Functional annotation of creeping bentgrass protein sequences based on convolutional neural network

Han-Yu Jiang
Nanjing Normal University
Jun He (junhe@njnu.edu.cn)
Nanjing Normal University

Research Article

Keywords: Functional annotation, protein sequences, creeping bentgrass, convolutional neural network

Posted Date: March 5th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-268906/v1

Functional annotation of creeping bentgrass protein sequences based on convolutional neural network

Han-Yu Jiang[a,b], Jun He[a*]

[a] Department of Physics and Institute of Theoretical Physics, Nanjing Normal University, Nanjing, Jiangsu 210097, China

[b] Sino-U.S. Center for Grazingland Ecosystem Sustainability/Pratacultural Engineering Laboratory of Gansu Province/ Key Laboratory of Grassland Ecosystem, Ministry of Education/College of Pratacultural Science, Gansu Agricultural University, Lanzhou, Gansu 730070, China

Abstract

Background:

Creeping bentgrass (Agrostis soionifera) is a perennial grass of Gramineae, belonging to cold season turfgrass, but has shallow adventitious roots, poor disease-resistance. Little is known about the ISR mechanism of turfgrass and the signal transduction involved in disease-resistance induction, especially the function of a large number of disease-resistance related proteins are urgent to be explored.

Results:

In this work, the protein sequences of creeping bentgrass were measured and annotated by a functional prediction model based on convolutional neural network. Creeping bentgrass seedlings were grown with BDO treatment, and the ISR response was induced by infecting *Rhizoctonia solani*. We preformed the transcriptome analysis by Illumina Sequencing and high-quality unigenes were obtained. A minority of assembled unigenes were functionally annotated according to the database alignment while a large part of the obtained amino acid sequences was left non-annotated. To treat the non-annotated sequences, a prediction model was established by training the data set from GO families in three domains to acquire good performance, especially the higher false positive control rate. With such model, we analyzed the non-annotated protein sequences of creeping bentgrass transcriptome, and annotated the disease-resistance response and signal transduction related proteins.

Conclusions:

The results provide good candidates of the proteins with certain functions. With the results in this work, the waste of transcriptome sequencing data of creeping bentgrass can be avoided, and research time and labor for the analysis of ISR characteristics of creeping bentgrass will be saved in further research. It also provides reference for the sequence analysis of turfgrass disease-resistance research.

Keyword: Functional annotation, protein sequences, creeping bentgrass, convolutional neural network

* Corresponding author, email: junhe@njnu.edu.cn

1. Introduction

Creeping bentgrass (*Agrostis soionifera*) is a perennial grass of Gramineae, belonging to cold season turfgrass. Due to its excellent characteristics, such as, strong adaptability, good ornamental, it is preferred grass species in golf course, lawn tennis court, courtyard, park and other green areas. However, creeping bentgrass has shallow adventitious roots, poor disease-resistance. For example, it is susceptible to coin spot and brown spot. The innate immunity can be induced in plant, which relies on a surprisingly complex response mechanism to recognize and counteract different invaders. The induced physical and chemical barriers are activated to effectively combat invasion by microbial pathogens, as well as inducible defensive mechanisms upon attack [1,2]. Among them, the induced systemic resistance (ISR) is often activated by plant growth promoting bacteria in soil rhizosphere, and has broad-spectrum resistance to bacteria, fungi and pathogens [3,4].

Butanediol (BDO) is a new type of disease resistance-inducing factor, which provides durable disease resistance. ISR produced by BDO effectively inhibits grass leaf diseases [5]. By the roots application of *Pseudomonas fluoresces* WCS417r, the resistance of plants to *Pst DC3000* could not be induced, and there was no ISR system response [6]. Ethylene (ET) signaling pathway was closely related to the establishment of ISR mechanism. At present, several resistance genes relevant to ISR mechanism have been found, and the expression of these genes is often associated with the signal process conducted by ET / JA（Jasmonic acid）[7]. Studies have shown that many resistance proteins enter the nucleus to activate the immune response and triggers the signal transduction pathway, including resistance signal activation, transcription factor regulation and hormone signal pathway activation [8]. For instance, a number of preliminary proteome analyses in rice successfully identified some known pathogenesis-related proteins that accumulate abundantly after JA treatment or inoculation by the pathogenic fungus *M. grisea* [9,10]. Oh et al. [11] analyzed the secreted protein encoding the lipase with antimicrobial activity in Arabidopsis. However, except for a few preliminary studies, proteome-based gene identification approaches have not yet fulfilled their promise for discovery of new defense genes. One major reason is that many proteins identified and analyzed involving in signaling processes are below the threshold of detection [12]. At present, little is known about the ISR mechanism of turfgrass and the signal transduction involved in disease-resistance induction, especially the function of a large number of disease-resistance related proteins are urgent to be explored.

In our previous work, BDO was used to induce ISR resistance in creeping bentgrass infected with *Rhizoctonia solani*. We laid a foundation of creeping bentgrass genetic research by the transcriptome analysis and analyzed ethylene-dependent signal transduction pathways involved in ISR mechanisms, and compared a few differential expressions of ISR resistance genes and ET/JA signal transduction pathway node genes [13]. The research on molecular biology of creeping bentgrass is still in blank and scarcely reported in the literature. Hence, there are a large number of protein sequences,

which were not aligned in NR (Non -Redundant Protein Sequence Database), Swissprot library, or aligned but not annotated. After predicting ORF by using software *estscan*, the nucleic acid and amino acid sequences encoded by these genes were obtained. However, the functional annotation of proteins still needs to be further studied. It is very helpful for further study of the disease resistance of creeping bentgrass if we can annotate some of the protein sequences correctly even not exhaustively.

The function of protein is usually analyzed and annotated by biochemical experiments, which are time- and labor-consuming. At present, the number of protein sequences in UniPort database exceeds 100 million, and still increases rapidly [14]. The traditional methods are not enough to make up the increasing gap between the requirement and the speed of protein annotation by experimental means [15]. Therefore, the protein function prediction methods, such as machine learning, were proposed and widely adopted in the research [16]. However, traditional computational methods have disadvantages such as high false positive rate and low accuracy. In the recent year, the deep learning technology becomes an important method in the field of protein research [17,18]. Using deep learning technology, such as convolutional neural network (CNN), we can extract protein features, and build accurate and stable function prediction model, which may solve the shortcomings of traditional methods [19].

In the current work, with the deep learning algorithm, we performed following analysis about the data of protein sequences, (1) Based on CNN and protein binary encoding representation strategy, a functional prediction model was established and adjusted to achieve a high false positive control rate with the annotated protein data from some Gene Ontology (GO) families, which were collected from the Uniport database; (2) The established model was applied to the non-annotated part of creeping bentgrass transcriptome sequences that we measured to predict protein function and obtain functional classification; (3) The prediction model was further applied to functional annotation of disease-resistance and signal transduction related proteins of creeping bentgrass transcriptome sequences. Our work provides comprehensive analysis for functional annotation of new proteins. The annotation results supplement and improve the sequence data analysis of creeping bentgrass transcriptome, avoid waste of sequencing data. The research lays foundation for further mining ISR response proteins of creeping bentgrass and exploring the disease-resistance mechanism of turfgrass.

2. Materials and methods

2.1 Plant growth conditions and production of sequencing libraries

Seeds from creeping bentgrass ‘PennA-4’ (Chinese Academy of Agricultural Sciences) were grown by modified method of Kroes et al [20,13]. The surface of seedswas disinfected with 70% ethanol for 1 min, disinfected with 15% sodium hypochlorite for 5min, cleaned with sterile water for 10 min, and finally dried with filter paper. Seeds were sown in 50-mL culture flask with 10 mL of MS medium containing

100 μmol L^{-1} of BDO. About 20 seedlings per flask were cultured in a growth chamber at 22 ℃ under 100 μ$Em^{-2}s^{-1}$ light. The experimental materials were seedlings cultured for twelve-day-old under the above conditions. *Rhizoctonia solani* (#3.2888 from China General Microbiological Culture Collection Center) in PDO liquid medium (potato 200 $g{\cdot}L^{-1}$, glucose 20 $g{\cdot}L^{-1}$) was shaking culture for 2-3 day with 120 $r{\cdot}min^{-1}$ at 25℃. Concentration of the bacterial sample was a final OD_{340} of 0.8. Roots of seedlings were directly sprayed with 2 mL of the bacterial fluid. The brown blotch symptoms of creeping bentgrass seedlings were observed, and the mycelium began to grow after 3–5-day post-inoculation. After 24, 48 and 72 h post-inoculation, the seedlings with different treatment were removed, and the leaves were cut. The transcriptome of the treated materials was tested and analyzed, and all the samples were mixed and spliced (Eukaryotic Non-reference Transcriptome).

Sequencing libraries were produced by NEBNext UltraTM RNA Library Prep Kit (NEB, San Diego, CA, USA), and each sample attributes sequences for index codes. The data read by RNA-seq were uploaded on the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) under accession number SRR5658390.

2.2 Establishing of protein function prediction model

2.2.1 Constructing the data sets of training and testing

At first, we should construct database for establishing the prediction model under CNN frame. We considered some GO families from three GO domains. Some of the protein sequences of creeping bentgrass can be aligned in NR, Swissprot library. We chose the GO families of these protein sequences, i.e., 7 families in cellar component (CC) domain, 10 families in molecular function (MF) domain, 12 families in biological process (BP) domain. The annotated protein sequences with these GO IDs were collected from the UniPort database (see Fig. 1).

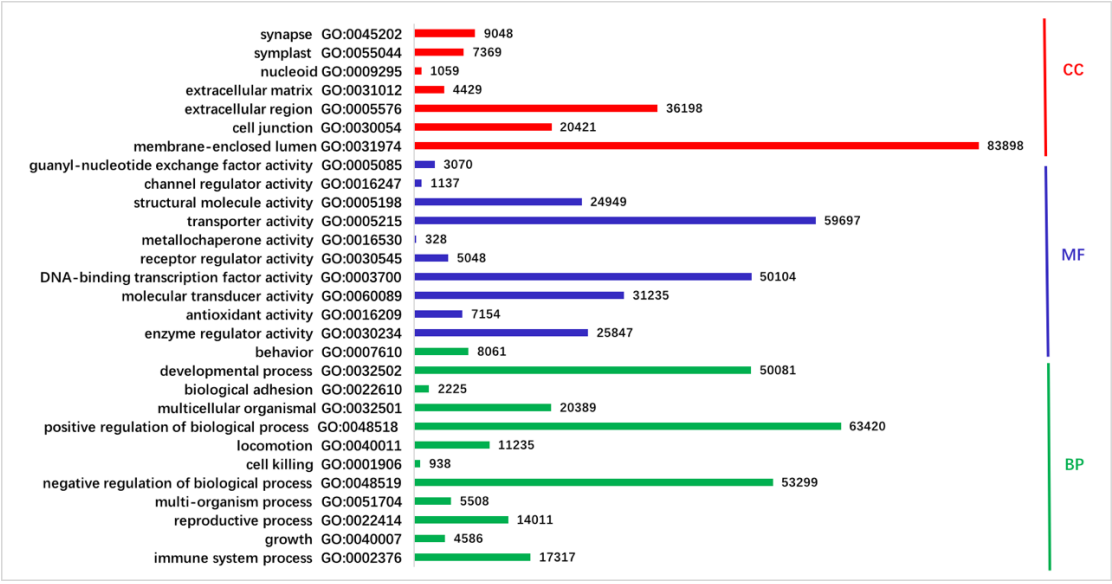

Fig. 1. The numbers of protein sequences in every GO IDs used in training.

With the annotated protein sequences collected, we constructed positive and

negative data for training by adopting binary classification. For a GO family studied, we considered the proteins in this family as positive data. The negative data were selected from the left GO families after removing the repeated sequences. To avoid overemphasize one of the left GO families, for a GO family studied with N sequences, we selected the sequences from the left GO families in order until 3N sequences were selected. A data set with $4N$ sequences was obtained. After shuffling, we used 60% of $4N$ proteins as training set, 20% as testing set, and 20% as valuation set. Here, we adopt imbalance binary classification to emphasize the negative data. Such treatment can improve the false positive control rate that we focused on, but lower sensitivity, which is less important in the present work.

2.2.2 Prediction model based on CNN

In the current work, we adopted a deep learning algorithm, convolutional neural network, to analyze the protein sequences. The protein is expressed as a one-dimensional sequence of amino acids, which is quite analogous to the sentence classification. We chose an explicit CNN frame, textCNN [21], which has been successfully applied to analyze the text, and to study the proteins [19]. The model was implemented with the Tensorflow3 library and the python programming language with some modifications to get best performance for the data sets considered in the current work. The binary cross-entropy loss function was adopted in all models training, and the Adam [22] optimizer with default parameters was used for the optimization during back-propagation. The weight parameters were initialized with the He initialization method [23], and biases were initialized to zero.

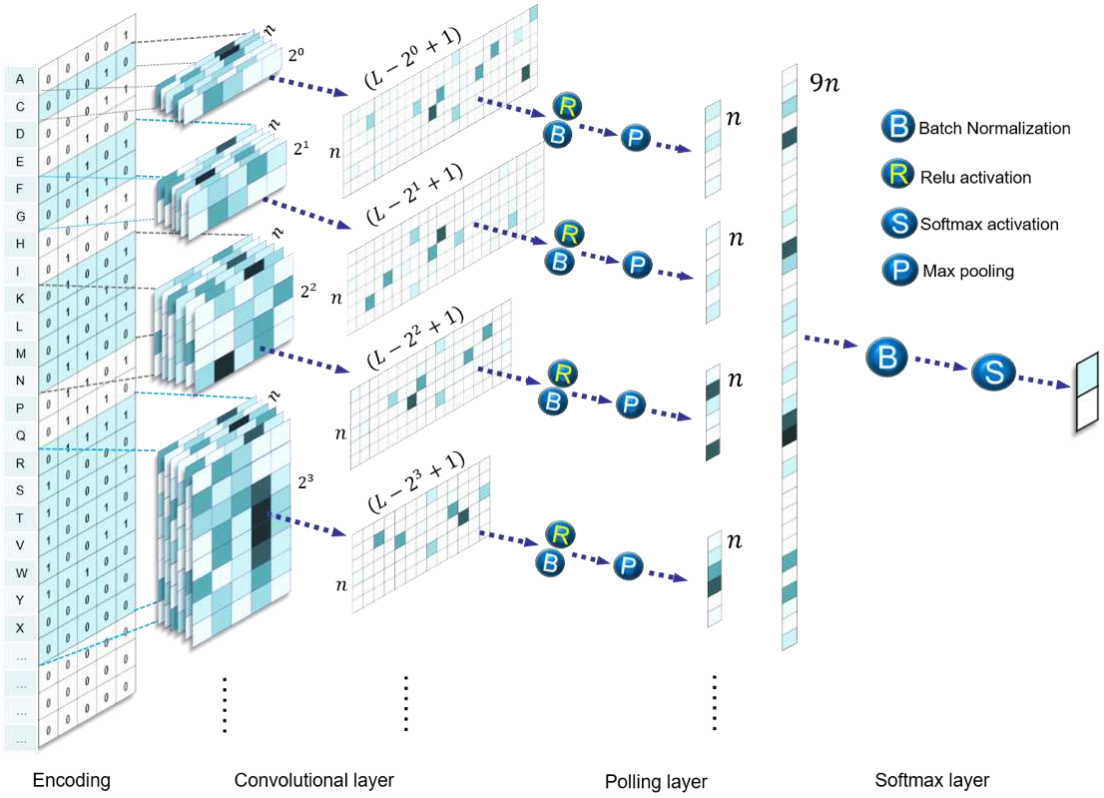

Fig. 2. The workflow of CNN adopted in the current work.

For a protein sequence, we need first encode the amino acids into binary vectors. Because only about twenty amino acids were discovered, we encoded an amino acid to a 5-bit binary vector as shown in Fig. 2. For example, the alanine is encoded as [0,0,0,0,1]. In the current work we did not distinguish the rare and undetermined amino acids, and encoded them all as [0,0,0,0,0]. The lengths of protein sequences are different while the CNN requires a fixed length. We considered the proteins of sequence length less than $L = 800$ amino acids, which constitute the majority (>95%) of the protein sequences in the studied GO families and 100% of non-annotated sequences. For the protein sequences less than 800 amino acids, the left positions were complemented by binary vector [0,0,0,0,0]. With such encoding, a protein sequence was converted to a $L \times 5$ matrix.

After encoding, the convolution layer was adopted to extract the information from the digitized protein sequences as shown in Fig. 2. To obtain the information in different length levels, convolution kernels with different sizes were adopted. In the current work, we chose $n = 120$ convolution kernels with size as $2^k \times 5$ with $k = 0, 1, 2, \ldots, 8$. After convolution, nine $(L - 2^k + 1) \times n$ arrays were obtained as

$$a_{ij}^k = \sum_{m=1}^{2^k} \sum_{l=1}^{5} \left(X_{(m+i-1)l} * W_{ml}^j \right) + b_i^j$$

where j is for the number of kernels with size as $2^k \times 5$. After batch normalization, the ReLu activation was applied. The max pooling was adopted by selecting the maximum in $(L - 2^k + 1)$ elements of a_{ij}^k for certain k and j. Then, the k vectors with same size n were concatenated to a vector with size $kn = 960$. Different from other work [19], fully connected layer was not applied here because it is found not helpful to improve the results and makes the model hard to converge. Instead, after batch normalization, we adopted the softmax activation directly to provide the classification probability.

3. Results and discussion

3.1 Model's performance with GO families in three domains

When establishing the prediction model above, we trained the model by the protein sequences collected from the Uniport database in 29 GO families, which numbers were shown in Fig. 1. The model and parameters were adjusted to obtain the best performance of false positive control rate because in the current work we want to establish a prediction model to select proteins with certain function correctly but not exhaustively. To evaluate the performance of the model, we introduce five widely-used measurements, sensitivity (SE), specificity (SP), precision (PR), accuracy (AC) and Matthews correlation coefficient (MCC). The explicit values of the measurements can be found in the Supplement 1. In Fig. 3, we presented the violin plots to show the overall picture of the results for three domains.

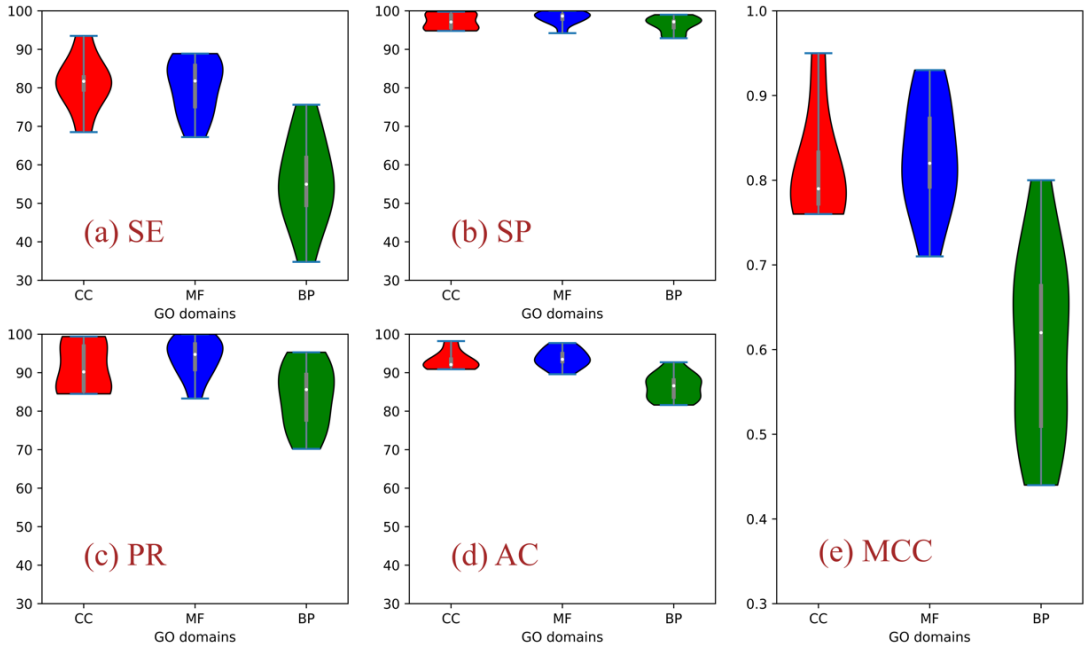

Fig. 3. Violin plots for the performance of prediction model in three domains.

Accessed by five measurements, the performance of our prediction model for the GO families in MF domain is comparable to the model in Ref. [19] where the studied GO families are also in MF domain. The model works well also for the GO families in CC domains. As expected, the functions in BP domain are more complex and harder to be learned by the prediction model, which is reflected by smaller MCC values for BP domain than other two domains. In current studies about disease-resistance response and signal transduction related proteins, it is more important to ensure the correctness of functional annotation of protein. One can find that the SP values for three domains are quite well (Fig. 3b), which satisfies our requirement.

3.2 Screening of non-annotated sequences

The non-annotated sequences used in our study referred to the sequences that was not aligned with NR and Swissprot library, or the sequences that was aligned but no predicted results. The ORF of these sequences was predicted by using estscan software, so as to obtain the amino acid sequences in the Supplement 2. The total number of transcripts was 338,453, and the total number of unigenes was 208,672. According to the database alignment and estscan prediction, 118,856 amino acid sequences were not annotated, which accounted for 56.9% of the total number of unigenes. The exploration of this part of protein function has great significance for further understanding the ISR mechanism of creeping bentgrass.

3.3 Functional annotation of non-annotated sequences by GO families in three domains

In the above, we established the prediction model with good performance, especially for the false positive control rate, by training the GO families in three domains. With such model, we can analyze non-annotated protein sequences of

creeping bentgrass and classify them into the GO families considered above. The protein sequences with the function of certain GO ID selected by prediction model were provided in the Supplement 3. In Table 1, the number of the proteins belonging to a GO family and the ratio of the number to the total number of the protein sequences are listed.

Table 1

The number and ratio of the proteins belonging to a GO family predicted from non-annotated protein sequences of creeping bentgrass.

	CC			MF			BP	
GO ID	Num	%	GO ID	Num	%	GO ID	Num	%
GO:0045202	10095	8.49	GO:0005085	1026	0.86	GO:0007610	1410	1.19
GO:0055044	1177	0.99	GO:0016247	24716	20.79	GO:0032502	4265	3.59
GO:0009295	5087	4.28	GO:0005198	9151	7.70	GO:0022610	1386	1.17
GO:0031012	1555	1.31	GO:0005215	2036	1.71	GO:0032501	8520	7.17
GO:0005576	43640	36.72	GO:0016530	6280	5.28	GO:0048518	2405	2.02
GO:0030054	7902	6.65	GO:0030545	10910	9.18	GO:0040011	4423	3.72
GO:0031974	25094	21.11	GO:0003700	4277	3.60	GO:0001906	31596	26.58
			GO:0060089	1628	1.37	GO:0048519	29664	24.96
			GO:0016209	11287	9.50	GO:0051704	10222	8.60
			GO:0030234	14411	12.12	GO:0022414	2495	2.10
						GO:0040007	2309	1.94
						GO:0002376	4514	3.80

In the CC domain, the number for GO: 0005576 (‘extracellular region’) was the largest, accounting for 36.72, while the number for GO:0055044 (‘symplast’) was the smallest, only 0.99%. In the MF domain, the number of proteins belonging to family GO:0016247 with function ‘channel regulator activity’ was the largest in the transcriptome of creeping bentgrass non-annotated protein sequences, accounting for 20.79%. The number for family GO: 0005085 (‘guanyl-nucleotide exchange factor activity’) is smallest, only 0.86%. From Ancestor Chart, the function ‘antioxidant activity’ (DO: 0016209) is a part of ‘cellular response to stimulus’ (GO: 0051716), which is closely related to the disease-resistance of creeping bentgrass. Its predicted number accounts for 9.50%. In the BP domain, family GO: 0001906 (‘Cell killing’) has the largest number, accounting for 26.58%, followed by DO:0048519 (‘negative regulation of biological process’), accounting for 24.96%. 3.8% of non-annotated protein sequences belongs to the family GO:0002376 (‘immune system process’).

3.4 Functional annotation of non-annotated sequences by GO families relevant to the disease-resistance response and signal transduction

In the above, the established prediction model was applied to select the protein sequences from non-annotated sequences belonging to the GO families chosen to train the model. In this subsection, we focused on 13 GO families with functions relevant to stimulus response and signal transduction related proteins. The performance of the prediction model for protein sequences of these GO families, which were also collected from the UniPort database, is presented in Table 2. It is well known that the disease-resistance response and signal transduction process belong to BP domains. One can expect that the results are analogous to the results in the BP domains in Fig. 1B. The SP values are considerable large, and spans from 96 up to 100%, which satisfies high false positive control rate required in the current work.

Table 2
The performance of prediction model and the number of the proteins for disease-resistance and signal transduction related GO families. The GO IDs, terms, and numbers N_{GO} of protein sequences in the GO families collected from the UniPort database are listed in the first to third columns. The measurements are listed in the fourth to eighth columns. The numbers of the proteins belonging to a GO family predicted from non-annotated protein sequences of creeping bentgrass are listed in last column.

DO ID	Function	N_{GO}	SE%	SP%	PR%	AC%	MCC	Num
GO:0009968	negative regulation of signal transduction	1730	42.3	97.3	84.4	83.0	0.52	148
GO:0032102	negative regulation of response to external stimulus	542	32.7	99.4	94.9	82.0	0.49	11555
GO:0044092	negative regulation of molecular function	1944	38.4	97.6	83.3	83.7	0.49	47437
GO:0032101	regulation of response to external stimulus	1486	41.9	98.6	90.3	84.8	0.55	792
GO:0002682	regulation of immune system process	2368	44.0	96.3	79.8	83.1	0.51	756
GO:0009607	response to biotic stimulus	3184	35.2	98.1	86.6	81.8	0.48	24287
GO:0006955	immune response	3214	34.7	97.7	83.1	82.4	0.46	1041
GO:0009719	response to endogenous stimulus	2334	44.8	96.8	81.4	84.6	0.53	1035
GO:0048585	negative regulation of response to stimulus	2341	35.9	97.4	83.7	80.9	0.46	402
GO:0002764	immune response-regulating signaling pathway	611	47.7	96.9	84.7	84.0	0.55	1260
GO:0044093	positive regulation of molecular function	2337	37.9	97.3	81.3	83.0	0.48	786
GO:0051606	detection of stimulus	913	75.7	99.6	98.5	94.1	0.83	12
GO:0080135	regulation of cellular response to stress	3819	68.4	96.4	85.7	89.8	0.70	3022

The explicit results of the protein sequences annotated are provided in the Supplement 4. In Table 2, we listed the number of protein sequences with certain function. The number of protein sequences with function of 'negative regulation of molecular function' (GO:0044092) is the largest, accounting for 47,437. In Ancestor Chart, 'negative regulation of molecular function' is the biological regulation process. There were 24,287 annotated proteins with function of 'response to biological stimulus' (GO:0009607), 11,555 annotated proteins with function of 'negative regulation of response to external stimulus' (GO:0032102), and 3022 annotated proteins with function of 'regulation of cellular response to stress' (GO:0080135). These protein functions are closely related to the disease-resistance response of creeping bentgrass, and also reflect the positive response of creeping bentgrass to external stimulation after being infected by *Rhizoctonia solani* by inducing the expression of a large number of disease-resistance related proteins. Besides, there were 1260 annotation proteins with function of 'immune response regulating signaling pathway' (GO:0002764) and 148 annotation proteins with function of 'negative regulation of signal transduction proteins' (GO:0009968), both of which were closely related to signal transduction process. The annotation of the above protein functions is of the great significance for further explore the disease-resistance and signal transduction of creeping bentgrass.

4. Conclusions

In this work, creeping bentgrass seedlings were grown with BDO treatment, and induced the ISR response by infecting *Rhizoctonia solani*. We preformed the

transcriptome analysis and obtained the high-quality unigenes. Assembled unigenes were functionally annotated according to the database alignment. However, a large part of the obtained amino acid sequences was not annotated. The exploration of this part of protein function has great significance for further understanding the ISR mechanism of creeping bentgrass. A functional prediction model was established based on convolutional neural network to annotate the disease-resistance response and signal transduction related proteins of the creeping bentgrass transcriptome. The model was trained by the data sets collected from UniPort database in GO families in three domains with emphasizing high false positive control rate. The established prediction model performs well for the GO families in three domains, especially high SP.

With such model, the non-annotated protein sequences of the creeping bentgrass were analyzed. The transcripts sequences were annotated as different protein functions, which mainly involve 'response to biological stimulus', 'negative regulation of response to external stimulus', 'negative regulation of molecular function', 'regulation of cellular response to stress' and 'immune response regulating signaling'. These protein molecules play different roles in the disease-resistance process of creeping bentgrass. The results provide good candidates of the proteins with certain functions. In further research, these suggested protein sequences can be studied by molecular biology technology. Primers were designed to amplify the disease-resistance related sequences, then transformed and induced gene expression, so as to further analyze the disease-resistance characteristics of the expressed products and verify the accuracy of the predicted protein function. Overall, with the results in this work the waste of transcriptome sequencing data of creeping bentgrass can be avoided, and the experiment consumption of time and labor can be saved. A new thought is promoted for the analysis of ISR characteristics of creeping bentgrass. Our findings also provide reference for the sequence analysis of turfgrass disease-resistance research.

Declarations

Ethics approval and consent to participate

The plant material used in this study was creeping bentgrass, which was gowned in the Laboratory of Gansu Agricultural University, Lanzhou, China, and no permits are required for the collection of plant samples. This study did not require ethical approval or consent as did not involve any endangered or protected species.

Consent for publication

Not applicable.

Availability of data and material

All data generated or analysed during this study are included in this published article and its supplementary information files. The code and relavant data downloaded from UniPort database are available from the corresponding author on reasonable request.

Competing interest

The authors declare no conflict of interest.

Funding

This project is partially supported by the National Natural Science Foundation of China

under Grant No. 11675228 and China postdoctoral Science Foundation under Grant No. 2015M572662XB.

Author contributions

Han-Yu Jiang and Jun He contributed to the conception of the study; Han-Yu Jiang performed the experiment to obtain the protein sequences of the creeping bentgrass; Jun He contributed significantly to establish the prediction model; Han-Yu Jiang and Jun He performed the data analyses and wrote the manuscript.

Acknowledgements

Not Applicable.

Declaration of competing interest

The authors declare no conflict of interest.

References

[1] D. Walters, A. Newton, G. Lyon, Induced resistance for plant defence. A sustainable approach to crop protection, Oxford: Blackwell Publishing (2007)

[2] C. X. Zhao, H. Y. Jiang, W. K. Dong, H. Chen, Y. X. Fanf, L. P. Xie, H. L. Ma. Effects of composite exogenous hormone application on induction of systemic resistance to Rhizoctonia solani in creeping bentgrass, Acta Prataculturae Sinica 27 (2018) 120-130.

[3] P.A.H.M. Bakker, R.V. Vanpeer, Suppression of soil-borne plant pathogens by fluorescent pseudomonads: Mechanisms and prospects. In biotic interactions and soil-borne diseases; A.B.R. Beemster, G.J.Bollen, Eds.; Elsevier Scientific: Amsterdam, the Netherlands (1991) 217–230.

[4] L.C. VanLoon, P.A.H.M. Bakker, Signaling in rhizobacteria-plant interactions. In root ecology, J. DeKroon, E.J.W. Visser, Eds. Springer: Berlin, Germany (2003) 287–330.

[5] A.M. Cortes-Barco, T. Hsiang, P.H. Goodwin, Induced systemic resistance against three foliar diseases of agrostis stolonifera by (2R, 3R)-Butanediol or an isoparaffin mixture, Ann. of Applied Biol. Plant Pathol. 157 (2010) 179-189.

[6] M. Knoester, C.M.J. Pieterse, J.F. Bol, L.C. Van Loon, Systemic resistance in Arabidopsis induced by rhizobacteria requires ethylene-dependent signaling at the site of application, Mol. Plant-Microbe In. 12 (1999) 720-727.

[7] H. Takahashi, T. Ishihara, S. Hase, A. Chiba, K. Nakaho, T. Arie, T. Teraoka, M. Iwata, T. Tugane, D. Shibata, S. Takenaka, Beta-cyanolanina synthase as a molecular marker for induced resistance by fungal glycoprotein elicitor and commercial plant activators, Phytopathology 96 (2006) 908-916.

[8] L. D. Bruyne, M. Höfte, D. D. Vleesschauwer. Connecting growth and defense: the emerging roles of brassinosteroids and gibberellins in plant innate immunity, Mol. Plant 7 (2014) 943–959.

[9] S. Kim, I.P. Ahn, Y.H. Lee, Analysis of genes expressed during rice - Magnapor the grisea interactions, Mol. Plant Microbe Interact. 14 (2001) 1340-1346.

[10] S.T. Kim, K.S. Cho, S. Yu, S.G. Kim, J.C. Hong, C.D. Han, D.W. Bac, M.H. Nam, K.Y. Kang, Proteomic analysis of differentially expressed proteins induced by rice blast fungus and clicitor in suspension-cultured rice cells, Proteomics 3 (2003) 2368-2378.

[11] I.S. Oh, A.R. Park, M.S. Bae, S.J. Kwon, Y.S. Kim, J.E. Lee, N.Y. Kang, S. Lee, H. Cheong, O.K. Park, Secretome analysis reveals an Arabidopsis lipase involved in defence against Alternaria brassicicola, The Plant Cell 17 (2005) 2832-2847.

[12] L.F. Thatcher，J.P. Anderson, K.B. Singh, Plant defence responses: what have we learnt from Arabidopsis? Funct. Plant Biol. 32 (2005) 1-19.

[13] H.Y. Jiang, J.L. Zhang, J.W. Yang, H.L. Ma, Transcript profiling and gene identification involved in the ethylene signal transduction pathways of creeping bentgrass (*Agrostis stolonifera)* during ISR response induced by butanediol, Molecules 13 (2018) 706.

[14] The Uniprot C. UniProt: the Universal Protein Knowledgebase, Nucleic Acids Res. (2017) 45(D1): D158-D69.

[15] M. Frasca, N. C. Bianchi, Multitask protein function prediction through task dissimilarity, Ieee Acm. T. Comput. Bi. 16 (2019) 1550-60.

[16] Y.X. Jiang，T.O. Ronnen，T. R. Wyatt Clark， B. Asma，R. Predrag, An expanded evaluation of protein function prediction methods shows an improvement in accuracy, Genome Biol. 17 (2016) 184.

[17] X. Pan, Y. Yang, C. Q. Xia, A.H. Mirza, H.B. Shen, Recent methodology progress of deep learning for RNA-protein interaction prediction, Wiley Interdiscip Rev. Rna 10 (2019) 1544.
[18] L. J. Colwell, Statistical and machine learning approaches to predicting protein-ligand interactions, Curr. Opin. Struct. Biol. 49 (2018) 123-8.
[19] J. Hong, Y. Luo, Y. Zhang, J.Ying, W. Xue, T. Xie, L. Tao, F. Zhu, Protein functional annotation of simultaneously improved stability, accuracy and false discovery rate achieved by a sequence-based deep learning. Brief. Bioinform. 21 (2019)1437-1447
[20] G.M.L.W Kroes, E. Sommers, Two in vitro assays to evaluate resistance in Linum usitatissimum to Fusarium wilt disease. Eur. J. Plant Pathol. 104 (1998) 561–568.
[21] Y. Kim, Convolutional neural networks for sentence classification, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2014) 1746–1751, arXiv:1408.5882
[22] Hamm CA, Wang CJ, Savic LJ, et al. Deep learning for liver tumor diagnosis part I: development of a convolutional neural network classifier for multi-phasic MRI. Eur Radiol, 29 (2019) 3338–47.
[23] K. He, X. Zhang, S. Ren, and J. Sun, Delving deep into rectifiers: surpassing human-level performance on imageNet classification, Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV) (ICCV '15). IEEE Computer Society, USA (2015) 1026–1034, arXiv: 1502. 01852

Figures

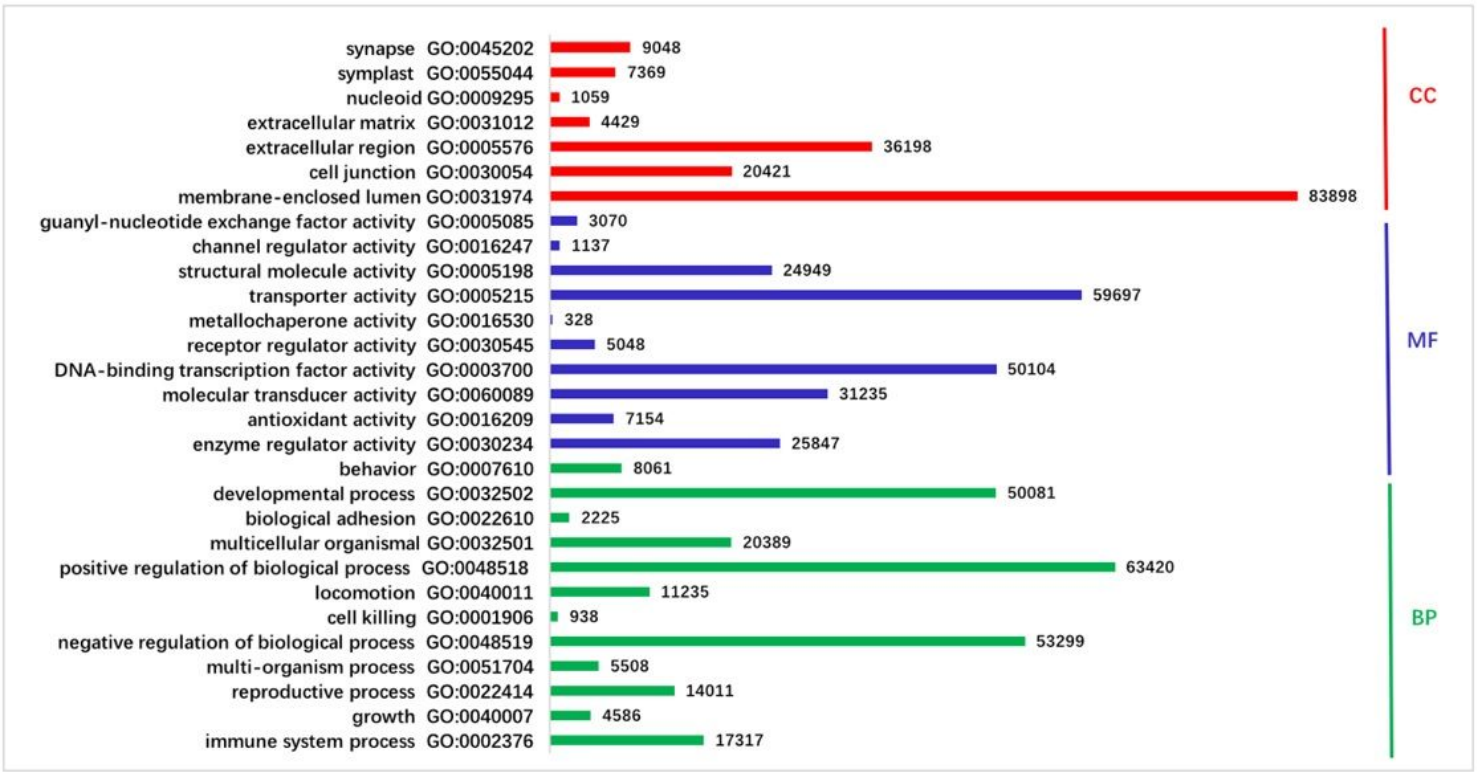

Figure 1

The numbers of protein sequences in every GO IDs used in training.

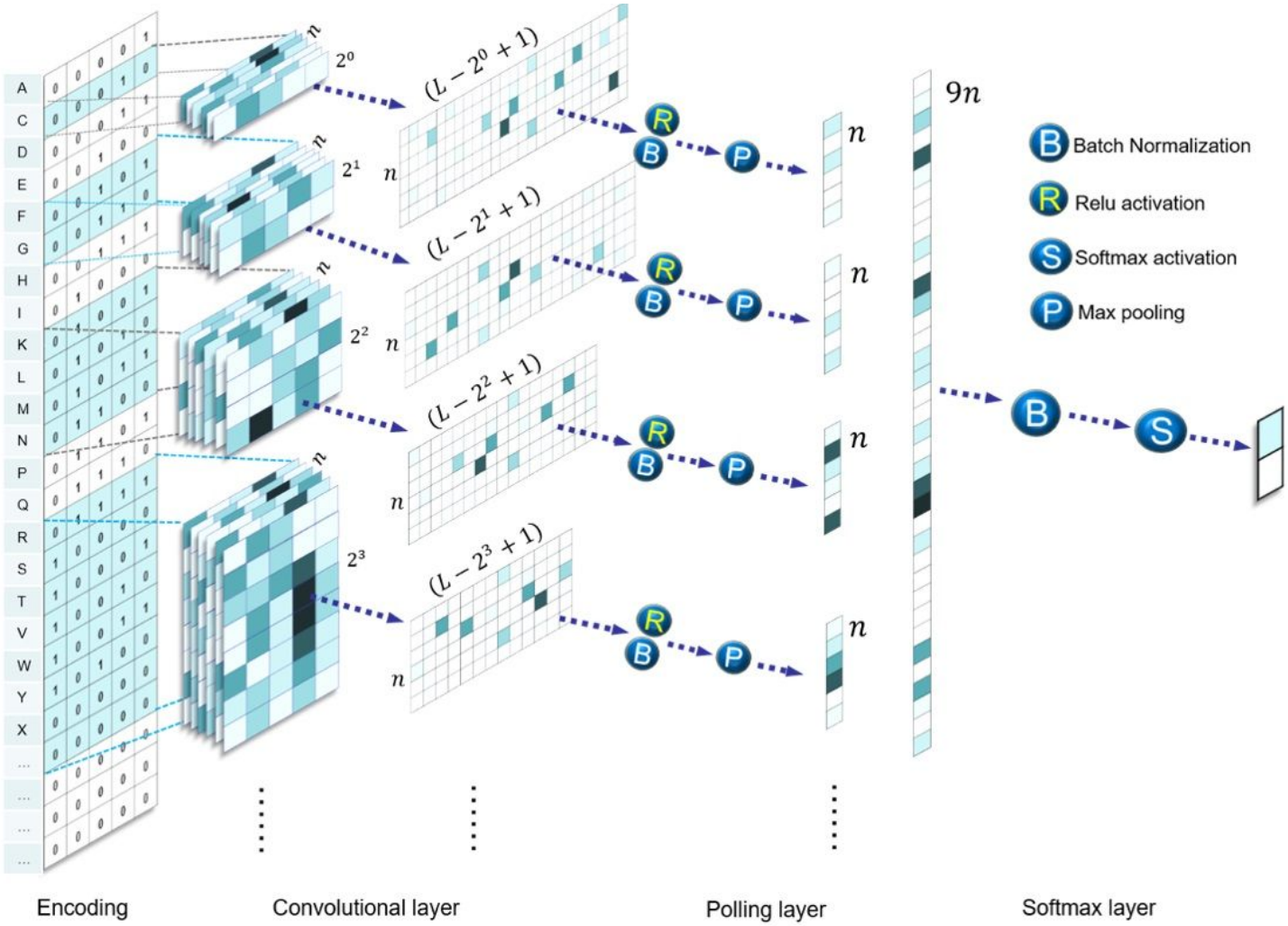

Figure 2

The workflow of CNN adopted in the current work.

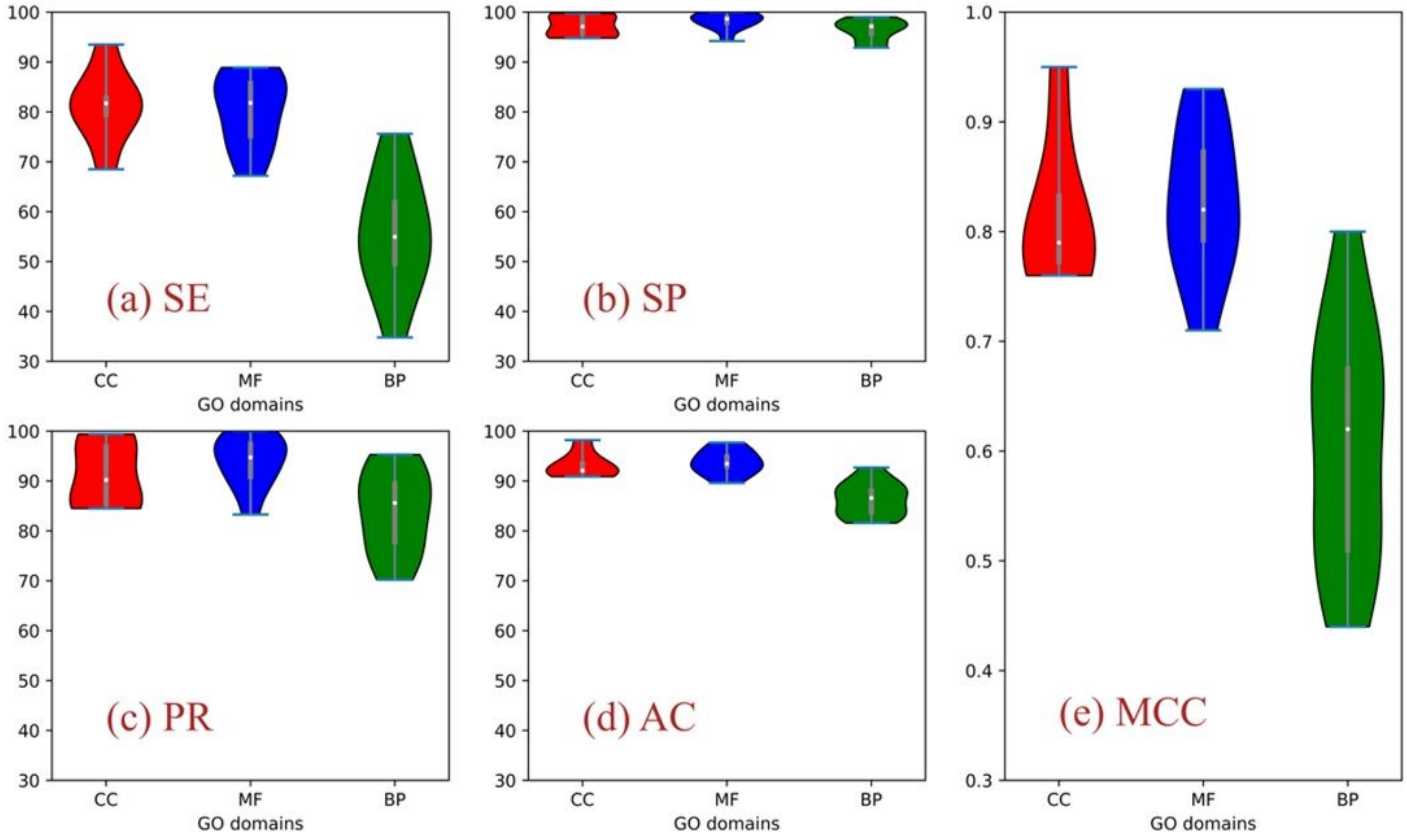

Figure 3

Violin plots for the performance of prediction model in three domains.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplement1measuments.docx
- Supplement2nonannotated.fasta.zip
- Supplement3CCMFBP.txt
- Supplement4immunesignal.txt