BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 23 2013, pages 3094–3096
doi:10.1093/bioinformatics/btt518

Sequence analysis Advance Access publication September 23, 2013

CPPpred: prediction of cell penetrating peptides

Thérèse A. Holton[1,2,3,4], Gianluca Pollastri[1,5], Denis C. Shields[1,2,3,*] and Catherine Mooney[1,2,3]

[1]Complex and Adaptive Systems Laboratory, [2]Conway Institute of Biomolecular and Biomedical Science, [3]School of Medicine and Medical Science, [4]Food For Health Ireland and [5]School of Computer Science and Informatics, University College Dublin, Belfield, Dublin 4, Ireland

Associate Editor: Martin Bishop

ABSTRACT

Summary: Cell penetrating peptides (CPPs) are attracting much attention as a means of overcoming the inherently poor cellular uptake of various bioactive molecules. Here, we introduce CPPpred, a web server for the prediction of CPPs using a N-to-1 neural network. The server takes one or more peptide sequences, between 5 and 30 amino acids in length, as input and returns a prediction of how likely each peptide is to be cell penetrating. CPPpred was developed with redundancy reduced training and test sets, offering an advantage over the only other currently available CPP prediction method.

Availability and Implementation: CPPpred is freely available to non-commercial users at http://bioware.ucd.ie/cpppred.

Contact: Denis.Shields@ucd.ie

Supplementary Information: Supplementary data are available at Bioinformatics online.

Received on May 17, 2013; revised on July 29, 2013; accepted on August 15, 2013

1 INTRODUCTION

Cell penetrating peptides (CPPs) demonstrate the ability to pass through cellular membranes. Accordingly, CPPs are emerging as potential tools for aiding the transport of therapeutic molecules that are not inherently bioavailable, and moreover as possible bioactive agents in their own right (Johansson *et al*., 2007). Although a number of broad classification systems for CPPs have been proposed (Heitz *et al*., 2009; Madani *et al*., 2011, Milletti, 2012), in general, CPPs are typified by a length of 5–30 amino acids, an enrichment in basic residues, therefore a net positive charge, and are hydrophobic in part (Lundberg and Langel, 2003). Several *in silico* CPP prediction algorithms have been developed to facilitate high-throughput screening of peptides. Initially, methods focused on the use of scales of chemical properties, known as *z*-descriptors (Sandberg *et al*., 1998), for the prediction of CPPs (Hällbrink *et al*., 2005). Other approaches applied to CPP prediction include quantitative structure–activity relationship models (Dobchev *et al*., 2010) and support vector machines (Gautam *et al*., 2013; Sanders *et al*., 2011), although only one of these methods, CellPPD (Gautam *et al*., 2013), is publicly available. Here, we present CPPpred, a CPP prediction method that uses a neural network (NN) approach. Although others have previously applied NN techniques to the prediction of CPPs (Dobchev *et al*., 2010), CPPpred represents the only publicly available NN-based method.

2 METHODS

2.1 Training and Test Datasets

We generated our training dataset using 111 CPPs and 34 experimentally validated non-CPPs from Sanders *et al*. (2011). We selected peptides 5–30 amino acids in length and internally redundancy reduced these sets independently to <80% sequence similarity using BLAST (Altschul *et al*., 1997), leaving 74 CPPs and 30 non-CPPs. We supplemented the negative set with 70 peptides randomly selected from a 30% internally redundancy reduced set of 1445 bioactive peptides of 5–30 amino acids in length retrieved from BIOPEP (Dziuba *et al*., 1999), PeptideDB (Liu *et al*., 2008), CAMP (Thomas *et al*., 2010) and APD2 (Wang *et al*., 2009). Only peptides that did not have the same sequence as a CPP in our training set and were not a subsequence of a CPP, or vice versa, were retained (Supplementary Table S1).

2.2 Independent Test Datasets

In all, 843 CPPs were retrieved from CPPsite (Gautam *et al*., 2012). From this set we selected peptides of 5–30 amino acids in length, labeled as having either ‘Greater’, ‘High’ or ‘Higher’ uptake efficiency (according to the classification of CPPsite) and redundancy reduced this set with respect to our training set to <80% sequence similarity. This left only 47 CPPs from the larger dataset, which we matched with 47 randomly selected bioactive peptides, as described previously, to create our negative set (Supplementary Table S1).

2.3 Predictive Architecture

N-to-1 neural networks, or N1-NNs, have been successfully used to predict the subcellular location of protein sequences (Mooney *et al*., 2011) and bioactive peptides (Mooney *et al*., 2012). Here, we apply this model to the prediction of CPPs. The aim of the model is to map a peptide sequence of variable length *N* into a single property i.e. CPP or non-CPP. Other models tackle this problem at the source, i.e. they transform/compress the sequence into a fixed number of descriptors (or into descriptors of pairwise relations between sequences) beforehand, and then map these descriptors into the property of interest. These descriptors are typically frequencies of residues or k-mers, sometimes computed separately on different parts of the sequence (Gautam *et al*., 2012; Sanders *et al*., 2011). In N1-NNs we do not compress the peptide in advance; instead, we decide beforehand only how many features we want to compress a peptide into.

N1-NNs are particularly suited to the problem of CPP prediction, as they do not rely on frequency counts, but look at whole motifs, and do so using a small number of free parameters. Therefore, they have an

*To whom correspondence should be addressed.

advantage over frequency-based methods, in that they incorporate information about relative positions of amino acids in a sequence, whilst also minimizing the risk of overfitting. All possible overlapping motifs in a sequence are processed by the same basic unit, using the same parameters. For this reason, they are advantageous over algorithms that look at a fixed input window, as they overcome the problem of choosing an arbitrary reference frame for peptides of different sizes. They also have a potential advantage over methods that do not rely on fixed windows (e.g. support vector machines (SVM) based on spectral or alignment kernels), in that N1-NNs learn a transform of the sequence that is informative to predict the target property, rather than relying on any predefined transform or set thereof. If training is successful, the feature vector will retain what is useful in the input and discard what is not. This is subtly different from compressing the sequence into some code first (potentially erasing useful information) and then optimizing the map between the code and the target.

2.4 Training and Ensembling

Each training was conducted in 5-fold cross-validation, i.e. five different sets of training runs were performed in which a different fifth of the overall set was reserved for testing and another fifth was reserved for validating. During training the networks that performed best on the validation set were saved; these models were then averaged over the ensemble and evaluated on the corresponding test set. The final result for the 5-fold cross-validation is the average of the results on each test set. When testing on the independent test set, we ensemble-combined *all* the models from all cross-validation folds of the best architecture.

3 RESULTS AND DISCUSSION

The only other publicly available CPP predictor that we are aware of is CellPPD (Gautam *et al*., 2013). Owing to the short nature of CPPs and the impact that single residue changes can have on the efficiency of penetration, or non-penetration, the authors did not internally redundancy reduce their training set, or redundancy reduce their independent test sets with respect to their training set. We were concerned that this could have potentially led to overfitting of their models. Because the datasets of penetrating peptides are relatively dominated by cationic and amphiphilic peptides, we considered that the redundancy reduction to avoid overfitting would eliminate most of the data. However, we were pleased to note that training was still possible even after 80% redundancy reduction (i.e. no more than 80% identity between any two sequences) (see Table 1). The predictive power of CPPpred is demonstrated in Figure 1. We note that the performance is good on both the redundancy reduced training set (Fig. 1a) and independent test set (Fig. 1b), which was redundancy reduced with respect to the training set, indicating that our method is not showing very obvious signs of extensive overfitting, despite the relatively small training set of 74 CPPs.

More extensive redundancy reduction in the training set would be likely to reduce the ability to train a useful predictor (Supplementary Material). However, it is worth noting that after 30% redundancy reduction of the test set, CPPpred correctly predicted 17 of the 29 CPPs and 27 of the 29 non-CPPs (Q = 75.86%).

3.1 Implementation

CPPpred has been implemented as a web server. The user submits a list of peptides, and CPPpred predicts the probability that each of these peptides will be cell penetrating. When the user is interpreting the results, it is important to note that the server predicts how likely the peptide is to be cell penetrating, and does not predict the degree of uptake efficiency. CPPpred was trained at a threshold of 0.5, i.e. any peptide predicted over a 0.5 threshold is labeled as cell penetrating. Roughly, one-fifth of peptides above 0.5 are false positives, whereas less than a tenth of peptides above 0.7 are false positives (Fig. 1a). However, the user may decide to choose a higher threshold to reduce the false-positive rate. Training and testing was performed using amino acid sequences without any modifications. After using CPPpred to discover CPPs *in silico*, various peptide modifications may be experimented with *in vitro* to increase the uptake efficiency of the peptide.

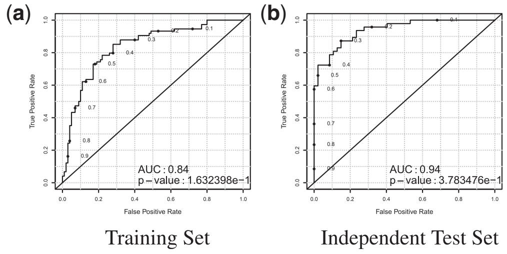

Fig. 1. Receiver operating characteristic plot of CPP predictor performance on (**a**) the 80% redundancy reduced training set in 5-fold cross-validation and (**b**) the independent test set that is 80% redundancy reduced with respect to the training set

Table 1. Results for training and independent test sets

	Training set				Independent test set			
	Spec	Sen	MCC	FPR	Spec	Sen	MCC	FPR
Non-CPP	81.02	81.00	0.54	27.05	75.41	97.87	0.69	31.91
CPP	73.82	72.95	0.54	19.00	96.97	68.09	0.69	2.13
		Q	77.60			Q	82.98	

Note: Specificity (spec), sensitivity (sen), matthews correlation coefficient (MCC) and false positive rate (FPR) (see Supplementary Material for more details).

3.2 CPPpred—a worked example

CPPpred provides researchers interested in bioactive peptides and nutraceuticals with an important resource for evaluating the cell penetrating potential of candidate peptides of interest. To demonstrate the application of CPPpred to empirical data, we selected a known bioactive peptide, lactoferricin B, from bovine milk for analysis. Lactoferricin B, consisting of 25 amino acids, exhibits antimicrobial activity (Bellamy *et al*., 1993) and has been shown experimentally to be cell penetrating (Haukland *et al*., 2001). Importantly, this peptide (or any associated subsets) is not found in either our training or test sets. When CPPpred is applied to lactoferricin B, a prediction score of 0.725 is achieved, demonstrating the predictive power of CPPpred.

Further to this, an 11 amino acid subsequence of lactoferricin B, also with known antimicrobial activity (Strøm *et al.*, 2002), attains a CPPpred score of 0.815. Although the cell penetrating capabilities of this undecapeptide have not yet been validated experimentally, it is widely appreciated that cell penetration is a feature of antimicrobial peptides (Splith and Neundorf, 2011). We feel that this example highlights the potential utility of CPPpred in assisting researchers in the selection of peptides for chemical synthesis to experimentally verify their cell penetrating potential.

Funding: Science Foundation Ireland (08/IN.1/B1864, 10/RFP/GEN2749) and by Enterprise Ireland (CC20080001) (Food for Health Ireland).

Conflict of Interest: none declared.

REFERENCES

Altschul,S. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Bellamy,W. *et al.* (1993) Killing of candida albicans by lactoferricin b, a potent antimicrobial peptide derived from the n-terminal region of bovine lactoferrin. *Med. Microbiol. Immun.*, **182**, 97–105.

Dobchev,D. *et al.* (2010) Prediction of cell-penetrating peptides using artificial neural networks. *Curr. Comput. Aided Drug. Des.*, **6**, 79–89.

Dziuba,J. *et al.* (1999) Database of biologically active peptide sequences. *Nahrung*, **43**, 190–195.

Gautam,A. *et al.* (2013) In silico approaches for designing highly effective cell penetrating peptides. *J. Transl. Med.*, **11**, 1–12.

Gautam,A. *et al.* (2012) CPPsite: a curated database of cell penetrating peptides. *Database*, **2012**: bas015.

Hällbrink,M. *et al.* (2005) Prediction of cell-penetrating peptides. *Int. J. Pept. Res. Ther.*, **11**, 249–259.

Haukland,H. *et al.* (2001) The antimicrobial peptides lactoferricin b and magainin 2 cross over the bacterial cytoplasmic membrane and reside in the cytoplasm. *FEBS Lett.*, **508**, 389–393.

Heitz,F. *et al.* (2009) Twenty years of cell-penetrating peptides: from molecular mechanisms to therapeutics. *Brit. J. Pharmacol.*, **157**, 195–206.

Johansson,H.J. *et al.* (2007) Characterization of a novel cytotoxic cell-penetrating peptide derived from p14ARF protein. *Mol. Ther.*, **16**, 115–123.

Liu,F. *et al.* (2008) The construction of a bioactive peptide database in metazoa. *J. Proteome. Res.*, **7**, 4119–4131.

Lundberg,P. and Langel,Ü. (2003) A brief introduction to cell-penetrating peptides. *J. Mol. Recognit.*, **16**, 227–233.

Madani,F. *et al.* (2011) Mechanisms of cellular uptake of cell-penetrating peptides. *J. Biophys.*, **2011**, 414729.

Milletti,F. (2012) Cell-penetrating peptides: classes, origin, and current landscape. *Drug Discov. Today*, **17**, 850–860.

Mooney,C. *et al.* (2011) SCLpred: protein subcellular localization prediction by N-to-1 neural networks. *Bioinformatics*, **27**, 2812–2819.

Mooney,C. *et al.* (2012) Towards the improved discovery and design of functional peptides: common features of diverse classes permit generalized prediction of bioactivity. *PLoS One*, **7**, e45012.

Sandberg,M. *et al.* (1998) New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. *J. Med. Chem.*, **41**, 2481–2491.

Sanders,W.S. *et al.* (2011) Prediction of cell penetrating peptides by support vector machines. *PLoS Comput. Biol.*, **7**, e1002101.

Splith,K. and Neundorf,I. (2011) Antimicrobial peptides with cell-penetrating peptide properties and vice versa. *Eur. Biophys. J.*, **40**, 387–397.

Strøm,M.B. *et al.* (2002) The effects of charge and lipophilicity on the antibacterial activity of undecapeptides derived from bovine lactoferricin. *J. Pept. Sci.*, **8**, 36–43.

Thomas,S. *et al.* (2010) CAMP: a useful resource for research on antimicrobial peptides. *Nucleic Acids Res.*, **38**, D774.

Wang,G. *et al.* (2009) APD2: the updated antimicrobial peptide database and its application in peptide design. *Nucleic Acids Res.*, **37**, D933.