to input data, only those DMSs reporting targeted sequences that are identical to the sequences of the structures that we obtained from AlphaFold DB[51] were considered for model comparison (N = 48). Since experimental DMS scores covering the same genes cannot be easily compared,[6] standard model evaluation procedures such as cross validation or bootstrap cannot be reliably carried out using DMS data pooled from different sources. To overcome this issue, and to correct for the "optimism" stemming from fitting a model to the same data used to test it, for each DMS and for each model, we evaluated and compared the predictive accuracy (measured in R^2 or AUC) of the two models after averaging across 1,000 bootstrap samples from the same dataset. The EvoRator2 + ETEVE model slightly but significantly outperformed the ETEVE model for the majority of proteins, both in terms of optimism corrected R^2 (Wilcoxon signed-rank test, $p = 5.4 \times 10^{-7}$) and AUC (Wilcoxon signed-rank test, $p = 3.5 \times 10^{-6}$), ranging from small to large gains in these metrics for most and some DMSs, respectively (Figure S2). We hypothesized that this variation reflects differences in the quality of the input MSA. One such quality measure is the number of effective sequences in the MSA (N_{eff}),[52] which estimates the information content of a given MSA. To test our hypothesis, we compared the relationship of MSA quality to performance in the two models by plotting N_{eff} against the optimism corrected R^2 and AUC that were obtained by each model (Figure S4). We observed that EvoRator2 + ETEVE outperformed ETEVE across a wide range of N_{eff} values, and that substantial gains in optimism corrected R^2 and AUC tend to be concentrated at lower N_{eff} levels values. Notably, EvoRator2 + ETEVE also provides narrower confidence intervals compared to ETEVE across a wide range of N_{eff} values, which is always desirable.

Taken together, these results suggest that a high-quality structure can serve as an effective alternative to MSA-based methods when few or no homologs can be found, and that an integrated structure-MSA based prediction should be preferred over MSA-based or structure-based prediction.

Discussion

An MSA with many diverged sequences may well capture the structural constraints that drive the protein evolution. However, a large number of factors may introduce errors and biases in such inference. First, MSAs are error prone and some regions within the MSA are less reliably aligned compared with others.[55] Second, the demand for a high number of diverged sequences is often unattainable. For example, the number of protein structures in the PDB that have few or no sequence homologs is constantly rising, requiring the development of structure-based protein models. Third, in positions that experienced insertion and deletion events, especially those that arise due to insertions in lineages leading to a single protein, there is no information in the MSA to infer the selective forces. Fourth, sequences within MSAs are not random samples from the space of protein sequences. Rather, they are connected by an underlying phylogenetic tree. Some sequences are sampled from closely related species, while others from diverged ones. This sampling bias may be corrected by accounting for the tree topology and its associated branches while computing the selective constraints. However, the phylogeny is also subjected to uncertainty, and thus possible errors in the reconstructed phylogeny may lead to erroneous inference of the selective forces. While EvoRator2 is trained on MSA-derived scores, we expect that averaging over many unrelated examples attenuates these biases.

Here we present EvoRator2, a web server that implements a neural network that was trained over thousands of protein structures to predict the distribution of substitutions at the residue level, without the need for an input MSA. EvoRator2 exploits a rich structural signature consisting of physicochemical, geometrical, and graph-based features, which capture the various constraints that act on a protein 3D structure. When MSA information is available, contrasting the two types of predictions may provide additional information regarding the evolutionary constraints, e.g., selective forces that stem from functional rather than structural constraints.

EvoRator2′s predictions are in good agreement with experimental DMS data. We have shown that DMS profiles can be well predicted by integrating the predictions of EvoRator2 with those of state-of-the-art MSA-based and sequence-based methods. It is possible that the accuracy of the prediction would further increase with more data and other deep-learning models, particularly graph neural networks.[56] However, the accuracy of the alignment and 3D structure, as well as the accuracy of the effect of substitution patterns on fitness (as determined by DMS or other methods), can also affect the accuracy of the prediction. The relevant contribution of each factor awaits further characterization.

Structure-based sequence generative models such as ESM-IF1[57] and ProteinMPNN[58] can score substitutions, with some success. These models generate sequences that agree with the structure, and thus, can predict allowable substitutions. Our methodology, in contrast, is trained to predict PSSMs rather than sequences. A single PSSM is far more informative than a single sequence, and PSSM prediction can be readily compared to MSA-derived PSSMs. More generally, protein structure and sequence data are combined in methods such as 3DCOFFEE[59] to generate high-quality MSAs. Moreover, the accuracy of the MSA increases as the number of combined structures increases. Such structure-aware MSAs can potentially further improve the performance of MSA-based methods in scoring substitutions. Our results demonstrate that for scoring substitutions at the residue level, a single, high-quality structure can sometimes be as informative as an MSA, which typically considers hundreds of sequences. A further gain in performance would likely be obtained by combining multiple structures/conformations, raw sequence data, and MSAs for developing the next generation of protein models.

The EvoRator2 approach is combined within our EvoRator web server, which is freely available for the scientific community at https://evorator.tau.ac.il. The user interface is intuitive and provides both visual and tabular outputs.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This study was supported by the Israel Science Foundation (ISF) [2818/21 to T.P. and 1764/21 to N. B.-T.], the Edmond J. Safra Center for Bioinformatics at Tel Aviv University (fellowship, N.N.), the Human Frontier Science Program (J.T., cross-disciplinary postdoctoral fellowship LT001058/2019-C), Len Blavatnik and the Blavatnik Family Foundation (H.J. W.). T.P.'s research is supported in part by the Edouard Seroussi Chair for Protein Nanobiotechnology, Tel Aviv University. N.B.-T.'s research is supported in part by the Abraham E. Kazan Chair in Structural Biology, Tel Aviv University.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2023.168155.

Received 9 March 2023;
Accepted 17 May 2023;
Available online xxxx

Keywords:
protein evolution;
protein structure;
protein function;
mutation;
deep learning

References

1. Echave, J., Spielman, S.J., Wilke, C.O., (2016). Causes of evolutionary rate variation among protein sites. *Nature Rev. Genet.* **17**, 109–121.
2. Kessel, A., Ben-Tal, N., (2018). Introduction to proteins: structure, function, and motion. Taylor & Francis LLC, Boca Raton, Florida.
3. Pearce, R., Yang, Z., (2021). Deep learning techniques have significantly impacted protein structure prediction and protein design. *Curr. Opin. Struct. Biol.* **68**, 194–207.
4. Katsonis, P., Wilhelm, K., Williams, A., Lichtarge, O., (2022). Genome interpretation using in silico predictors of variant impact. *Hum. Genet.* **141**, 1549–1577.
5. Starr, T.N., Greaney, A.J., Hilton, S.K., Ellis, D., Crawford, K.H.D., Dingens, A.S., Navarro, M.J., Bowen, J.E., et al., (2020). Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell* **182**, 1295–1310.
6. Dunham, A.S., Beltrao, P., (2021). Exploring amino acid functions in a deep mutational landscape. *Mol. Syst. Biol.* **17**, e10305.
7. Schwersensky, M., Rooman, M., Pucci, F., (2020). Large-scale in silico mutagenesis experiments reveal optimization of genetic code and codon usage for protein mutational robustness. *BMC Biol.* **18**, 1–17.
8. Vaser, R., Adusumalli, S., Leng, S.N., Sikic, M., Ng, P.C., (2016). SIFT missense predictions for genomes. *Nature Protoc.* **11**, 1–9.
9. Hopf, T.A., Green, A.G., Schubert, B., Mersmann, S., Schärfe, C.P.I., Ingraham, J.B., Toth-Petroczy, A., Brock, K., et al., (2019). The EVcouplings Python framework for coevolutionary sequence analysis. *Bioinformatics* **35**, 1582–1584.
10. Adzhubei, I., Jordan, D.M., Sunyaev, S.R., (2013). Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **1**, 7–20.
11. Riesselman, A.J., Ingraham, J.B., Marks, D.S., (2018). Deep generative models of genetic variation capture the effects of mutations. *Nature Methods* **15**, 816–822.
12. Frazer, J., Notin, P., Dias, M., Gomez, A., Min, J.K., Brock, K., Gal, Y., Marks, D.S., (2021). Disease variant prediction with deep generative models of evolutionary data. *Nature* **599**, 91–95.
13. Laine, E., Yasaman, K., Alessandra, C., (2019). GEMME: a simple and fast global epistatic model predicting mutational effects. *Mol. Biol. Evol.* **36**, 2604–2619.
14. Notin, P., Dias, M., Frazer, J., Marchena-Hurtado, J., Gomez, A., Marks, D.S., Gal, Y., (2022). Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. *Proc. 39th Int. Conf. Mach. Learn.*, 16990–17017.
15. Kimura, M., Ota, T., (1974). On some principles governing molecular evolution. *PNAS* **71**, 2848–2852.
16. Kleinman, C.L., Rodrigue, N., Lartillot, N., Philippe, H., (2010). Statistical potentials for improved structurally constrained evolutionary models. *Mol. Biol. Evol.* **27**, 1546–1560.
17. Wilke, C.O., Sydykova, D.K., Jack, B.R., Spielman, S.J., (2017). Measuring evolutionary rates of proteins in a structural context. *F1000Research* **6**, 1845.
18. Yeh, S.-W., (2014). Local packing density is the main structural determinant of the rate of protein sequence evolution at site level. *Biomed Res. Int.* **2014**, 572409
19. Franzosa, E.A., Xia, Y., (2009). Structural determinants of protein evolution are context-sensitive at the residue level. *Mol. Biol. Evol.* **26**, 2387–2395.
20. Perez-Riverol, Y., Csordas, A., Bai, J., Bernal-Llinares, M., Hewapathirana, S., Kundu, D.J., Inuganti, A., Griss, J., et al., (2019). The PRIDE database and related tools and resources in 2019: Improving support for quantification data. *Nucleic Acids Res.* **47**, D442–D450.
21. Panjkovich, A., Daura, X., (2010). Assessing the structural conservation of protein pockets to study functional and allosteric sites: Implications for drug discovery. *BMC Struct. Biol.* **10**, 1–14.
22. Beltrao, P., Bork, P., Krogan, N.J., Van Noort, V., (2013). Evolution and functional cross-talk of protein post-translational modifications. *Mol. Syst. Biol.* **9**, 714.
23. Mintseris, J., Weng, Z., (2005). Structure, function, and evolution of transient and obligate protein-protein interactions. *PNAS* **102**, 10930–10935.
24. Chen, L., Perlina, A., Lee, C.J., (2004). Positive Selection Detection in 40,000 HumanImmunodeficiency Virus (HIV) Type 1 Sequences Automatically IdentifiesDrug Resistance and Positive Fitness Mutations in HIV Protease and Reverse Transcriptase. *J. Virol.* **78**, 3722–3732.
25. Duvvuri, V.R.S.K., Duvvuri, B., Cuff, W.R., Wu, G.E., Wu, J., (2009). Role of Positive Selection Pressure on the Evolution of H5N1 Hemagglutinin. *Genomics, Proteomics Bioinforma.* **7**, 47–56.
26. Weinberger, H., Moran, Y., Gordon, D., Turkov, M., Kahn, R., Gurevitz, M., (2010). Positions under positive selection-key for selectivity and potency of scorpion α-toxins. *Mol. Biol. Evol.* **27**, 1025–1034.
27. Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., Serrano, L., (2005). The FoldX web server: An online force field. *Nucleic Acids Res.* **33**, W382–W388.
28. Kellogg, E.H., Leaver-Fay, A., Baker, D., (2011). Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins Struct. Funct. Bioinforma.* **79**, 830–838.
29. Nagar, N., Ben Tal, N., Pupko, T., (2022). EvoRator: Prediction of Residue-level Evolutionary Rates from Protein Structures Using Machine Learning. *J. Mol. Biol.* **434**, 167538
30. Pupko, T., Bell, R.E., Mayrose, I., Glaser, F., Ben-Tal, N., (2002). Rate4Site: An algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics* **18**, S71–S77.
31. Celniker, G., Nimrod, G., Ashkenazy, H., Glaser, F., Martz, E., Mayrose, I., Pupko, T., Ben-Tal, N., (2013). ConSurf:

Using evolutionary data to raise testable hypotheses about protein function. *Isr. J. Chem.* **53**, 199–206.
32. Glaser, F., Pupko, T., Paz, I., Bell, R.E., Bechor-Shental, D., Martz, E., Ben-Tal, N., (2003). ConSurf: Identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics* **19**, 163–164.
33. Ashkenazy, H., Abadi, S., Martz, E., Chay, O., Mayrose, I., Pupko, T., Ben-Tal, N., (2016). ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res.* **44**, W344–W350.
34. Landau, M., Mayrose, I., Rosenberg, Y., Glaser, F., Martz, E., Pupko, T., Ben-Tal, N., (2005). ConSurf 2005: The projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res.* **33**, W299–W302.
35. Ashkenazy, H., Erez, E., Martz, E., Pupko, T., Ben-Tal, N., (2010). ConSurf 2010: Calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res.* **38**, W529–W533.
36. Ben Chorin, A., Masrati, G., Kessel, A., Narunsky, A., Sprinzak, J., Lahav, S., Ashkenazy, H., Ben-Tal, N., (2020). ConSurf-DB: An accessible repository for the evolutionary conservation patterns of the majority of PDB proteins. *Protein Sci.* **29**, 258–267.
37. Goldenberg, O., Erez, E., Nimrod, G., Ben-Tal, N., (2009). The ConSurf-DB: Pre-calculated evolutionary conservation profiles of protein structures. *Nucleic Acids Res.* **37**, D323–D327.
38. Fu, L., Niu, B., Zhu, Z., Wu, S., Li, W., (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152.
39. Tubiana, J., Schneidman-Duhovny, D., Wolfson, H.J., (2022). ScanNet: A Web Server for Structure-based Prediction of Protein Binding Sites with Geometric Deep Learning. *J. Mol. Biol.* **434**, 167758
40. Tubiana, J., Schneidman-Duhovny, D., Wolfson, H.J., (2022). ScanNet: an interpretable geometric deep learning model for structure-based protein binding site prediction. *Nature Methods* **19**, 730–739.
41. Pedregosa, F., Grisel, O., Weiss, R., Passos, A., Brucher, M., Varoquax, G., Gramfort, A., Michel, V., et al., (2011). Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830.
42. Rumelhart, D.E., Hinton, G.E., Williams, R.J., (1986). Learning representations by back-propagating errors. *Nature* **323**, 533–536.
43. Kullback, S., Leibler, R.A., (1951). On Information and Sufficiency. *Ann. Math. Stat.* **22**, 79–86.
44. François, C., (2015). Keras: The Python Deep Learning library. https://keras.io.
45. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., et al., (2016). TensorFlow: A system for large-scale machine learning (chairs Keeton, K. & Roscoe, T.). *In: OSDI'16: Proc. 12th USENIX Conf. Operating Systems Design and Implementation*. USENIX Association, pp. 265–283.
46. Huang, G.B., (2003). Learning capability and storage capacity of two-hidden-layer feedforward networks. *IEEE Trans. Neural Netw.* **14**, 274–281.
47. Pearl, F., Todd, A., Sillitoe, I., Dibley, M., Redfern, O., Lewis, T., Bennett, C., Marsden, R., et al., (2005). The CATH domain structure database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. *Nucleic Acids Res.* **33**, D247–D251.
48. Pearl, F.M.G., Lee, D., Bray, J.E., Sillitoe, I., Todd, A.E., Harrison, A.P., Thornton, J.M., Orengo, C.A., (2000). Assigning genomic sequences to CATH. *Nucleic Acids Res.* **28**, 277–282.
49. Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., Thornton, J.M., (1997). CATH - A hierarchic classification of protein domain structures. *Structure* **5**, 1093–1109.
50. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., et al., (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589.
51. Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., et al., (2022). AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444.
52. Hopf, T.A., Ingraham, J.B., Poelwijk, F.J., Schärfe, C.P.I., Springer, M., Sander, C., Marks, D.S., (2017). Mutation effects predicted from sequence co-variation. *Nature Biotechnol.* **35**, 128–135.
53. Boucher, J., Bolon, D.N., Tawfik, D.S., (2016). Quantifying and understanding the fitness effects of protein mutations: laboratory versus nature. *Protein Sci.* **25**, 1219–1226.
54. Harrell, F.E.J., (2016). Package "rms". *Compr. R Arch Netw.*.
55. Penn, O., Privman, E., Ashkenazy, H., Landan, G., Graur, D., Pupko, T., (2010). GUIDANCE: a web server for assessing alignment confidence scores. *Nucleic Acids Res.* **38**, W23–W28.
56. Xu, K., Jegelka, S., Hu, W., Leskovec, J., (2019). How powerful are graph neural networks? *In: 7th Int Conf. Learn. Represent.* ICLR, p. 2019.
57. Hsu, C., Verkuil, R., Liu, J., Lin, Z., Hie, B., Sercu, T., Lerer, A., Rives, A., (2022). Learning inverse folding from millions of predicted structures. *BioRxiv.*. 2022.04.10.487779.
58. Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R.J., Milles, L.F., Wicky, B.I.M., Courbet, A., et al., (2022). Robust deep learning–based protein sequence design using ProteinMPNN. *Science* **378**, 49–56.
59. O'Sullivan, O., Suhre, K., Abergel, C., Higgins, D.G., Notredame, C., (2004). 3DCoffee: Combining protein sequences and structures within multiple sequence alignments. *J. Mol. Biol.* **340**, 385–395.