Bioinformatics, 35(9), 2019, 1503–1512
doi: 10.1093/bioinformatics/bty813
Advance Access Publication Date: 20 September 2018
Original Paper

OXFORD

Structural bioinformatics

High precision protein functional site detection using 3D convolutional neural networks

Wen Torng [1] **and Russ B. Altman**[1,2,*]

[1]Department of Bioengineering and [2]Department of Genetics, Stanford University, Stanford, CA 94305, USA

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on January 31, 2018; revised on August 14, 2018; editorial decision on September 15, 2018; accepted on September 19, 2018

Abstract

Motivation: Accurate annotation of protein functions is fundamental for understanding molecular and cellular physiology. Data-driven methods hold promise for systematically deriving rules underlying the relationship between protein structure and function. However, the choice of protein structural representation is critical. Pre-defined biochemical features emphasize certain aspects of protein properties while ignoring others, and therefore may fail to capture critical information in complex protein sites.

Results: In this paper, we present a general framework that applies 3D convolutional neural networks (3DCNNs) to structure-based protein functional site detection. The framework can extract task-dependent features automatically from the raw atom distributions. We benchmarked our method against other methods and demonstrate better or comparable performance for site detection. Our deep 3DCNNs achieved an average recall of 0.955 at a precision threshold of 0.99 on PROSITE families, detected 98.89 and 92.88% of nitric oxide synthase and TRYPSIN-like enzyme sites in Catalytic Site Atlas, and showed good performance on challenging cases where sequence motifs are absent but a function is known to exist. Finally, we inspected the individual contributions of each atom to the classification decisions and show that our models successfully recapitulate known 3D features within protein functional sites.

Availability and implementation: The 3DCNN models described in this paper are available at https://simtk.org/projects/fscnn.

Contact: rbaltman@stanford.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Accurate annotation of protein functions is key to understanding cellular physiology at the molecular level. Structural genomics efforts have enabled an exponential growth in the determination of novel protein structures (Grabowski *et al.*, 2009). However, biochemical characterization of protein function does not scale as our database of structures increases (Liolios *et al.*, 2010). Computational methods for protein function annotation are therefore urgently needed. Traditionally, sequence similarities are used to infer function for newly identified proteins. One common approach is to search for local sequence motifs within a functional family; sequence motif databases include PROSITE (Sigrist *et al.*, 2012), PRINTS (Attwood, 2002), BLOCKS (Henikoff *et al.*, 2000), InterPro (Hunter *et al.*, 2009) and PANTHER (Mi *et al.*, 2005). However, as sequences diverge, the conserved signals become increasingly weak. Sequence-based methods can therefore have high false negative (FN) rates for novel proteins (Xiong, 2006). They also have limited ability to capture physiochemical properties and 3D orientation essential for a functional site.

An alternative approach is to seek -3D motifs within protein sites—when a structure is available. Early template-based methods such as FFFs (Fetrow and Skolnick, 1998) and TESS (Wallace *et al.*, 1997), use information about key site residues and local structural alignments to identify consensus 3D motifs within functionally similar proteins. More recently, machine learning algorithms have

© The Author(s) 2018. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/bioinformatics/article/35/9/1503/5104336 by guest on 08 November 2020

enabled automated discovery of 3D motifs in functional sites. GASPS (Polacco and Babbitt, 2006) uses a genetic algorithm strategy to create 3D templates within a protein family to best identify family members from the background. GASS (Izidoro *et al.*, 2015; Moraes *et al.*, 2017), on the other hand, employs genetic algorithms to search for similar active sites in proteins, given active site templates. Structurally Aligned Local Sites of Activities (Wang *et al.*, 2013) combines predicted functional residues from POOL (Somarowthu *et al.*, 2011) with local structural alignments to create characteristic structural patterns within a functional family. These methods provide a natural way of identifying key residues in active sites. However, in cases where functional residues are not as precisely oriented, incorporation of higher-level physiochemical features can help performance.

Property-based methods use a pre-defined set of biochemical or structural properties to describe a protein site microenvironment. For example, FEATURE (Bagley and Altman, 1995) defines a protein site by its 3D local neighborhood, consisting of six 1.25 Å-thick concentric shells, and evaluates 80 physicochemical properties within each shell. It has been applied to the detection of functional sites (Buturovic *et al.*, 2014), the characterization of protein pockets (Liu and Altman, 2011) and prediction of pocket–ligand interactions (Tang and Altman, 2014). However, FEATURE is limited by the high dimensionality, the inhomogeneity of features and the loss of orientation information. The need to define input biochemical features prior to using machine learning algorithms can lead to loss of critical information relevant to protein functional mechanisms.

The emergence of deep learning has enabled development of methods that extract task-specific features directly from raw data. Deep learning algorithms have been applied to small molecule representations (Duvenaud *et al.*, 2015; Kearnes *et al.*, 2016), protein contact prediction (Skwark *et al.*, 2014) and protein–ligand interaction prediction (Gomes *et al.*, 2017; Ragoza *et al.*, 2017). The strength of deep learning lies in its ability to learn useful representations directly from raw data (LeCun *et al.*, 2015). Convolutional neural networks (CNNs) (Krizhevsky *et al.*, 2012) are a subclass of deep learning networks that specialize in extracting spatial features in data. CNNs search for recurring spatial patterns and compose them into complex features in a hierarchical manner. Biochemical interactions start between atoms and can extend over space to form complex interactions. We have previously applied 3D convolutional neural networks (3DCNNs) to amino acid similarity analysis and showed deep learning framework outperformed conventional feature-based algorithms (Torng and Altman, 2017).

In this paper, we develop a general framework that applies 3DCNNs for protein functional site annotation. We represent protein structures as 3D images; analogous to red, green, blue channels in images, a protein site is represented as four atom 'channels' (corresponding to carbon, oxygen, nitrogen and sulfur) in a 20-Å box surrounding a location within the protein site. Driven by supervised labels, the developed pipeline automatically extracts task-specific features from the raw atom distribution. We perform head-to-head comparisons of prediction performances between our 3DCNNs, SVM models trained with raw atom distributions (Voxel-SVM) and SVM and 1DCNN classifiers that utilize the FEATURE descriptors. Our 3DCNNs achieve an average prediction recall of 0.955 at the precision threshold of 0.99 on PROSITE functional families, compared to recalls of 0.883, 0.857 and 0.754 of the Voxel-SVM, FEATURE-SVM and FEATURE-1DCNN models, respectively. We characterized performance of the models on challenging cases where PROSITE motifs miss or falsely detect functional signals and additionally benchmarked our performance with GASS (Izidoro *et al.*, 2015) on enzyme site detection tasks. Finally, we visualized individual contributions of each atom to the classification decision and show that our networks recognize meaningful biochemical features within protein functional sites.

2 Materials and methods

2.1 Datasets

2.1.1 PROSITE functional families

To demonstrate the advantages of 3DCNNs over conventional models, we focus on 10 of the 20 functional sites where models in our previous work performed least well (Buturovic *et al.*, 2014) (Supplementary Table S1). Each of the 10 functional sites was defined using sequence motifs annotated in the PROSITE database. Each PROSITE pattern comprises multiple conserved residues, each of which can be used as a reference residue to train a 'residue model' for the overall functional site. In this study, to simplify the procedure, for each functional site, we choose a single conserved residue in the PROSITE pattern, and a key functional atom selected based on its chemical properties. Each site is then defined around this functional atom of the selected residue (Supplementary Table S2).

To train and validate our models, we used the PROSITE database to construct the training and independent test datasets for each functional site. Specifically, for each functional family, the PROSITE database provides (i) PROSITE true positive (PROSITE TP) sequences: when PROSITE motif successfully detects a true site. (ii) PROSITE false negative (PROSITE FN) sequences: when the PROSITE motif is absent but the function is known to exist. (iii) PROSITE false positive (PROSITE FP) sequences: when the PROSITE motif is present but the function is not. We trained our models on examples of each functional site, using the PROSITE TP sites as positive training examples, and randomly sampled PDB (Berman *et al.*, 2000) structures as negative training examples. To independently validate our trained models, we used PROSITE FN sites as the positive test set and PROSITE FP sites as the negative test set. Since only PROSITE TP sites and randomly sampled negative sites were used in training, none of our models have seen these test sites.

2.1.1.1 Training dataset. We used the same true positive and negative structures as in Buturovic *et al.*, 2014. For each true positive protein structure, we mapped the PROSITE pattern to the protein sequence, identified our target residue within the mapped subsequence(s), and extracted the coordinate(s) of the functional atom of the residue(s)—thus defining the true positive sites. For the negative structures, we extract coordinates of all atoms with the same (residue, atom) types as the (target residue, functional atom) and sampled 50 000 atom coordinates per functional family. Number of true positive and negative examples are summarized in Supplementary Table S3.

2.1.1.2 Independent test dataset. For each functional site, we collected the FN Uniprot (Consortium, 2014) sequences from PROSITE. For each mapped structure, we identified key subsequences or catalytic residues associated with the function (Supplementary Note S1). We then extracted coordinates of the functional atoms of the identified catalytic residues or of all target residues within the key subsequences—thus defining the PROSITE FN sites. To construct the PROSITE FP set, for each functional site, we downloaded structures that map to the PROSITE FP Uniprot sequences, and removed structures that do not conform to the PROSITE pattern.

Downloaded from https://academic.oup.com/bioinformatics/article/35/9/1503/5104336 by guest on 08 November 2020

For the remaining, we extracted coordinates of the functional atoms of all target residues. The final PROSITE FP and FN datasets comprise a set of Uniprot IDs with their PDB structures and coordinates (Supplementary Tables S4 and S5).

2.1.2 Protein enzyme site detection

2.1.2.1 NOS dataset. We collected our nitric oxide synthase (NOS) structure set following procedures for DS 1 described in Izidoro *et al.* (Izidoro *et al.*, 2015), resulting in 138 NOS structures, 5000 negative train structures and 139 negative test structures (Supplementary Note S2.) Using Catalytic Site Atlas (CSA) (Furnham *et al.*, 2014) annotation of structure 3NOS, we identified four key residue types: arginine (ARG), cysteine (CYS), glutamic acid (GLU), tryptophan (TRP), and constructed four positive and negative training sets, each for one of the key residue types. For each key residue type, (i) we identified CSA residue(s) that matches the residue type in the positive NOS structures to form the positive training set. (ii) We identified all residues that match the key residue type in the negative train structures, and sampled 50 000 sites to form the negative training set. The test sites comprise all residues that match the key residue types in the negative test structures and positive NOS structures. Sites in the positive NOS structures are evaluated by the corresponding test fold model trained by 5-fold cross-validation (Sections 2.4.3.1 and 2.4.3.2).

2.1.2.2 TRYPSIN-like dataset. The TRYPSIN-like dataset is constructed following procedures for DS 2 in Izidoro *et al.* (Izidoro *et al.*, 2015). We first identified PDBs under the same superfamily as structure 1A0J from SCOP (Chandonia *et al.*, 2017), and removed the structures that lack CSA annotation, resulting in 1447 structures. Coordinates of the functional atoms of all histidine (HIS) and serine (SER) residues in all 1447 structures are used as the test set.

2.1.3 Input featurization and processing

2.1.3.1 Atom-channel dataset. For each of the atom coordinate extracted in Sections 2.1.1 and 2.1.2, we define a local 20-Å cubical box using orthogonal axes defined by the backbone geometry of the parent amino acid. The positive z-axis is chosen such that it is orthogonal to the x–y plane defined by N-C_α-C, and has a positive dot product with the C_α-C_β bond. Using the defined orientation, we extract a 20 Å box around the C_β atom of the residue (Fig. 1).

Each local 20-Å box is then divided into square 3D voxels with 1-Å dimension. Within each voxel, we record the presence of carbon, oxygen, sulfur and nitrogen atoms in a corresponding atom type channel (Fig. 2). To approximate atom connectivity and electron delocalization, we apply Gaussian filters to the discrete counts, using the average Van der Waals radii of the atom types as the SD. The resulting four 3D-matrices are then stacked together as different input channels.

2.1.3.2 FEATURE dataset. For each functional site, we generated the true positive, true negative, PROSITE FP and PROSITE FN datasets by applying FEATURE (Fig. 3) (Supplementary Table S6) to each recorded functional atom location (Section 2.1.1).

2.2 Model design

To perform comparisons between the deep learning framework and conventional machine learning models, we benchmarked performances of models that use combinations of two different input representations: 3D Voxels versus FEATURE descriptors, and two

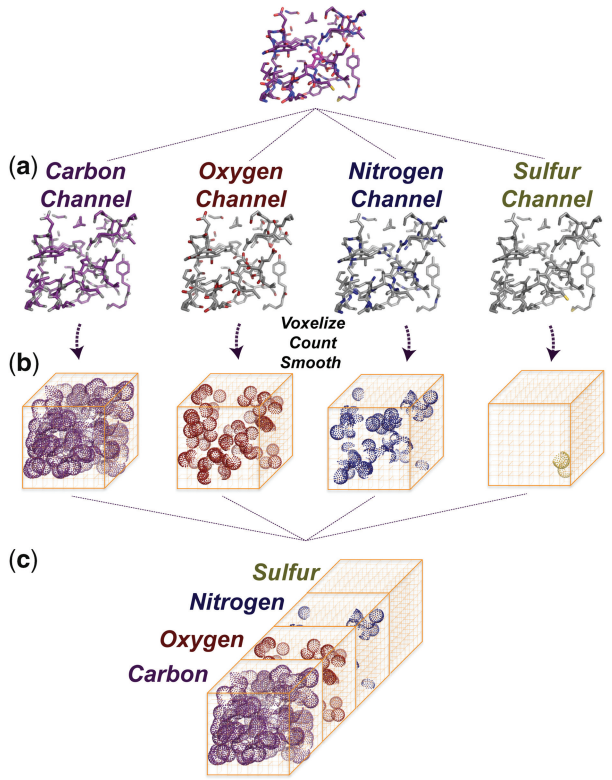

Fig. 2. Local box featurization. (**a**) Structure in each local 20 Å box is decomposed into Carbon, Oxygen, Nitrogen, and Sulfur channels. (**b**) Each atom type channel structure is further divided into 3D 1-Å voxels, within which the presence of atom of the corresponding atom type is recorded. Gaussian filters are applied to the discrete counts within each channel. (**c**) The resulting numerical 3D matrices of the four atom types are then stacked together as different input channels

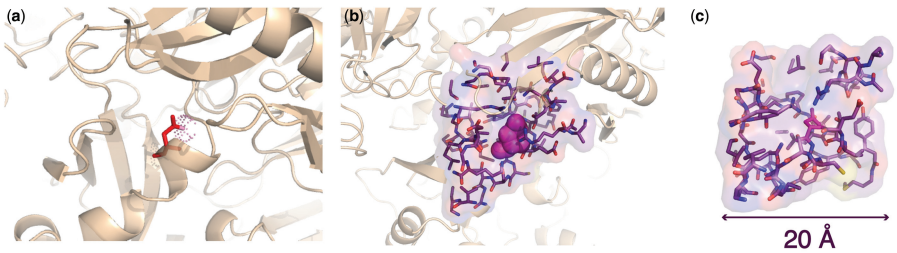

Fig. 1. Local box extraction. (**a**) For each recorded functional atom coordinate (dotted sphere), the amino acid which this atom belongs to is identified (highlighted in red) and assigned as the central amino acid. (**b**) Backbone atoms of the selected amino acid are used to calculate the orthogonal axes for box extraction. (**c**) Using the defined orientation, a local box of 20 Å is extracted around the selected amino acid, centering on the C_β atom

Downloaded from https://academic.oup.com/bioinformatics/article/35/9/1503/5104336 by guest on 08 November 2020

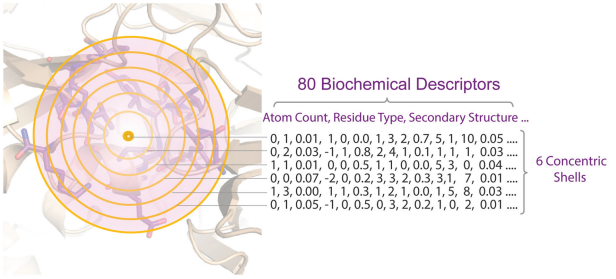

Fig. 3. The FEATURE program. FEATURE characterizes a specified location in protein structure by dividing the local environment into six concentric shells, each of 1.25 Å in thickness. Within each shell, 80 different physicochemical properties are evaluated, resulting in a numeric vector of length 480

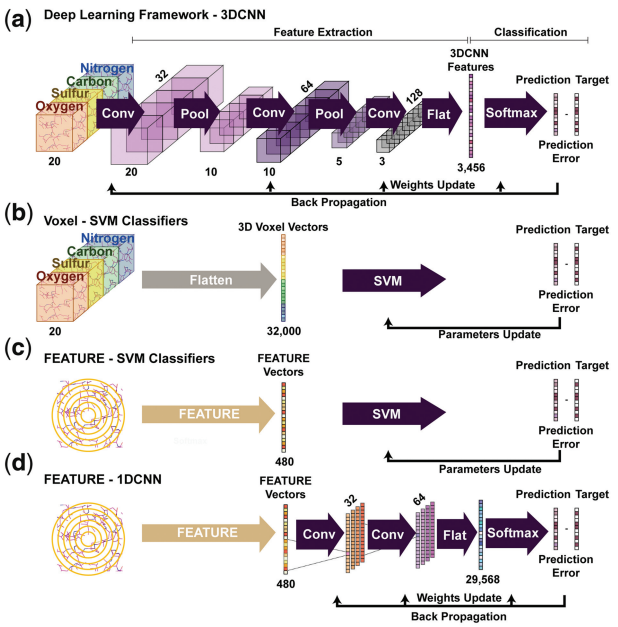

Fig. 4. Schematic diagram of the models. (**a**) 3DCNNs take in voxel representation of local protein boxes. The 3D convolutional and max-pooling layers extract 3D biochemical features at different spatial scales and the Softmax classifier calculates the class probabilities. Prediction error drives parameter updates in the Softmax layer and convolutional filters to achieve optimal performance. (**b**) The Voxel-SVM models take in flattened voxel representation and calculate class probabilities via the SVM classifier. (**c**) The FEATURE-SVM models use FEATURE vectors as input, and calculate class probabilities via the SVM classifier. In contrast to 3DCNNs, the input features are fixed during training. (**d**) The FEATURE-1DCNNs extract 1D features from FEATURE vectors and use the Softmax classifier to make predictions

different learning algorithms: SVM models versus CNNs. Thus, we construct the following four models: 3DCNN, Voxel-SVM Classifier, FEATURE-SVM Classifier and FEATURE-1DCNN (Fig. 4).

3DCNNs directly take in voxel representation of local protein boxes, and use three sequential alternating 3D Convolutional and max-pooling layers to search for biochemical features at different spatial scales. The Softmax classifier layer uses the extracted features and calculates the probability of the function of interest. Model parameters are summarized in Supplementary Table S7.

The Voxel-SVM and FEATURE-SVM classifiers take in voxel and FEATURE representation of local protein environment, respectively, and employ SVM classifiers to make classifications. FEATURE-1DCNNs take in FEATURE representation of local protein structure as input, followed by two 1D convolutional layers with 32 and 64 filters, respectively (each with filter size of 10), and end with a Softmax classifier layer.

2.3 Model training

For each functional site, we used 5-fold cross-validation to train our models. The folds were created using stratification by class label. Within each training fold, we up-sampled the positive examples such that the final training examples are balanced. The same train/validation/test splits were used to train models of the four methods. Implementation and training procedures of the models are summarized in Supplementary Note S3.

2.4 Evaluation

2.4.1 Five-fold cross-validation on PROSITE TP and TN sites

We evaluated performance by examining recalls of models on PROSITE true positive and negative datasets from our cross-validation procedure. To minimize FP predictions, we evaluated our models at high precision of 0.99. For each site, we treated all five test folds as a single continuous experiment and evaluated the aggregate precision and recall values (Supplementary Note S4). We then chose the threshold that results in a precision value of 0.99 or above, and evaluated the recall value.

Additionally, for each functional site, we computed the mean and SDs of the precision and recall between the fold models for each method. To evaluate the statistical significance of our results, we followed procedures outlined in (Demšar, 2006) for comparing performances of multiple classifiers over multiple data sets. Specifically, we performed Friedman test (Friedman, 1937, 1940) followed by *post-hoc* two-tailed Nemenyi test (Nemenyi, 1963).

2.4.2 Independent test: PROSITE FN and FP examples

For each functional site model, we then evaluated the PROSITE FN and FP examples, using the same probability threshold determined in Section 2.4.1. Specifically, representative probability estimate of a test site is produced by selecting the maximum probability estimate generated among the 5-fold models. The test site is then assigned as a positive site if the representative probability score is higher than the specified threshold, or negative otherwise.

A PROSITE FN site is predicted as positive if any of the 5-fold models assigns a probability estimate higher than the threshold. If a site is predicted as positive, we confirm by examination of the 3D local structure. A PROSITE FN PDB structure or Uniprot sequence is considered correctly predicted if any of its true sites are detected. On the other hand, a PROSITE FP site is predicted to be negative only if all probability estimates from the 5-fold models are below the threshold. A PROSITE FP PDB structure or Uniprot sequence is considered correctly predicted only if all examined sites are negative. For both datasets, our final evaluation metric evaluates the number of PDBs and Uniprot IDs predicted correctly, divided by the total number of PDBs and Uniprot IDs in the datasets.

2.4.3 Benchmarking protein enzyme site detection with GASS

We benchmarked our 3DCNNs against GASS (Izidoro *et al.*, 2015) on the three protein enzyme site detection and classification tasks:

2.4.3.1 Detecting catalytic sites within a family. Our first experiment tests our ability to detect catalytic sites within the NOS family. We trained four residue-3DCNNs on our CYS, ARG, TRP, GLU training datasets, each using 5-fold cross-validation. Using the corresponding test fold model, we examined all residues that match the

Downloaded from https://academic.oup.com/bioinformatics/article/35/9/1503/5104336 by guest on 08 November 2020

key residue types in the positive structures. For each residue model, we evaluate (1) The percentage of true sites detected (2) Number of FP sites detected. The overall performance was also evaluated by considering sites with any positive key residue predictions as positive.

2.4.3.2 Classifying sites within a family against random structures. The second experiment tests our models on 139 non-NOS family structures previously unseen by our models. Combining probability scores of the negative structures (Supplementary Note S5) and of the NOS structures from Section 2.4.3.1, we evaluate precision and recall for each residue model for classifying NOS structures against random structures, and area under the curve (AUC) scores at the residue-level and structure-level (Supplementary Note S5).

2.4.3.3 Detecting catalytic sites within less-controlled datasets. In addition to the tests that benchmarked the methods on small and controlled sets of enzymes, we compared performances of 3DCNNs and GASS on detecting sites in 1447 Trypsin-like enzymes. We employed the TRYPSIN_HIS and TRYPSIN_SER models from our PROSITE experiment to examine all HIS and SER residues in the test structures. For each of the two models, we evaluate (i) the percentage of true sites detected (ii) number of FP sites detected. We also evaluate the overall performance by considering sites with any positive residue predictions as positive. Details of the above three experiments are summarized in Supplementary Note S5.

2.5 Network visualization: atom importance map

The atom importance map shows the contribution of each atom to the final classification decisions by displaying the importance score (0–100) in color. Importance scores are calculated following the procedure described previously (Torng and Altman, 2017). We additionally use mTM-align (Dong *et al.*, 2018) to identify conserved structural features among local site boxes. For each functional site, we randomly sampled 30 positive site boxes to perform the alignment.

3 Results

3.1 Performances of models on PROSITE functional families

Precision and recall values of 3DCNNs, Voxel-SVM, FEATURE-SVM and FEATURE-1DCNN models for the true positive and true negative datasets are summarized in Table 1. Means and SDs of precision and recall between the fold models are summarized in Supplementary Table S8.

Our statistical analysis indicates that the performance of the four methods are not drawn from the same distribution ($P < 0.05$) and that 3DCNN models significantly outperformed FEATURE-SVM and FEATURE-1DCNN models ($P < 0.05$). Also, the Voxel-SVM models performed significantly better than the FEATURE-1DCNN models ($P < 0.05$) (details provided in Supplementary Note S6 and Supplementary Table S9). Performance statistics of the models on the independent test set: PROSITE FP and FN datasets are summarized in Table 2.

3.2 Performances on enzyme site detection tasks

3.2.1 Detecting catalytic sites within the NOS family

Among the 138 NOS structures, there are a total of 1066 residue sites, and 270 overall catalytic sites. At the individual residue-level, our 3DCNNs detected 1060 out of 1066 residue sites, missing 3 CYS sites from structures 1TLL and 1F20, and 3 GLU sites from structures 2ORQ, 2ORP, and 1NOC. No FP site was detected among the positive NOS structures. At the NOS site level, we detected 267 out of 270 NOS sites (98.89%) annotated in CSA. Detailed results are summarized in Supplementary Note S7 and Supplementary Table S11.

The percentage of sites detected using the 3DCNN models and GASS are summarized in Table 3. Although there is a slight difference in the number of NOS structures between the two datasets, we believe this would not result in significant biases. An example visualization of detecting NOS sites using 3DCNNs is shown in Supplementary Figure S1.

3.2.2 Classifying catalytic NOS sites against random structures

Using the 5-fold models trained in Section 2.4.3.1, no FP sites were detected among the negative test PDBs at the default probability

Table 1. Precision and recall values for PROSITE TP and TN examples

PROSITE site	Input	Method	Precision	Recall	Recall SD between folds
EGF_1	Voxels	3DCNN	0.992	**0.836**	0.105
		SVM	0.99	0.693	0.097
	FEATURE	SVM	0.988	0.571	0.122
		1DCNN	0.984	0.450	0.285
TRYPSIN_SER	Voxels	3DCNN	0.994	**0.991**	0.013
		SVM	0.990	0.984	0.020
	FEATURE	SVM	0.990	0.962	0.019
		1DCNN	0.990	0.968	0.027
RNASE_ PANCREATIC	Voxels	3DCNN	0.992	**0.985**	0.010
		SVM	0.990	0.982	0.030
	FEATURE	SVM	0.992	0.900	0.042
		1DCNN	0.992	0.926	0.042
EF_HAND_1	Voxels	3DCNN	0.996	**0.899**	0.051
		SVM	0.990	0.888	0.060
	FEATURE	SVM	0.990	0.840	0.089
		1DCNN	0.991	0.851	0.056
IG_MHC	Voxels	3DCNN	0.990	**0.915**	0.035
		SVM	0.990	0.741	0.053
	FEATURE	SVM	0.990	0.610	0.070
		1DCNN	0.982	0.304	0.227
PROTEIN_ KINASE_TYR	Voxels	3DCNN	0.997	**0.993**	0.008
		SVM	0.993	0.94	0.025
	FEATURE	SVM	0.993	0.913	0.059
		1DCNN	0.989	0.63	0.096
TRYPSIN_HIS	Voxels	3DCNN	0.993	**0.998**	0.005
		SVM	0.993	0.998	0.004
	FEATURE	SVM	0.991	0.957	0.033
		1DCNN	0.990	0.935	0.036
INSULIN	Voxels	3DCNN	0.993	0.954	0.013
		SVM	0.991	**0.972**	0.019
	FEATURE	SVM	0.989	0.858	0.058
		1DCNN	0.991	0.916	0.024
PROTEIN_ KINASE_ST	Voxels	3DCNN	0.992	**0.977**	0.023
		SVM	0.988	0.641	0.101
	FEATURE	SVM	0.991	0.939	0.018
		1DCNN	0.990	0.587	0.125
ADH_SHORT	Voxels	3DCNN	1.0	**0.995**	0.006
		SVM	0.995	0.992	0.016
	FEATURE	SVM	0.992	0.987	0.017
		1DCNN	0.989	0.976	0.021

The bold values highlight the best performing model(s) for each task.

Downloaded from https://academic.oup.com/bioinformatics/article/35/9/1503/5104336 by guest on 08 November 2020

Table 2. Performance statistics of PROSITE FP and PROSITE FN sites

PROSITE site	Method	PROSITE FN Uniprot	PROSITE FN PDB	PROSITE FP Uniprot	PROSITE FP PDB
EGF_1	3DCNN	**6/15**	**58/90**	3/3	19/19
	Voxel-SVM	5/15	57/90	3/3	19/19
	FF-SVM	6/15	34/90	3/3	19/19
	1DCNN	4/15	31/90	3/3	19/19
TRYPSIN_SER	3DCNN	**6/7**	**9/12**	**1/1**	**1/1**
	Voxel-SVM	**6/7**	**9/12**	0/1	0/1
	FF-SVM	6/7	9/12	**1/1**	**1/1**
	1DCNN	6/7	9/12	0/1	0/1
EF_HAND_1	3DCNN	**12/16**	**34/48**	**23/25**	**125/128**
	Voxel-SVM	11/16	30/48	22/25	**125/128**
	FF-SVM	11/16	28/48	22/25	106/128
	1DCNN	11/16	27/48	21/25	77/128
IG_MHC	3DCNN	**5/7**	8/47	10/10	31/31
	Voxel-SVM	4/7	**9/47**	10/10	31/31
	FF-SVM	4/7	8/47	10/10	31/31
	1DCNN	3/7	8/47	10/10	31/31
PROTEIN_	3DCNN	1/1	3/3	5/5	20/20
KINASE_	Voxel-SVM	1/1	3/3	5/5	20/20
TYR	FF-SVM	1/1	3/3	5/5	20/20
	1DCNN	1/1	3/3	5/5	20/20
TRYPSIN_HIS	3DCNN	**5/6**	**10/16**	2/2	4/4
	Voxel-SVM	**5/6**	**10/16**	2/2	4/4
	FF-SVM	2/6	3/16	2/2	4/4
	1DCNN	2/6	3/16	2/2	4/4
PROTEIN_	3DCNN	**19/20**	**268/271**	—	—
KINASE_	Voxel-SVM	16/20	226/271	—	—
ST	FF-SVM	18/20	264/271	—	—
	1DCNN	15/20	235/271	—	—
ADH_	3DCNN	5/8	7/14	8/9	32/33
SHORT	Voxel-SVM	5/8	7/14	**9/9**	**33/33**
	FF-SVM	6/8	8/14	**9/9**	**33/33**
	1DCNN	**8/8**	**12/14**	**9/9**	**33/33**

Note: For each functional site and each method, the entry records (the number of Uniprot IDs and PDBs correctly predicted)/(total number of Uniprot IDs and PDBs in the PROSITE FP or FN dataset) A '–' indicates that no PROSITE FP or FN structure was available for the functional site. FF-SVM refers to FEATURE-SVM models. The bold values highlight the best performing model(s) for each task.

Table 3. Percentage of NOS sites detected using 3DCNNs and GASS

Method	Residue model/template	Percentage of sites detected (%)	
3DCNN	CYS	98.89	Out of 270 sites in
	ARG	98.89	138 PDBs
	TRP	98.89	
	GLU	95.93	
	Overall	98.89	
GASS	Best template—3NOS	100	Out of 248 sites in
	All structures against all	94.49	125 PDBs

Table 4. 3DCNN and GASS AUC scores for NOS structure classification

Method	AUC	Total	Positive
3DCNN—residue level	0.998	42 556 sites	1066 sites
3DCNN—structure level	0.986	277 structures	138 structures
GASS—structure level	0.97	251 structures	125 structure

threshold 0.5. Precisions and recalls of each residue model for classifying NOS structures against random structures are summarized in Supplementary Table S12. GASS reported the AUC considering the distance threshold as the performance metric. To compare performance, we report the residue-level and structure-level AUC measures while varying our probability threshold (Supplementary Note S5). The AUC scores are summarized in Table 4.

3.2.3 Detecting catalytic sites within TRYPSIN-like enzymes

3DCNN models detected catalytic sites in 1330 out of 1447 enzymes by using our TRYPSIN_HIS models only. Using TRYPSIN_SER model only, we detected catalytic sites in 1287 out 1447 of enzymes. Using both models, we detected catalytic sites in 1344 out 1447 of enzymes. Percentage of sites detected using the TRYPSIN_HIS and TRYPSIN_SER 3DCNN models, and average results of GASS running nine templates against 1085 enzymes are summarized in Table 5. Nine 'false positive' residue sites are detected by our 3DCNNs but are not annotated in CSA. Further examination showed that they are correct predictions from our models that are not documented in CSA. For example, CSA annotation of structure 2LPR list ASP (A, 102), GLY (A, 193), and HIS (A, 57) as key residues. Our 3DCNN models additionally predicted SER (A, 195) as part of the catalytic site, which agrees with *Bone et al.* (Bone *et al.*, 1991). These nine predictions are summarized in Supplementary Table S13.

Downloaded from https://academic.oup.com/bioinformatics/article/35/9/1503/5104336 by guest on 08 November 2020

Table 5. Percentage of sites detected in TRYPSIN-like enzymes

Method	Model/rank size	Percentage of sites detected (%)	
3DCNN	TRYPSIN_HIS	91.90	Out of 1447 enzymes
	TRYPSIN_SER	88.94	
	Both	92.88	
GASS	Rank size—1	82.85	Out of 1085 enzymes
	Rank size—5	90.94	
	Rank size—10	93.52	

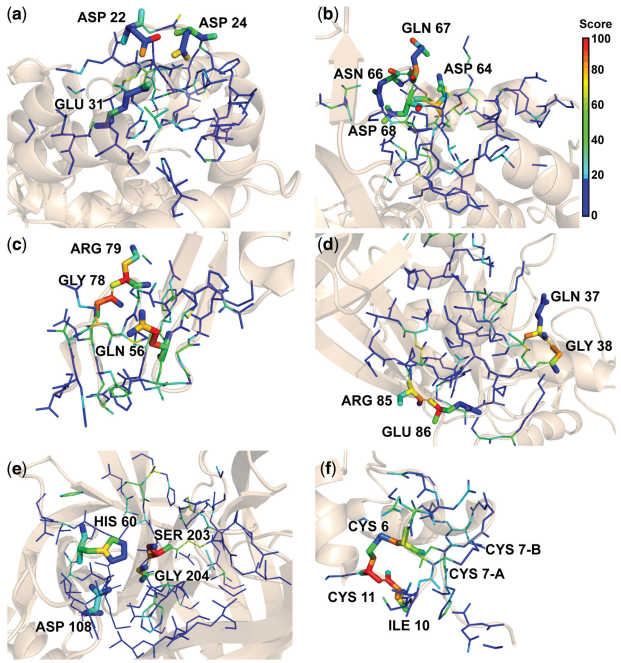

Fig. 5. Importance visualization of local functional site microenvironments. The atom importance map displays the importance score (0–100) of each input atom in heat map colors, from the most important (red) to the least important (blue). (**a**) Microenvironment surrounding a conserved ASP residue in the EF_HAND motif (PDB: 1CKK, ASP 20). (**b**) Microenvironment surrounding a ASP residue in a PROSITE false negative EF_HAND site (PDB: 2LE9, ASP 64). (**c**) Microenvironment surrounding a conserved CYS residue in the EGF_1 motif (PDB: 1BF9, CYS 81). (**d**) Microenvironment surrounding a conserved LYS residue in the RNASE_PANCREATIC motif (PDB: 11BG, LYS 41). (**e**) Microenvironment surrounding a TRYPSIN_SER motif, centered on a conserved SER residue (PDB: 1ELD, SER 203). (**f**) Microenvironment surrounding an INSULIN motif, centered on a conserved CYS residue (PDB: 1IOG, CYS 7-chain A)

3.3 Network visualization

Example visualizations for true positive examples of functional sites EGF_1, EF_HAND_1, RNASE_ PANCREATIC, TRYPSIN_SER and INSULIN, and a PROSITE FN example of functional site EF_HAND_1 are shown in Figure 5. The color shows how each atom contributes to the decision. The red to blue heat map spectrum highlights the most important to the least important atoms.

4 Discussion

4.1 Cross-validation performances on PROSITE TP and TN sites

As shown in Table 1, 3DCNN models achieve better or comparable recall compared to the other models for all the 10 functional sites—sites chosen because they were challenging for both Naïve Bayes and SVM classifiers (Buturovic *et al.*, 2014). On average, at the precision threshold of 0.99, 3DCNN models achieved a recall of 0.955, whereas Voxel-SVM models, FEATURE-SVM models and FEATURE-1DCNN models attained recalls of 0.883, 0.857 and 0.754, respectively. Among the four methods, the 3DCNN models significantly outperformed FEATURE-SVM and FEATURE-1DCNN models ($P < 0.05$) and the Voxel-SVM models performed significantly better than the FEATURE-1DCNNs ($P < 0.05$). We observe better performances of 3DCNNs over the Voxel-SVMs and improved performance of Voxel-SVMs over the FEATURE-SVMs although the comparisons were not statistically significant.

3DCNN models achieved the best improvements over the other three methods on sites EGF_1.10, IG_MHC.3 and PROTEIN_KINASE_ST. These sites generally have a higher degree of local structural variations and thus are particularly challenging. The success of 3DCNNs on these challenging sites suggests that 3DCNNs have stronger ability to capture features that are generalizable to sites with local structural variations.

Although FEATURE employ biochemical features with different levels of details, larger perturbations of local structure may cause a biochemical feature to trans-locate to a different shell and therefore significantly change its input representation. The loss of orientation-specific information may also affect performances of the models. FEATURE-SVM models performed particularly well on site PROTEIN_KINASE_ST, but poorly on sites EGF_1.10 and IG_MHC.3. FEATURE-1DCNN models generally do not perform well. This is not surprising because CNNs generally thrive when local spatial correlations are present in the input representation while attributes in FEATURE vectors typically do not have such properties.

Overall, 3DCNNs and Voxel-SVMs demonstrated better performances over the FEATURE-based methods. The similar performance of 3DCNNs and Voxel-SVMs on some PROSITE functional sites may be due to the low structural diversity of these sites. In the PROSITE dataset, since the positive training examples are defined around selected conserved residues in sequence motifs, local site microenvironments are inherently more conserved, and information of conserved atom positions can contribute significantly to model performances. To investigate the effects of 3D site diversity of functional families on model performances, we additionally characterized the performance of different models on predicting adenosine triphosphate (ATP) binding residues. ATP binding sites are ubiquitous, diverse and critical to drug development (Maxwell and Lawson, 2003).

We trained and validated 3DCNN, Voxel-SVM and FEATURE-SVM models on the PATP-388 and PATP-TEST datasets described in Hu *et al.*, 2018, and show that 3DCNNs demonstrate strong advantages over both FEATURE-SVM and Voxel-SVM models on the ATP binding site prediction task. 3DCNNs achieved an AUC of 0.906 separating ATP binding residues from non-binding residues, whereas FEATURE-SVM and Voxel-SVM attained AUC scores of 0.809 and 0.764, respectively. In this case, Voxel-SVMs performed significantly worse than both 3DCNNs and FEATURE-SVM models, since for sites with diverse local microenvironments, there is less conservation of atom position and orientation. Experimental designs, model performance and an example visualization for ATP site detection are summarized in Supplementary Note S8, Supplementary Tables S14–S16 and Supplementary Figure S2.

4.2 Performances on PROSITE FP and FN sites

As reported in Table 2, on the independent test set, we achieved better or comparable performance for 7 out of 8 functional sites with

Downloaded from https://academic.oup.com/bioinformatics/article/35/9/1503/5104336 by guest on 08 November 2020

available PROSITE FN structures, and demonstrated better or comparable ability ruling out PROSITE FP examples for 6 out of 7 functional sites. 3DCNN models achieved the best improvements in detecting PROSITE FN sites on sites EF_HAND_1, TRYPSIN_HIS, PROTEIN_KINASE_ST and EGF_1, on average detecting 1.6-fold more PROSITE FN structures compared to the other three methods.

FEATURE-based models achieved better performance on site ADH_SHORT, which has the most structurally conserved positive training sites among the ten sites. Because 3DCNNs rely on variations within training data to learn features that are robust to noise, they might fail to capture the key invariant features within the sites when they are highly conserved. In this case, FEATURE-based models detected more signals in PROSITE FN sites by using high-level and shell-based descriptors.

4.3 Protein enzyme site detection

4.3.1 Detecting NOS catalytic sites

3DCNNs and GASS have key differences and strengths. For each run of GASS, only a single template is required and used to search for similar active sites in the test proteins. GASS thus has advantages when only few known active sites are available, and each of them can be used as a template to search for active sites in query structures. However, the results could be sensitive to the choice of the template. Using the best template, GASS achieved 100% detection rate whereas the average detection rate using all available templates reduced to 94%. 3DCNNs, on the other hand, require a set of positive structures for training. The method is less applicable when only a single structure is available. However, the models integrate information from all training sites and thus are relatively robust. GASS reported 2BHJ as a particular challenging case, while 3DCNNs detected all catalytic residues in 2BHJ with high confidence. CYS 194, ARG 197, TRP 366 and GLU 371 received probability scores of 0.989, 0.999, 1.0 and 0.933, respectively. Importantly, because of the high precision of our residue models, 3DCNNs made no FP predictions in all the NOS structures.

We further looked at the NOS sites that our 3DCNN models missed. Structures 1TLL and 1F20, which our models failed to detect, have very different sets of catalytic residues compared to those of the other NOS structures. While typical NOS structures have catalytic residues CYS, ARG, TRP and ARG, the two structures have SER, ASP and CYS. Furthermore, 1TLL and 1F20 are under different SCOP Superfamilies from the other NOS proteins: while the template structure 3NOS belongs to Superfamily d.174.1: NOS oxygenase domain, the two chains of structure 1F20 belong to Superfamily b.43.4 and Superfamily c.25.1, respectively. The significantly different site microenvironments may have caused the models to miss the functional signals.

For the task of classifying NOS structures against random structures, we also achieved AUCs comparable with GASS, as reported in Table 4.

4.3.2 Detecting catalytic sites within TRYPSIN-like enzymes

As shown in Table 5, we achieved comparable performances with GASS on detecting catalytic sites in TRYPSIN-like enzymes. Further analysis on catalytic sites that we missed showed that the majority of structures that our models failed to detect are structures under SCOP family b.47.1.4: viral CYS protease of trypsin fold. These structures are CYS proteases that use CYS and HIS as main catalytic residues, instead of the ASP, SER, HIS triad observed in SER proteases. Our models were not able to capture strong signals from these sites.

As described in Section 3.2.3, nine 'false positive' residue sites are detected by our 3DCNNs but are not annotated in CSA. However, further examination of these sites show that they are true sites that are missed in the CSA annotation (Supplementary Table S13).

4.4 Network visualization

To visualize site-specific information captured by 3DCNNs, we present examples of atom importance maps of functional site microenvironments, highlighting the key features contributing to the 3DCNN classifications. We additionally compared our key features to conserved residues identified by local structural alignments using mTM-align.

4.4.1 EF_HAND_1.1

Figure 5a shows a true positive site surrounding a conserved ASP residue in the EF_HAND motif [PDB: 1CKK (Osawa *et al.*, 1999), ASP 20]. The 3DCNN correctly identifies the positive site and the atom importance map shows that the decision depends on the oxygen atoms in the side-chains of ASP 22, ASP 24 and GLU 31, known residues involved in the calcium-binding (Moncrief *et al.*, 1990). On the other hand, mTM-align identifies LYS 21 and LEU 18 in 1CKK as the core common region among the aligned EF_HAND boxes. The two residues lie in close proximity to the key residues identified by the 3DCNN, but are not in direct contact with the calcium ion (Supplementary Fig. S3).

Figure 5b shows an EF_HAND site surrounding a ASP residue [PDB: 2LE9 (Rani *et al.*, 2014), ASP 64, chain B], which is a PROSITE FN. The site has alternative residue SER at PROSITE motif position 6 and LYS at position 8. Nevertheless, the 3DCNN correctly classifies the site with high confidence. The importance map shows that the decision relies on oxygen atoms in ASP 64, ASN 66, GLN 67, ASP 68 and SER 69. The successful detection suggests that 3DCNNs can capture similar physiochemical and structural features in sites with local structural variations. This PROSITE FN site was not detected by the other classifiers.

4.4.2 EGF_1.10

Figure 5c shows a true positive site surrounding a conserved CYS residue in the EGF_1 motif [PDB: 1BF9 (Muranyi *et al.*, 1998), CYS 81]. 3DCNN correctly classifies the site using key residues GLY 78, ARG 79, GLN 56, along with four conserved CYS residues that form two disulfide bonds (CYS 55, CYS 70, CYS 72, CYS 81). GLY 78 and the CYS residues are conserved in the PROSITE motif. The most weighted residue, ARG 79, is crucial in the formation of the epidermal growth factor receptor-ligand complex (Engler *et al.*, 1990). Local structural alignments on EGF_1.10 site boxes identified a single residue GLY 59 as the structurally conserved residue, and did not identify the residues highlighted by the 3DCNN.

4.4.3 RNASE_PANCREATIC.2

Figure 5d shows a microenvironment surrounding a conserved LYS residue in the RNASE_PANCREATIC motif [PDB: 11BG (Vitagliano *et al.*, 1999), LYS 41]. The importance map shows that the correct prediction depends on GLU 86 and GLY 38. GLY 38 is critical for the ribonucleolytic activity of human pancreatic ribonuclease on double-stranded RNA (Gaur *et al.*, 2002). The center catalytic LYS is not highlighted, likely because a significant number of both positive and negative examples have LYS in similar conformation. On the other hand, mTM-align identified ASN 44 and

Downloaded from https://academic.oup.com/bioinformatics/article/35/9/1503/5104336 by guest on 08 November 2020

THR 45 as the core common region among the site boxes, which are not heavily used by the 3DCNN for classifications.

4.4.4 TRYPSIN_SER

Figure 5e shows the TRYPSIN_SER motif, centered on a conserved SER [PDB: 1ELD (Mattos *et al.*, 1995), SER 203]. The importance map highlights GLY 204, SER 203, HIS 60, which are consistent with the key biochemical features identified by Bagley and Altman, 1996. SER 203, HIS 60 and ASP 108 form a SER-HIS-ASP catalytic triad (Rawlings and Barrett, 1994), although ASP 108 received a lower importance score here. LEU, PHE and TYR residues around the active site show moderate importance, and may facilitate non-specific polypeptide ligand binding and stabilization. mTM-align similarly identified HIS 60 and ASP 108. However, it additionally identified 13 other residues as structurally conserved: THR 57, ALA 58, ALA 59, CYS 61, VAL 62, THR 44, CYS 45, GLY 46, GLY 47, LEU 33, THR 145, GLY 146 and GLY 148. These residues do not contribute significantly to the 3DCNN classification.

4.4.5 INSULIN

Figure 5f shows a PROSITE INSULIN motif, centered on a conserved CYS residue [PDB: 1IOG (Olsen *et al.*, 1998), CYS 7-chain A]. The importance map shows that the prediction depends on ILE 10, CYS 6 and CYS 11. The two CYSs form a disulfide bond and are conserved in the INSULIN motif (Blundell and Humbel, 1980). mTM-align identified two CYS residues CYS 6-chain A, CYS 7-chain A and nearby residues THR 8 and SER 9 as the conserved region among the site local boxes. The 3DCNN uses some but not all of these residues to make the classification.

4.5 Input featurization and network architecture

In this study, we employed 3DCNNs with three alternating 3D convolutional and max-pooling layers to learn features that are less sensitive to translational variance within inputs. Due to its robustness to noise and ability to extract task-specific features, the same framework can be directly applied to different functional sites without explicit tuning of model parameters or architecture. Therefore, our proposed framework is general-purposed and is applicable to any functional site given available data.

Two important hyper-parameters in our system are the dimension of the site bounding box and the voxel size. Larger boxes increase marginal information accessible by the network but are more computationally expensive. We choose to extract local protein boxes of 20 Å based on our previous experience with the FEATURE program (Bagley and Altman, 1995) and our experience of applying 3DCNNs to protein structures (Torng and Altman, 2017). Beyond a 16–20 Å cutoff, the atomic details do not provide additional information. The choice of 1 Å voxels together with small filter sizes allow the models to extract features with fine spatial resolution. Because the grid voxel system is not rotational invariant, we calibrate all boxes using backbone atoms to ensure similar orientation.

Although we focused on 10 representative functional sites in this study, since our framework can be applied to different functional sites without manual adjustments, the framework can be easily applied to datasets with large amount of functional sites. Furthermore, because our models exploit variations within the training data to learn features that are robust to noise, our method especially thrives when a large number of training examples are available. Once the models have been trained, evaluation at test time is efficient, which enables our models to be applied to large datasets at test time.

The success of 3DCNNs on difficult annotation tasks suggests that this framework is well-suited for protein structural analysis and can discover features from raw data that outperform pre-defined features. As more structural data become available, deep learning models hold promise for advanced protein engineering applications.

Acknowledgements

Computation for this study was performed on the Sherlock cluster. We would like to thank Stanford University and the Stanford Research Computing Center for providing computational resources and support that contributed to these research results. This work also used the XStream computational resource, supported by the National Science Foundation Major Research Instrumentation program (ACI-1429830).

Funding

This work was supported by the National Institutes of Health [GM102365, LM05652 and HL117798].

Conflict of Interest: none declared.

References

Attwood,T.K. (2002) The PRINTS database: a resource for identification of protein families. *Brief. Bioinform.*, **3**, 252–263.

Bagley,S.C. and Altman,R.B. (1995) Characterizing the microenvironment surrounding protein sites. *Protein Sci.*, **4**, 622–635.

Bagley,S.C. and Altman,R.B. (1996) Conserved features in the active site of nonhomologous serine proteases. *Fold. Des.*, **1**, 371–379.

Berman,H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Blundell,T. and Humbel,R. (1980) Hormone families: pancreatic hormones and homologous growth factors. *Nature*, **287**, 781.

Bone,R. *et al.* (1991) Structural basis for broad specificity in. alpha.-lytic protease mutants. *Biochemistry*, **30**, 10388–10398.

Buturovic,L. *et al.* (2014) High precision prediction of functional sites in protein structures. *PLoS One*, **9**, e91240.

Chandonia,J.-M. *et al.* (2017) SCOPe: manual curation and artifact removal in the structural classification of proteins–extended database. *J. Mol. Biol.*, **429**, 348–355.

Consortium,T.U. (2014) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.

Demšar,J. (2006) Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, **7**, 1–30.

Dong,R. *et al.* (2018) mTM-align: a server for fast protein structure database search and multiple protein structure alignment. *Nucleic Acids Res.*, **46**, W380–W386.

Duvenaud,D.K. *et al.* (2015) Convolutional Networks on Graphs for Learning Molecular Fingerprints. Paper presented at: Advances in neural information processing systems.

Engler,D.A. *et al.* (1990) Human epidermal growth factor. Distinct roles of tyrosine 37 and arginine 41 in receptor binding as determined by site-directed mutagenesis and nuclear magnetic resonance spectroscopy. *FEBS Lett.*, **271**, 47–50.

Fetrow,J.S. and Skolnick,J. (1998) Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to Glutaredoxins/Thioredoxins and T1Ribonucleases1. *J. Mol. Biol.*, **281**, 949–968.

Friedman,M. (1937) The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Am. Stat. Assoc.*, **32**, 675–701.

Friedman,M. (1940) A comparison of alternative tests of significance for the problem of m rankings. *Ann. Math. Stat.*, **11**, 86–92.

Furnham,N. *et al.* (2014) The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes. *Nucleic Acids Res.*, **42**, D485–D489.

Downloaded from https://academic.oup.com/bioinformatics/article/35/9/1503/5104336 by guest on 08 November 2020

Gaur,D. *et al*. (2002) Glycine 38 is crucial for the ribonucleolytic activity of human pancreatic ribonuclease on double-stranded RNA. *Biochem. Biophys. Res. Commun*., **297**, 390–395.

Gomes,J. *et al*. (2017) Atomic convolutional networks for predicting protein-ligand binding affinity. *arXiv*, 170310603.

Grabowski,M. *et al*. (2009) Benefits of structural genomics for drug discovery research. *Infect. Disord. Drug Targets*, **9**, 459–474.

Henikoff,J.G. *et al*. (2000) Blocks-based methods for detecting protein homology. *Electrophoresis*, **21**, 1700–1706.

Hu,J. *et al*. (2018) ATPbind: accurate Protein–ATP Binding Site Prediction by Combining Sequence-Profiling and Structure-Based Comparisons. *J. Chem. Inf. Model*., **58**, 501–510.

Hunter,S. *et al*. (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res*., **37**, D211–D215.

Izidoro,S.C. *et al*. (2015) GASS: identifying enzyme active sites with genetic algorithms. *Bioinformatics*, **31**, 864–870.

Kearnes,S. *et al*. (2016) Molecular graph convolutions: moving beyond fingerprints. *J. Comput. Aided Mol. Des*., **30**, 595–608.

Krizhevsky,A. *et al*. (2012) Imagenet Classification with Deep Convolutional Neural Networks. Paper presented at: Advances in neural information processing systems.

LeCun,Y. *et al*. (2015) Deep learning. *Nature*, **521**, 436.

Liolios,K. *et al*. (2010) The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res*., **38**, D346–D354.

Liu,T. and Altman,R.B. (2011) Using multiple microenvironments to find similar ligand-binding sites: application to kinase inhibitor binding. *PLoS Comput. Biol*., **7**, e1002326.

Mattos,C. *et al*. (1995) Structural analysis of the active site of porcine pancreatic elastase based on the X-ray crystal structures of complexes with trifluoroacetyl-dipeptide-anilide inhibitors. *Biochemistry*, **34**, 3193–3203.

Maxwell,A. and Lawson,D.M. (2003) The ATP-binding site of type II topoisomerases as a target for antibacterial drugs. *Curr. Top. Med. Chem*., **3**, 283–303.

Mi,H. *et al*. (2005) The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res*., **33**, D284–D288.

Moncrief,N.D. *et al*. (1990) Evolution of EF-hand calcium-modulated proteins. I. Relationships based on amino acid sequences. *J. Mol. Evol*., **30**, 522–562.

Moraes,J.P. *et al*. (2017) GASS-WEB: a web server for identifying enzyme active sites based on genetic algorithms. *Nucleic Acids Res*., **45**, W315–W319.

Muranyi,A. *et al*. (1998) Solution structure of the N-terminal EGF-like domain from human factor VII. *Biochemistry*, **37**, 10605–10615.

Nemenyi,P. (1963) Distribution-free multiple comparisons. Doctoral Dissertation, Princeton University, Dissertation Abstracts International 25, 1233.

Olsen,H.B. *et al*. (1998) The relationship between insulin bioactivity and structure in the NH2-terminal A-chain helix1. *J. Mol. Biol*., **284**, 477–488.

Osawa,M. *et al*. (1999) A novel target recognition revealed by calmodulin in complex with Ca 2+-calmodulin-dependent kinase kinase. *Nat. Struct. Mol. Biol*., **6**, 819.

Polacco,B.J. and Babbitt,P.C. (2006) Automated discovery of 3D motifs for protein function annotation. *Bioinformatics*, **22**, 723–730.

Ragoza,M. *et al*. (2017) Protein–Ligand scoring with Convolutional neural networks. *J. Chem. Inf. Model*., **57**, 942–957.

Rani,S.G. *et al*. (2014) Interaction of S100A13 with C2 domain of receptor for advanced glycation end products (RAGE). *Biochim. Biophys. Acta*, **1844**, 1718–1728.

Rawlings,N.D. and Barrett,A.J. (1994) Families of serine peptidases. *Methods Enzymol*., **244**, 19–61.

Sigrist,C.J. *et al*. (2012) New and continuing developments at PROSITE. *Nucleic Acids Res*., **41**, D344–D347.

Skwark,M.J. *et al*. (2014) Improved contact predictions using the recognition of protein like contact patterns. *PLoS Comput. Biol*., **10**, e1003889.

Somarowthu,S. *et al*. (2011) High-performance prediction of functional residues in proteins with machine learning and computed input features. *Biopolymers*, **95**, 390–400.

Tang,G.W. and Altman,R.B. (2014) Knowledge-based fragment binding prediction. *PLoS Comput. Biol*., **10**, e1003589.

Torng,W. and Altman,R.B. (2017) 3D deep convolutional neural networks for amino acid environment similarity analysis. *BMC Bioinformatics*, **18**, 302.

Vitagliano,L. *et al*. (1999) A potential allosteric subsite generated by domain swapping in bovine seminal ribonuclease1. *J. Mol. Biol*., **293**, 569–577.

Wallace,A.C. *et al*. (1997) TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Sci*., **6**, 2308–2323.

Wang,Z. *et al*. (2013) Protein Function Annotation with Structurally Aligned Local Sites of Activity (SALSAs). *BMC Bioinformatics*, **14**, S13.

Xiong,J. (2006) *Essential Bioinformatics*. Cambridge University Press, Cambridge.

Downloaded from https://academic.oup.com/bioinformatics/article/35/9/1503/5104336 by guest on 08 November 2020