A machine learning approach for the prediction of prokaryotes coding gene functions

Introduction:

With the development of the field of bioinformatics, many different kinds of tools have been developed for the prediction of gene functions. Some of the methods are based on the homology of the proteins (Loewenstein et al., 2009), some of the methods are structure-based (Sael et al., 2012), and some of the methods are based on the conserved co-expression,(van Noort et al., 2003) etc. In my research, I will discuss the feasibility of developing a method to predict the gene functions of some not well-studied prokaryotes by machine learning on the genomes of some well-studied and annotated prokaryotes.

As we all known, E. coli (*Escherichia coli str. K-12 substr.)* is one of the prokaryotes that we have done most research on. According to the data from NCBI, among the ~4400 of the coding genes of E. coli genome, more than 3/4 of the genes have been clearly annotated with particular functions. Therefore, for my study, I will use the annotated genome of E. coli as my training group. After the training, I will use the annotated genome of another prokaryote *Granulosicoccus antarcticus IMCC3135* as the test group, to see if the method works. By the testing results, the prediction accuracy is pretty high.

Methods:

The central idea of the method is to train a set of classifiers for some particular functional categories to make predictions on whether a gene belongs to this functional category. In this project, I will choose 7 different functional categories to work on and use both support vector machine (SVM) and logistic regression methods to train the classifier and compare their prediction accuracy on the test results. The number of times each amino acid k-mer (3-mer for my project) occurs in the protein sequence of each gene will be used as features for the SVM classifiers training (Leslie et al., 2002) and for the logistic regression classifiers training.

The first step of my work is to find out which functional categories I want to train with. In order to do that, I need to count which key words appear for the most times in the descriptions of the genes. The purpose of this is to make sure that there are enough number of positively labeled samples for the training of each classifier. By the result, I decided to work with the functional categories: transcriptional, transporter, membrane, dehydrogenase, kinase, prophage and synthase. A gene is assumed to belong to this particular category if this keyword appears in the description of the gene. Each of the category has around 150-300 positively labeled genes.

Then, for the training procedure of each functional category classifier, it is done in following way: for each gene, the k-spectrum of the gene will be the x-value; if the keyword of the functional category is in the description of the gene, the y-value of the gene will be 1, if not, the y-value of the gene will be 0.

The k-spectrum is a vector of the number of times each amino acid k-mer of a fixed length occurs. For the 3-mer case I use, there are totally 20\*\*3 = 8000 kinds of k-mers, so the length of the k-spectrum vector is 8000. The kernel function used in the SVM training will simply be the dot product of two k-spectrum vectors.

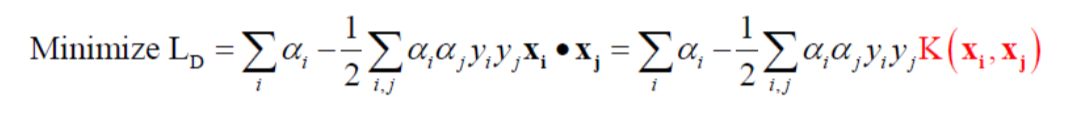


Figure 1. The formula for the training of support vector machine, K(xi, xj) denotes the kernel function.

One potential problem for the training is that there are always less training samples labeled as 1 than labeled as 0, so we need to find a way to make the sample size in both groups equal. Use the category “transcriptional” as the example, there are 291 genes annotated as “transcriptional” regulators (activators, repressors, etc.), 4066 not annotated as “transcriptional” in the description. Then, when we are creating the training set, if the keyword “transcriptional” is not in the description, I would generate a random number between 0 and 1. Only if the number < 0.072 (291/4066), the gene will be put into the training set. In this way, approximately same number of samples labeled as 1 or 0 in y-value.

After the X and Y sets are all ready to be used, an SVM and a logistic regression are trained for each functional category.

A screenshot of a cell phone

Description automatically generated

Figure 2. The python code used for the training the classifier of “transcriptional” related genes with SVM method.

Results:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | SVM, with keyword (sensitivity) | SVM, without keyword (specificity) | Logistic, with keyword (sensitivity) | Logistic, without keyword (specificity) |
| Transcriptional | 0.78 | 0.72 | 0.735 | 0.71 |
| Transporter | 0.81 | 0.81 | 0.79 | 0.8 |
| Dehydrogenase | 0.75 | 0.74 | 0.83 | 0.71 |
| Membrane | 0.64 | 0.68 | 0.72 | 0.69 |
| Kinase | 0.65 | 0.63 | 0.44 | 0.69 |
| Synthase | 0.61 | 0.65 | 0.75 | 0.55 |
| Prophage | 0 | 0.65 | 0 | 0.68 |

Table 1. Results of prediction accuracy for the test of the classifiers on *Granulosicoccus antarcticus* genome. “With keyword” means the prediction result should be 1, “without keyword” means the prediction result should be 0.

In the table, the sensitivity means the proportion of genes that got prediction y-value “1” and their expected y-value are “1”, and the specificity is the proportion of genes that got prediction y-value “0” and their expected y-value are “0”.

From the prediction result table, we can see the prediction for transcriptional, transporter, and dehydrogenase are pretty accurate. The sensitivities and specificities are all pretty high by using both methods. It is reasonable because inside each of these categories, the proteins would share many similar sequences to do same work. For example, for all the transcriptional factor proteins, there are sequences that are responsible for DNA binding (Calvin et al., 1987); for all the “transporter” protein, there are sequences that are responsible for the transmembrane pump.

The accuracy of the prediction of membrane proteins is slightly lower than those three categories, that may because of the fact that some proteins are related to inner membrane and some are related to outer membrane (Nakae, 1976). That means, if we have enough samples for the training, they should actually belong to two categories.

The prediction accuracy of kinase and synthase are close, and lower than the others. That may because that the kinase and synthase are some wide ranges of different proteins. It is harder to regard them as one single category to train and classify.

The sensitivity for the prophage is 0. That is because there is no protein annotated as prophage in the test group.

Compare the training method SVM with the method logistic regression, the prediction accuracy results are similar within each category. That means for this task, there no significant difference between using SVM and logistic regression.

Discussion:

By testing the classifiers on the test group, the prediction accuracy results show that with enough annotated genes in the training group, using the number of amino acid k-mers in protein sequences as the feature to do supervised machine learning may be a feasible way to predict the function of the proteins of the genes. However, different functional categories of proteins may also have different prediction accuracies.

Besides training the classifiers by the features of amino acid 3-mers, I also tried to use 2-mers and 4-mers as the features, but the results were not ideal.

If there is no enough annotated genes for a specific functional categories in a single prokaryote genome, we can try to combine different prokaryotes genome together to do training. By doing this, we can make classifiers for some smaller categories. For example, if we combine two to three genome annotations together, we can try to train the classifier for inner membrane and outer membrane proteins separately, instead of a classifier just for membrane proteins.

However, the insufficiency of the method is, although the sensitivity and specificity are high, the PPV (positive predictive value) is low. Which means, if a gene is predicted as positive by a classifier, only small proportion of them are really positive. That is because of the fact that, among to the entire set of coding genes there are always only a small proportion of proteins are belong to a specific functional category. To improve the prediction accuracy and PPV, maybe we can try to use some other more complex model to train the classifier, like gapped k-mer (Lee et al., 2015) or some deep learning methods.

In conclusion, this research may provide some new idea for future study about how to utilize the knowledge we’ve got to do supervised machine learnings then make predictions.

References

Calvin B. Harley, Robert P. Reynolds, *Analysis of E.Coli Pormoter sequences*, Nucleic Acids Research, Volume 15, Issue 5, Pages 2343–2361 (1987).  <https://doi.org/10.1093/nar/15.5.2343>

Leslie C., Eskin E. & Noble W.S. *The Spectrum Kernel: A String Kernel for SVM Protein Classification*. Proceedings of the Pacific Symposium on Biocomputing, pages 564-575 (2002). Doi: [10.1142/9789812799623\_0053](https://www.researchgate.net/deref/http%3A%2F%2Fdx.doi.org%2F10.1142%2F9789812799623_0053)

Lee D., Gorkin D.U., Baker M. et al. *A method to predict the impact of regulatory variants from DNA sequence.* [Nat Genet.](https://www.ncbi.nlm.nih.gov/pubmed/26075791) 47(8):955-61 (2015). Doi: 10.1038/ng.3331.

Loewenstein, Y., Raimondo, D., Redfern, O.C. et al. *Protein function annotation by homology-based inference*. Genome Biol 10, 207 (2009). <https://doi.org/10.1186/gb-2009-10-2-207>

Nakae T. *Identification of the outer membrane protein of E. coli that produces transmembrane channels in reconstituted vesicle membranes.* [Biochemical and Biophysical Research Communications](https://www.sciencedirect.com/science/journal/0006291X) Volume 71, Issue 3, 11, Pages 877-884 (1976). <https://doi.org/10.1016/0006-291X(76)90913-X>

Sael, L., Chitale, M. & Kihara, D. *Structure- and sequence-based function prediction for non-homologous proteins.* J Struct Funct Genomics 13, 111–123 (2012). <https://doi.org/10.1007/s10969-012-9126-6>

Van Noort V., Snel B. & Huynen M.A. *Predicting gene function by conserved co-expression.* Trends in Genetics Volume 19, Issue 5, Pages 238-242 (2003). <https://doi.org/10.1016/S0168-9525(03)00056-8>