An SVM based Tool for the Prediction of Nitrogen Fixing Proteins

Deepa Rajan S, Rejimoan R
Department of Computer Science and Engineering
SreeChitraTirunal College of Engineering, Kerala
India

Sivakumar KC
Bioinformatics facility,
Rajiv Gandhi Centre for Biotechnology,
Thiruvanathapuram
Kerala, India

Sathish Mundayoor
Bioinformatics facility,
Rajiv Gandhi Centre for Biotechnology,
Thiruvanathapuram
Kerala, India

ABSTRACT
The symbiotic nitrogen fixing metabolic capacity is known to present in several prokaryotic bacterium across taxonomic groups. Experimental detection of nitrogen fixation in microbes requires species-specific environment, making it complex to achieve a widespread survey of this attribute. Rhizobia legume symbiosis is an attractive research field because of its importance in agriculture. Rhizobia interact with host legume plants in soil to develop root nodules, which convert atmospheric nitrogen into ammonia, a form of nitrogen used by plants as nutrients. Experimental identification of nitrogen fixing proteins (nifu) is labor- as well as time-intensive. In this work, we present a Support Vector Machine (SVM) based method for the prediction of nifu and nifu-like proteins. The SVM models were trained PSI-BLAST derived PSSM matrices. The best classifiers are based on compositional properties as well as PSSM and yield an overall accuracy of 98.16%. This work will aid rapid and rational identification of nifu, expedite the pace of experimental characterization of novel nifu proteins and enhance our knowledge about role of Rhizobia –legume interaction.

Keywords
Rhizobium, legumes, genes, SVM

1. INTRODUCTION
Biological nitrogen fixation is the major route for the conversion of atmospheric nitrogen gas (N2) to ammonia [1]. However, this process is thought be limited to a small subset of prokaryotes named diazotrophs, which have been identified in diverse taxonomic groups [2]. This biochemical pathway is only manifested when species-specific metabolic and environmental conditions are met, thus making it difficult to develop a standard screen for detection of this biological reaction [3,4]. The complications in experimentally detecting nitrogen fixation may be a reason for the relatively low number and relatively sparse distribution of known diazotrophic species.

All known diazotrophs contain at least one of the three closely related sub-types of nitrogenase: Nif, Vnf, and Anf. Despite differences in their metal content, these nitrogenase sub-types are structurally, mechanistically, and phylogenetically related. Their catalytic components include two distinct proteins: dinitrogenase (comprising the D and K component proteins) and dinitrogenase reductase (the H protein) [1,2]. The only known exception to this rule is the superoxide-dependent nitrogenase from Streptomyces thermoautotrophicus, whose protein sequence is unknown [5].

In the last few years, significant advances have been made in the functional assignment of individual gene products implicated in the biosynthesis of FeMoco in Azotobacter vinelandii [6,7,8]. The current biosynthetic system involves a conglomerate of proteins that assembles the individual apparatus, iron and sulfur, into Fe-S cluster modules for successive change into precursors of higher nuclearity, and addition of the heteroatom (Mo) and organic component (homocitrate). The creation of FeMoco is completed in a so-called scaffold protein, NifEN, and shuttled to the concluding target by cluster carrier proteins. Interestingly, the scaffold NifEN has amino acid sequence similarity to NifDK [9].

The current growth of genomic databases now including nearly 2,000 completed microbial genomes motivated us to re-evaluate the diversity of species capable of nitrogen fixation. Identification of co-occurrence of nitrogen fixing genes in species known to fix nitrogen enabled us to identify novel potential diazotrophs based on their genetic makeup. Our findings expand the expected occurrence of nitrogen fixation and the biodiversity of diazotrophs. In addition we have identified a large number of phylogenetically diverse nitrogenase-proteins that may represent ancestral forms of the enzyme and may have evolved to perform other metabolic functions.

2. MATERIALS AND METHODS

2.1 Datasets for SVM training
Different keywords like ‘nifu’, ‘symbiosis’ with the limiting filter of taxonomy as prokaryotes were used to compile a raw pool of nifu sequences from UniProtKB(http://www.uniprot.org/uniprot). Proteins with known intracellular locations, such as nucleus, cytoplasm, mitochondria, endoplasmic reticulum etc. were collected and assigned to the non-nifu set. Both the sets were filtered for hypothetical proteins and protein fragments and the redundancy were removed. Hereupon, we had two sets containing full-length and well-annotated sequences of 3224 nifu proteins and 3224 non-nifu sequences.

2.2 Benchmark dataset for testing
In order to examine the unbiased prediction efficiency of our best SVM models, we tested their performance on independent datasets not used in training or testing cycles. While one test dataset consisted of 599 nifu, the other had 310 non-nifu negative dataset.

2.3 PSSM (Position Specific Scoring Matrices)

A PSSM is a Position Specific Scoring Matrix and is a commonly used for representing the position of biological sequence. Biological sequences are converted into machine readable form by generating PSSM. This method is commonly used for predicting the sequence position. PSI-BLAST (Position Specific Iterative –BLAST) derives a Position Specific Scoring Matrix (PSSM) or profile from the multiple sequence alignment of sequences detected above a given score threshold using protein-protein BLAST. This PSSM is used to further search the database for new matches, and is updated for subsequent iterations with these newly detected sequences. Thus PSI-BLAST provides a means of detecting distant relationships between proteins. PSI-BLAST is one of the most powerful and popular homology search programs currently available. The position specific scoring matrices or profile it is used in Protein Blast and obtained amino acid substitution scores which is given separately for each position in a protein multiple sequence alignment. Alignment means extract a segment from each sequence, if sequence length is smaller than the other then add gap symbols to each segment to create equal length sequence and place one padded segment over the other. The PSSM is used to get numerical value, if the numerical value is high from the previous one then that is the better alignment from the previous ones. PSSM scores are normally positive or negative integers. Positive scores indicate that the given amino acid substitution occurs more frequently in the alignment than expected by chance. PSSMs are generated by using PSI-BLAST, which finds similar protein sequences from the query sequences and then construct PSSM from the resulting alignment [5]. The dimensionality of the PSSM is multidimensional data, but here consider only one feature of PSSM.

2.3.1 PSIBLAST Algorithm

1. Perform initial alignment with BLAST using BLOSUM 62 substitution matrix.
2. Construct a multiple alignment from hits.
3. Prepare a position specific scoring matrix (PSSM).
4. Use PSSM profile as the scoring matrix for a second BLAST (run against database).
5. Repeat steps 2-4 until convergence.

2.3.2 Constructing a Position Specific Scoring Matrix (PSSM)

Dimension of a PSSM: lq × 20, where lq is the length of the query protein.

1. Run BLAST against the database (local alignment).
2. Collect database sequence segments with E-value below threshold (default is 0.01).
3. Remove similar sequences.
   - Remove sequence segments identical to a query segment.
   - Retain one copy for any rows that are >98% identical to one another.
4. Construct the multiple alignment block M with the remaining segments (length M = lq)
   - Ignore pair wise alignment columns that involve gap characters inserted into the query.
5. For each column C:
   a. Reduce M to MC (1 · C · query length)
      - Let R be the set of sequences with a residue in C.
      - Columns of MC are columns of M with all sequences in R. In other words, MC only contains those database sequences in R. Therefore, MC contains a subset of M's columns and rows (see the figure below).
   b. Compute weights for each sequence in R
   c. Compute Pi, the background frequency of residue i over MC.
      - Compute weighted frequency fi for each residue i.
   d. Estimate the relative number of independent observations NC as the mean number of different residue including gap characters.
   e. Compute pseudo count gi for each residue i (expectation based on score gi matrix).

$$gi = \sum_j (fi \times pj) \times qij \qquad (2.2.2.1)$$

$$qij = PiPje^{\gamma u Sij} \qquad (2.2.2.2)$$

Where the target frequencies are implicit in the substitution matrix, sij is the substitution matrix score for aligning each pair of amino acids i and j, and ¸u is a constant parameter for ungapped alignments.

   f. Compute Qi as the weighted sum of fi and gi.

$$Qi = \frac{\alpha fi + \beta gi}{\alpha + \beta} \qquad (2.2.2.3)$$

$$\alpha = Nc - 1 \qquad (2.2.2.4)$$

$$B=10(\text{empirically}) \qquad (2.2.2.5)$$

2.3.3 Matrices Reported in a PSSM Output File

The PSSM can be saved to a file by using the -Q switch of blastpgp. A PSSM file contains two matrices. The first one is the regular PSSM that contains the log-odds ratios rounded down to the nearest integer. This matrix is the one that is computed in the last PSIBLAST iteration. The second matrix is the weighted observed percentages rounded down to the nearest integer (i.e., 100 × fi values).

2.3.4 Composition of Position specific Scoring Matrix (PSSM 400)

For better accuracy and to get correct position of amino acid convert the PSSM into PSSM 400 units. In PSSM 400, row contains 20 amino acids and 20 amino acids in column. Each and every element in this vector was divided by the length of sequence. The resultant matrix with 400 elements was used as input feature of SVM. In this work performance can be increased with more metrics using physiochemical properties and other amino acid compositions. But consider these properties we get more reliable results in PSSM and PSSM 400.Therefore, PSSM 400 as input feature of our machine learning technique

2.4 SVMs and SVM^{light}

First pioneered by Vapnik in 1995, SVM is a supervised machine learning method which delivers state-of-the-art performance in recognition and discrimination of cryptic patterns in complex datasets [11]. SVM is used in conjunction

with kernel functions which implicitly map input data to high dimensional non-linear feature space. SVM then constructs a hyper plane separating the positive examples from the negative ones in the new space representation. To avoid over fitting, SVM chooses the Optimal Separating Hyper plane (OSH) that maximizes the margin i.e. the minimal distance between the hyper plane and the training examples [12]. The selected data points supporting the hyper plane are called support vectors. We implemented SVM using SVMlight package (http://svmlight.joachims.org) which allows us to choose a number of parameters and kernels (e.g. linear, polynomial, radial basis function, sigmoid or any user-defined kernel).

In this study we used the RBF kernel. For detailed descriptions of SVM please refer [12]. In this work the positive class for building SVM models implies nifu proteins while the negative class signifies localization proteins. We performed training testing cycles using in-house shell and PERL scripts. We used RBF kernel to train and test our SVM models. The values of d, g and regularization parameter C were optimized on the training datasets by cross validation. The overall strategy was to choose the best parameters in a way so as to maximize accuracy along with nearly equal sensitivity and specificity, wherever possible.

To evaluate the accuracy of SVM classifiers developed in cross validation cycles, we used the following four measures:

1) Sensitivity: percentage of nifu protein sequences that are correctly predicted as nifu.

2) Specificity: percentage of non-nifu protein sequences that are correctly predicted as non-nifu.

3) Accuracy: percentage of correct predictions out of total number of predictions.

4) Matthews correlation coefficient (MCC): a measure of both sensitivity and specificity, MCC= 0 indicates completely random prediction, while MCC= 1 indicates perfect prediction.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \times 100$$

$$\text{Specificity} = \frac{TN}{TN + FP} \times 100$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

3. RESULTS

3.1 Performance of similarity-based searches

Position-Specific Iterative-Basic Local Alignment Search Tool (PSI-BLAST) is usually the first method of choice for the functional annotation of proteins. We carried out the PSI-BLAST analysis on the non-redundant positive dataset of nifu proteins in a manner like leave-one-out cross-validation (LOO CV), with the cut-off E-value (-e option of blastpgp) of 0.001 and the number of iterations as 3. Each sequence was used as the query sequence once with the rest forming the target database, thus iterating, for each sequence. Herein, no significant hits were obtained for 2998 out of 3224 sequences, which signify that homology-based searches alone are not sufficient to identify these proteins. The brief flow chart of the prediction procedure implemented in nifu classification tool is show in Figure 1.

3.2 PSSM profile based SVM classifiers

Apart from encapsulating residue composition, the PSSM profiles capture useful information about conservation of residues at crucial positions within the protein sequence, because in evolution the amino acid residues with similar physio-chemical properties tend to be highly conserved due to selective pressure. PSSM profiles have been employed for training SVMs for a legion of classification problems, like prediction of cyclins [13], nucleic acid binding residues [14], protein subcellular localization [15] etc. For the model generated with PSSM profiles normalized using the logistic function (PSSM-a), we got a maximum accuracy of 98.16%.

3.3 Performance on benchmarking datasets

Table 1 lists the performance of the three classifiers on the independent positive and negative test datasets. This was assessed at the default thresholds obtained by cross-validation studies; however for practical purposes, the higher the scores, the higher is the confidence level of prediction. The remarkably fair accuracies of the three classifiers for both the datasets demonstrate its efficiency and justify its use for practical application. The values of d, g and regularization parameter C were optimized on the training datasets by cross validation using in-house shell and PERL script. This strategy helped to find the best parameters in a way so as to maximize the accuracy along with nearly equal sensitivity and specificity in Radial and polynomial kernel functions (Table 2).

4. DISCUSSION

The positive dataset used in the study represents nifu from 29 different species with diverse taxonomic positions; however this certainly does not represent nifu from all prokaryotes. The reason for the successful performance of the models on sequences of species not included in training, gather sufficient information to create classification model based on only a small set of the training examples. Though we have tested the sensitivity of the approach on species not represented in training sequences, the true sensitivity towards extremely divergent species may only be tested when such sequences are available in future. The prediction method developed in the study can expedite the discovery of nitrogen fixing proteins and needs to be judiciously used, keeping the SVM scores as well as other complementary evidence into consideration. Thus SVM based nifu prediction system has the potential to be used for scanning nifu-like properties in proteomes. In future, availability of additional nifu sequences with a better representation of different symbiotic species and inclusion of more functional properties would further enhance the accuracy of the program.

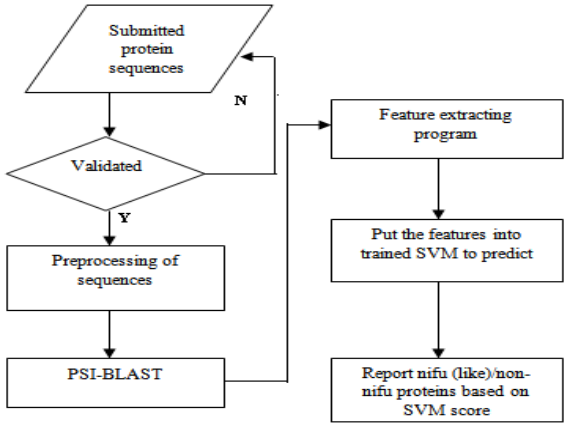
Table -1 Performance of various SVM models

Performance Parameter	Linear model	Polynomial	Radial
Accuracy	98.16%	98.16%	98.33%
Sensitivity	96.33	96.33%	96.66%
Specificity	100%	100%	100%
MCC	0.963	0.963	0.967

Table -2 Maximum accuracy of models along with nearly equal sensitivity and specificity in Radial and polynomial kernel functions by optimizing the values of C, d and g.

Performance Parameter	Polynomial (C=.70,d=2)	Radial(C=.70,g=2.5)
Accuracy	99%	98.66%
Sensitivity	98%	98%
Specificity	100%	99.33%
MCC	0.980	0.973

Figure 1: Flow chart of the algorithm implemented in nifu classification tool



5. ACKNOWLEDGMENTS

The authors acknowledge to BTISNet, Department of Biotechnology, and Government of India for the Bioinformatics Facility.

6. REFERENCES

[1] Seefeldt, L.C., Hoffman, B.M., Dean, D.R., 2009, Mechanism *of Mo-dependent nitrogenase*, Annu Rev Biochem, vol. 78, pp. 701–722.

[2] Hartmann, L.S., Barnum, S.R., Inferring the evolutionary history of Mo-dependent nitrogen fixation from phylogenetic studies of nifK and nifDK, 2010, J Mol Evol, vol. 71, pp. 70–85.

[3] O'Carroll, I.P., Dos, Santos, P.C., 2011, *Genomic analysis of nitrogen fixation*, Methods Mol Biol, vol. 766, pp. 49–65.

[4] Zehr, J.P., Jenkins, B.D., Short, S.M., Steward, G.F., 2003, *Nitrogenase gene diversity and microbial community structure: a cross-system comparison***.** Environ Microbiol, vol. 5**,** pp. 539–554.

[5] Ribbe, M., Gadkari, D., Meyer, O., 1997, N2 fixation by Streptomyces thermoautotrophicus involves a molybdenum- dinitrogenase and a manganese-superoxide oxidoreductase that couple N2 reduction to the oxidation of superoxide produced from O2 by a molybdenum-CO dehydrogenase, J Biol Chem, vol. 272 pp. 26627–26633.

[6] Rubio, L.M., Ludden, P.W., 2008, *Biosynthesis of the iron-molybdenum cofactor of nitrogenase*, 2008, Annu Rev Microbiol, vol. 62, pp. 93–111.

[7] Hu, Y., Ribbe, M.W., 2011, *Biosynthesis of Nitrogenase FeMoco*, Coord Chem Rev, vol. 255, pp. 1218–1224.

[8] Kaiser, J.T., Hu, Y., Wiig, J.A., Rees, D.C., Ribbe, M.W., 2011, *Structure of precursor-bound NifEN:a nitrogenase FeMo cofactor maturase/insertase.* Science, vol. 331, pp. 91–94.

[9] Brigle, K.E., Weiss, C.M., Newton, W.E., Dean, D.R., 1987, Products of the iron-molybdenum cofactor-specific biosynthetic genes, nifE and nifN, are structurally homologous to the products of the nitrogenase molybdenum-iron protein genes, nifH and nifK, J Bacteriol, vol. 169, pp.1547–1553.

[10] Schaffer A.A., Aravind L., Madden T.L., Shavirin S., Spouge J.L., Wolf Y.I., Koonin E.V., Altschul S.F., 2001, *Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements*, Nucleic Acids Res., vol. 15, pp. 2994-3005.

[11] Vapnik, N.V., 1998, *Statistical Learning Theory*. New York: Wiley-Interscience.

[12] Ben-Hur, A., Ong, C.S., Sonnenburg, S., Scholkopf, B, Ratsch, G., 2008, *SupportVector Machines and Kernels for Computational Biology*. PLoS Comput Biol, vol. 4 pp. e1000173

[13] M.K. Kalita, U.K. Nandal, A. Pattnaik, A. Sivalingam, G. Ramasamy, M. Kumar, G.P.S. Raghava, and D.Gupta, 2008, *CyclinPred: A SVM-Based Method for Predicting Cyclin Protein Sequences*, PLoS ONE, vol. 3, pp. e2605.

[14] Manish K., Gromiha M.M., Raghava G.P.S., 2008, *Prediction of RNA binding sites in a protein using SVM and PSSM profile*, Proteins: Structure, Function, and Bioinformatics, vol. 71, pp. 189–194.

[15] Guo J., Lin Y., 2006, TSSub: eukaryotic protein subcellular localization by extracting features from profiles, Bioinformatics, vol. 22, pp. 1784–1785.