Predicting the binding-site of a protein for druggable ligands from sequence-based features using Deep Learning

Vineeth R. Chelur and U. Deva Priyakumar*

Center for Computational Natural Sciences & Bioinformatics, IIIT-H, Hyderabad

E-mail: deva@iiit.ac.in

Abstract

Motivation: Protein-drug interactions play important roles in many biological processes. The prediction of the active binding site of a protein helps discover such interactions. The tertiary structure of a protein determines the binding sites available to the drug molecule. But the methods for structure determination are labour-intensive and time-consuming. Hence it becomes important to make predictions using the sequence alone. Deep Learning has been used in a variety of biochemical tasks and has been hugely successful. In this paper, a residual neural network is implemented to predict a protein's most active binding site using features extracted from just the sequence.

Results: The model achieved a Matthew's Correlation Coefficient of 0.54 and an accuracy of 93% on the validation set. On the test set, an MCC of 0.51 and an accuracy of 91% was obtained.

Implementation: https://github.com/crvineeth97/protein-binding-site-prediction

Introduction

The rapid speed of sequencing attained with modern DNA sequencing technology has been instrumental in the sequencing of complete DNA sequences, which leads to faster indirect sequencing of proteins. Although there have been improvements in the determination of the protein three-dimensional structure by techniques such as X-ray crystallography and NMR Spectroscopy, the gap between the number of known sequences (333,201,385 as of July, 2020) and the number of known structures (154,706 as of July, 2020) is increasing rapidly. Proteins perform a vast array of functions within organisms and the tertiary structure of a protein can provide important clues about these functions.

Deep Learning is a subfield of machine learning based on artificial neural networks with feature learning. When a deep learning model is fed large amounts of data, it can automatically discover the representations needed for feature detection or classification. Deep learning has been hugely successful in all fields of Natural Sciences, including, but not limited to, binding affinity predictions, protein contact map predictions, and protein-structure predictions. For example, Google's DeepMind team designed a deep learning model, called AlphaFold,[1] which represents a considerable advance in protein-structure prediction from sequence.

In the process of drug design, new medications are found based on the knowledge of a biological target such as a protein. The identification of the potential active binding site of a protein is an important step in drug design. Predicting the binding site of a protein based on sequence alone becomes critical when the three-dimensional structure of the protein is not available.

In this paper, a deep residual neural network is used to make the binary prediction of whether an amino acid residue of the sequence belongs to the binding site or not.

Dataset

For the training and validation of the model, the sc-PDB[2] dataset (v 2017) is used. The database consists of druggable binding sites of the Protein Data Bank along with prepared protein structures. Thus each sample in the dataset contains the three-dimensional structure of one ligand, one protein, and one site, all stored in mol2 format.

Typically, the complete structure of a protein is unavailable due to missing residues and hence the full sequence of the protein is obtained from RCSB[3] website*. A one-to-one mapping of the amino acid in the sequence to the one in the protein mol2 file is required, to know which amino acid belongs to a binding site. This is done by first extracting the protein sequence from the 3D structure. Next, Needleman-Wunsch dynamic programming algorithm is used to align the sequence extracted from the structure file to the downloaded RCSB sequence. This algorithm is implemented using a modified version of Zhanglab's NW-Align program.[4] The protein structure file is then reindexed, based on this alignment, to match the indexing of the RCSB sequence. This way, the specific binding residues can be labelled in the RCSB sequence.

The training set consists of 17,594 PDB structures with 28,959 sequences (9519 unique sequences), originating from 1240 organisms, the most abundant being human(28.26%), (Add the rest here). The dataset was diverse and contained proteins from 1996 different PFAM families (Most abundant needs to be added) and 856 PFAM clans.

This data is split into 10-folds (each containing 1586 structures), based on Uniprot ID, exactly like the paper on Kalasanty.[5] This ensures that there is no data leakage between the validation and training set, by putting all structures of a single protein in the same fold.

The test set is constructed using all PDBs from 2018 onwards, till 28th February 2020. All PDBs available during this period and having at least one ligand are considered. These are then run through the IChem Toolkit[6] to generate a dataset similar to the training set.

---

*Some PDBs in the dataset were obsoleted, and hence the sequences were manually tracked on RCSB, and the corresponding sequences were used. List of obseleted PDBs is provided in Supporting Information

The test set consists of 2,274 PDB structures with 3,434 sequences (1889 unique sequences), originating from 548 organisms, the most abundant being human(23.76%). The test set contained proteins from 882 PFAM families and 452 PFAM clans.

Methods

MSA Generation

Collections of multiple homologous sequences (called Multiple Sequence Alignments or MSAs) can provide critical information for the modelling of the structure and function of unknown proteins. DeepMSA[7] is an open-source method for sensitive MSA construction, which has homologous sequences and alignments created from multiple sources of databases through complementary hidden Markov model algorithms.

The search is done in 2 stages. In stage 1, the query sequence is searched against the UniClust30[8] database using HHBlits from HH-suite[9] (v2.0.16). If the number of effective sequences is < 128, Stage 2 is performed where the query sequence is searched against the Uniref50[10] database using JackHMMER from HMMER[11] (v3.1b2). Full-length sequences are extracted from the JackHMMER raw hits and converted into a custom HHBlits format database. HHBlits is then applied to jump-start the search from Stage 1 sequence MSA against this custom database.

Feature Extraction

There are 9519 unique protein sequences in the sc-PDB dataset and 1889 unique protein sequences in the test set. The MSAs are generated for these sequences using the method described above and stored in PSICOV[12] .aln format. The following features are extracted from the generated MSAs.

PSSM and IC

Position Specific Scoring Matrix (PSSM) is a commonly used representation of patterns in biological sequences. The MSA is converted into a position probability matrix which tells the probability of an amino acid occurring at a specific position in the protein sequence. (Check what Easel does for making the PSSM)

The information content (IC) of a PSSM gives an idea about how different the PSSM is from a uniform distribution. Note: Can give details of the math

Secondary Structure and Solvent Accessibility

The secondary structure is defined by the pattern of hydrogen bonds formed between the amino hydrogen and carboxyl oxygen atoms in the peptide backbone. This gives an idea about the three-dimensional structure of the protein. The secondary structural elements are alpha helices, beta sheets and turns. PSIPRED (v4.0)[13] is used to predict the probability of each state of the 3-state secondary structure for every amino acid in the sequence.

The solvent-accessible surface area is the surface area of a biomolecule that is accessible to a solvent. SOLVPRED from MetaPSICOV 2.0[14] is used to predict the relative solvent accessibility (RSA) of every amino acid in the sequence. RSA can be calculated as $RSA = ASA/MaxASA$, where ASA is the solvent accessible surface area and MaxASA is the maximum possible solvent accessible surface area for the amino acid residue.

SPOT-1D Features

As a means to provide better features, SPOT-1D[15] was used to generate the following features: solvent accessibility, half-sphere exposure, contact number, 3-state secondary structure, 8-state secondary structure, and phi, psi, theta and tau torsion angles.

The first step in the prediction pipeline was to get the ASCII PSSM file in PSI-BLAST format. Then, hhmake was used to generate the HHM file from the MSA. SPIDER3, DCA and CCMPRED predictions were made and stored.

The second step was to predict the contact map using SPOT-Contact, which used the previous steps predictions.

Finally, SPOT-1D was used to make the final predictions using all the previous files as input.

Deep Learning Model

Residual neural networks are immensely popular in image recognition and also have been gaining traction in the field of computational chemistry as well. A variety of deep learning models were tested but 1D Resnets proved to be the best model. A Resnet consists of multiple basic blocks stacked on top of each other

Results

Table of results here

Evaluation Metrics

Accuracy

Precision, Recall and F1-score

MCC

DCC

Experiments

test

Discussion

Case Studies

test

Areas for Improvement

test

Flaws

test

Acknowledgement

Supporting Information Available

This will usually read something like: "Experimental procedures and characterization data for all new compounds. The class will automatically add a sentence pointing to the information on-line:

References

(1) Senior, A. W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Žídek, A.; Nelson, A. W.; Bridgland, A., et al. Improved protein structure prediction using potentials from deep learning. *Nature* **2020**, *577*, 706–710.

(2) Desaphy, J.; Bret, G.; Rognan, D.; Kellenberger, E. sc-PDB: a 3D-database of ligand-able binding sites—10 years on. *Nucleic acids research* **2015**, *43*, D399–D404.

(3) Burley, S. K.; Berman, H. M.; Bhikadiya, C.; Bi, C.; Chen, L.; Di Costanzo, L.; Christie, C.; Dalenberg, K.; Duarte, J. M.; Dutta, S., et al. RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic acids research* **2019**, *47*, D464–D474.

(4) Zhang, Y. `http://zhanglab.ccmb.med.umich.edu/NW-align`.

(5) Stepniewska-Dziubinska, M. M.; Zielenkiewicz, P.; Siedlecki, P. Improving detection of protein-ligand binding sites with 3D segmentation. *Scientific reports* **2020**, *10*, 1–9.

(6) Da Silva, F.; Desaphy, J.; Rognan, D. IChem: A Versatile Toolkit for Detecting, Comparing, and Predicting Protein–Ligand Interactions. *ChemMedChem* **2018**, *13*, 507–510.

(7) Zhang, C.; Zheng, W.; Mortuza, S.; Li, Y.; Zhang, Y. DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. *Bioinformatics* **2020**, *36*, 2105–2112.

(8) Mirdita, M.; von den Driesch, L.; Galiez, C.; Martin, M. J.; Söding, J.; Steinegger, M. Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic acids research* **2017**, *45*, D170–D176.

(9) Remmert, M.; Biegert, A.; Hauser, A.; Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature methods* **2012**, *9*, 173–175.

(10) Suzek, B. E.; Wang, Y.; Huang, H.; McGarvey, P. B.; Wu, C. H.; Consortium, U. UniRef

clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **2015**, *31*, 926–932.

(11) Johnson, L. S.; Eddy, S. R.; Portugaly, E. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC bioinformatics* **2010**, *11*, 431.

(12) Jones, D. T.; Buchan, D. W.; Cozzetto, D.; Pontil, M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* **2012**, *28*, 184–190.

(13) Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of molecular biology* **1999**, *292*, 195–202.

(14) Jones, D. T.; Singh, T.; Kosciolek, T.; Tetchner, S. MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics* **2015**, *31*, 999–1006.

(15) Hanson, J.; Paliwal, K.; Litfin, T.; Yang, Y.; Zhou, Y. Improving prediction of protein secondary structure, backbone angles, solvent accessibility and contact numbers by using predicted contact maps and an ensemble of recurrent and residual convolutional neural networks. *Bioinformatics* **2019**, *35*, 2403–2410.