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

With improvements in sequencing methods, the number of protein sequences available is rapidly increasing. However, because of the high cost and labour-intensive nature of structural experiments, the gap between the number of protein sequences and solved structures is widening rapidly. One of the earliest steps in drug discovery is identifying the active binding site of the target protein. 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 primary sequence. (Add information of results)

Introduction

Dataset

For the training and validation of the model, the sc-PDB[1] 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 one ligand, one protein, and one site, all stored in mol2 format. Since the predictions are made from the sequence alone, the provided mol2 files are reindexed to match the sequence downloaded from RCSB. This way, the specific binding residues can be labelled in the sequence. Some PDB IDs are obsoleted, and hence the sequences were manually tracked on RCSB, and the corresponding sequences were used.

Needleman-Wunsch dynamic programming for pairwise protein sequence alignment implemented using a modified version of Zhanglab's NW-Align program[2] was used to reindex a protein according to its 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. The data from sc-PDB is split into 10-folds (each containing 1586 structures), based on Uniprot ID, exactly like Kalasanty.[3]

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 were then run through IChem Toolkit[4] to generate a dataset similar to the sc-PDB dataset. 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

As described in the introduction, the number of protein sequences is rapidly exploding. Collections of multiple homologous sequences (called Multiple Sequence Alignments or MSAs)

can provide critical information to the modelling of the structure and function of unknown proteins. DeepMSA[5] 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[6] database using HHBlits from HH-suite[7] (v2.0.16). If the number of effective sequences is < 128, Stage 2 is performed where the query sequence is searched against the Uniref50[8] database using JackHMMER from HMMER[9] (v3.1b2). Full-length sequences are extracted from the JackHMMER raw hits and converted into a custom HHBlits format database. HHBlits is 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 training + validation set and 1889 unique protein sequences in the test set. The MSAs are generated using the method described above and stored in PSICOV[10] .aln format. The following features are extracted using the MSAs.

PSSM and IC

Position Specific Scoring Matrix is a commonly used representation of patterns in biological sequences. The MSA is converted into a position probability matrix, and then the log-likelihoods of each element is taken. The information content 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. The two most common secondary structural elements are alpha helices and beta sheets. The secondary structure

gives an idea of the 3D structure of the protein. The solvent-accessible surface area is the surface area of a biomolecule that is accessible to a solvent. The PSICOV .aln file is first converted into PSI-BLAST[11] profile format (.mtx). PSIPRED (v4.0) and SOLVPRED (MetaPSICOV 2.0) were used to predict the 3-state secondary structure and relative solvent accessibility, respectively.

SPOT-1D Features

As a means to provide better features, SPOT-1D[12] was used to generate the following features: solvent accessibility, half-sphere exposure, contact number, 3-state secondary structure, 8-state secondary structure, phi, psi, theta, and tau.

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

Results

Table of results here

Experiments

test

Discussion

Case Studies

test

Areas for Improvement

test

Flaws

test

Acknowledgement

The author thanks Yashaswi Pathak for being a fruitful part of the project discussions.

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

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

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

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

(5) 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.

(6) 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.

(7) 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.

(8) 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.

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

(10) 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.

(11) Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.;

Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* **1997**, *25*, 3389–3402.

(12) 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.