Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Friday 27 March 2020
Manuscript Category

Bioinformatics (COMP0082) - Mini project

Predicting the subcellular location of eukaryotic proteins using a Random Forest classifier

Giorgos Felekis

Department of Computer Science, UCL, London, WC1E 6BS, UK

Abstract

Motivation: Protein function prediction is crucial to the proteomics field due to the increasing number of preotein sequences available and the costly exerimental process needed determine their function. However, function prediction is a highly complex problem and some simplifications such as subcellular location prediction can help orienting further research. The usage of machine learning methods has been applied in bioinformatics and computational biology with success in areas such as genomic analysis or protein structure prediction.
Results: In this report we present a Random Forest classifier which is an effective algorithm for predicting subcellular localization of eukaryotic proteins. After implementing a 7-fold cross validation to asses performance of the model we are get a 65% accuracy.
Availability: Our machine learning model together with the data pre-processing and feature generation are publicly available at the following Colab notebook link: https://colab.research.google.com/drive/1_I_XgxgO4fvF1SPdaEaqae6I6Mp1LFRC
Contact: georgios.felekis.19@ucl.ac.uk
Supplementary information: http://www0.cs.ucl.ac.uk/staff/D.Jones/coursework/

1 Introduction

Over the past decade, the demand for automated protein function prediction has risen due to the increasing number of protein sequences produced by high throughput sequencing methods. Proteins play a vital role in many processes of living organisms, including the structure, function, and regulation of the body's tissues and organs. Proteins are constituted of smaller units called amino acids, which attatched to one another in long chains, give the final protein sequence. Eukaryote cells use 20 amino acids that can be combined together to make a protein. However, some procariote cells also use selenocysteine (21st aminoacid) to manufacture some of their proteins. It's crucial to understand that each protein has a specific amino acid chain that defines the protein's unique 3-dimensional structure and specific function. Proteins can be described according to their large range of functions in the body as Antibody, Enzyme, Messenger, Structural component and Transport/storage. They can also be divided based on their location on the cell, also known as subcellular location. Specifically, we can find proteins in many parts of a cell, including nucleus, cytoplasm, extracellular, mitochondrion, cell membrane, endoplasmic reticulum, plastids, Golgi apparatus and a few more. One of the primary goals in cellular biology is to identify the functions of proteins in the context of compartments that organize them in the cellular environment. Although the information about protein subcellular localization can be determined by performing various experiments, they are costly, time-expensive and labor-intensive. From all the above we can understand that there is an urgent need for computationally efficient methods to tackle these issues. Many studies have shown that proteins may simultaneously locate in different cellular areas and be involved in different biological processes with different roles.

2 Related work

In recent years, the prediction of eukaryotic protein subcellular location is becoming a more and more well-studied topic in bioinformatics and computational biology due to its relevance in proteomics research. Moreover, due to the abundant amount of data which is available, many machine learning methods have been successfully applied in this task. Lately, various classifiers have been proposed, some of them approach the task with more classical machine learning algorithms like SVMs, see: Hua and Sun (2001), Mehedi and Hasan (2017) and Sarda et.al (2005), k-NN (see Chou and Shen, 2006) and Naïve Bayes (see Briesemeister et. al, 2010) while the more recent ones are mainly based on neural network architectures of either ConvNets (see Pang L. et al, 2018) models or Recurrent Neural Netwrok architectures (see Sønderby, et. al) and sometimes a combination of both (see Armenteros et.al, 2017). The

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

work that has been done in Wan and Mak (2015) provides a nice overall understanding of some of the methods that traditionally been used.

3 The data

The data that has been used is a collection of protein sequences that are available in the colab notebook alonside the code used to produce the results described in this paper. Protein data is provided in FASTA format which is a text-based format for representing either nucleotide sequences or peptide sequences, in which base pairs or amino acids are represented using single-letter codes. Our data is divided in four classes, based on the location of the protein. Specifically, we had Cytosolic, Secreted, Mitochondrial and Nuclear protein sequences.
Sequences were expected to be represented in the standard IUB/IUPAC amino acid and nucleic acid codes.

4 Methods

In this report we present an efficient approach to the problem of predicting the subcellular location of proteins even though we are using one of the most classic machine learning algorithms, Random Forest.

4.1 RandomForest

Random forests (see Breiman, 2001) are a prediction algorithm part of the wider family of decision trees algorithms. Specifically, they are an ensemble of Decision Trees. Each Decision Tree in the ensemble is built based on a different random subset of the input features, and the final prediction of the ensemble is most of the time decided by a simple rule amongst all the trees such as majority voting. Random forests are a reasonable choice for the certain task since they exhibit many features that re of interest in this problem: they are able to represent non-linear decision boundaries, they have been proven to support multi-label classification on thousands of labels quite efficiently, and they can be scaled up by parallelization of both training and inference. The classifier was implemented using scikit-learn library version 0.22.2, a free software machine learning library for the Python programming language.

4.2 Experiment set up

Our aim is to efficiently match proteins to their location on the cell. Specifically, given the amino acid sequence we want to identify the probability of a certain protein to be in one of the four following locations of the cell:

- Cytosolic - i.e. within the cell itself, but not inside any organelles
- Secreted - proteins which are transported out of a cell
- Nuclear - proteins found/used within the cell's nucleus
- Mitochondrial - proteins transported to the cell's mitochondria

We subsequently randomly divide the dataset into two parts: The training set used for training the classifier and the validation set which is used for the final performance estimation. The The training and validation subsets contain 80% and 20% of the whole data each. We used a blind-test dataset to showcase the value of predicting subcellular location of proteins.
Different feature groups were tested in combination to select the ones that gave the best classification performance. We also perform a 7-fold cross validation to estimate how the model is expected to perform in general when used to make predictions on unseen before data.

4.3 Model set up

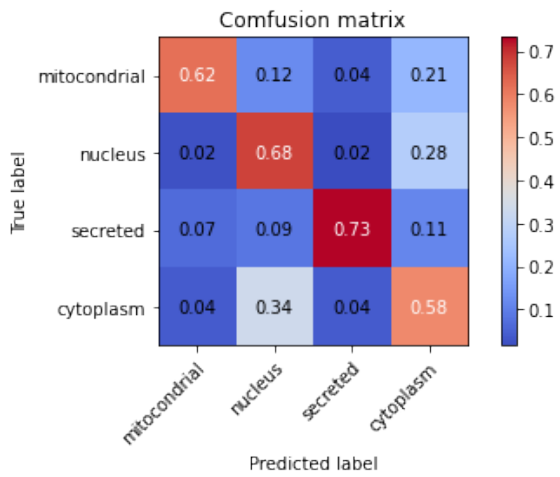
The final Random Forest selected after the cross-validation step was an ensemble of 100 decision trees, trained to minimize the gini coefficient and maximum tree depth was set to 27 to avoid overfititng. The minimum samples for any given split was set to 2 while the minimum number of samples for any leaf note was set to 1.
In order to decide the location, we used five sets of features regarding the sequence length and the global amino acid composition (i.e percentages of all 20 amino acids present in the whole sequence), the local amino acid composition (i.e. over first 50 amino acids or last 50 amino acids to local features near the end or the start of the protein), amino acid physicochemical properties (the isoelectric point, molecular weight, etc) and signal peptides (https://www.ncbi.nlm.nih.gov/pubmed/29958716) (i.e. short sequences of amino acids that have been proven to determine the subcellular location).

5 Results and Discussion

In our final experiment we got a 100.0% training accuracy and 65% validation accuracy. Given the fact that we used the baseline features test accuracy could be increased in significant way and match state-of-the-art results (which is around 80% accuracy) by:

- Addition of more data points e.g more protein sequences.
- Addition of more sophisticated features that might well correlate with the subcellular location other than the five sets we used here (domain identification, local features alongside the whole protein, etc). A few other features that have been used in different methods, e.g., phylogenetic profiling (see Marcotte, et.al), domain projection (see Mott, Schultz), sequence homology (see Yu, et.al), and compartment-specific features (see Su, et.al).
- Use of a neural network architecture like an LSTM model that would learn all those features from scratch and would take advantage of the sequential nature of proteins.

Overall, the key to enhancing the prediction quality for protein subcellular location is to find the most relevant features but also find the appropriate model that can appoint them.



Also after training the model to the full initial dataset we tested our method on a final "blinded" test set of mutated proteins to see how it performs on true unknown sequences. The results are presented in the following table and provide a mean confidence of 37.8%:

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(-35,0)(1,0)30 (0,35)(0,-1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,35)(0,-1)30 picture

Confidence Table of "Blind Test" Proteins			
SEQ677	Secr	Confidence	40.0%
SEQ231	Secr	Confidence	38.0%
SEQ871	Secr	Confidence	37.0%
SEQ388	Secr	Confidence	41.0%
SEQ122	Secr	Confidence	35.0%
SEQ758	Nucl	Confidence	33.0%
SEQ333	Nucl	Confidence	32.0%
SEQ937	Cyto	Confidence	41.0%
SEQ351	Cyto	Confidence	28.9%
SEQ202	Secr	Confidence	40.0%
SEQ608	Cyto	Confidence	35.0%
SEQ402	Secr	Confidence	42.0%
SEQ433	Secr	Confidence	42.0%
SEQ821	Secr	Confidence	35.0%
SEQ322	Secr	Confidence	43.0%
SEQ982	Nucl	Confidence	45.0%
SEQ951	Cyto	Confidence	34.0%
SEQ173	Cyto	Confidence	34.0%
SEQ862	Secr	Confidence	35.0%
SEQ224	Secr	Confidence	42.0%

6 Conclusion

Prediction of protein subcellular location is a very challenging and complicated problem. In the modern era, the amount of data is huge and the need to apply machine learning methods to exploit that seems more reasonable than ever. The more the subcellular locations covered, the lower the probability of making a correct prediction. Also, the more strictly the working dataset in excluding homologous sequences, the harder it becomes to get a higher success rate for cross-validation tests. Moreover, most of the existing prediction methods that had been proposed only cover 2 to 5 subcellular locations so it would be a challenge to increase this number. In this direction more complex and state-of-the-art models should be used. A good example is the model in Armenteros et.al (2017) which is a convolutional BLSTM neural network with attention mechanism. Overall, subcellular location prediction of proteins is an active area of researcg and many iprovements might be done in the following years.

References

Long Pang, Junjie Wang, Lingling Zhao, Chunyu Wang and Hui Zhan. (2018).A Novel Protein Subcellular Localization Method With CNN-XGBoost Model for Alzheimer's Disease.

Wan, Shibiao Mak, Man-Wai. (2015). Machine Learning for Protein Subcellular Localization Prediction.

Hua,S. and Sun,Z. (2001) Support vector machine approach for protein subcellular localization prediction.Bioinformatics,17,721–728

Mehedi Hasan A, Ahmad S, Molla K. Prediction of protein subcellular localization using support vector machine with the choice of proper kernel. BioTechnologia. 2017;98(2):85-96.

Sarda, D., Chua, G.H., Li, K. et al. pSLIP: SVM based protein subcellular localization prediction using multiple physicochemical properties. BMC Bioinformatics 6, 152 (2005). https://doi.org/10.1186/1471-2105-6-152

Chou, Kuo-Chen Shen, Hong-Bin. (2006). Predicting eukaryotic protein subcellular location by fusing optimized Evidence-Theoretic K-Nearest neighbor classifiers.. Journal of proteome research. 5. 1888-97. 10.1021/pr060167c.

Briesemeister S, Rahnenführer J, Kohlbacher O. Going from where to why–interpretable prediction of protein subcellular localization. Bioinformatics. 2010;26(9):1232–1238. doi:10.1093/bioinformatics/btq115

José Juan Almagro Armenteros, Casper Kaae Sønderby, Søren Kaae Sønderby, Henrik Nielsen, Ole Winther, DeepLoc: prediction of protein subcellular localization using deep learning, Bioinformatics, Volume 33, Issue 21, 01 November 2017, Pages 3387–3395

Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

Sønderby, S. K., Sønderby, C. K., Nielsen, H., Winther, O. (2015). Convolutional LSTM Networks for Subcellular Localization of Proteins. In A-H. Dediu, F. Hernández-Quiroz, C. Martín-Vide, D. A. Rosenblueth (Eds.), Algorithms for Computational Biology: Second International Conference, AlCoB 2015, Mexico City, Mexico, August 4-5, 2015, Proceedings (Vol. 9199, pp. 68-80). Springer. Lecture Notes in Computer Science

Marcotte EM, Xenarios I, van Der Bliek AM, Eisenberg D. Localizing proteins in the cell from their phylogenetic profiles. Proc Natl Acad Sci U S A. 2000;97(22):12115–12120. doi:10.1073/pnas.220399497

Mott R, Schultz J, Bork P, Ponting CP. Predicting protein cellular localization using a domain projection method. Genome Res. 2002;12(8):1168–1174. doi:10.1101/gr.96802

Yu, C.-S., Chen, Y.-C., Lu, C.-H. and Hwang, J.-K. (2006), Prediction of protein subcellular localization. Proteins, 64: 643-651. doi:10.1002/prot.21018

Su EC, Chiu HS, Lo A, Hwang JK, Sung TY, Hsu WL. Protein subcellular localization prediction based on compartment-specific features and structure conservation. BMC Bioinformatics. 2007;8:330. Published 2007 Sep 8. doi:10.1186/1471-2105-8-330

picture(0,0)(-35,0)(1,0)30 (0,-35)(0,1)30 picture

picture(0,0)(35,0)(-1,0)30 (0,-35)(0,1)30 picture