Interdisciplinary Sciences: Computational Life Sciences (2019) 11:108–114
https://doi.org/10.1007/s12539-018-0287-2

ORIGINAL RESEARCH ARTICLE

NeuroPP: A Tool for the Prediction of Neuropeptide Precursors Based on Optimal Sequence Composition

Juanjuan Kang[1] · Yewei Fang[1] · Pengcheng Yao[1] · Ning Li[1] · Qiang Tang[1] · Jian Huang[1,2]

Received: 3 January 2018 / Revised: 10 February 2018 / Accepted: 22 February 2018 / Published online: 10 March 2018

Abstract
Neuropeptides (NPs) are short secreted peptides produced mainly in the nervous system and digestive system. They activate signaling cascades to control a wide range of biological functions, such as metabolism, sensation, and behavior. NPs are typically produced from a larger NP precursor (NPP) which includes a signal peptide sequence, one or more NP sequences, and other sequences. With the drastic growth of unknown protein sequences generated in the post-genomic age, it is highly desired to develop computational methods for identifying NPP rapidly and efficiently. In this article, we developed a predictor for NPPs based on optimized sequence composition of single amino acid, dipeptide, and tripeptide. Evaluated with independent data set, the predictor showed excellent performance that achieved an accuracy of 88.65% with AUC of 0.95. The corresponding web server was developed, which is freely available at http://i.uestc.edu.cn/neuropeptide/neuropp/home.html. It can help relevant researchers to screen candidate NP precursor, shorten experimental cycle, and reduce costs.

Keywords Neuropeptide · Neuropeptide precursor · Predictor · Sequence composition · Recognition

1 Introduction

The discovery of neuropeptides (NPs) is due to the groundbreaking progress in physiology, endocrinology, and biochemistry during the last century. NPs are widely distributed in both the peripheral and central nervous system [1]. The functions which NPs mediated cover not only neural activity, but also various aspects of non-neuronal cells, including food uptake, energy consumption, and social behavior [2, 3]. Mature NPs are stored in dense-cored vesicles and controlled release upon a stimulus [4]. It activates a signaling cascade by binding G protein-coupled receptor commonly [5].

In general, short bioactive NPs are generated from a series of cleavages of a larger neuropeptide precursor (NPP) which rely on proteolytic enzymes and maturation events, such as C-terminal amidation, post-translational modifications. Notably, the cleavages mostly occur at basic residues (Gly, Lys, and Arg) motifs that flank the NPs [6, 7]. Meanwhile, signal peptides in N-terminal are important region which control the NP to the secretory pathway [8]. They are cleaved off during the translocation of NP through the endoplasmic reticulum membrane. The common feature of signal peptides is enrichment of hydrophobic residues.

The NP characterization depends on mass spectrometry which can provide high-quality data, but this approach is time-consuming and labor intensive [9–11]. As the complete genome sequence of many animals now becomes available, more effective and faster method is required to identify all potential NPs and their precursor. Several bioinformatics methods have been developed to identify NPPs based on sequence conservation traits [12, 13]. For most cases, due to the function of a particular peptide only depends on a short conserved motif, the peptide precursor sequence may show no significant sequence similarity [14]. In this study, we assume that specific monobasic, dibasic, or tribasic amino acid compositions which embody cleavage sites and signal peptides and other motifs will contribute to recognize NPPs. From this hypothesis, we aim to construct a predictor based on sequence compositions to identify NPPs and then provide a web server to make it easier to use.

✉ Jian Huang
hj@uestc.edu.cn

[1] Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 611731, China

[2] Key Laboratory for Neuroinformation of Ministry of Education, University of Electronic Science and Technology of China, Chengdu 611731, China

2 Materials and Methods

2.1 Data Sets

The NPP data set and protein unrelated to NPP (UnNPP) data set were requisite for training the model. The NPP data set was provided by SwissProt [15] and NeuroPedia [16]. We searched SwissProt with the "keyword" term "Neuropeptide (KW-0527)" but not "Receptor [KW-0675]" and collected all the results. We also downloaded all 270 human neuropeptide sequences from the NeuroPedia database. Then, the data from the two sources were merged into one data set. Those proteins whose sequence status is fragment were removed, due to they could be mature neuropeptide greatly.

For preparing a data set with high quality, the following procedure was executed: (1) Protein sequences including unclear residues ("B", "J", "X" etc.) were removed. (2) The CD-Hit software [17] was applied to keep the sequence similarity of each NPP sequence below 90%. (3) The protein should have a clear gene source, it means that the protein entry contains "GN" information.

We constructed candidate UnNPP pool through extracting UnNPPs from SwissProt by excluding the sequences related to neuropeptide. After excluding peptides containing ambiguous residues, the CD-Hit with identity of 0.9 was also performed. We randomly selected UnNPPs with the same number of NPP from candidate UnNPP pool as UnNPP data set. During the selection of UnNPP, it is ensured that the UnNPP data set has the same length distribution with the NPP data set.

2.2 Quantitative Features

Extracting a set of typical features is a crucial step in the process of pattern recognition. Single amino acid composition (AAC) [18], dipeptide composition (DPC) [19], and tripeptide composition (TPC) [20] have achieved excellent performances in the field of pattern classification. To establish the best model, each individual peptide sequence in data sets can be characterized by these three types of quantitative features. The AAC, DPC, and TPC defined as the following equations:

$$\mathrm{AAC}(i) = \frac{x(i)}{\sum_{i=1}^{20} x(i)}$$

$$\mathrm{DPC}(j) = \frac{y(j)}{\sum_{j=1}^{400} y(j)}$$

$$\mathrm{TPC}(n) = \frac{p(n)}{\sum_{n=1}^{8000} p(n)}$$

where i denote one of the 20 amino acids, j can be any one of the 400 dipeptides, and n represents one out of the 8000 tripeptides. $x(i)$, $y(j)$, and $p(n)$ are their counts in each sequence, respectively. Thus, each sequence in the data set is quantized by three feature encoding schemes, AAC, DPC, and TPC. We also constructed a combined peptide composition (CPC) including all 8420 features including AAC, DPC, and TPC. The selection of optimal combined peptide composition (OCPC) was accomplished as follows.

2.3 Selecting the Optimal Feature Set

In the model building process, existence of irrelevant and noisy features can result in poor model performance and increased computational complexity. To select the optimal reduced subsets, feature selection technique based on analysis of variance (ANOVA) [21] was performed. The following feature optimal steps [22] were conducted to construct OCPC against CPC: (1) sorted each feature based on F-score derived from ANOVA in descending order; (2) added a feature to the feature set one by one; (3) calculated accuracy for each new feature set using five-fold cross validation; and (4) selected the feature set with the highest accuracy as OCPC subset.

Based on the ANOVA theory, the significance of sequence compositions can be illustrated by calculating the F-score [23] which can be expressed by

$$F(\mu) = \frac{S_B^2(\mu)}{S_W^2(\mu)}$$

where (S_B^2) and (S_W^2) denote the inter-class and intra-class variance, respectively. They can be defined as

$$S_B^2(\mu) = \frac{1}{df_B} \sum_{i=1}^{K} m_i \left(\frac{\sum_{i=1}^{mi} f_\mu(i,j)}{m_i} - \frac{\sum_{i=1}^{K} \sum_{j=1}^{m_i} f_\mu(i,j)}{\sum_{i=1}^{K} m_i} \right)^2$$

$$S_W^2(\mu) = \frac{1}{df_W} \sum_{i=1}^{K} \sum_{j=1}^{m_i} \left(f_\mu(i,j) - \frac{\sum_{i=1}^{K} \sum_{j=1}^{m_i} f_\mu(i,j)}{\sum_{i=1}^{K} m_i} \right)^2$$

where df_B and df_W are degrees of freedom for sample variance between groups and within groups, defined as K-1 and M-K, respectively; K and M represent the number of groups and all samples; m_i stands for the number of samples in ith group; $f_\mu(i,j)$ means sequence composition frequency for the jth sample in the ith group; and μ ranges from 1 to 8420 for CPC. In our case, K and M are equal to 2 and 800, and both m_1 and m_2 are 300. The value of $F(\mu)$ shows the relevance between the μth feature and variable between groups. The

greater value of $F(\mu)$ is, the more importance it is to classify groups.

2.4 Constructing Support Vector Machine Models

Support vector machine (SVM) is an effective machine-learning methods for supervised pattern recognition, based on statistical learning theory, and has been widely used in the field of bioinformatics [24–32]. The basic idea of SVM is to map the low dimensional data into a high dimensional feature space through the kernel function, and then find the hyperplane with the largest separating distance between two groups. In general, four kernel functions, including radial basis function, polynomial function, sigmoid function, and linear function, will be selected to perform the prediction. Since the excellent effectiveness of radial basis function, we utilized it as kernel function in the current work. The two parameters as the kernel parameter γ and penalty parameter C were determined via grid search approach. In this report, the SVM model was implemented using the LibSVM software [33]. For the sake of the best optimal prediction, four models are trained with AAC, DPC, TPC, and OCPC, respectively.

2.5 Evaluating Performance

To evaluate the performance of model, four common metrics including sensitivity (Sn), specificity (Sp), accuracy (Acc), and Matthews correlation coefficient (MCC) were calculated and defined as follows:

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{TN} + \mathrm{FP}}$$

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

Here TP, FP, TN, and FN denote true positive, false positive, true negative, and false negative, respectively.

Actually, MCC is a correlation coefficient between the expectation and prediction. Its value varies between − 1 and + 1. The former represents an entirely opposite prediction, the latter indicates a perfect prediction, and 0 means no better than random prediction.

Area under receiver operating characteristic curve (ROC), named AUC was also applied to measure the quality of the binary classification. ROC is a graphical plot that indicates the performance of a two-class classifier as its probability threshold is varied. It relates sensitivity and 1-specificity. The machine-learning researchers usually use AUC for model comparison as its performance does not depend on the choice of the discrimination threshold. For good model performance, the AUC value should be close to 1, and a value of 0.5 means a random guess.

3 Results and Discussion

3.1 Collection of Data Set

The NPP and UnNPP data sets were constructed from SwissProt and NeuroPedia (Method 2.1). We removed repetitive sequences to construct a NPP union. After a serious of data cleaning, 407 NPPs were retained. The length distribution of NPP set showed that the lengths of majority NPPs were less than 500 aa, account for 96% (Fig. 1). To reduce the differences of sequence length, we removed the two NPPs longer than 1000 aa. The 405 UnNPPs less than 1000 aa was randomly selected to be UnNpp data set as described in "Materials and Methods" section.

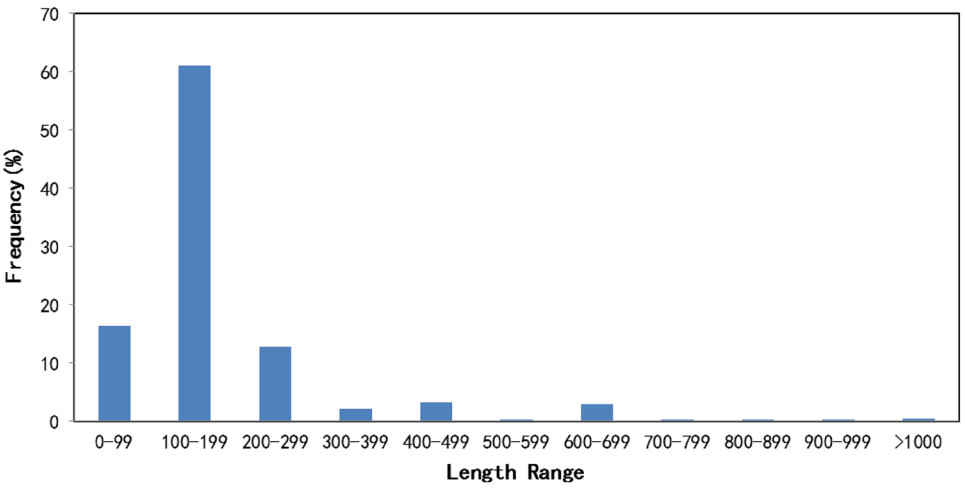

Fig. 1 Length distribution of the NPP set. The NPP set contains 407 NPPs. The lengths of majority NPPs were less than 500 aa account for 95.59%. Only two NPPs were longer than 1000 aa

Finally, the NPP and UnNPP data sets both have 405 sequences. For a comprehensive assessment, we divided all data into a training data set and an independent testing data set. We randomly selected 300 NPPs and 300 UnNPPs to constructed training data set. The testing data set was consisted of the rest 105 NPPs and 105 UnNPPs.

3.2 Optimization of Feature Set

Three feature encoding schemes were used in the current approach, including AAC, DPC, and TPC. In addition, each protein corresponds to a 20, 400, and 8000—dimension vector. The CPC feature set contains an 8420—dimension vector. As shown in Table 1, for five-fold cross validation, the model of SVM based on AAC reached accuracy of 88.83%; accordingly, those of DPC and TPC reached 91.83% and 93.66%. It shows that TPC-based models are superior to AAC and DPC-based models in the classification. In addition, the optimized reduced OCPC subset was obtained following the "Materials and Methods" section. The model based on OCPC feature set had the highest accuracy as 96.67% with the feature set which contains 1521 sequence compositions and achieved a better performance than NeuroPID [34]. Obviously, the feature selection technique can not only optimize the operation time, but also achieve better predictive performance.

3.3 Evaluation of Different Models

We applied four feature sets described in Table 1 to construct four models. Their performances were assessed in a rigorous way by the independence testing data set. No entry of the testing data set appeared in the training of the current model. The results are given in Table 2. The accuracy, MCC, and AUC obtained by OCPC are 88.62%, 0.78, and 0.95. They are slightly higher than corresponding values obtained by other models. Similarly, Fig. 2 shows that the AUC of red line which stands for OCPC model is higher than that of other three models. Finally, the OCPC feature set in training data set was chosen to construct model for further application.

Table 1 Accuracy of SVM-based models trained with different features via five-fold cross validation

Feature set	Feature size	Acc (%)
Single amino acid composition (AAC)	20	88.83
Dipeptide composition (DPC)	400	91.83
Tripeptide composition (TPC)	8000	93.66
Optimal combined peptide composition (OCPC)	1521	96.67

Table 2 Prediction performances of models with different feature sets for independent testing data set

Feature set	Sn (%)	Sp (%)	Acc (%)	MCC	AUC
AAC	89.52	84.76	87.14	0.7437	0.9403
DPC	83.81	90.48	87.14	0.7445	0.9453
TPC	83.81	92.38	88.10	0.7647	0.9452
OCPC	93.40	83.81	88.62	0.7759	0.9540

3.4 Analysis of Sequence Composition

We performed a feature analysis for sequence composition. Figure 3 shows a histogram for F-scores of AACs. The *x*-axis represents the 20 single amino acids, and the *y*-axis stands for the F-score for the corresponding AAC. As shown in Fig. 3, the residues I (Ile), V (Val), S (Ser), and R (Arg) have more variances between NPP and UnNPP. In those residues, I and V are hydrophobic residues which are enriched in signal peptide. In addition, R is the basic residue of cleavage sites [6].

The heat map analysis was also performed, as given in Fig. 4. The row of the heat map denotes the first amino acid of dipeptides, and the column represents the second one of that, respectively. Each square stands for one of the 400 dipeptide composition and the color is quantized according to its F-score. The features in blue boxes are different in NPPs and UnNPPs, while those in red boxes are the same in two classes. It was observed that most of the F-score for the dipeptide composition are near 0 (in red box), indicating that a large proportion of features is redundant and irrelevant for NPP predictions. The top four significant DPCs were LL

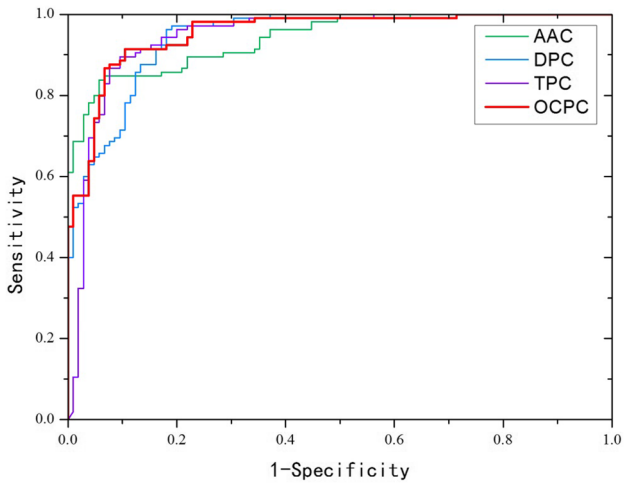

Fig. 2 ROC for models with different feature sets for independent testing data set. The curves with different colors stand for ROC of four SVM models. The AUC of red line which stands for OCPC model is 0.9540 and higher than that of other three models

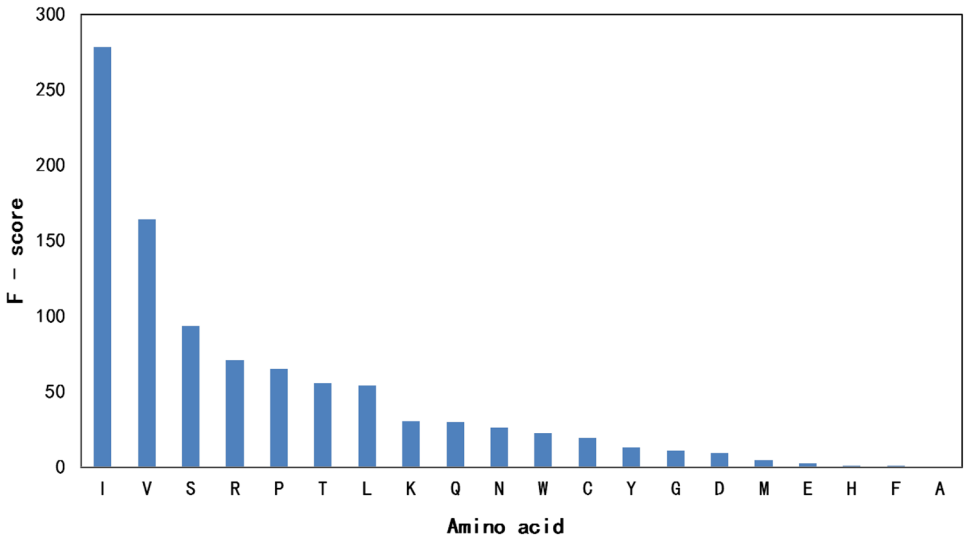

Fig. 3 Histogram for F-scores of 20 single amino acid compositions

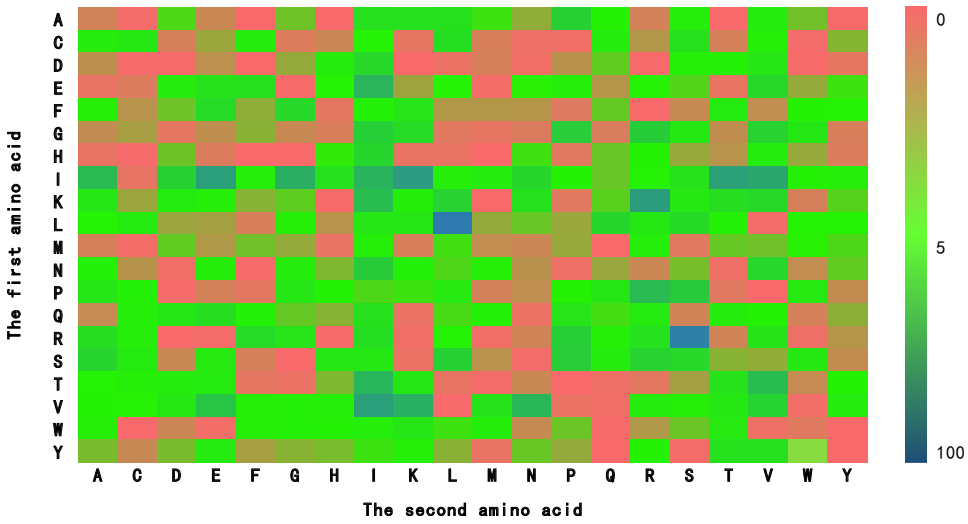

Fig. 4 Heat map for F-scores of 400 dipeptide compositions

(Leu–Leu), RS (Arg-Ser), IK (Ile-Lys), and KR (Lys-Arg). Interestingly, KR is apt to be cleavage site. For CPC features, the most significant feature was GKR (Gly-Lys-Arg), which is the most common known consensus cleavage site [6]. The significant features that are not basic residues may provide new idea about sequence characteristics in NPPs.

3.5 Web-Server Guide

For the convenience of other researchers, a web-server publicly accessible named NeuroPP has been developed. The web interface of NeuroPP was coded with Perl and is very friendly to use. The home page of web-server is shown in Fig. 5. In the prediction page, user can submit protein sequences in FASTA format in the textbox directly, or upload a local sequence file to the server. After clicking the predict button, the prediction results will be returned as an online table. The "view more" and "download" options can be chosen to obtain more information. The users should notice that the web server aims to recognize NPPs less than 1000 aa. At the result page, user can rank the results by length or probability to get more intuitive observation. The web server provides a useful interface to recognize unknown NPP.

4 Conclusions

To identify the NPPs from poorly annotated proteomes, several tools have been explored using machine-learning methods. However, only models and methods are far from satisfactory. User-friendly web servers or stand-alone programs are urgently needed. Dan et al. [34] had developed a predictor called NeuroPID to predict the NPPs from metazoan proteomes. This NeuroPID achieved 89–94% accuracy in cross validation, and this method did yield quite

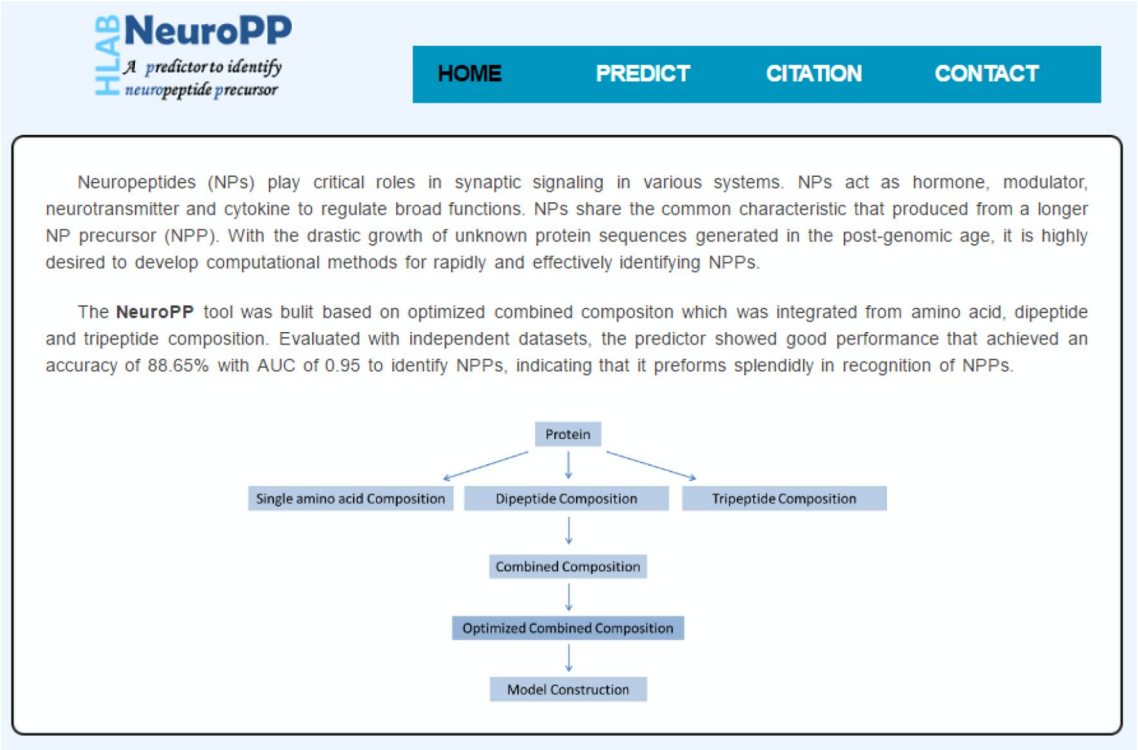

Fig. 5 Screenshot to show the home page of the NeuroPP web server

encouraging results. However, the prediction performance was not evaluated by an independent data set. Moreover, no online web server of the predictor is available now. In this study, we develop a predictor—NeuroPP to identify NPPs. In cross validation, NeuroPP achieved a higher accuracy with 96.67%. More than that, it showed better performance with accuracy of 88.62% and AUC of 0.9540 for an independent testing data set. The tripeptide composition which is not considered in NeuroPID may contribute to slightly increased accuracy for identifying NPPs. In brief, NeuroPP can perform splendidly in recognition NPPs, save time and cost for relevant experimental biologist, and improve the annotation of proteomes.

Acknowledgements The authors are grateful to the anonymous reviewers for their valuable suggestions and comments, which have led to the improvement of this paper. This work was supported by the National Natural Science Foundation of China [61571095] and the Fundamental Research Funds for the Central Universities of China [ZYGX2015Z006].

References

1. Brain SD, Cox HM (2006) Neuropeptides and their receptors: innovative science providing novel therapeutic targets. Br J Pharmacol 147(Suppl 1):S202–S211. https://doi.org/10.1038/sj.bjp.0706461
2. Insel TR, Young LJ (2000) Neuropeptides and the evolution of social behavior. Curr Opinion Neurobiol 10(6):784–789
3. Hokfelt T, Broberger C, Xu ZQ, Sergeyev V, Ubink R, Diez M (2000) Neuropeptides—an overview. Neuropharmacology 39(8):1337–1356
4. Funkelstein L, Beinfeld M, Minokadeh A, Zadina J, Hook V (2010) Unique biological function of cathepsin L in secretory vesicles for biosynthesis of neuropeptides. Neuropeptides 44(6):457–466. https://doi.org/10.1016/j.npep.2010.08.003
5. Jekely G (2013) Global view of the evolution and diversity of metazoan neuropeptide signaling. Proc Natl Acad Sci USA 110(21):8702–8707. https://doi.org/10.1073/pnas.1221833110
6. Rholam M, Brakch N, Germain D, Thomas DY, Fahy C, Boussetta H, Boileau G, Cohen P (1995) Role of amino acid sequences flanking dibasic cleavage sites in precursor proteolytic processing. The importance of the first residue C-terminal of the cleavage site. Eur J Biochem 227(3):707–714
7. von Eggelkraut-Gottanka R, Beck-Sickinger AG (2004) Biosynthesis of peptide hormones derived from precursor sequences. Curr Med Chem 11(20):2651–2665
8. von Heijne G (1990) The signal peptide. J Membr Biol 115(3):195–201
9. Svensson M, Skold K, Svenningsson P, Andren PE (2003) Peptidomics-based discovery of novel neuropeptides. J Proteome Res 2(2):213–219
10. Che FY, Biswas R, Fricker LD (2005) Relative quantitation of peptides in wild-type and Cpe(fat/fat) mouse pituitary using stable isotopic tags and mass spectrometry. J Mass Spectrom 40(2):227–237. https://doi.org/10.1002/jms.742
11. Baggerman G, Boonen K, Verleyen P, De Loof A, Schoofs L (2005) Peptidomic analysis of the larval Drosophila melanogaster central nervous system by two-dimensional capillary liquid

chromatography quadrupole time-of-flight mass spectrometry. J Mass Spectrom 40(2):250–260. https://doi.org/10.1002/jms.744
12. Loewenstein Y, Raimondo D, Redfern OC, Watson J, Frishman D, Linial M, Orengo C, Thornton J, Tramontano A (2009) Protein function annotation by homology-based inference. Genome Biol 10(2):207. https://doi.org/10.1186/gb-2009-10-2-207
13. Hummon AB, Richmond TA, Verleyen P, Baggerman G, Huybrechts J, Ewing MA, Vierstraete E, Rodriguez-Zas SL, Schoofs L, Robinson GE, Sweedler JV (2006) From the genome to the proteome: uncovering peptides in the Apis brain. Science 314(5799):647–649. https://doi.org/10.1126/science.1124128
14. Liu F, Baggerman G, Schoofs L, Wets G (2006) Uncovering conserved patterns in bioactive peptides in Metazoa. Peptides 27(12):3137–3153. https://doi.org/10.1016/j.peptides.2006.08.021
15. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bansal P, Bridge AJ, Poux S, Bougueleret L, Xenarios I (2016) UniProtKB/Swiss-Prot, the manually annotated section of the UniProt knowledgebase: how to use the entry view. Methods Mol Biol 1374:23–54. https://doi.org/10.1007/978-1-4939-3167-5_2
16. Kim Y, Bark S, Hook V, Bandeira N (2011) NeuroPedia: neuropeptide database and spectral library. Bioinformatics 27(19):2772–2773. https://doi.org/10.1093/bioinformatics/btr445
17. Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28(23):3150–3152. https://doi.org/10.1093/bioinformatics/bts565
18. Tang Q, Nie F, Kang J, Ding H, Zhou P, Huang J (2015) NIEluter: predicting peptides eluted from HLA class I molecules. J Immunol Methods 422:22–27. https://doi.org/10.1016/j.jim.2015.03.021
19. He B, Kang J, Ru B, Ding H, Zhou P, Huang J (2016) SABinder: a web service for predicting streptavidin-binding peptides. BioMed Res Int 2016:9175143. https://doi.org/10.1155/2016/9175143
20. Ding C, Yuan LF, Guo SH, Lin H, Chen W (2012) Identification of mycobacterial membrane proteins and their types using over-represented tripeptide compositions. J Proteomics 77:321–328. https://doi.org/10.1016/j.jprot.2012.09.006
21. Ding H, Feng PM, Chen W, Lin H (2014) Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis. Mol Biosyst 10(8):2229–2235. https://doi.org/10.1039/c4mb00316k
22. Ru B, t Hoen PA, Nie F, Lin H, Guo FB, Huang J (2014) PhD7Faster: predicting clones propagating faster from the Ph.D.-7 phage display peptide library. J Bioinform Comput Biol 12(1):1450005. https://doi.org/10.1142/S021972001450005X
23. Lin H, Ding H (2011) Predicting ion channels and their types by the dipeptide mode of pseudo amino acid composition. J Theor Biol 269(1):64–69. https://doi.org/10.1016/j.jtbi.2010.10.019
24. Lin H, Deng EZ, Ding H, Chen W, Chou KC (2014) iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. Nucleic Acids Res 42(21):12961–12972. https://doi.org/10.1093/nar/gku1019
25. Chen W, Feng PM, Deng EZ, Lin H, Chou KC (2014) iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. Anal Biochem 462:76–83. https://doi.org/10.1016/j.ab.2014.06.022
26. Ding H, Deng EZ, Yuan LF, Liu L, Lin H, Chen W, Chou KC (2014) iCTX-type: a sequence-based predictor for identifying the types of conotoxins in targeting ion channels. BioMed Res Int 2014:286419. https://doi.org/10.1155/2014/286419
27. Chen W, Feng P, Ding H, Lin H, Chou KC (2015) iRNA-Methyl: Identifying N(6)-methyladenosine sites using pseudo nucleotide composition. Anal Biochem 490:26–33. https://doi.org/10.1016/j.ab.2015.08.021
28. Chen W, Feng PM, Lin H, Chou KC (2014) iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition. BioMed Res Int 2014:623149. https://doi.org/10.1155/2014/623149
29. Zhu PP, Li WC, Zhong ZJ, Deng EZ, Ding H, Chen W, Lin H (2015) Predicting the subcellular localization of mycobacterial proteins by incorporating the optimal tripeptides into the general form of pseudo amino acid composition. Mol Biosyst 11(2):558–563. https://doi.org/10.1039/c4mb00645c
30. Tang H, Chen W, Lin H (2016) Identification of immunoglobulins using Chou's pseudo amino acid composition with feature selection technique. Mol Biosyst 12(4):1269–1275. https://doi.org/10.1039/c5mb00883b
31. Ding H, Lin H, Chen W, Li ZQ, Guo FB, Huang J, Rao N (2014) Prediction of protein structural classes based on feature selection technique. Interdiscip Sci Comput Life Sci 6(3):235–240. https://doi.org/10.1007/s12539-013-0205-6
32. Li N, Kang J, Jiang L, He B, Lin H, Huang J (2017) PSBinder: a web service for predicting polystyrene surface-binding peptides. BioMed Res Int 2017:5. https://doi.org/10.1155/2017/5761517
33. Chang C-C, Lin C-J (2011) LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol (TIST) 2(3):27
34. Ofer D, Linial M (2014) NeuroPID: a predictor for identifying neuropeptide precursors from metazoan proteomes. Bioinformatics 30(7):931–940. https://doi.org/10.1093/bioinformatics/btt725