Mathematical Biology and Bioinformatics
2020. V. 15. № 2. P. 429–440. doi: 10.17537/2020.15.429

========================== BIOINFORMATICS ==========================

Conserved Peptides Recognition by Ensemble of Neural Networks for Mining Protein Data – LPMO Case Study

Dotsenko G.S.[*], Dotsenko A.S.[**]

Federal Research Center «Fundamentals of Biotechnology» of the Russian Academy of Sciences, Moscow, Russian Federation

Abstract. Mining protein data is a recent promising area of modern bioinformatics. In this work, we suggested a novel approach for mining protein data – conserved peptides recognition by ensemble of neural networks (CPRENN). This approach was applied for mining lytic polysaccharide monooxygenases (LPMOs) in 19 ascomycete, 18 basidiomycete, and 18 bacterial proteomes. LPMOs are recently discovered enzymes and their mining is of high relevance for biotechnology of lignocellulosic materials. CPRENN was compared with two conventional bioinformatic methods for mining protein data – profile hidden Markov models (HMMs) search (HMMER program) and peptide pattern recognition (PPR program combined with Hotpep application). The maximum number of hypothetical LPMO amino acid sequences was discovered by HMMER. Profile HMMs search proved to be the more sensitive method for mining LPMOs than conserved peptides recognition. Totally, CPRENN found 76 %, 67 %, and 65 % of hypothetical ascomycete, basidiomycete, and bacterial LPMOs discovered by HMMER, respectively. For AA9, AA10, and AA11 families which contain the major part of all LPMOs in the carbohydrate-active enzymes database (CAZy), CPRENN and PPR + Hotpep found 69–98 % and 62–95 % of amino acid sequences discovered by HMMER, respectively. In contrast with PPR + Hotpep, CPRENN possessed perfect precision and provided more complete mining of basidiomycete and bacterial LPMOs.

Keywords: *mining protein data, conserved peptides recognition, ensemble of neural networks, lytic polysaccharide monooxygenases.*

INTRODUCTION

Mining protein data is a recent promising area of modern bioinformatics. Due to the accelerating progress of next generation sequencing, protein databases are continuously replenished with new amino acid sequences. However, many of the deposited sequences are not annotated or existing annotation is not reliable [1]. In this situation, bioinformatic methods for mining protein data become important tools for comprehensive use of the accumulated information.

Apart from data mining, several principal techniques are available for protein data analysis and search. Similarity searching, including sequence comparison, is widely used by computational biologists for screening protein databases. The most popular tool for this purpose is BLAST (basic local alignment search tool) [2, 3]. Since proteins with closely related amino acid sequences (identity above 40–70 %) typically possess the same functional properties [4], BLAST is usually applied either for functional annotation of unknown protein sequences or for screening functional analogs of known proteins.

[*] Corresponding author, gsdotsenko@gmail.com
[**] a.dotsenko@fbras.ru

Hidden Markov models (HMMs) are general statistical models widely applied in bioinformatics for various pattern recognition problems [5]. Application of profile HMMs is one of the most reliable methods for finding relationship between evolutionary distant proteins [6]. Amino acid sequences of such proteins typically share low identity but still include similar conserved regions. Profile HMMs describe this information as a probabilistic pattern of considered amino acid sequences [5, 6]. The most popular tool for profile HMMs construction and subsequent protein search is HMMER [7]. A large collection of profile HMMs for different proteins is available in the Pfam database [8].

There are several methods for finding protein relationship based on recognition of conserved regions (signatures) in protein sequences. PROSITE recognizes two types of signatures: generalized profiles that describe protein families and modular protein domains and patterns that describe short sequence motifs often corresponding to functionally or structurally important residues [9]. Peptide pattern recognition (PPR) clusters protein sequences into groups that share a set of short conserved peptide motifs [10]. Identified sets of peptide motifs can be further applied for mining protein data independently or using the homology to peptide pattern (Hotpep) method [11]. Hotpep for carbohydrate-active enzymes is available as a stand-alone application [12]. A large collection of conserved domains for different proteins is available in the conserved domain database (CDD) [13].

PPR and Hotpep efficiency for mining protein data was demonstrated in several publications [10–12, 14, 15]. Authors declared that PPR provides functionally meaningful subdivision of glycoside hydrolases sharing as low as 20 % identity [10]. PPR consists of two steps: (i) finding a limited number of *n*-mer short sequences that are highly conserved in a collection of protein sequences and (ii) selecting protein sequences that contain more than a threshold number of the *n*-mer short sequences [10]. PPR is able to deal with conserved peptides of only one fixed length at a time (in one run). Authors demonstrated that six amino acids constitute the optimal length of conserved peptides for the best performance of PPR [10].

In this work, we developed and tested an alternative approach for mining protein data – conserved peptides recognition by ensemble of neural networks (CPRENN). In contrast with PPR, CPRENN performs simultaneous recognition of conserved peptides of six different lengths (from trimers to octamers). Unlike Hotpep, CPRENN is based on artificial intelligence rather than deterministic algorithm. The hypothesis for the present work was that CPRENN may perform better than PPR + Hotpep for mining protein data. We applied CPRENN for mining lytic polysaccharide monooxygenases (LPMOs) and compared data obtained with HMMER and PPR + Hotpep results. LPMOs are recently discovered enzymes that carry out oxidative degradation of polysaccharides. The discovery of LPMOs is a breakthrough in the biotechnology of lignocellulosic materials [16, 17] and finding novel enzymes of this class is of high relevance.

METHODOLOGY

Protein data

LPMO amino acid sequences belonging to seven families (AA9, AA10, AA11, AA13, AA14, AA15, AA16) of the carbohydrate-active enzymes database (CAZy) [18] were retrieved from the NCBI protein database [19] in May 2020. Fungal and bacterial proteomes were retrieved from the UniProt database [20] at the same time.

Training data consisted of positive (LPMO amino acid sequences) and negative (not LPMO amino acid sequences) sets. A list of LPMOs belonging to seven CAZy families (6214 sequences) was used as the positive set. A list of *Candida albicans* (UP000000559), *Saccharomyces cerevisiae* (UP000002311), and *Wickerhamomyces anomalus* (UP000094112) proteins (18490 sequences) was used as the negative set because proteomes of these organisms were reported to be free of LPMOs [15].

CPRENN implementation

CPRENN was implemented using six independent artificial neural networks. Each neural network was designed for recognition of short peptides of a particular length (3–8 amino acid residues). The final result was calculated by averaging outputs of all neural networks. All scripts were written in Python 3.7.1, SciPy and NumPy extensions were applied.

Neural networks configuration

Six groups of short peptides (3–8 amino acid residues) occurring in the positive training set were extracted to form six peptide libraries (Fig. 1). Then a fully connected (dense) feedforward artificial neural network consisting of three layers (input layer, hidden layer, output layer) was created for each peptide library. Number of neurons in the input layer was determined by the number of peptides in the corresponding peptide library. Number of neurons in the hidden layer was empirically chosen to be approx. 0.001 of the input neurons number but not less than 100. Number of neurons in the output layer was determined by the number of considered LPMO families (7).

Each neural network was initialized with random synaptic weights belonging to the standard normal distribution with the standard deviation calculated as (number of neurons in the layer)$^{-0.5}$. The logistic activation function was applied for all neural networks.

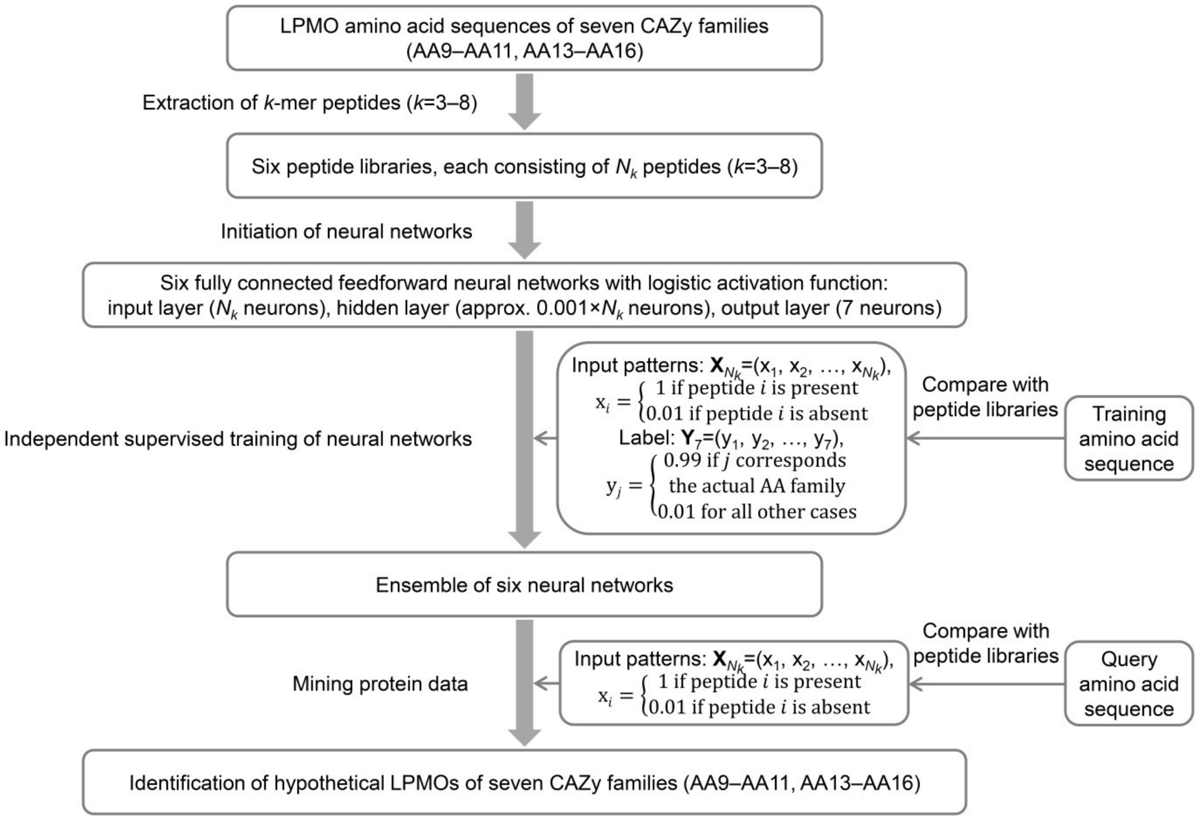
LPMO amino acid sequences of seven CAZy families (AA9–AA11, AA13–AA16)
Extraction of k-mer peptides (k=3–8)
Six peptide libraries, each consisting of N_k peptides (k=3–8)
Initiation of neural networks
Six fully connected feedforward neural networks with logistic activation function: input layer (N_k neurons), hidden layer (approx. 0.001×N_k neurons), output layer (7 neurons)
Independent supervised training of neural networks
Input patterns: $\mathbf{X}_{N_k}=(x_1, x_2, \ldots, x_{N_k})$, $x_i = \begin{cases} 1 \text{ if peptide } i \text{ is present} \\ 0.01 \text{ if peptide } i \text{ is absent} \end{cases}$
Label: $\mathbf{Y}_7=(y_1, y_2, \ldots, y_7)$, $y_j = \begin{cases} 0.99 \text{ if } j \text{ corresponds the actual AA family} \\ 0.01 \text{ for all other cases} \end{cases}$
Compare with peptide libraries
Training amino acid sequence
Ensemble of six neural networks
Mining protein data
Input patterns: $\mathbf{X}_{N_k}=(x_1, x_2, \ldots, x_{N_k})$, $x_i = \begin{cases} 1 \text{ if peptide } i \text{ is present} \\ 0.01 \text{ if peptide } i \text{ is absent} \end{cases}$
Compare with peptide libraries
Query amino acid sequence
Identification of hypothetical LPMOs of seven CAZy families (AA9–AA11, AA13–AA16)

Fig. 1. Schematic representation of CPRENN principle demonstrated for mining LPMOs.

Neural networks training

Each neural network was independently trained using backpropagation with the stochastic gradient descent and the quadratic loss function. The input pattern of a given amino acid sequence was formed by comparison of its peptide composition with the corresponding peptide library. Each item of the input pattern was either 1 (if library peptide occurred in a given amino acid sequence) or 0.01 (if library peptide did not occur in a given amino acid sequence). Label (the expected output signal) of a given amino acid sequence was formed using the *one-hot* encoding with 0.99 for *hot* value and 0.01 for *cold* values. The *hot* value indicated the actual AA family of the CAZy database for a given amino acid sequence (Fig. 1).

Each neural network was trained for five epochs consisting of 5000 iterations. Learning schedule was based on the linear decreasing learning rate. Type of training amino acid sequence (positive or negative) was randomly chosen for each iteration. In case of the positive type, the particular amino acid sequence was randomly selected from the positive training set using the stratified sampling. In case of the negative type, the particular amino acid sequence was randomly selected from the negative training set.

Mining protein data

For mining protein data by CPRENN, six independently trained neural networks were joined into an ensemble returning their average result as a tuple consisting of seven components. This result was analyzed using the *one-versus-all* strategy. If the greatest component of the tuple exceeded the decision threshold, empirically chosen to be 0.4, then the query amino acid sequence was classified as a hypothetical LPMO belonging to the corresponding AA family of the CAZy database.

Multiple alignments of amino acid sequences belonging to each of seven considered LPMO families were performed using Clustal Omega 1.2.4 software [21]. Profile HMMs for these families were constructed using HMMER 3.2.1 software [7]. Profile HMMs were applied for mining protein data using the same software with the E-value reporting threshold of 10^{-15}.

Statistical analysis

CPRENN performance was evaluated using cross-validation with 80 % data applied for training and 20 % data applied for testing. Sensitivity, precision, and F_1 score (the harmonic mean of sensitivity and precision) were calculated according to the following formulas:

$$\text{Sensitivity} = \frac{\text{True positives}}{\text{True positives+False negatives}},$$

$$\text{Precision}=\frac{\text{True positives}}{\text{True positives+False positives}},$$

$$\text{F}_1\ \text{score}=\frac{2}{1/\text{Sensitivity} + 1/\text{Precision}}.$$

Completeness of mining LPMOs by CPRENN (PPR + Hotpep) programs in comparison with HMMER program was calculated as

$$\frac{\text{Number of sequences found by CPRENN } (\text{PPR } + \text{ Hotpep})}{\text{Number of sequences found by HMMER}}\times 100\%$$

When calculating completeness of mining LPMOs by PPR + Hotpep, PPR + Hotpep false positive recognitions were not considered.

RESULTS AND DISCUSSION

CPRENN implementation

A common problem for large data analysis is a large number of data features to be taken into consideration. This problem is especially important for neural networks since its strong effect on computational complexity. As it follows from Fig. 1, all peptides occurred in the positive training set were used to form six peptide libraries. An obvious solution for reduction the input dimensionality would be to discard peptides that occurred in the positive training set only once. This solution seems to be reasonable for abundant datasets containing enough protein sequences for each class (family). In this work, we considered seven LPMO families of the CAZy database of which only AA10 family seemed to be abundant (approx. 5000 enzymes). Since other families contained much fewer enzymes (approx. 20–500), we used all occurred peptides for peptide libraries construction.

We intentionally applied a very simple configuration of neural networks. Therefore, each neural network contained only one hidden layer and utilized the logistic activation function. However, implementation of deeper neural networks with different activation functions is interesting for further studies.

CPRENN performance

CPRENN performance was evaluated using cross-validation with 80 % data applied as a training set and 20 % data applied as a test set. As can be seen from Table 1, sensitivity of LPMOs recognition was very high (0.9–1.0) for five considered families (AA9, AA10, AA11, AA13, AA15). Sensitivity of LPMOs recognition for two other families (AA14, AA16) was lower (0.6–0.8). The latter result is most likely due to the limited number of protein sequences in these families. Remarkably, perfect precision was observed for all repeats of cross-validation due to absence of false positive recognition.

Table 1. CPRENN performance for LPMOs recognition in a cross-validation procedure (80 % data applied as a training set and 20 % data applied as a test set). Sensitivity, precision, and accuracy (measured as F_1 score) are presented as mean ± standard deviation for five repeats unless standard deviation is zero

CAZy family	Approx. number of amino acid sequences		Sensitivity	Precision	Accuracy (F_1 score)
	Training set	Test set			
AA9	437	110	0.95 ± 0.02	1.00	0.98 ± 0.01
AA10	4163	1041	0.99 ± 0.01	1.00	0.99 ± 0.01
AA11	100	26	0.97 ± 0.03	1.00	0.98 ± 0.02
AA13	18	5	1.00	1.00	1.00
AA14	16	4	0.6 ± 0.3	1.00	0.7 ± 0.4
AA15	205	52	0.90 ± 0.05	1.00	0.95 ± 0.03
AA16	29	8	0.78 ± 0.07	1.00	0.88 ± 0.05
Total	4968	1246	0.98 ± 0.01	1.00	0.99 ± 0.01

As already mentioned, PPR combined with Hotpep is an efficient method for mining protein data. Busk et al. [12] reported that PPR + Hotpep sensitivity, precision, and accuracy (measured as F_1 score) for a wide range of proteins were 0.77–0.88, 0.84–0.88, and 0.82–0.86, respectively. In contrast with PPR + Hotpep, CPRENN possessed perfect precision. At the same time, sensitivity of CPRENN was generally higher than sensitivity of PPR + Hotpep. Accuracy of CPRENN (measured as F_1 score) was also generally higher than accuracy of PPR + Hotpep.

For most of practical situations, the perfect separation of samples belonging to positive and negative classes is not possible. In such cases, shifting the decision threshold can either increase sensitivity or precision but not both (sensitivity/precision trade-off). The correct balance between sensitivity and precision is usually chosen based on the specificity of a concrete problem. Mining protein data typically implies analysis of large data with a rare occurrence of target proteins. In this situation, even a slight rate of false positive classification (precision < 1) finally results in a considerable number of false positive recognitions. Therefore, in this work, we prioritized perfect precision, rather than high sensitivity.

As described above, CPRENN demonstrated perfect precision in the cross-validation procedure (Table 1). Nevertheless, we decided to further verify this property using a larger dataset. For this purpose, we selected 25 proteomes (3 fungi and 22 bacteria) that previously were reported to be free of LPMOs [15]: *Batrachochytrium dendrobatidis* (UP000077115), *Mucor circinelloides* (UP000014254), *Spizellomyces punctatus* (UP000053201); *Agrobacterium radiobacter* (UP000001600), *Azospirillum brasilense* (UP000007319), *Bacteroides fragilis* (UP000006731), *Bifidobacterium breve* (UP000003191), *Bradyrhizobium japonicum* (UP000193335), *Butyrivibrio fibrisolvens* (UP000182584), *Clostridium thermocellum* (UP000002145), *Cyanobacterium stanieri* (UP000010483), *Escherichia coli* (UP000019194), *Eubacterium eligens* (UP000001476), *Fervidobacterium pennivorans* (UP000007384), *Fibrobacter succinogenes* (UP000000517), *Lactobacillus amylovorus* (UP000007033), *Microbacterium testaceum* (UP000008975), *Mycobacterium chelonae* (UP000180043), *Prevotella ruminicola* (UP000184130), *Propionibacterium acnes* (UP000008987), *Rhizobium etli* (UP000248982), *Ruminococcus albus* (UP000004259), *Selenomonas ruminantium* (UP000182958), *Sinorhizobium meliloti* (UP000009045), *Slackia heliotrinireducens* (UP000002026). The applied dataset consisted of 125056 amino acid sequences. Again, no false positive recognitions were registered and CPRENN demonstrated perfect precision.

Mining protein data

Since CPRENN demonstrated a near-perfect performance for considered test datasets, it was of high interest to compare this approach with other bioinformatic methods. For this purpose, we selected two programs – HMMER [7] and PPR (in combination with Hotpep application) [10–12]. CPRENN and HMMER were applied for mining LPMOs in 19 ascomycete, 18 basidiomycete, and 18 bacterial proteomes. Data obtained were compared with PPR + Hotpep results previously reported by Busk and Lange [15] (Tables 2–4). As it followed from the obtained results, the maximum number of hypothetical LPMO amino acid sequences was discovered by HMMER. Profile HMMs search proved to be more sensitive method for mining LPMOs than conserved peptides recognition. This result is consistent with the methodology of considered methods. Theoretically, profile HMMs can be applied for mining proteins containing single conserved amino acid residues separated by variable regions [5–7]. Obviously, recognition of conserved peptides is relevant only for mining proteins containing several conserved amino acid residues in a row. Therefore, completeness of mining LPMOs by CPRENN and PPR + Hotpep was further analyzed in relation to HMMER results.

Unfortunately, Busk and Lange [15] reported only total number of hypothetical LPMOs found by PPR + Hotpep in the selected organisms for each of three considered families (AA9, AA10, and AA11), while neither identified sequences nor their accession numbers were published. Having compared the reported data with HMMER results, we identified several false positive recognitions done by PPR + Hotpep for the following organisms and families: *Chaetomium thermophilum* AA9, *Hypocrea jecorina* AA11, *Metarhizium anisopliae* AA11, *Ceriporiopsis subvermispora* AA11, *Enterobacter cloacae* AA10 (Tables 2–4). The observed discrepancies may be a matter of balance between the decision threshold of CPRENN, the E-value reporting threshold of HMMER, and the applied PPR + Hotpep parameters. However, no false positive recognitions were observed for CPRENN in comparison with HMMER (Supplementary Tables S1–S3).

Table 2. Numbers of hypothetical LPMOs found in 19 ascomycete proteomes by CPRENN, HMMER, and PPR + Hotpep programs. Distribution of hypothetical LPMOs among different CAZy families is presented in parentheses

Ascomycete	**Proteome accession number (UniProt)**	**Number of hypothetical LPMOs**		
		CPRENN	**HMMER**	**PPR + Hotpep**
		(AA9, AA10, AA11, AA13, AA14, AA15, AA16)		**(AA9, AA10, AA11)**
Arthroderma gypseum	UP000002669	4 (0, 0, 4, 0, 0, 0, 0)	7 (0, 1, 5, 1, 0, 0, 0)	5 (0, 0, 5)
Arthroderma otae	UP000002035	3 (0, 0, 3, 0, 0, 0, 0)	6 (0, 1, 4, 1, 0, 0, 0)	5 (0, 1, 4)
Aspergillus nidulans	UP000000560	14 (9, 0, 2, 2, 0, 0, 1)	20 (10, 0, 2, 5, 0, 2, 1)	12 (10, 0, 2)
Aspergillus niger	UP000006706	11 (7, 0, 3, 0, 0, 0, 1)	12 (7, 0, 3, 1, 0, 0, 1)	10 (7, 0, 3)
Ceratocystis fimbriata	UP000222788	7 (4, 0, 3, 0, 0, 0, 0)	10 (4, 0, 3, 2, 0, 0, 1)	4 (3, 0, 1)
Chaetomium globosum	UP000001056	47 (39, 0, 5, 2, 0, 0, 1)	56 (43, 0, 6, 3, 0, 1, 3)	46 (40, 0, 6)
Chaetomium thermophilum	UP000008066	18 (16, 0, 2, 0, 0, 0, 0)	21 (16, 0, 2, 1, 2, 0, 0)	20 (17, 0, 3)
Coccidioides immitis	UP000054565	2 (0, 0, 2, 0, 0, 0, 0)	4 (0, 0, 3, 1, 0, 0, 0)	3 (0, 0, 3)
Cyphellophora europaea	UP000030752	7 (6, 0, 1, 0, 0, 0, 0)	13 (6, 0, 3, 1, 3, 0, 0)	8 (5, 0, 3)
Hypocrea jecorina	UP000024376	6 (3, 0, 3, 0, 0, 0, 0)	10 (3, 1, 3, 1, 2, 0, 0)	6 (2, 0, 4)
Metarhizium anisopliae	UP000054544	7 (1, 0, 6, 0, 0, 0, 0)	12 (1, 1, 6, 1, 3, 0, 0)	10 (2, 0, 8)
Myceliophthora thermophila	UP000007322	30 (22, 0, 4, 1, 0, 0, 3)	33 (22, 0, 4, 3, 1, 0, 3)	26 (22, 0, 4)
Neurospora crassa	UP000001805	19 (14, 0, 4, 1, 0, 0, 0)	22 (14, 0, 4, 3, 1, 0, 0)	18 (14, 0, 4)
Sordaria macrospora	UP000001881	25 (19, 0, 3, 1, 0, 0, 2)	31 (19, 0, 4, 4, 1, 1, 2)	23 (19, 0, 4)
Talaromyces marneffei	UP000029285	2 (1, 0, 1, 0, 0, 0, 0)	5 (1, 0, 1, 3, 0, 0, 0)	3 (1, 0, 2)
Talaromyces stipitatus	UP000001745	5 (1, 0, 4, 0, 0, 0, 0)	9 (1, 0, 4, 4, 0, 0, 0)	5 (1, 0, 4)
Thielavia terrestris	UP000008181	24 (18, 0, 5, 0, 0, 0, 1)	30 (20, 0, 5, 3, 1, 0, 1)	22 (17, 0, 5)
Trichophyton rubrum	UP000008864	4 (0, 0, 4, 0, 0, 0, 0)	7 (0, 1, 5, 1, 0, 0, 0)	5 (0, 0, 5)
Uncinocarpus reesii	UP000002058	3 (0, 0, 3, 0, 0, 0, 0)	4 (0, 0, 3, 1, 0, 0, 0)	3 (0, 0, 3)
Total		238 (160, 0, 62, 7, 0, 0, 9)	312 (167, 5, 70, 40, 14, 4, 12)	234 (160, 1, 73)

Table 3. Numbers of hypothetical LPMOs found in 18 basidiomycete proteomes by CPRENN, HMMER, and PPR + Hotpep programs. Distribution of hypothetical LPMOs among different CAZy families is presented in parentheses

Basidiomycete	Proteome accession number (UniProt)	Number of hypothetical LPMOs		
		CPRENN	HMMER	PPR + Hotpep
		(AA9, AA10, AA11, AA13, AA14, AA15, AA16)		(AA9, AA10, AA11)
Botryobasidium botryosum	UP000027195	33 (32, 1, 0, 0, 0, 0, 0)	46 (33, 1, 0, 2, 8, 0, 2)	15 (15, 0, 0)
Ceriporiopsis subvermispora	UP000016930	10 (9, 0, 0, 0, 1, 0, 0)	15 (9, 0, 0, 4, 2, 0, 0)	7 (6, 0, 1)
Coprinopsis cinerea	UP000001861	26 (26, 0, 0, 0, 0, 0, 0)	41 (34, 0, 0, 2, 5, 0, 0)	18 (18, 0, 0)
Cryptococcus neoformans	UP000002149	3 (1, 0, 1, 0, 1, 0, 0)	4 (1, 0, 1, 0, 2, 0, 0)	1 (0, 0, 1)
Fibroporia radiculosa	UP000006352	2 (2, 0, 0, 0, 0, 0, 0)	6 (2, 0, 0, 1, 3, 0, 0)	1 (1, 0, 0)
Fomitopsis pinicola	UP000015241	4 (4, 0, 0, 0, 0, 0, 0)	10 (4, 0, 0, 2, 4, 0, 0)	2 (2, 0, 0)
Galerina marginata	UP000027222	19 (18, 1, 0, 0, 0, 0, 0)	26 (18, 1, 0, 3, 3, 0, 1)	14 (13, 1, 0)
Gloeophyllum trabeum	UP000030669	5 (4, 0, 0, 0, 1, 0, 0)	8 (4, 0, 0, 2, 2, 0, 0)	3 (3, 0, 0)
Heterobasidion irregulare	UP000030671	11 (10, 0, 0, 0, 1, 0, 0)	17 (10, 0, 0, 3, 3, 0, 1)	8 (8, 0, 0)
Jaapia argillacea	UP000027265	14 (14, 0, 0, 0, 0, 0, 0)	24 (15, 0, 0, 3, 2, 0, 4)	7 (7, 0, 0)
Laccaria bicolor	UP000001194	3 (3, 0, 0, 0, 0, 0, 0)	12 (7, 0, 0, 1, 4, 0, 0)	12 (12, 0, 0)
Phanerochaete carnosa	UP000008370	10 (10, 0, 0, 0, 0, 0, 0)	17 (11, 0, 0, 3, 3, 0, 0)	9 (9, 0, 0)
Pleurotus ostreatus	UP000027073	28 (28, 0, 0, 0, 0, 0, 0)	35 (28, 0, 0, 3, 3, 0, 1)	17 (17, 0, 0)
Postia placenta	UP000194127	2 (1, 0, 0, 0, 1, 0, 0)	6 (2, 0, 0, 1, 3, 0, 0)	2 (2, 0, 0)
Pycnoporus cinnabarinus	UP000029665	19 (15, 0, 0, 0, 4, 0, 0)	21 (16, 0, 0, 1, 4, 0, 0)	13 (13, 0, 0)
Schizophyllum commune	UP000007431	28 (19, 0, 8, 0, 0, 0, 1)	36 (22, 0, 8, 1, 3, 0, 2)	21 (14, 0, 7)
Ustilago maydis	UP000000561	1 (0, 1, 0, 0, 0, 0, 0)	1 (0, 1, 0, 0, 0, 0, 0)	1 (0, 1, 0)
Wolfiporia cocos	UP000218811	2 (2, 0, 0, 0, 0, 0, 0)	6 (2, 0, 0, 1, 3, 0, 0)	1 (1, 0, 0)
Total		220 (198, 3, 9, 0, 9, 0, 1)	331 (218, 3, 9, 33, 57, 0, 11)	152 (141, 2, 9)

Table 4. Numbers of hypothetical LPMOs found in 18 bacterial proteomes by CPRENN, HMMER, and PPR + Hotpep programs. Distribution of hypothetical LPMOs among different CAZy families is presented in parentheses

Bacterium	**Proteome accession number (UniProt)**	**Number of hypothetical LPMOs**		
		CPRENN	**HMMER**	**PPR + Hotpep**
		(AA9, AA10, AA11, AA13, AA14, AA15, AA16)		**(AA9, AA10, AA11)**
Actinoplanes missouriensis	UP000007882	7 (0, 7, 0, 0, 0, 0, 0)	14 (0, 12, 0, 2, 0, 0, 0)	7 (0, 7, 0)
Bacillus thuringiensis	UP000011719	3 (0, 3, 0, 0, 0, 0, 0)	4 (0, 4, 0, 0, 0, 0, 0)	3 (0, 3, 0)
Brevibacillus laterosporus	UP000005850	2 (0, 2, 0, 0, 0, 0, 0)	2 (0, 2, 0, 0, 0, 0, 0)	2 (0, 2, 0)
Enterobacter cloacae	UP000007838	0 (0, 0, 0, 0, 0, 0, 0)	0 (0, 0, 0, 0, 0, 0, 0)	1 (0, 1, 0)
Herpetosiphon aurantiacus	UP000000787	1 (0, 1, 0, 0, 0, 0, 0)	2 (0, 1, 0, 1, 0, 0, 0)	1 (0, 1, 0)
Klebsiella oxytoca	UP000236461	1 (0, 1, 0, 0, 0, 0, 0)	1 (0, 1, 0, 0, 0, 0, 0)	1 (0, 1, 0)
Lactococcus lactis	UP000002196	1 (0, 1, 0, 0, 0, 0, 0)	2 (0, 2, 0, 0, 0, 0, 0)	1 (0, 1, 0)
Listeria monocytogenes	UP000000817	1 (0, 1, 0, 0, 0, 0, 0)	2 (0, 2, 0, 0, 0, 0, 0)	2 (0, 2, 0)
Nocardiopsis dassonvillei	UP000002219	4 (0, 4, 0, 0, 0, 0, 0)	5 (0, 5, 0, 0, 0, 0, 0)	1 (0, 1, 0)
Photorhabdus asymbiotica	UP000002747	1 (0, 1, 0, 0, 0, 0, 0)	1 (0, 1, 0, 0, 0, 0, 0)	1 (0, 1, 0)
Pseudomonas aeruginosa	UP000002438	1 (0, 1, 0, 0, 0, 0, 0)	2 (0, 2, 0, 0, 0, 0, 0)	1 (0, 1, 0)
Saccharophagus degradans	UP000001947	1 (0, 1, 0, 0, 0, 0, 0)	4 (0, 4, 0, 0, 0, 0, 0)	1 (0, 1, 0)
Salinispora arenicola	UP000001153	4 (0, 4, 0, 0, 0, 0, 0)	6 (0, 5, 0, 1, 0, 0, 0)	4 (0, 4, 0)
Serratia marcescens	UP000050507	3 (0, 3, 0, 0, 0, 0, 0)	4 (0, 4, 0, 0, 0, 0, 0)	2 (0, 2, 0)
Stenotrophomonas maltophilia	UP000006955	2 (0, 2, 0, 0, 0, 0, 0)	2 (0, 2, 0, 0, 0, 0, 0)	2 (0, 2, 0)
Streptomyces viridochromogenes	UP000011205	6 (0, 6, 0, 0, 0, 0, 0)	8 (0, 8, 0, 0, 0, 0, 0)	5 (0, 5, 0)
Vibrio cholerae	UP000000584	2 (0, 2, 0, 0, 0, 0, 0)	3 (0, 3, 0, 0, 0, 0, 0)	2 (0, 2, 0)
Xylanimonas cellulosilytica	UP000002255	2 (0, 2, 0, 0, 0, 0, 0)	3 (0, 3, 0, 0, 0, 0, 0)	2 (0, 2, 0)
Total		42 (0, 42, 0, 0, 0, 0, 0)	65 (0, 61, 0, 4, 0, 0, 0)	39 (0, 39, 0)

Totally, CPRENN found 76 %, 67 %, and 65 % of hypothetical ascomycete, basidiomycete, and bacterial LPMOs discovered by HMMER, respectively (Table 5). For AA9, AA10, and AA11 families which contain the major part of all LPMOs in the CAZy database, CPRENN and PPR + Hotpep found 69–98 % and 62–95 % of amino acid sequences discovered by HMMER, respectively. As it followed from the obtained results, CPRENN provided more complete mining of basidiomycete and bacterial LPMOs than PPR + Hotpep.

Table 5. Completeness of mining LPMOs by CPRENN and PPR + Hotpep programs in comparison with HMMER program. Data were obtained for 19 ascomycete, 18 basidiomycete, and 18 bacterial proteomes

CAZy family	**Completeness of mining LPMOs, %**					
	Ascomycetes		**Basidiomycetes**		**Bacteria**	
	CPRENN	**PPR + Hotpep**	**CPRENN**	**PPR + Hotpep**	**CPRENN**	**PPR + Hotpep**
AA9	95.8	95.2	90.8	64.7	–[b]	–[b]
AA10	0.0	20.0	100.0	66.7	68.9	62.3
AA11	88.6	100.0	100.0	100.0	–[b]	–[b]
AA13	17.5	nd[a]	0.0	nd[a]	0.0	nd[a]
AA14	0.0	nd[a]	15.8	nd[a]	–[b]	nd[a]
AA15	0.0	nd[a]	–[b]	nd[a]	–[b]	nd[a]
AA16	75.0	nd[a]	9.1	nd[a]	–[b]	nd[a]
AA9, AA10, AA11	91.7	95.0	97.8	66.1	68.9	62.3
Total	76.3	nd[a]	66.5	nd[a]	64.6	nd[a]

[a]no data available, [b]no hypothetical LPMOs found by HMMER

CONCLUSIONS

In this work, we suggested a novel approach for mining protein data – conserved peptides recognition by ensemble of neural networks (CPRENN). This approach was compared with profile hidden Markov models (HMMs) search (HMMER program) and peptide pattern recognition (PPR program combined with Hotpep application) for mining lytic polysaccharide monooxygenases (LPMOs). The maximum number of hypothetical LPMO amino acid sequences was discovered by HMMER. Profile HMMs search proved to be more sensitive method for mining LPMOs than conserved peptides recognition. Totally, CPRENN found 76 %, 67 %, and 65 % of hypothetical ascomycete, basidiomycete, and bacterial LPMOs discovered by HMMER, respectively. For AA9, AA10, and AA11 families which contain the major part of all LPMOs in the carbohydrate-active enzymes database (CAZy), CPRENN and PPR + Hotpep found 69–98 % and 62–95 % of amino acid sequences discovered by HMMER, respectively. In contrast with PPR + Hotpep, CPRENN possessed perfect precision and provided more complete mining of basidiomycete and bacterial LPMOs.

REFERENCES

1. Ijaq J., Chandrasekharan M., Poddar R., Bethi N., Sundararajan V.S. Annotation and curation of uncharacterized proteins – challenges. *Frontiers in Genetics.* 2015. V. 6. Article No. 119.
2. Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. Basic local alignment search tool. *Journal of Molecular Biology.* 1990. V. 215. P. 403–410.
3. Pertsemlidis A., Fondon III J.W. Having a BLAST with bioinformatics (and avoiding BLASTphemy). *Genome Biology.* 2001. V. 2. Article No. reviews2002.

4. Tian W., Skolnick J. How well is enzyme function conserved as a function of pairwise sequence identity? *Journal of Molecular Biology.* 2003. V. 333. P. 863–882.
5. Yoon B.-J. Hidden Markov models and their applications in biological sequence analysis. *Current Genomics.* 2009. V. 10. P. 402–415.
6. Choo K.H., Tong J.C., Zhang L. Recent applications of hidden Markov models in computational biology. *Genomics, proteomics and bioinformatics.* 2004. V. 2. P. 84–96.
7. *HMMER: Biosequence Analysis Using Profile Hidden Markov Models.* URL: http://hmmer.org/ (accessed 01.09.2020).
8. El-Gebali S., Mistry J., Bateman A., Eddy S.R., Luciani A., Potter S.C., Qureshi M., Richardson L.J., Salazar G.A., Smart A., Sonnhammer E.L.L., Hirsh L., Paladin L., Piovesan D., Tosatto S.C.E., Finn R.D. The Pfam protein families database in 2019. *Nucleic Acids Research.* 2019. V. 47 (Database Issue). P. D427–D432.
9. Sigrist C.J.A., de Castro E., Cerutti L., Cuche B.A., Hulo N., Bridge A., Bougueleret L., Xenarios I. New and continuing developments at PROSITE. *Nucleic Acids Research.* 2013. V. 41 (Database Issue). P. D344–D347.
10. Busk P.K., Lange L. Function-based classification of carbohydrate-active enzymes by recognition of short, conserved peptide motifs. *Applied and Environmental Microbiology.* 2013. V. 79. P. 3380–3391.
11. Busk P.K., Lange M., Pilgaard B., Lange L. Several genes encoding enzymes with the same activity are necessary for aerobic fungal degradation of cellulose in nature. *PLoS ONE.* 2014. V. 9. Article No. e114138.
12. Busk P.K., Pilgaard B., Lezyk M.J., Meyer A.S., Lange L. Homology to peptide pattern for annotation of carbohydrate-active enzymes and prediction of function. *BMC Bioinformatics.* 2017. V. 18. Article No. 214.
13. Lu S., Wang J., Chitsaz F., Derbyshire M.K., Geer R.C., Gonzales N.R., Gwadz M., Hurwitz D.I., Marchler G.H., Song J.S., Thanki N., Yamashita R.A., Yang M., Zhang D., Zheng C., Lanczycki C.J., Marchler-Bauer A. CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Research.* 2020. V. 48 (Database Issue). P. D265–D268.
14. Agger J.W., Busk P.K., Pilgaard B., Meyer A.S., Lange L. A new functional classification of glucuronoyl esterases by peptide pattern recognition. *Frontiers in Microbiology.* 2017. V. 8. Article No. 309.
15. Busk P.K., Lange L. Classification of fungal and bacterial lytic polysaccharide monooxygenases. *BMC Genomics.* 2015. V. 16. Article No. 368.
16. Hemsworth G.R., Johnston E.M., Davies G.J., Walton P.H. Lytic polysaccharide monooxygenases in biomass conversion. *Trends in Biotechnology.* 2015. V. 33. P. 747–761.
17. Johansen K.S. Lytic polysaccharide monooxygenases: the microbial power tool for lignocellulose degradation. *Trends in Plant Science.* 2016. V. 21. P. 926–936.
18. *CAZy, carbohydrate-active enzymes database.* URL: http://www.cazy.org/ (accessed 01.09.2020).
19. *NCBI protein database.* URL: https://www.ncbi.nlm.nih.gov/protein/ (accessed 01.09.2020).
20. *UniProt Database.* URL: https://www.uniprot.org/ (accessed 01.09.2020).
21. Sievers F., Wilm A., Dineen D.G., Gibson T.J., Karplus K., Li W., Lopez R., McWilliam H., Remmert M., Söding J., Thompson J.D., Higgins D.G. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology.* 2011. V. 7. Article No. 539.

Received 24.09.2020. Revised 28.11.2020. Published 22.12.2020.

Распознавание консервативных пептидов ансамблем нейронных сетей для глубинного анализа белковых данных на примере LPMO

Доценко Г.С., Доценко А.С.

Федеральный исследовательский центр «Фундаментальные основы биотехнологии» Российской академии наук, Москва, Российская Федерация

Аннотация. Глубинный анализ белковых данных – это новое перспективное направление современной биоинформатики. В этой работе мы предложили новый подход для глубинного анализа белковых данных – распознавание консервативных пептидов ансамблем нейронных сетей (CPRENN). Этот подход был применён для поиска литических полисахаридмонооксигеназ (LPMO) в протеомах 19 аскомицетов, 18 базидиомицетов и 18 бактерий. LPMO – это недавно открытые ферменты, и их поиск имеет большое значение для биотехнологии лигноцеллюлозных материалов. CPRENN был сопоставлен с двумя стандартными биоинформатическими методами для глубинного анализа белковых данных – поиском по скрытым марковским моделям (HMM, программа HMMER) и распознаванием пептидных мотивов (программа PPR совместно с приложением Hotpep). Максимальное число аминокислотных последовательностей гипотетических LPMO было обнаружено с помощью программы HMMER. Метод HMM оказался более чувствительным для поиска LPMO, чем распознавание консервативных пептидов. В целом, с помощью CPRENN было найдено 76 %, 67 % и 65 % гипотетических аскомицетных, базидиомицетных и бактериальных LPMO, обнаруженных HMMER, соответственно. Для AA9, AA10 и AA11 семей, содержащих основную часть всех LPMO в базе данных CAZy, с помощью CPRENN и PPR + Hotpep было найдено 69–98 % и 62–95 % аминокислотных последовательностей, обнаруженных HMMER, соответственно. В отличие от PPR + Hotpep, CPRENN обладал идеальной точностью и обеспечивал более полный поиск базидиомицетных и бактериальных LPMO.

Ключевые слова: *глубинный анализ белковых данных, распознавание консервативных пептидов, ансамбль нейронных сетей, литические полисахаридмонооксигеназы.*