Bioinformatics, 32(8), 2016, 1223–1225
doi: 10.1093/bioinformatics/btv714
Advance Access Publication Date: 9 December 2015
Applications Note

Sequence analysis

NBSPred: a support vector machine-based high-throughput pipeline for plant resistance protein NBSLRR prediction

Sandeep K. Kushwaha[1,2,3,*], Pallavi Chauhan[1], Katarina Hedlund[1] and Dag Ahrén[1,3]

[1]Department of Biology, Lund University, Ecology Building, Lund 22363, Sweden, [2]PlantLink, Department of Plant Protection, Swedish University of Agricultural Sciences, Alnarp, Sweden and [3]Bioinformatics Infrastructure for Life Sciences (BILS), Lund University, Lund, Sweden

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on 24 July 2015; revised on 13 November 2015; accepted on 2 December 2015

Abstract

Summary: The nucleotide binding site leucine-rich repeats (NBSLRRs) belong to one of the largest known families of disease resistance genes that encode resistance proteins (R-protein) against the pathogens of plants. Various defence mechanisms have explained the regulation of plant immunity, but still, we have limited understanding about plant defence against different pathogens. Identification of R-proteins and proteins having R-protein-like features across the genome, transcriptome and proteome would be highly useful to develop the global understanding of plant defence mechanisms, but it is laborious and time-consuming task. Therefore, we have developed a support vector machine-based high-throughput pipeline called NBSPred to differentiate NBSLRR and NBSLRR-like protein from Non-NBSLRR proteins from genome, transcriptome and protein sequences. The pipeline was tested and validated with input sequences from three dicot and two monocot plants including *Arabidopsis thaliana, Boechera stricta, Brachypodium distachyon Solanum lycopersicum* and *Zea mays.*

Availability and implementation: The NBSPred pipeline is available at http://soilecology.biol.lu.se/nbs/.

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Contact: sandeep.kushwaha@biol.lu.se

1 Introduction

R-proteins are well known as plant defence proteins against pathogens that contain different domains like nucleotide binding domain (NB-ARC), leucine-rich repeat (LRR), Toll-interleukin-like receptor (TIR), Coiled-Coiled (CC) and kinase (KIN). Five different kinds of R-protein domain association categories (Sanseverino and Ercolano, 2012) were found as majority including TIR-NBS-LRR (TNL), CC-NBS-LRR (CNL), Receptor-like kinase, Receptor-like proteins and last category Others. A number of hypotheses have been proposed for explaining the regulation of plant immunity against pathogens such as the guard hypothesis, the receptor-ligand hypothesis (Marone *et al.*, 2013; Soosaar *et al.*, 2005) and the structure-based functional hypothesis (Takken and Goverse, 2012). But several questions are still unanswered: How many R-genes need to be expressed to acquire resistance against micro-organisms acting as pathogens in plants? How many other proteins are involved directly or indirectly in the resistance process (Takken and Goverse, 2012)? How is the dynamic intracellular localization achieved, as the majority of nucleotide binding site leucine-rich repeat (NBSLRR) proteins lack signal peptides (Rafiqi *et al.*, 2009)? A limited number of known R-proteins could be one possible

reason for reduced understanding of multi-layered innate immune system of plant defence. Exploration of R-proteins and proteins having R-protein-like features could provide useful links to develop the global understanding of plant defence against different pathogens.

Currently, sequence and motif similarity, domain matching and domain association-based methods are in use, each demanding a lot of data processing and time for genome-wide identification of R-proteins (Sanseverino and Ercolano, 2012; Shang *et al.*, 2009; Tan and Wu, 2012). Disease Resistance Analysis and Gene Orthology (DRAGO) pipeline of the PRG database is also based on BLAST search and domain analysis (Sanseverino *et al.*, 2013), whereas the NLR-parser is based on MAST motif search (Steuernagel *et al.*, 2015). Prediction of R-proteins on the basis of sequence and domain similarity with a small set of reference R-genes is challenging due to the high level of diversity, as R-genes are under high selection pressure to adapt their immunity to the rapidly evolving effector genes in the pathogens (Marone *et al.*, 2013). Experimental screening of R-genes in plants would be difficult to perform at large scale. But presently, a large number of plant genomes and transcriptomes have been sequenced and assembled. Despite data availability, identification of NBSLRR sequences in sequenced plants is still limited at large scale with present available tools. Nowadays, machine-learning techniques are used much more efficiently for answering biological questions. As a solution, the NBSPred an automated pipeline have been proposed by using support vector machine (SVM) technique. NBSPred is facilitating the high-throughput identification of NBSLRR sequences from genome, transcripts and protein sequences as input.

2 Methods

R-protein and non-R-protein sequences were retrieved from the GenBank database. Redundancy removal was performed through clustering. A domain-based approach was used for the final selection of sequences in both datasets. The sequences having both NB-ARC and LRR domain together along with other additional domains like Pkinase, TIR, CC and so forth were selected in the positive dataset, whereas the sequences that had domains of positive dataset were removed from negative dataset. Six types of sequence compositional frequencies (amino acid frequency, dipeptide frequency, tripeptide frequency, multiplet frequency, charge and hydrophobicity composition) were calculated for each sequence, and a numerical feature vector was created for each sequence of positive and negative datasets (Chaudhuri *et al.*, 2011; Ramana and Gupta, 2010). SVM_learn and SVM_classify modules from the SVM^{light} package (Joachims, 1999) were used to generate an SVM classifier for NBSLRR protein prediction. Best classifiers were identified through 5-fold cross validation technique (Supplementary Sections S1 and S2). Augustus 2.7 was used in the pipeline for the annotation of plant genome (Stanke and Morgenstern, 2005). TransDecoder (Grabherr *et al.*, 2011) was used for the protein Open Reading Frame generation from transcripts. The Pfam database was used for the domain identification in predicted NBSLRR proteins. The flowchart of the pipeline is given in Figure 1.

3 Implementation

The NBSPred pipeline was hosted on a Dell PowerEdge T320 Server E5-2430 with 12 core processors of 2.2 GHz, running on CentOS 7.0 64 bits. The pipeline is freely accessible as a web interface which

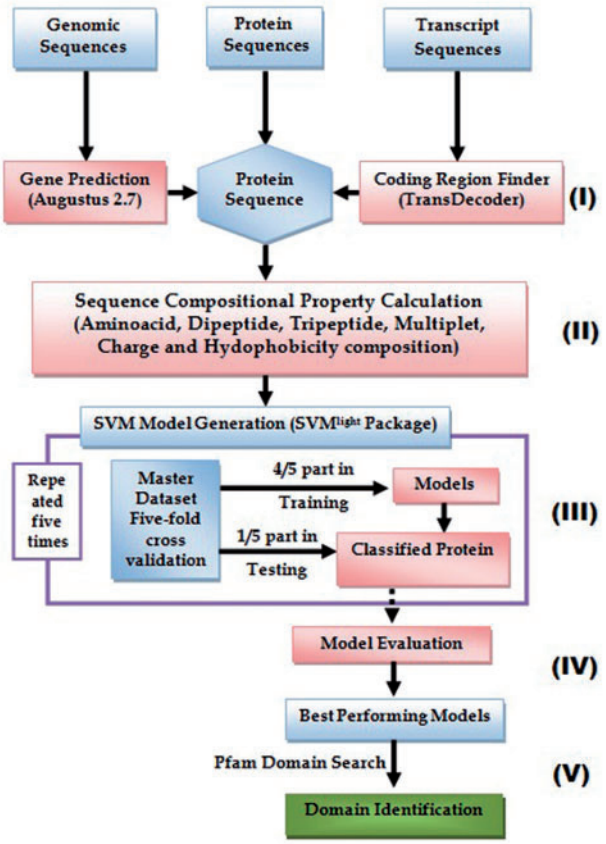

Fig. 1. Data processing workflow for the NBSPred pipeline

was developed in PHP version 5.5.10 along with other freely available academic software. Moreover, the NBSPred pipeline also provides non-interactive options for batch submission and user notification via email.

4 Results and discussion

A total of 974 sequences were involved in the training, which have NB-ARC and LRR domain together (Supplementary Table S1). Composition-based amino-acid frequencies were used for prediction after the analysis of training sequences (Supplementary Section S1.4). Five-hundred eighty-eight models were generated through variable input of different kernel function and kernel-associated parameters. For the polynomial kernel, values of d and C were increased stepwise through a combination of 1, 2, 3, $4 \ldots$ to $\ldots 9$ for the d and 10^{-5}, $10^{-6} \ldots$ to $\ldots 10^{15}$ for C. For the Radial basis function kernel, the gamma (g) was incremented stepwise $10^{-15} \ldots$ to $\ldots 10^{3}$ and parameter C from $10^{-5} \ldots$ to $\ldots 10^{15}$. The mean Matthews correlation coefficient and prediction accuracy of the best performed model, kernel type and kernel-associated values are given in Supplementary Table S2. NBSPred achieved ~99% and 83% prediction accuracy for training and independent dataset, respectively (Supplementary Tables S2 and S3). NBSPred has shown high exploration potential than NLR-parser in comparison of different datasets (Supplementary Tables S4–S9). NLRParser has detected 91.12% sequences as R-protein from independent dataset but only 49.72% sequences have class assignment. Prediction of R-protein classes (CNL and TNL) through NLRParser are varies like many sequences are predicted as CNL/TNL but all the domains does not appear in Pfam search.

Moreover, NLR-parser is not compatible for prediction from genomic assemblies. All the processed files for prediction and comparison are available at help section of NBSPred website.

4.1 High-throughput prediction and validation

As proof, we demonstrated the efficiency of the NBSPred prediction and explorative capability through screening of five plants (three dicots and two monocots) for Phytozome data repository (Goodstein *et al.*, 2012). Number of detected sequences through NBSPred for transcriptome and proteome are given, respectively, for *Boechera stricta* (722, 699), *Arabidopsisthaliana* (604, 595), *Solanum lycopersicum* (564, 577), *Zea Mays* (477, 428) and *Brachypodium distachyon* (752, 738). Three plant genomes (*A.thaliana*, *Solanum lycopersicum* and *Zea mays*) were scanned for R-protein domains from genomic assemblies through Augustus trained gene prediction models (Supplementary Tables S7–S9). NBSpred have long-term objective to provide good gene prediction models for crop plants. NBSPred-predicted sequences with high score, with or without NB-ARC and LRR domain are strong R-gene candidates which possess high level of sequence composition similarity like R-genes, but these sequences are required careful annotation and experimental validation.

Acknowledgements

The authors would like to thank Department of Biology, Lund University and PlantLink, Department of Plant Protection, Swedish University of Agricultural Sciences, Alnarp, Sweden, for providing much needed resources for carrying out this research.

Funding

This work was supported by The Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning (FORMAS) and Bioinformatics Infrastructure for Life Sciences (BILS).

Conflict of Interest: none declared.

References

Chaudhuri,R. *et al.* (2011) FungalRV: adhesin prediction and immunoinformatics portal for human fungal pathogens. *BMC Genomics*, **12**, 192.

Goodstein,D.M. *et al.* (2012) Phytozome: a comparative platform for green plant genomics,. *Nucleic Acids Res.*, **40**, D1178–D1186.

Grabherr,M.G. *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotech.*, **29**, 644–652.

Joachims,T. (1999) Making large-scale support vector machine learning practical. In: Schölkopf,B. *et al.* (eds) *Advances in Kernel Methods*. MIT Press, Cambridge, MA, USA, pp. 169–184.

Marone,D. *et al.* (2013) Plant nucleotide binding site–leucine-rich repeat (NBS-LRR) genes: active guardians in host defense responses. *Int. J. Mol. Sci.*, **14**, 7302–7326.

Rafiqi,M. *et al.* (2009) In the trenches of plant pathogen recognition: role of NB-LRR proteins. *Semin. Cell Dev. Biol.*, **20**, 1017–1024.

Ramana,J. and Gupta,D. (2010) FaaPred: a SVM-based prediction method for fungal adhesins and adhesin-like proteins. *PLoS One*, **5**, e9695

Sanseverino,W. and Ercolano,M.R. (2012) In silico approach to predict candidate R proteins and to define their domain architecture. *BMC Res. Notes*, **5**, 678

Sanseverino,W. *et al.* (2013) PRGdb 2.0: towards a community-based database model for the analysis of R-genes in plants. *Nucleic Acids Res.*, **41**, D1167–D1171.

Shang,J. *et al.* (2009) Identification of a new rice blast resistance gene, Pid3, by genomewide comparison of paired nucleotide-binding site—leucine-rich repeat genes and their pseudogene alleles between the two sequenced rice genomes. *Genetics*, **182**, 1303–1311.

Soosaar,J.L.M. *et al.* (2005) Mechanisms of plant resistance to viruses. *Nat. Rev. Microbiol.*, **3**, 789–798.

Stanke,M. and Morgenstern,B. (2005) AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.*, **33**, W465–W467.

Steuernagel,B. *et al.* (2015) NLR-parser: rapid annotation of plant NLR complements. *Bioinformatics*, **31**, 1665–1667.

Takken,F.L.W. and Goverse,A. (2012) How to build a pathogen detector: structural basis of NB-LRR function. *Curr. Opin. Plant Biol.*, **15**, 375–384.

Tan,S. and Wu,S. (2012) Genome wide analysis of nucleotide-binding site disease resistance genes in Brachypodium distachyon. *Comp. Funct. Genomics*, **2012**, 12,