BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–7

PeptideLocator: Prediction of Bioactive Peptides in Protein Sequences

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

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Peptides play important roles in signalling, regulation, and immunity within an organism. Many have successfully been used as therapeutic products often mimicking naturally occurring peptides. Here we present PeptideLocator for the automated prediction of functional peptides in a protein sequence.

Results: We have trained a machine learning algorithm to predict bioactive peptides within protein sequences. PeptideLocator performs well on training data achieving an AUC of 0.92 when tested in five-fold cross-validation on a set of 2202 redundancy reduced peptide containing protein sequences. It has predictive power when applied to antimicrobial peptides, cytokines, growth factors, peptide hormones, toxins, venoms and other peptides. It can be applied to refine the choice of experimental investigations in functional studies of proteins.

Availability: PeptideLocator is freely available for academic users at http://bioware.ucd.ie/.

Contact: denis.shields@ucd.ie

1 INTRODUCTION

Peptides, which are often generated by cleavage from precursor proteins, are important molecules found in all organisms and play a role in a wide range of biological processes. It is this universal distribution, along with the diversity of functions that they undertake, that make peptides interesting as putative therapeutic agents, for example, as antimicrobials (Hancock and Sahl, 2006; Fjell *et al.*, 2012) or analgesics (Diochot *et al.*, 2012). Peptides can function as hormones and immunomodulating agents, through interactions with cytokines, receptors and other signalling proteins (Möller *et al.*, 2008), and have attracted significant interest with recent appreciation of their ability to act as inhibitors or antagonists of protein-protein interaction networks (Boonen *et al.*, 2009). Other diverse functions include quorum sensing in the regulation of gut bacteria populations (Swift *et al.*, 2000); inhibition or activation of enzymes (e.g. ACE inhibitors), membrane-bound protein channels, transporters and receptors by toxin or venom peptides (Lewis *et al.*, 2003); and the disruption of cell membranes to facilitate the destruction of microbial cells by antimicrobial peptides.

*To whom correspondence should be addressed

While most bioactive peptides have been defined with a view to being therapeutic, not all are so, including for example toxins produced by one organism that may limit the growth of another organism, or peptides found to inhibit an enzyme in an in vitro assay. The definition of "bioactive" used here is a general catch-all term for peptides having biological effects, and is representative of the concept that there may be some commonalities shared among diverse bioactive peptides derived from larger precursor proteins.

Several health benefits have been associated with bioactive peptides, particularly those derived from food sources such as plants and bovine milk (Clare and Swaisgood, 2000). They have been shown to be functionally active, from simple dipeptides in ACE inhibition (Norris *et al.*, 2012), to larger milk peptides important in immunogenicity and nutrition in early development (Newburg and Walker, 2007). Specific health promoting properties of food peptides include: antithrombotic, antimicrobial, mineral binding, opiod, immune and cytomodulatory, and blood pressure reduction, with many peptides exhibiting more than one function (Hartmann and Meisel, 2007). Functional additives in food is an increasingly interesting area of development given the potential to deliver improvements in diet, health and resistance to infection. Peptides offer a naturally derivable source of such functional foods, and as a result there has been considerable interest in the identification of bioactive peptides in food products, particularly milk (Clare and Swaisgood, 2000). Several peptides have already been commercialised as nutraceuticals (Korhonen and Pihlanto, 2006) demonstrating the need for more methods for the detection and characterisation of novel bioactive peptides in this area (Khaldi, 2012).

Many databases have been created cataloguing instances of bioactive peptides. These include PepBank (Shtatland *et al.*, 2007), PeptideDB (Liu *et al.*, 2008), BIOPEP (Dziuba *et al.*, 1999) and the antimicrobial peptide databases, APD2 (Wang *et al.*, 2009) and CAMP (Thomas *et al.*, 2010). The peptide activity classes found in PeptideDB and BIOPEP include antimicrobial peptides, cytokines and growth factors, peptide hormones and toxin/venom peptides, whereas APD2 and CAMP are limited to antimicrobial peptides such as antiviral, antifungal, antibacterial and antiparasitic peptides. Such databases have facilitated the bioinformatic analysis of peptides promoting the creation of tools to aid in the identification of bioactive peptides.

The identification of bioactive peptides within protein sequences is the first step in the development of a therapeutic product. In food-based protein precursors bioactive peptides can be isolated by bacterial fermentation, or through the use of gastrointestinal or alternative proteolytic enzymes (Korhonen and Pihlanto, 2006; Vijayakumar *et al.*, 2012). *In silico* strategies used to detect novel bioactive peptides include the examination of sequence and structure homology as well as the emerging approaches of peptidomics (Sasaki *et al.*, 2010; Khaldi, 2012). Homology based prediction of bioactivity has successfully identified many instances of antimicrobial peptides (Lynn *et al.*, 2004). Other bioinformatic methods have been used to predict peptide modulators of human platelet function (Edwards *et al.*, 2007). As in many applications of bioinformatics, tools for the prediction of bioactivity have employed propensity scales which capture information about the statistical preferences of different amino acids to be bioactive peptides or not (Torrent *et al.*, 2011). Other prediction tools have been based on Support Vector Machines (SVM) (Lata *et al.*, 2010; Thomas *et al.*, 2010), hidden Markov models (HMMs) (Fjell *et al.*, 2007) or sequence alignments and feature selection (Wang *et al.*, 2011). Most tools focus on a particular class of bioactive peptide prediction, for example antimicrobial peptides, however, PeptideRanker (Mooney *et al.*, 2012), which has been developed recently, is a general bioactive peptide predictor. These tools predict the bioactivity of stand-alone peptide sequences but they are not optimised for the discovery of bioactive peptides within larger parent protein sequences, ignoring the protein context of the bioactive peptide which may provide additional information.

Here, we have combined many classes of bioactive peptides that are functionally distinct, but nevertheless have shared properties. Using the precursor protein sequences of these peptides, retrieved from PeptideDB, we have trained bidirectional recursive neural networks (BRNN), which take as input the protein sequence and a number of predicted structural features of the protein, and predict bioactive peptides within the protein sequence. As far as we are aware only one other method exists for the identification of potential bioactive peptide regions within a protein sequence and this method is specialised for the prediction of antimicrobial peptides (Torrent *et al.*, 2012). We have tested the performance of PeptideRanker on an independent test set of antimicrobial peptides from APD2 and followed up with an analysis of 661 food proteins from the BIOPEP database.

2 METHODS

2.1 Training and Test Datasets

Proteins with bioactive peptides of length less than or equal to 50 residues were retrieved from PeptideDB (Liu *et al.*, 2008). From 6,145 peptide precursor protein sequences we generated two internally redundancy reduced sets: PeptideDB.90 which was redundancy reduced to less than 90% sequence similarity leaving 4,505 sequences; and PeptideDB.30 which was redundancy reduced to less than 30% sequence similarity leaving 2,202 protein sequences. Fig. S1 shows the length of the bioactive peptides plotted against the length of the precursor proteins in both of these datasets.

To generate the PeptideLocator training dataset we split the PeptideDB.30 dataset into five equally sized test sets. The training set for each test set was created by running a BLAST search (Altschul *et al.*, 1997) (e-value $<$ 0.001) of each sequence in the PeptideDB.90 dataset against that test set. If a hit was found then that sequence was removed. This resulted in five 90% internal redundancy reduced training sets with no BLAST hit with an e-value $<$ 0.001 to any sequence in the corresponding test set. In summary, no sequence in a test set had $>$ 30% sequence similarity to any other sequence in that or any other test set and no sequence was used in training a model of the network that has a BLAST hit with an e-value $<$ 0.001 to a sequence used for testing that model.

	PeptideDB.30	PeptideDB.90	APD2.30
Non-bioactive residues	62,734	192,744	302
Bioactive residues	45,139	112,794	509
Total residues	107,873	305,538	811
Total sequences	2,202	4,505	15

Table 1. The number of protein sequences and the number of residues per class and in total for training and test datasets, and the independent test dataset.

Each residue of every sequence in the datasets was labelled as either a bioactive peptide residue or non-bioactive peptide residue (Table 1). Secondary structure, solvent accessibility, structural motifs and disorder were predicted for the full protein sequence using Porter (Pollastri and McLysaght, 2005), PaleAle (Pollastri *et al.*, 2007), Porter+ (Mooney *et al.*, 2006) and IUPred (Dosztányi *et al.*, 2005) respectively. Each residue was labelled with the predicted probability of being in each of three classes in the case of secondary structure, four classes for solvent accessibility, fourteen classes for structural motifs and a single label representing the predicted probability of the residue being disordered. We included solvent accessibility, secondary structure and structural motifs as they provide information on the potential structure of the peptide. We included predicted disorder as this may provide information not only on the peptides structural state, but also the context of the peptide which may facilitate biological release by proteolytic cleavage or biological activity as a self-contained signalling motif unconstrained by adjacent ordered regions.

The sequences were then split into domain/non-domain sections using SMART (Letunic *et al.*, 2009). The input to the BRNN for each domain/non-domain protein section was the length of the section, the bioactive/non-bioactive labels per residue, the three predicted structural features (secondary structure, solvent accessibility and structural motifs), the predicted disorder per residue, and a single extra input representing if the section was, or was not, predicted to be a domain by SMART.

2.2 Independent Test Datasets

To create an independent test set we retrieved any peptide from the Antimicrobial Peptide Database (APD2) (Wang *et al.*, 2009) which had information on the original protein sequence from which it was derived. We then searched Swiss-Prot (The UniProt Consortium, 2012) for the protein sequence. If the protein sequence was not longer than the peptide sequence it was discarded. This resulted in 132 unique protein sequences. We redundancy reduced this set to less than 30% sequence similarity to any sequences in PeptideDB.90. This left us with 15 protein sequences. Table 1 shows the number of residues per class (bioactive/non-bioactive).

2.3 Predictive Algorithms and Implementation

We use a BRNN to learn the mapping between inputs $\mathcal{I}$ and outputs $\mathcal{O}$ (protein sequence to bioactive peptide residue). BRNN have been used successfully for the prediction of secondary structure (Pollastri and McLysaght, 2005), solvent accessibility (Pollastri *et al.*, 2007) and structural motifs (Mooney *et al.*, 2006) amongst other things and have the advantage over standard feed-forward neural networks that they can automatically find the optimal context on which to base a prediction, i.e. the number of residues that are informative to determine a property. Because of their recursive nature, BRNN also have a relatively low number of free parameters

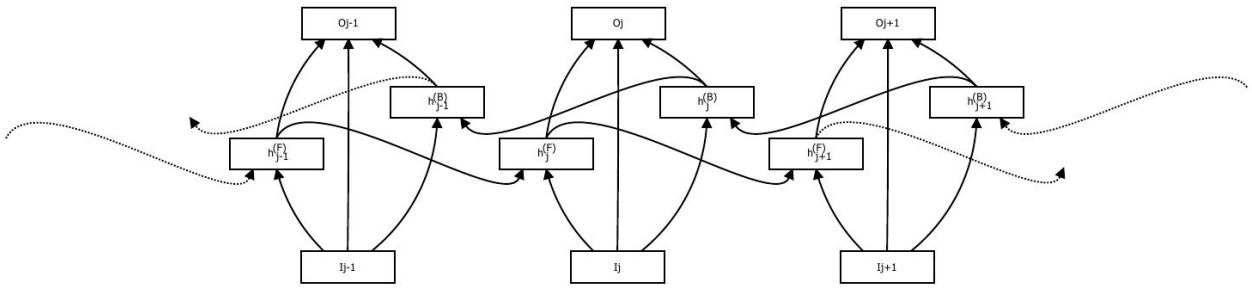

Fig. 1: **Graphical representation of a BRNN.**

Rectangles represent input, hidden and output vectors. Arrows represent functional dependencies, for example o_j is a function of i_j, $h_j^{(F)}$ and $h_j^{(B)}$; $h_j^{(F)}$ is a function of i_j and $h_{j-1}^{(F)}$; etc. Terminal states $h_0^{(F)}$ and $h_{N+1}^{(F)}$ (not represented) complete the graphical model. Notice that any input can, in principle, affect any output.

compared to other neural networks with similar input size. See Baldi *et al.* (1999) for a detailed explanation of the BRNN model and Fig. 2.3 which illustrates the topology.

These networks take the form:

$$\begin{aligned} o_j &= \mathcal{N}^{(O)}\left(i_j, h_j^{(F)}, h_j^{(B)}\right) \\ h_j^{(F)} &= \mathcal{N}^{(F)}\left(i_j, h_{j-1}^{(F)}\right) \\ h_j^{(B)} &= \mathcal{N}^{(B)}\left(i_j, h_{j+1}^{(B)}\right) \\ & j = 1, \ldots, N \end{aligned}$$

where i_j (respectively o_j) is the input (respectively output) of the network in position j, and $h_j^{(F)}$ and $h_j^{(B)}$ are forward and backward chains of hidden vectors with $h_0^{(F)} = h_{N+1}^{(B)} = 0$. We parametrise the output update, forward update and backward update functions (respectively $\mathcal{N}^{(O)}$, $\mathcal{N}^{(F)}$ and $\mathcal{N}^{(B)}$) using three two-layered feed-forward neural networks.

Encoding sequence and structural information Input i_j associated with the j-th residue contains primary sequence information and predicted structural information:

$$i_j = (i_j^{(E)}, i_j^{(T)})$$

where, assuming that e units are devoted to sequence, and t to structural information:

$$i_j^{(E)} = (i_{j,1}^{(E)}, \ldots, i_{j,e}^{(E)})$$

and:

$$i_j^{(T)} = (i_{j,1}^{(T)}, i_{j,t}^{(T)})$$

Hence i_j contains a total of $e + t$ components.

We use $e = 26$: beside the 20 standard amino acids, B (aspartic acid or asparagine), U (selenocysteine), X (unknown), Z (glutamic acid or glutamine) and . (gap) are considered, while the 26th input encodes the length of the sequence. The frequency of gaps in a column provides information about the conservation of a site, and has proven helpful in preliminary tests. We use $t = 23$ for representing structural information. The first three structural input units contain the predicted three-class secondary structure representing the predicted probability of the j-th residue belonging to either helix, strand or coil. The next four input units contain the predicted probability of the j-th residue belonging to one of four solvent accessibility classes. The following 14 input units contain the predicted probability of the j-th residue belonging to one of 14 structural motifs classes, and the final two inputs are the predicted probability of the residue occurring in a disordered region or a predicted domain. Hence the total number of inputs for a given residue is $e + t = 49$. The output is the predicted probability of the j-th residue belonging to a bioactive peptide.

Training, Ensembling Training is conducted by five-fold cross-validation, i.e. five sets of training are performed in which a different fifth of the overall set is reserved for testing. Before being split into the five test sets the test dataset was sorted by PeptideDB ID. We then picked instance 1 for test set 1, instance 2 for test set 2 .. etc., in an interleaved or stratified manner. As the PeptideDB IDs tend to be clustered by peptide function this ensures that all types of peptide function are distributed similarly across and present in each of the test sets.

The training set is used to learn the free parameters of the network by gradient descent. Five models are trained independently for each of the five folds i.e. 25 models in total. Differences among models in each fold are introduced by varying the architectural parameters of the network. 250 pass through the entire training set (epochs) of training are performed for each model and the learning rate (which controls how fast the algorithm converges) is halved every time we do not observe a reduction of the error for more than 50 epochs.

By the end of training 22 of the 25 models of the network have errors of less than 5%. We can therefore assume that the networks are converging to find good local optima. To test the final predictor we ensemble the five models in each of the five folds by averaging their results on their respective test sets, and then combining these five results to get the overall five-fold cross-validation result. To build the final predictor and when testing on an entirely independent set from the one used during training we ensemble-combine (average) the results across all 25 models in one step.

2.4 Evaluating Performance

To evaluate the performance of PeptideLocator we measure the true positive rate (TPR) and false positive rate (FPR) as we increase the discrimination threshold from 0 to 1. The results are shown as a Receiver Operating Characteristic (ROC) curve where TPR is plotted against FPR, which are

calculated as follows:

$$TPR = \frac{TP}{TP+FN}$$
$$FPR = \frac{FP}{FP+TN}$$

where:

- True positives (TP): the number of residues predicted in a class that are observed in that class.
- False positives (FP): the number of residues predicted in a class that are not observed in that class.
- True negatives (TN): the number of residues predicted not to be in a class that are not observed in that class.
- False negatives (FN): the number of residues predicted not to be in a class that are observed in that class.

The area under the curve, AUC, which is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance (Fawcett, 2006) is also shown. The AUC is a number between 0 and 1 where 0.5 indicates a random model and 1 is perfect. R is used to plot the curves and calculate the AUC (R Development Core Team, 2008). Specificity (Spec), sensitivity (Sen), Matthews Correlation Coefficient (MCC) and the accuracy (Q) at a 0.5 threshold are measured as follows:

$$Spec = 100\frac{TP}{TP+FP}$$
$$Sen = 100\frac{TP}{TP+FN}$$
$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$
$$Q = 100\frac{TP+TN}{TP+TN+FP+FN}$$

MCC measures the correlation coefficient between the observed and predicted classifications. A value of 1 represents a perfect prediction, 0 a random prediction and -1 an inverse prediction and is a good indicator of the overall performance of the predictive methods for both bioactive peptide and non-bioactive peptide residues.

3 RESULTS AND DISCUSSION

For every residue PeptideLocator predicts the probability between 0 and 1 of that residue being part of a bioactive peptide. The closer the probability is to 1, the more confident PeptideLocator is that the residue is part of a bioactive peptide. The results in five-fold cross-validation for the test dataset, PeptideDB.30, are shown in Table 2 and in Fig. S2(A) as a ROC curve with thresholds from 0 to 1, i.e. the cut-off above which a residue is considered to be predicted as a bioactive peptide residue. Specificity, sensitivity and the accuracy (Q) are over well 80%, with an AUC of 92% and MCC of 0.71 showing a strong correlation between predicted and the observed classifications.

We chose secondary structure, solvent accessibility, structural motifs and disorder as additional inputs into the network as we considered these factors may contribute to the identification of a bioactive peptide within a protein sequence. We trained the networks leaving out each of these in turn and found that the results were similar in all cases (Q 85.1 - 85.8%, AUC 0.91 - 0.92) except for when solvent accessibility was excluded (Q 80%, AUC 0.87). This would seem to suggest that solvent accessibility contributes most to the accuracy of the BRNN, however a model of the network trained only including solvent accessibility resulted in a Q of 83.7%, and an AUC of 0.90, compared to a Q of 85.8 and an AUC of 0.92 when all four features are included. Given this contribution to performance, we chose to retain all four features in the predictor. Unfortunately, there is no easy way to reverse engineer how the model decides upon a prediction because it compresses the input sequence into many hidden states/features for each residue, which do not correspond directly to any known property of the protein.

	Spec	Sen	MCC	FPR	Q	AUC
PeptideDB.30	83.50	82.33	0.71	11.70	85.80	0.92
APD2.30	83.39	90.77	0.63	30.36	82.86	0.89

Table 2. Results for PeptideLocator on the PeptideDB.30 test dataset and the APD2.30 independent test datasets at a threshold of 0.5.

Although the APD2.30 dataset is very small (15 sequences), and is restricted to antimicrobial peptides only, we can see from Table 2 that PeptideLocator is able to predict which class (bioactive/non-bioactive) residues fall into with an accuracy of over 82%. 90% of the bioactive peptide residues are predicted correctly. However, the FPR of 30.36% (i.e. the percentage of non-bioactive peptide residues incorrectly predicted as bioactive) is high. This could be corrected for example by increasing the threshold to 0.8 which would decrease the FPR to approximately 20% without reducing the TPR (see ROC curve Fig. S2(B)). For this reason we have chosen a threshold of 0.8 for the food protein examples which we present below. Improved larger databases of bioactive peptides linked to precursor proteins will permit more accurate evaluation of the performance of the method we have presented, in addition to providing larger training sets for future methods development.

3.1 Peptide Activity Classes

To assess the ability of PeptideLocator to predict bioactive peptides in other classes (not only antimicrobial peptides as in APD2.30) we split the PeptideDB.30 results into six peptide activity subsets: antifreeze, antimicrobial, cytokines and growth factor, peptide hormones, toxins and venom, and unique/other. The number of residues per class are shown in Table 3 and the results are shown as ROC curves (Fig. 2). The classes for which PeptideLocator is most accurate are antifreeze, antimicrobial and toxins and venom, which all have AUCs $\geq$ 95%. PeptideLocator performs worst on the cytokines and growth factor peptide activity subset, however an AUC of 86% is still high.

3.2 Predicted Bioactive Peptides in Food Proteins

To illustrate the functionality of PeptideLocator, the sequences of all proteins stored in BIOPEP were retrieved from UniProt (The UniProt Consortium, 2012). This amounted to 661 unique proteins, primarily from food, which were redundancy reduced with respect to the training set so that any protein with a BLAST alignment with more than 30% sequence similarity to any protein in the training set was removed leaving 615 protein sequences (Fig. S3). The

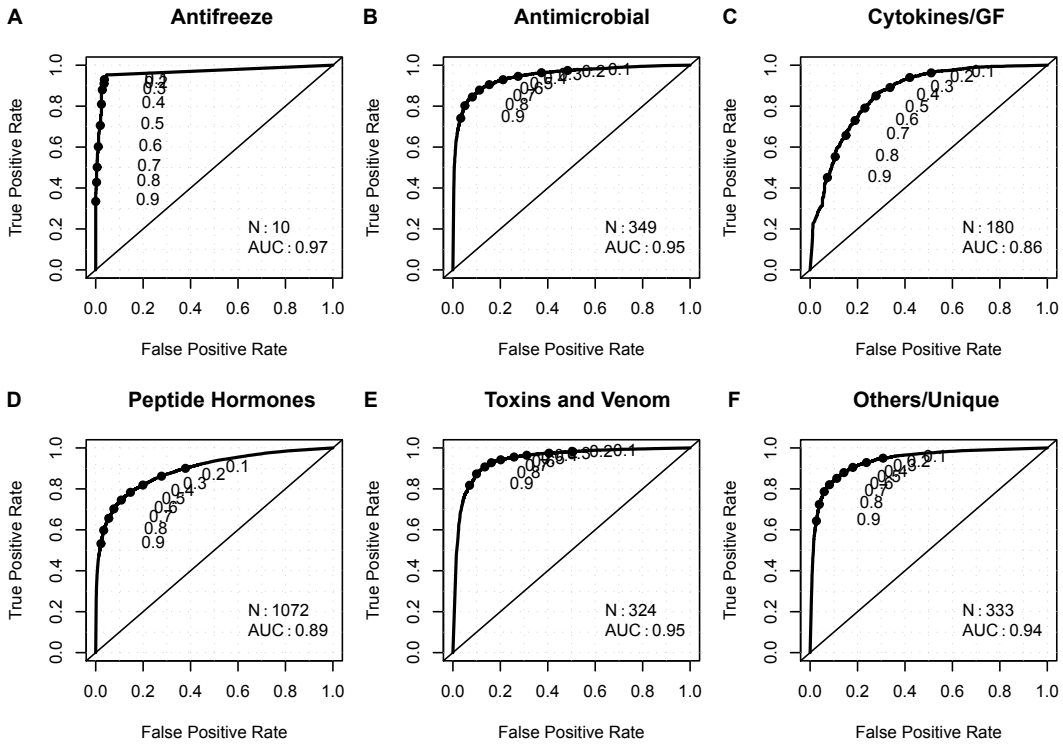

Fig. 2: **Peptide activity classes.** ROC curves for PeptideDB.30 split into six peptide activity subsets; tested in five-fold cross-validation with the threshold (labelled) for a positive prediction increasing from 0 to 1.

	Antifreeze	Antimicrobial	Cytokines and Growth Factor	Peptide Hormones	Toxins and Venom	Unique/Other
Non-bioactive	824	5665	576	43922	3367	587
Bioactive	299	7757	4825	17724	8258	6574
Total	1123	13422	5401	61646	11625	17161

Table 3. Number of bioactive and non-bioactive residues in each of the six peptide activity classes.

sequences were submitted to PeptideLocator and the results were searched for proteins with peptide regions with y residues over a threshold x. We selected the 25 proteins which had at least 16 consecutive residues predicted over a threshold of 0.8 for further analysis (Table S1). None of these predicted peptide regions are found in our training set.

Twelve of the proteins are less than 40 residues in length. For most of these the whole protein (or protein fragment) is labelled as bioactive (using a threshold of 0.8), and in all cases at least 80% of the residues are predicted over the 0.8 threshold. A number of these predicted peptides are alpha-amylase inhibitors from Rye (*Secale cereale*); alpha-amylase inhibitors are being investigated as food additives to aid in weight loss (Barrett and Udani, 2011). The sheep lactoferrin fragment (33 residue N-terminal peptide without signal peptide) has been shown to be an inhibitor of platelet aggregation (Qian *et al.*, 1995). The wheat allergen peptide (ALCC_WHEAT) associated with Bakers asthma (Amano *et al.*, 1998) and wheat gluten, important in coeliac disease, were also identified.

We found that for the medium length proteins (40 - 100 residues) the predicted peptide was either at the N- or C-terminus, with the exception of TIM13_ORYSJ; three of the proteins (DF39_PEA, Q9FUP3_PHACN and NO14_PEA) had signal peptides which were not predicted as bioactive. For example, PeptideLocator correctly predicted the *Pisum sativum* (Garden pea) bioactive peptide from Defensin-like protein 39 and has not predicted the signal peptide as bioactive (residues 1 to 28, Fig. 3). The experimentally validated peptide from Defensin-1 (UniProt:P81929) can be found in the CAMP database (Thomas *et al.*, 2010) and differs from our peptide sequence only at the N-terminal lysine. It has been shown to have anti-fungal activity upon contact with a number of plant pathogen fungi (Almeida *et al.*, 2000). Another example is the Kunitz trypsin inhibitor protein (Q9FUP3_PHACN) from *Phaseolus coccineus* (Scarlet runner bean). Although not annotated as such by UniProt,

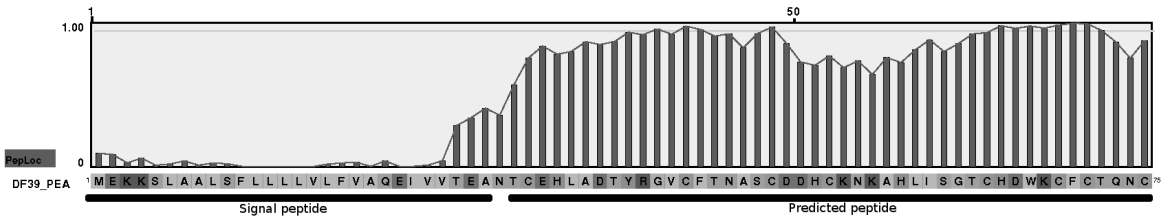

Fig. 3: **DF39_PEA Defensin-like protein 39.** The predicted peptide is underlined (PeptideLocator score > 0.5) and differs from the experimentally validated peptide in only two places, an extra N-terminal lysine and an asparagine (N) instead of an aspartic acid (D) at position 36 in the predicted peptide; Dark grey bars: PeptideLocator prediction.

there is a signal peptide predicted by SignalP (Petersen *et al.*, 2011) at the start of this protein sequence which is not predicted as bioactive by PeptideLocator. The predicted peptide (residues 28 to 73) aligns with a peptide in the CAMP database with 54% sequence identity; this peptide is found in a *Capsicum annuum* secreted protein (UniProt: Q43413) and has been experimentally validated to have anti-fungal activity against *Fusarium oxysporum* (Thomas *et al.*, 2010).

Three of the longer proteins (> 100 residues) are of particular interest (Q43673_VICFA, CVCA_PEA and GLCAP_SOYBN). Fig. S4 shows a strongly predicted bioactive peptide region between residues 333 and 348 in Legumin (Q43673_VICFA), a *Vicia faba* (Broad bean) 11S seed storage protein. Convicilin (CVCA_PEA) from *Pisum sativum* (Garden pea) is a 7S seed storage protein. Fig. S5 shows a strongly predicted region of bioactivity between residues 74 and 90 and another shorter peptide regions between residues 58 to 63. These proteins have also been investigated for potential ACE inhibitory activity (Vermeirssen *et al.*, 2004). A potential bioactive peptide of *Glycine max* (Soybean) Beta-conglycinin 7S seed storage protein (GLCAP_SOYBN) is predicted to be in the propeptide region of the protein sequence (residues 23 - 62). This protein has been shown to be a major food allergen (Krishnan *et al.*, 2009).

The appropriate test of these predictive methods is a large scale experimental screen that employs the prediction method developed here, ideally contrasted with a randomly selected set of peptides which ignores these predictions. Carrying out a large experiment would allow a quantification of whether the predictive power of the method is supported in the laboratory as well as *in silico*.

4 CONCLUSION

PeptideLocator can accurately identify peptide regions in protein sequences according to the likelihood that they are bioactive. We have demonstrated the use of PeptideLocator as part of a targeted identification pipeline for food peptides with beneficial properties, for example, antimicrobial peptides. It can be used in conjunction with other tools and data, for example, food hydrolysates that provide information on the proteolyic cleavage of proteins. By combining cleavage patterns with the bioactivity prediction from PeptideLocator it is possible to identify bioavailable peptides from common food, or other, proteins. Similarly, peptides predicted by PeptideLocator for a given precursor protein can act as a guide for the selection of specific enzymes that will result in their release. The utility of such an approach is demonstrated in our initial analysis of the BIOPEP dataset, showing the ability of PeptideLocator to identify peptides from a set of food proteins. There is a growing need for the identification of novel bioactive peptides in proteins and this web server is a useful addition to the repertoire of tools for investigating bioactivity.

PeptideLocator is available as part of our web server for bioactive peptide discovery and annotation. The web server implementation of the algorithm requires as input a UniProt accession number which the server uses to retrieve the sequence from UniProt (The UniProt Consortium, 2012). A graphical representation of the score per residue at a given sequence position is included in the results page, along with IUPred disorder prediction scores. The server builds an alignment of the orthologues of the protein sequence in order to provide the relative local conservation (RLC) score (Davey *et al.*, 2009) as an additional plot in the output for metazoan sequences. The PeptideLocator score is unaffected by the presence of the RLC score for the protein; however, this is instructive in the interpretation of results. Similarly, interpretation of the likely biological context of the peptide is aided by the inclusion of disorder predictions from IUPred (Dosztányi *et al.*, 2005).

PeptideLocator is designed to allow fast and reliable annotation of protein sequences and is freely available for academic users at http://bioware.ucd.ie/. A packaged version of the PeptideLocator code is available by contacting the authors directly or through the Bioware Users Group (details on the PeptideLocator web page). A web service version of the software has been registered on biocatalogue.org (Bhagat *et al.*, 2010) together with examples of how to run PeptideLocator programmatically by submitting jobs in an automated fashion.

ACKNOWLEDGEMENT

The authors acknowledge the Research IT Service at University College Dublin for providing high performance computing resources that have contributed to the research results reported within this paper. We thank the authors of PeptideDB for providing peptide precursor protein sequences from their PeptideDB database.

Funding: This work was supported by Science Foundation Ireland [grant numbers 08/IN.1/B1864, 10/RFP/GEN2749] and by Enterprise Ireland (Food for Health Ireland).

REFERENCES

Almeida, M., Cabral, K., Zingali, R., and Kurtenbach, E. (2000). Characterization of two novel defense peptides from pea *pisum sativum* seeds. *Arch Biochem Biophys*, **378**(2), 278–286.

Altschul, S., Madden, T., Schäffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25**(17), 3389–3402.

Amano, M., Ogawa, H., Kojima, K., Kamidaira, T., Suetsugu, S., Yoshihama, M., Satoh, T., Samejima, T., and Matsumoto, I. (1998). Identification of the major allergens in wheat flour responsible for baker's asthma. *Biochem J*, **330**(Pt 3), 1229–1234.

Baldi, P., Brunak, S., Frasconi, P., Soda, G., and Pollastri, G. (1999). Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics*, **15**(11), 937.

Barrett, M. and Udani, J. (2011). A proprietary alpha-amylase inhibitor from white bean (phaseolus vulgaris): a review of clinical studies on weight loss and glycemic control. *Nutr J*, **10**, 24.

Bhagat, J., Tanoh, F., Nzuobontane, E., Laurent, T., Orlowski, J., Roos, M., Wolstencroft, K., Aleksejevs, S., Stevens, R., Pettifer, S., *et al.* (2010). BioCatalogue: a universal catalogue of web services for the life sciences. *Nucleic Acids Res*, **38**(suppl 2), W689–W694.

Boonen, K., Creemers, J., and Schoofs, L. (2009). Bioactive peptides, networks and systems biology. *BioEssays*, **31**(3), 300–314.

Clare, D. and Swaisgood, H. (2000). Bioactive milk peptides: a prospectus. *J Dairy Sci*, **83**(6), 1187–1195.

Davey, N. E., Shields, D. C., and Edwards, R. J. (2009). Masking residues using context-specific evolutionary conservation significantly improves short linear motif discovery. *Bioinformatics*, **25**(4), 443–50.

Diochot, S., Baron, A., Salinas, M., Douguet, D., Scarzello, S., Dabert-Gay, A.-S., Debayle, D., Friend, V., Alloui, A., Lazdunski, M., and Lingueglia, E. (2012). Black mamba venom peptides target acid-sensing ion channels to abolish pain. *Nature*.

Dosztányi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, **21**(16), 3433–3434.

Dziuba, J., Minkiewicz, P., Nałecz, D., and Iwaniak, A. (1999). Database of biologically active peptide sequences. *Nahrung*, **43**(3), 190–195.

Edwards, R., Moran, N., Devocelle, M., Kiernan, A., Meade, G., Signac, W., Foy, M., Park, S., Dunne, E., Kenny, D., and Shields, D. (2007). Bioinformatic discovery of novel bioactive peptides. *Nat Chem Biol*, **3**(2), 108–112.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recogn Lett*, **27**(8), 861–874.

Fjell, C., Hancock, R., and Cherkasov, A. (2007). AMPer: a database and an automated discovery tool for antimicrobial peptides. *Bioinformatics*, **23**(9), 1148–1155.

Fjell, C., Hiss, J., Hancock, R., and Schneider, G. (2012). Designing antimicrobial peptides: form follows function. *Nat Rev Drug Discov*, **11**(1), 37–51.

Hancock, R. and Sahl, H. (2006). Antimicrobial and host-defense peptides as new anti-infective therapeutic strategies. *Nat Biotechnol*, **24**(12), 1551–1557.

Hartmann, R. and Meisel, H. (2007). Food-derived peptides with biological activity: from research to food applications. *Curr Opin Biotech*, **18**(2), 163–169.

Khaldi, N. (2012). Bioinformatics approaches for identifying new therapeutic bioactive peptides in food. *Functional Foods in Health and Disease*, **2**, 325–338.

Korhonen, H. and Pihlanto, A. (2006). Bioactive peptides: production and functionality. *Int Dairy J*, **16**(9), 945–960.

Krishnan, H., Kim, W., Jang, S., and Kerley, M. (2009). All three subunits of soybean β-conglycinin are potential food allergens. *J Agr Food Chem*, **57**(3), 938–943.

Lata, S., Mishra, N., and Raghava, G. (2010). AntiBP2: improved version of antibacterial peptide prediction. *BMC Bioinformatics*, **11**(Suppl 1), S19.

Letunic, I., Doerks, T., and Bork, P. (2009). SMART 6: recent updates and new developments. *Nucleic Acids Res*, **37**(suppl 1), D229–D232.

Lewis, R., Garcia, M., *et al.* (2003). Therapeutic potential of venom peptides. *Nat Rev Drug Discovery*, **2**(10), 790–802.

Liu, F., Baggerman, G., Schoofs, L., and Wets, G. (2008). The construction of a bioactive peptide database in metazoa. *J Proteome Res*, **7**(9), 4119–4131.

Lynn, D., Higgs, R., Gaines, S., Tierney, J., James, T., Lloyd, A., Fares, M., Mulcahy, G., and OFarrelly, C. (2004). Bioinformatic discovery and initial characterisation of nine novel antimicrobial peptide genes in the chicken. *Immunogenetics*, **56**(3), 170–177.

Möller, N., Scholz-Ahrens, K., Roos, N., and Schrezenmeir, J. (2008). Bioactive peptides and proteins from foods: indication for health effects. *Eur J Nutr*, **47**(4), 171–182.

Mooney, C., Vullo, A., and Pollastri, G. (2006). Protein structural motif prediction in multidimensional ø-ψ space leads to improved secondary structure prediction. *J Comput Biol*, **13**(8), 1489–1502.

Mooney, C., Haslam, N. J., Pollastri, G., and Shields, D. C. (2012). Towards the improved discovery and design of functional peptides: Common features of diverse classes permit generalized prediction of bioactivity. *PLoS ONE*, **7**(10), e45012.

Newburg, D. and Walker, W. (2007). Protection of the neonate by the innate immune system of developing gut and of human milk. *Pediatr Res*, **61**(1), 2–8.

Norris, R., Casey, F., FitzGerald, R., Shields, D., and Mooney, C. (2012). Predictive modelling of angiotensin converting enzyme inhibitory dipeptides. *Food Chem*, **133**(4), 1349–1354.

Petersen, T., Brunak, S., von Heijne, G., and Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods*, **8**(10), 785–786.

Pollastri, G. and McLysaght, A. (2005). Porter: a new, accurate server for protein secondary structure prediction. *Bioinformatics*, **21(8)**, 1719–20.

Pollastri, G., Martin, A., Mooney, C., and Vullo, A. (2007). Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information. *BMC Bioinformatics*, **8**, 201.

Qian, Z., Jollès, P., Migliore-Samour, D., and Fiat, A. (1995). Isolation and characterization of sheep lactoferrin, an inhibitor of platelet aggregation and comparison with human lactoferrin. *BBA-Gen Subjects*, **1243**(1), 25–32.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Sasaki, K., Takahashi, N., Satoh, M., Yamasaki, M., and Minamino, N. (2010). A peptidomics strategy for discovering endogenous bioactive peptides. *J Proteome Res*, **9**(10), 5047.

Shtatland, T., Guettler, D., Kossodo, M., Pivovarov, M., and Weissleder, R. (2007). PepBank–a database of peptides based on sequence text mining and public peptide data sources. *BMC Bioinformatics*, **8**, 280.

Swift, S., Vaughan, E., and de Vos, W. (2000). Quorum sensing within the gut ecosystem. *Microb Ecol Health D*, **12**(2), 81–92.

The UniProt Consortium (2012). Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res*, **40**(D1), D71–D75.

Thomas, S., Karnik, S., Barai, R., Jayaraman, V., and Idicula-Thomas, S. (2010). CAMP: a useful resource for research on antimicrobial peptides. *Nucleic Acids Res*, **38**(suppl 1), D774.

Torrent, M., Andreu, D., Nogués, V. M., and Boix, E. (2011). Connecting peptide physicochemical and antimicrobial properties by a rational prediction model. *PLoS ONE*, **6**(2), e16968.

Torrent, M., Di Tommaso, P., Pulido, D., Nogués, M., Notredame, C., Boix, E., and Andreu, D. (2012). AMPA: an automated web server for prediction of protein antimicrobial regions. *Bioinformatics*, **28**(1), 130–131.

Vermeirssen, V., Van Der Bent, A., Van Camp, J., Van Amerongen, A., and Verstraete, W. (2004). A quantitative in silico analysis calculates the angiotensin I converting enzyme (ACE) inhibitory activity in pea and whey protein digests. *Biochimie*, **86**(3), 231–239.

Vijayakumar, V., Guerrero, A., Davey, N., Lebrilla, C., Shields, D., and Khaldi, N. (2012). EnzymePredictor: a tool for predicting and visualizing enzymatic cleavages of digested proteins. *J Proteome Res*, **11**(12), 6056–6065.

Wang, G., Li, X., and Wang, Z. (2009). APD2: the updated antimicrobial peptide database and its application in peptide design. *Nucleic Acids Res*, **37**(Database issue), D933.

Wang, P., Hu, L., Liu, G., Jiang, N., Chen, X., Xu, J., Zheng, W., Li, L., Tan, M., Chen, Z., Song, H., Cai, Y., and Chou, K. (2011). Prediction of antimicrobial peptides based on sequence alignment and feature selection methods. *PLoS ONE*, **6**(4), e18476.