MACHINE LEARNING ALGORITHMS FOR PROTEIN FUNCTION PREDICTION FROM PRIMARY STRUCTURE

UNIVERSITY OF LINCOLN

Patrick Ellison, Vassilis Cutsuridis

School of Computer Science, University of Lincoln, Lincoln, U.K.
Email: vcutsuridis@lincoln.ac.uk

Background

Proteins are complex macromolecules made up of often hundreds of amino acids, each attached to one another in long, un-branched chains. The amino acid sequence determines each protein's unique 3-dimensional structure and its specific function. One such function is the specific read-out of DNA sequences, e.g. for genome editing via suitable DNA-binding proteins.

DNA binding proteins that target a specific site in the genome are able of selecting the correct binding site out of a vast number of potential sites on the DNA using various interactions which promote molecular recognition (Halford and Marko, 2004). These interaction modes are associated with a multitude of functionally distinct proteins that determine the morphology, function of cells, tissues and whole organisms giving rise to unique phenotypes of living individuals.

Genome editing via DNA-binding domains such as mega-nucleases, zinc finger nucleases, transcription activator–like effector nucleases and the clustered regularly interspaced short palindromic repeat (CRISPR)-associated nuclease Cas9 is accomplished by the introduction of DNA double-strand breaks at specific genomic sites that activate DNA repair via mutations, insertions, replacements and chromosomal rearrangements (Chen and Gao, 2013). Such technologies are opening up new possibilities of therapeutic genome editing in diseased cells and tissues through the removal, correction and insertions of mutations (Turitz-Cox et al., 2015).

Over the years, various statistical and machine learning techniques including neural networks have been employed to understand the sequence-structure-function relationship of proteins and uncover the mechanisms of their evolution.

Objective

Compare and contrast the performance of machine learning classifiers

- ❑ kNN
- ❑ SVM (support vector machine)
- ❑ ANN (deep feedforward neural network)
- ❑ CNN (comvolutional neural network)

in predicting the function of a protein from its primary amino-acid structure.

Datasets and Protein Coding

- ❑ Gene Ontology-based search against the UniProt/SwissProt database for "DNA sequence-specific binding proteins" retrieved 6492 sequences of amino acids.
- ❑ Complement search set returned 524406 sequences.
- ❑ Sequences less than 150 amino acids and over 1000 amino acids in length were thrown out, whereas the remaining ones were made equal-in-length by padding them with "Xs" till their length was equal to 1000.
- ❑ Every amino acid was converted into its corresponding integer value using MATLAB's *aa2int* function.

Amino Acid	Code	Integer	Amino Acid	Code	Integer
Alanine	A	1	Methionine	M	13
Arginine	R	2	Phenylalanine	F	14
Asparagine	N	3	Proline	P	15
Aspartic acid (Aspartate)	D	4	Serine	S	16
Cysteine	C	5	Threonine	T	17
Glutamine	Q	6	Tryptophan	W	18
Glutamic acid (Glutamate)	E	7	Tyrosine	Y	19
Glycine	G	8	Valine	V	20
Histidine	H	9	Asparagine or Aspartic acid (Aspartate)	B	21
Isoleucine	I	10	Glutamine or Glutamic acid (Glutamate)	Z	22
Leucine	L	11	Unknown amino acid (any amino acid)	X	0
Lysine	K	12			

- ❑ Any instance of the amino acid 'X' was replaced with '0'.
- ❑ Every AA sequence was further appended with a '1' if 'DNA binding' and '0' if 'non-DNA binding'.

Training and Testing

- ❑ Different size datasets were created: (1) Small dataset: 100 sequences (50 DNA binding and 50 non-DNA binding); (2) Large dataset: 10500 sequences (5250 DNA binding, 5250 non-DNA binding)).
- ❑ Amino acid sequences were randomly shuffled, so the classifiers did not learn the order of presentation.
- ❑ Using the 'train_test_split' function from the SciKitLearn library 90% of each dataset was used training and 10% for testing.
- ❑ To avoid overfitting, 10-fold cross validation was used.

Performance Measures

Accuracy measured the proportion of correct predictions:

$$Ac = \frac{TP}{Total\ number\ of\ proteins}$$

Sensitivity measured the proportion of actual positives that were correctly identified:

$$S = \frac{TP}{TP + FN}$$

Specificity measured the proportion of actual negatives that were correctly identified:

$$Sp = \frac{TN}{TN + FP}$$

Results

Classifier	Accuracy	Sensitivity	Specificity
ANN	0.79	0.89	0.70
CNN	0.82	0.80	**0.84**
kNN	**0.84**	0.88	0.80
SVM	0.74	**0.91**	0.58

References

Chen K, Gao C. (2013). TALENs: Customizable molecular DNA scissors for genome engineering of plants. J Gen Genomics, 40: 271-279

Halford SE, Marko JF. (2004). How do site-specific DNA-binding proteins find their targets? Nuclei Acids Research 32(10): 3040-3052

Turitz-Cox DB, Platt RJ, Zhang F. (2015). Therapeutic genome editing: prospects and challenges. Nat Medicine 21(2): 121-131