BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 9 2007, pages 1148–1155
doi:10.1093/bioinformatics/btm068

Databases and ontologies

AMPer: a database and an automated discovery tool for antimicrobial peptides

Christopher D. Fjell[1,*], Robert E.W. Hancock[2] and Artem Cherkasov[1]

[1]Division of Infectious Diseases, Department of Medicine, Faculty of Medicine, University of British Columbia, 2733 Heather street, Vancouver, BC, Canada, V5Z 3J5 and [2]Centre for Microbial Diseases and Immunity Research, University of British Columbia, #2259 Lower Mall Research Station, Vancouver, British Columbia, V6T 1Z3, Canada

Received on May 29, 2006; revised on January 23, 2007; accepted on February 22, 2007

Advance Access publication March 6, 2007

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Increasing antibiotics resistance in human pathogens represents a pressing public health issue worldwide for which novel antibiotic therapies based on antimicrobial peptides (AMPs) may offer one possible solution. In the current study, we utilized publicly available data on AMPs to construct hidden Markov models (HMMs) that enable recognition of individual classes of antimicrobials peptides (such as defensins, cathelicidins, cecropins, etc.) with up to 99% accuracy and can be used for discovering novel AMP candidates.

Results: HMM models for both mature peptides and propeptides were constructed. A total of 146 models for mature peptides and 40 for propeptides have been developed for individual AMP classes. These were created by clustering and analyzing AMP sequences available in the public sources and by consequent iterative scanning of the Swiss-Prot database for previously unknown gene-coded AMPs. As a result, an additional 229 additional AMPs have been identified from Swiss-Prot, and all but 34 could be associated with known antimicrobial activities according to the literature. The final set of 1045 mature peptides and 253 propeptides have been organized into the open-source AMPer database.

Availability: The developed HMM-based tools and AMP sequences can be accessed through the AMPer resource at http://www.cnbi2.com/cgi-bin/amp.pl

Contact: cfjell@interchange.ubc.ca

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Antimicrobial peptides (AMPs) represent a diverse class of natural peptides that form a part of the innate immune system of mammalians, insects, amphibians and plants among others (for example, Sima *et al*., 2003a, b). In the face of increasing antibiotic resistance in pathogenic microorganisms, AMPs have drawn significant scientific attention as a novel class of prospective antimicrobial therapeutics as both antibacterial drugs and modulators of innate immunity (Finlay and Hancock, 2004; Hamilton-Miller, 2004; Koczulla and Bals, 2003; Levy and Marshall, 2004). Although the AMPs exhibit relatively lower potency against susceptible bacterial targets compared to conventional low-molecular-weight antibiotic compounds, they hold several compensatory advantages including fast target killing, broad range of activity, low toxicity and minimal development of resistance in target organisms (Hancock, 2001; Yount and Yeaman, 2003).

Despite the fact that a broad spectrum of AMPs have been identified and discussed in the literature, their structure-activity relationships are not well understood, largely because of substantial sequence and structure diversity of AMPs. Examples include the alpha-helical cecropins and magainins and the beta-sheet structure of beta-defensins among others. It should be mentioned, however, that AMP 3D structures are often dependent on binding to membrane or lipopolysaccharide, and in solution many AMPs may exist in different, and/or non-ordered configuration (Chapple *et al*., 2004; Yount and Yeaman, 2003). Thus, the general views on the AMP characteristic features typically involve their cationic character, relatively high hydrophobicity and short length (Powers and Hancock, 2003; Yount and Yeaman, 2003).

The mechanisms of peptide antimicrobial action are also under debate; while membrane disruption has been a common theme, other evidence suggests that peptides transit into the cytosol and disrupt intracellular targets and that the membrane effects are distinct from (and not always crucial to) the killing effects (Hancock and Rozek, 2002; Patrzykat *et al*., 2002). In addition, the relative importance of direct killing versus immunomodulatory effects of mammalian AMPs is not obvious since some peptides generally considered as AMPs do not appear to have direct microbe-killing effects *in vivo* (Bowdish *et al*., 2005; Brogden, 2005).

All the above-mentioned controversies make '*in silico*' discovery and/or modeling of AMPs an important but challenging Bioinformatics task. Currently, sequence analysis for AMP discovery has been done on a limited number of AMPs: the beta-defensins and other cysteine-containing peptides. A number of novel beta-defensins in mouse and human were identified by analysis of a specific exon of beta-defensins followed by scanning of genomic sequence (Scheetz *et al*., 2002; Schutte *et al*., 2002). Manual identification of a predictive motif, GXC, for cysteine-containing AMPs was

*To whom correspondence should be addressed.

also used to find novel AMPs of that type (Yount and Yeaman, 2004). However, these efforts were applicable only to a small number of AMP types.

We decided to conduct a more generalized study of AMP sequences using profile-based hidden Markov models (HMMs) in combination with sequence clustering and protein structure annotation. The major objective of the study was to produce HMM models for the existing AMP types such as defensins, cathelicidins and histatins among others, and to apply these methods to create a more consistent classification of antimicrobial sequences. This new resource is available as an online database, for investigation of AMP sequence diversity, and as a set of HMM files for the discovery of novel gene-coded AMP candidates.

2 RESULTS AND DISCUSSION

The analysis of the AMPs proceeded as described next and is summarized in Supplementary Figure 1.

2.1 Database of AMPs

Initially, we used the set of known gene-coded AMPs from the AMSDb collection at the University of Trieste to compile a generalized set of known AMP sequences (see the 'Web Resources' section for more details about the source of AMP sequences). This resulting set of confirmed AMPs contained 890 sequences and encompassed all major AMP classes including defensins, cathelicidins and granulin among others. These peptides are available as entire holopeptides, containing both mature functional peptides as well as prosequences.

Some of these proteins were found to contain obsolete annotations and refer to obsolete Uniprot IDs. Since we are interested in analyzing the mature and prosequence regions separately, we required the proteins be present in the current version (August 2006) of the Uniprot database. To associate the proteins in AMSDb to the current Uniprot we performed a pairwise similarity comparison using blastp of the BLAST tool (Altschul *et al*., 1990). We considered a match to be made where the AMSDb protein has at least 99% sequence identity over at least 99% of the length of the smaller sequence of the pair. We tried relaxing the criteria to 95% for each parameter — this resulted in only two more matches, which we did not consider significant to justify the additional risk of incorrect assignment. In addition, 33 proteins were identified based on sequence ID that were the same proteins between AMSDb and Uniprot, but the sequence was <99% similar. These 33 Uniprot proteins were used. Of the 890 original AMSDb proteins, 741 proteins were matched in Uniprot (661 from Swiss-Prot and 80 from TrEMBL).

The peptide location annotations were used from Uniprot to identify mature peptide and propeptide regions. A total of 679 Uniprot proteins were found to have suitable annotation for mature peptides, yielding 767 mature peptides. Most proteins contributed one mature peptide while one protein, human Histatin-3 (HIS3_HUMAN), contributed 26 peptides, the highest number. A total of 238 Uniprot proteins had annotations for propeptides, yielding 316 propeptides. Most proteins contributed 1 propeptide, but up to 7 (for AMP_IMPBA from Balsam plant) were contributed for a single protein.

Table 1. Effect of similarity threshold on clustering of mature peptides

Threshold (%)	Number of clusters	Clustered fraction (%)
10	136	94
20	142	92
30	149	90
40	148	84
50	151	80
60	158	75
70	153	66
80	136	56
90	120	42

The original set of mature peptides were clustered for several values of the minimum global percent similarity (Threshold). The clustered fraction is the fraction of the original set of mature peptides that were placed in clusters for the given threshold.

2.2 Clustering of the AMPs

As it has already been mentioned, AMPs are very diverse in their sequences and fall into classification of a small number of secondary structures (Hwang *et al*., 1998; Powers and Hancock, 2003). However, our objective in clustering was to group similar peptides for later analysis by HMMs. For this purpose, we wanted to capture in a single cluster the diversity of sequences that likely corresponded to single type of peptide. While a large number of AMP groups can be defined based on descriptions in the literature (such as defensins, magainins, cathelicidins), this nomenclature is not amenable to specification for automated grouping, due to the large diversity in sequence as well as length for a given protein name or description. Since no classification scheme was found that was suitable for our purpose, we chose to group AMPs by sequence analysis using custom sequence similarity.

In short, clusters were constructed to have a minimum amount of similarity between all peptides in the cluster (see Methods section for details). Two sets of clusters were constructed, for mature peptides and propeptides. Each peptides was compared to the peptides in existing clusters and a minimum 'global' sequence identity was calculated as the number of matching amino acids divided by the length of the shorter peptide using the most significant alignment given by the blastp algorithm. A peptide was placed in an existing cluster based on the minimum global sequence identity for any peptide in the cluster. The peptide was placed in the cluster giving the highest minimum match, if the minimum was greater than a given minimum identity threshold. Peptides not placed in existing clusters were used to start new clusters.

Minimum similarity thresholds in the range 10–90% were used to evaluate the resulting clusters. Decreasing the threshold to a minimum of 10% global similarity gives the maximum number of peptides placed in clusters. However, when we examined the multiple alignments of these clusters for low thresholds we found problems: Many contained two or more sets of closely related peptides that were more appropriately

separated into distinct clusters. As well, short peptides were found to be inserted into clusters where the matching amino acids in the multiple alignment were interspersed with gaps between matching positions of only one or two amino acids. However, for higher thresholds, dramatically lower coverage of the peptides was represented in the clusters, with a 90% threshold yielding clusters for only 42% of the starting peptides.

Therefore, we decided to use an intermediate threshold of 30% global sequence and manually correct the clusters by removing short peptides having poor alignment, and by splitting clusters into additional clusters where the peptides consisted of two or more highly similar sets of peptides. In total, 20 peptides were removed from 19 clusters; 3 clusters were split into 6; and 6 clusters composed of 2 clones each were removed. There were 146 resulting clusters for mature peptides, containing 655 peptides. The propeptide clusters were treated similarly using a threshold of 30% global identity. There were 207 clustered propeptides in 42 clusters before manual edits. Four propeptides were removed from 4 clusters; 5 clusters were removed; and 3 clusters were split into 6. There were 40 resulting clusters containing 192 propeptides.

As anticipated, such classification approach allowed grouping together all related peptides as in the conventional classes such as beta-defensins, cecropins, magainins etc. Peptides of a particular class such as the beta-defensins were also separated into multiple clusters, indicating subclasses of these peptides. We did not try to reduce the number of clusters, for example, to produce a single cluster for each type of defensin. We considered that the larger number of clusters with more highly similar peptides in each is beneficial for model building, as the more specific models may reflect important sequence motifs that may be lost if the clusters contain too much variation.

2.3 HMM modeling

Once we had created the initial clusters, we created profile HMMs for the clusters to be used to search for additional members of the AMP groups that were not present in the original AMP dataset. The HMMER software package (Eddy, 1998; http://hmmer.wustl.edu/) has been utilized to create one profile HMM for each AMP cluster. ClustalW was used to generate the multiple alignments used by HMMER. The HMMER package was chosen over other tools, because it is considered to be less sensitive to small misalignments in the multiple sequence alignments and to report reliable E-values (Madera and Gough, 2002).

2.4 Iterative enhancement of clusters

To enhance our initial clusters, we identified AMP sequences from Swiss-Prot and used these to enrich the initial clusters of the AMPs by iteratively applying the corresponding HMM models to Swiss-Prot sequences. For the current work, we considered only the Swiss-Prot database as it contains confirmed and relatively well-studied peptide sequences to allow validation of the process to be done.

We found that it was not possible to use a specific threshold for significance of match (such as expectation value, E-value, from BLAST or HMMER) to distinguish between hits to AMPs and non-AMPs. In an attempt to identify an E-value threshold that will distinguish significant matches from matches due to chance when searching the Swiss-Prot database, we evaluated the clustered peptides with the models specific for their cluster specifying the size of the dataset as the number of peptides in Swiss-Prot. When these E-values were plotted against the length of the model, it becomes clear that there is no E-value that can distinguish significant matches from random matches for short peptides (Fig. 1). (Note that the length of the HMM is approximately the length of the peptides upon which it was trained.)

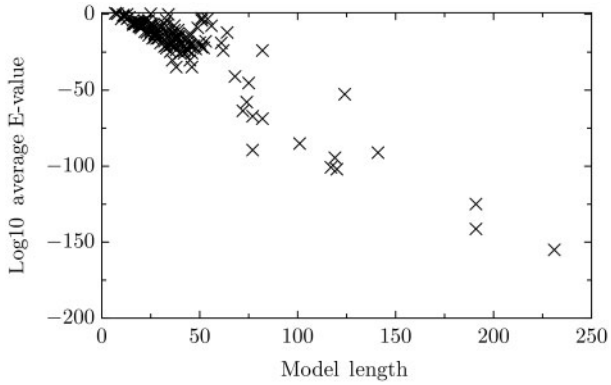

Fig. 1. The relationship between E-value and model length. The peptides in each cluster were scanned with the model corresponding to the cluster. For the shortest models (created from the shortest peptides) the E-values are greater than one.

Since E-values alone are not sufficient to identify significant matches, we decided to use additional information from the Swiss-Prot database to determine significance. For each Swiss-Prot protein, the model giving an HMM match with the lowest E-value was identified. The annotations for the Swiss-Prot protein were used to identify any protein regions overlapping with the region matched by a model. The Swiss-Prot peptide with highest mutual overlap with the region matched by the model was identified. This peptide was also compared to all peptides in the model's cluster to determine its similarity to a listed AMP. To be considered as a significant match, the mutual overlap between the region matched by a model and the annotated peptide was at least 90%. In addition, the blastp match between the Swiss-Prot peptide and the best-matching clustered peptide was at least 50% identity over 90% of the peptide length.

Those Swiss-Prot entries that produced a significant match to any of the 186 HMMs (146 for mature peptides and 40 for propeptides) were added into the existing AMP clusters. After peptides were added to a cluster, a new multiple alignment and HMM were constructed as described above. The new model, based on a larger number of sequences, was then used to scan Swiss-Prot. This was repeated until no additional peptides had a significant match: there were five iterations for the mature peptide models, and only one for the propeptide models.

The iterative scanning of the Swiss-Prot database (containing 230 133 peptides) resulted in an additional 389 mature peptides

Table 2. Changing consensus sequence with number of iterations for mature peptides in cluster 137. *N* is the iteration number with $N=0$ the initial data from AMSDb

N	Consensus
0	GlLDtLKnlAktagKGalqslLntaSCKLsgqC
1	GiLDtlKnlAkgvaKgvaqsLLdklsCKlskgC
2	GiLDtlKnlAkgaAKgvaqsLLdtlkCKltggC
3	GiLDtlKnlAkgaaKgaaqsLLdtlsCKlsggC
4	GiLDtlKglAknaGKGvaqsLLdtlsCKisggC
5	GiLDtlKnlAkgaAKgaAqsLLdtlsCKisggC

from 229 Swiss-Prot proteins being added to the AMP dataset as candidate AMPs, for a total of 1045 peptide from 970 Uniprot proteins. Sixty-one propeptides were also added for a total of 253 propeptides from 223 proteins. Peptides were considered to be properly included as AMPs where the annotations included reference to antimicrobial activity or the protein belonged to the same family as a known AMP already in the database (see Methods for details).

The utility of a selection process that does not rely on the E-value can be seen in Cluster 1 (see on-line Supplementary tables) for the mature peptides. Starting with an initial 2 AMPs, an additional 9 peptides are added to the cluster. Despite the high E-values (up to 5.9), all peptides were found to have annotations that demonstrate antimicrobial activity.

The relationship between the mature peptides and propeptides from the same protein is shown in Figure 2. In Figure 2A, mature clusters are joined to propeptide clusters where the propeptides are derived from the same protein as a mature peptide in the cluster. Only the mature peptide clusters of at least ten peptides. Similarly, Figure 2B shows links from the largest propeptide sequences to mature peptide clusters. These figures suggest there is greater conservation of propeptide sequence, since a greater proportion of propeptide clusters have links to multiple mature clusters. A full mapping between clusters is available as Figure 4 (Supplementary Material).

Of the 229 proteins added, 34 either did not have annotation for antimicrobial activity, or annotation specifically stated that they were not antimicrobial. Among these are two groups of peptides that have AMPs in the same family: 9 Dahlein peptides are annotated as inactive (2 other Dahleins are active, DAH11_LITDA and DAH12_LITDA), and 8 Aurein peptides are annotated as inactive while 6 are active. An additional 17 peptides are peptide hormones such as cholecystokinin that do not have annotations for antimicrobial activity. However, there is considerable controversy surrounding whether certain peptides should be considered antimicrobial or not; in particular, differing assay conditions used by different investigators lead to differing results. For this reason, these peptides were left in the AMPer database and it is left to the investigators to review the relevant literature provided through links from the AMPer system.

The physico-chemical properties of the mature peptides vary dramatically between clusters. As can be seen in Table 3A (Supplementary Material) for the largest AMP clusters (size >10 peptides), the net charge depends strongly on the type of AMP. As expected, the median charges typically exceed +2 but one class is negative. Except for one cluster, the median hydrophobicity is above 40% with a maximum of 77%. There are 5 clusters of propeptides size 10 or greater, shown in Table 3B (Supplementary Material). These tend to be strongly negative and much less hydrophobic than the mature clusters.

2.5 Accuracy of models

The 186 final clusters were produced with high stringency requirements for matches to HMMs. Such stringency explains the relatively large number of identified clusters containing similar annotation: for example, there are 22 clusters of defensins which are split along the defensin subclasses (including several subclasses of alpha- and beta-defensins, cryptdins and other enteric defensins). Further investigation of the effect of using lower stringency thresholds for the initial clustering and for addition of peptides to clusters might allow these clusters to be merged, and a more representative model to be produced. However, performing additional merges may also lead to incorrect merges that give less-accurate models. We consider that the presence of multiple clusters of similar peptides reflects subclasses of these peptides, and that the larger number of higher accuracy models may be beneficial for further work on mechanisms of action of AMPs that differ between subclasses.

To assess the expected performance of the system to identify previously unknown AMPs from proproteins, we performed an ~10-fold cross-validation on the AMP identification procedure as described in detail below. Since we were interested in the capacity of the system to identify AMPs in proproteins, we performed the testing steps of the validation on full proproteins from Swiss-Prot rather than simply the peptide comprising the clusters. The presence of another peptide from the same protein in both testing and training sets severely complicates interpretation of the results. The current pipeline is intended to identify proteins that contain additional AMPs and will not properly handle recognition of additional peptides of the same cluster type. For this reason, only the 105 mature peptides and 29 propeptide clusters that did not contain more than one peptide from the same proprotein were considered. In addition, for creation of HMMs, at least 2 peptides are required; to select a test peptide from the set, therefore, a minimum cluster size of 3 is needed. This left a total of 81 mature and 20 prosequence clusters used for cross-validation.

The results of the cross-validation shows great variation in performance for recognizing additional AMPs. The cross-validation sensitivity varied from 0% for one mature cluster containing three peptides, to 100% for 36 mature clusters. The average sensitivity of all mature clusters was 82% (the SD of the cluster mean sensitivities was 23%). The specificity and accuracy were both 99.2% (SD 1.3%). For the prosequence clusters, the sensitivity also varied between 0% for two clusters of 3 peptides, and 100% for 9 clusters with average 81% (SD 30%); the average specificity for the prosequence clusters was 98.8% (SD 2.7%) and accuracy was 98.8% (SD 2.7%). The values for each cluster are available in Tables 4A and B (Supplementary Material).

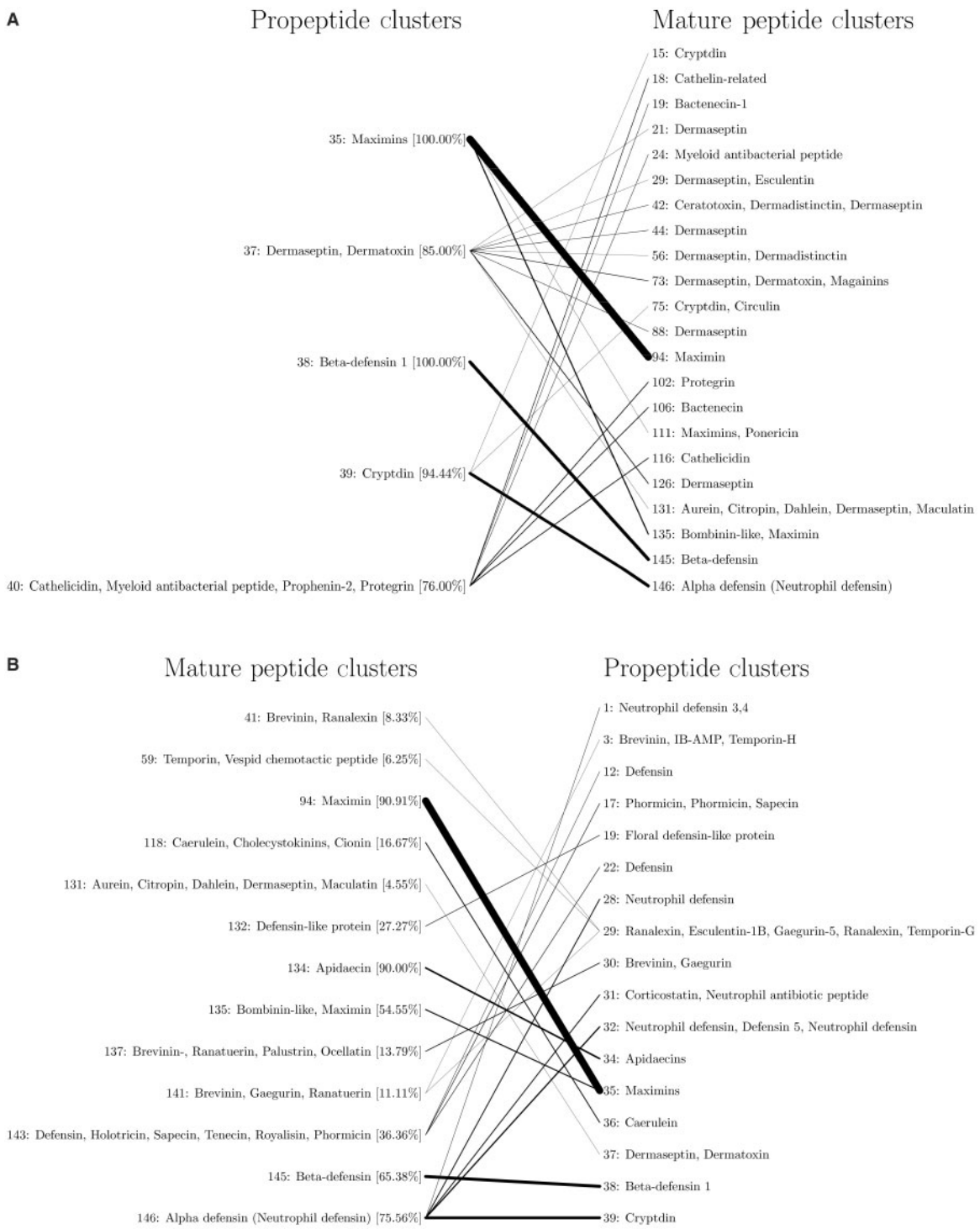

Fig. 2. (**A**) Cluster linkage from largest mature peptide clusters. (**B**) Cluster linkage from largest propetide clusters. Relationship between mature peptides and propeptides from the same protein. (A) For mature peptide clusters of 10 or more peptides, the corresponding propeptide clusters are indicated by a line joining the clusters. The width of the line indicates the number of propeptides in that cluster that are from the same protein IDs as the mature peptides. Similarly, (B) shows the linkage from propeptide clusters with ten or more propeptides. Percentage values following the left clusters are the fraction of peptides with links to the right clusters.

It should be noted that the specificity is conservatively based on distinguishing a class of AMPs from other possibly very similar AMPs (such as one class of defensins from several other classes of defensins). As well, the accuracy is dominated by the number of negatives, since the number of actual negatives is much larger than the number of actual positives. In scanning a large database of unrelated proteins such as Swiss-Prot, the specificity and accuracy is expected to be significantly better since the number of false positives will be much lower, as demonstrated by the low number of total positive matches found for all of Swiss-Prot. The low sensitivity of some clusters is thought to be due to the relatively large variation in sequence in these clusters, especially for clusters containing few peptides. A variety of technical reasons were found for why peptides were missed: the HMM search did not give a significant match (E-value >10), or the HMM match did not align well with the Uniprot feature list, or the BLAST match to the closest training peptide was too poor (data not shown). This suggests that a simple tweaking of system parameters will not lead to a dramatic increase in sensitivity without undesired decrease in specificity; therefore, a search for better search parameters was not pursued in this study.

2.6 Online tools

All materials described here have been made available online (http://www.cnbi2.com/cgi-bin/amp.pl). All AMP sequences and final clusters are available for download. In addition, utilities are provided on-line to scan sequence provided by the user to categorize the sequence according to these models. The HMMER HMM files used to predict and classify AMPs are available for researchers to download and use to scan sequence files using the HMMER package independently. This is a unique contribution to the community: one other site, ANTIMIC (Brahmachary *et al*., 2004; http://research.i2r.a-star.edu.sg/Templar/DB/ANTIMIC), provides some limited search against a few specific models but does not categorize submitted sequence, and does not provide for download of the sequences or the few HMM models it contains.

Web pages are available for viewing the AMPs and corresponding properties. The initial page (http://www.cnbi2.com/cgi-bin/amp.pl) provides links to lists of the AMP clusters and the peptides themselves. In addition to properties such as peptide length, charge and hydrophobicity, the consensus sequence is given as well as links to navigate to the list of AMPs in each cluster. For each peptide, there are clickable links to the Swiss-Prot web site and to the Swiss-Prot records for the version used in this study. The iteration number ('round') is indicated for each peptide with round zero indicating the peptide is from the original set from AMSDb database (a link is also given to AMSDb).

Several properties of the peptide subsequence matched by the HMM model are also given: amino acid sequence, length, charge, hydrophobicity (as hydrophobic fraction — fraction of amino acids that are hydrophobic), position of the subsequence within the main protein as well as the E-value of the model match for this peptide. Additionally, values used for analysis are also given: the coverage of the best-matching peptide by the region matched by the HMM and vice versa, and the best-matching (by blastp) previously clustered peptide with percent identity and alignment length.

In summary, we utilized a set of documented AMPs to collect additional known gene-coded AMPs into a single database using a hybrid method for identifying AMPs. We clustered the peptides and enriched the clusters with peptides from Swiss-Prot, which could be matched by the trained HMM at high confidence by integrating additional information using pair-wise sequence comparison and annotations of peptide positions. The HMM models and sequence files are made available to the public from the AMPer website. We anticipate that these will be useful for discovering novel AMPs from unannotated sequence.

3 METHODS

3.1 Initial peptide set

The initial set of gene-coded AMP sequences was obtained from the Biochemistry Department University of Triest, Italy (http://www.bbcm.units.it/~tossi/pag5.htm). These peptides were compared to the current Uniprot (Swiss-Prot and TrEMBL) databases (downloaded from http://www.pir.uniprot.org/ on 4 August 2006) to determine the current naming and annotation of the initial AMPs. Pairwise comparison was done using the blastp algorithm of the BLAST package with no filtering (parameters -F F). We considered a match to be positive when there was at least 99% identity of amino acids over a match length of at least 99% of the length of the AMP in the initial set.

For AMSDb proteins with current Uniprot IDs but where the sequence was significantly different, the current Uniprot record was used. Mature peptides and propeptides were identified for each protein using the feature annotations available from Uniprot. For proteins with multiple mature peptides, those peptides annotated as antimicrobial were kept for analysis. Peptides were required to have definite start and end positions (records with '?' were rejected).

3.2 Clustering

Pairwise similarity between peptides was calculated using blastp (BLAST package, Altschul *et al*., 1990) with filtering off (-F F) and word size of 2 (-W 2). Clusters of similar peptides were constructed based on the pairwise alignments using a percentage match defined as the number of amino acids identical between the two peptide in the most significant alignment (highest bit score) divided by the length of the shorter of the two peptides. Clusters were built by successively adding peptides to a cluster where the percentage match was greater than threshold for every peptide in the cluster.

The percentage match threshold was varied between 10 and 90% for clustering mature peptides. Multiple alignments were created for each cluster using ClustalW (Thompson *et al*., 1994). The alignments of mature peptide clusters resulting from several thresholds were examined. Low thresholds produced clusters containing similar peptides mixed with smaller peptides that were aligned at widely spaced intervals to the longer peptides.

The clusters from a 30% threshold were manually edited for both mature peptides and propeptides. Peptides were removed that aligned with a large number of widely spaced inserts, and clusters containing two groups of highly similar peptides were split into two clusters.

3.3 Iterative enhancement of clusters

At the start of an iteration, multiple sequence alignments were built for each cluster using ClustalW (as above). The HMMER software

package (Eddy, 1998; http://hmmer.wustl.edu/) was used to create one HMM for each cluster from the multiple alignment, using the utility, hmmbuild. Default parameters were used except for '–f' parameter, used to create local models. The Swiss-Prot database was scanned using the HMMER utility, hmmsearch, for each model file. Custom Java, Python and BASH shell code were used to execute hmmsearch and parse resulting output.

Scanning of Swiss-Prot was performed for all models. For each Swiss-Prot protein matched, the information for the most significant match (lowest E-value) for any model was stored. Sequence regions matched by the HMMs were then compared to the annotated feature regions from Swiss-Prot. The annotated region (mature peptides or propeptides) having the greatest overlap with the HMM match region were stored. As an additional check, the clustered sequences were aligned to the full Swiss-Prot proteins matched by the HMMs using blastp. The best-matching clustered peptide was determined based on highest bit score. Swiss-Prot peptides were considered positive matches and added to the clusters if the regions matched by the HMMs and the feature annotation agreed to at least 90% of the their length, and the best-matching peptide from the same cluster had at least 50% identity to the Swiss-Prot protein.

Positive matches were then added to the clusters for mature peptides and propeptides if they were not already present in any cluster. A new multiple alignment was then created using ClustalW, and a new model file was created using HMMER as described above. The Swiss-Prot sequences were scanned again using the new model files, and any additional matching peptides were added to the clusters. The process of scanning Swiss-Prot, adding matching peptide to clusters, and rebuilding the model files was repeated until no additional Swiss-Prot peptides were found. Consensus sequence was obtained using, the utility, hmmemit, with the '-c' option. Mature peptide clusters were mapped to propeptide clusters by identifying clusters containing peptides from the same Uniprot protein. Graphics were created with PyX (http://pyx.sourceforge.net/) and ImageMagick (http://www.imagemagick.org).

3.4 Accuracy of models

An ~10-fold cross-validation was performed to estimate the expected performance of the models. Cross-validation was performed for each cluster independent of the others. Testing and training sets of peptides were created by randomly assigning peptides in a cluster to a number of sets of approximately equal size. Where the cluster had 10 or more peptides, 10% of the peptides were assigned to each set. Where the number of peptides in a cluster was not evenly divisible by 10, additional peptides were randomly assigned to sets (allowing only one additional peptide per set) until all peptides were assigned to exactly one set. Where a cluster had <10 peptides, one peptide was assigned to each of N sets, where N is the number of peptides in the cluster. By selecting one set, in turn, as the positive data for the test set and the other sets as positive data for the training sets, the sets of data were prepared to give an ~10-fold cross-validation for clusters having more than ~10 peptides, and leave-one-out cross-validation for clusters having less than 10 peptides. In all cases, peptides from all other clusters were taken as negative test data (HMMs do not use negative training data). Since the software system was intended to identify unrecognized AMPs from proteins, the system will not attempt to recognize additional peptides from a protein already known to contain AMPs. Therefore, performing a cross-validation was done using only clusters where each peptide was derived from unique proteins. This avoids the situation where a test peptide is automatically considered a positive match since it belongs to the same protein as a training peptide. In addition, for HMMs to be created, at least two peptides are required; therefore, only clusters of size three or greater were evaluated (so that one peptide would be available for the test set).

The same procedure was used during validation as was used in identifying additional AMPs from Swiss-Prot. For each cluster, the training peptides were used to create an HMM. Since the purpose of the method is to identify AMPs from within full proteins, the HMM was used to scan the full Swiss-Prot protein corresponding to the test peptides. A BLAST search was performed between the training peptides and the corresponding Swiss-Prot proteins. As before, positives were defined when proteins passed the conditions that the region matched by the HMMs and the feature annotation agreed to at least 90% of the their length, and the best-matching peptide had at least 50% identity to the Swiss-Prot protein.

3.5 Online tools

The web site uses a Perl CGI script running on an Apache Linux server with a MySQL RDBMS. Online sequence analysis uses utilities from the HMMER package.

ACKNOWLEDGEMENTS

CDF is supported by a Doctoral Research Award from the Canadian Institutes for Health Research.

Conflict of Interest: none declared.

WEB RESOURCES

Biochemistry Department University of Triest, Italy: http://www.bbcm.units.it/~tossi/pag5.htm
HMMER: http://hmmer.wustl.edu/
Uniprot database: http://www.pir.uniprot.org/

REFERENCES

Altschul,S.F. *et al*. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.

Brahmachary,M. *et al*. (2004) ANTIMIC: a database of antimicrobial sequences. *Nucleic Acids Res*., **32**, 1–589.

Bowdish,D.M. *et al*. (2005) A re-evaluation of the role of host defence peptides in mammalian immunity. *Curr. Protein Pept. Sci*., **6**, 35–51.

Brogden,K.A. (2005) Antimicrobial peptides: pore formers or metabolic inhibitors in bacteria? *Nat. Rev. Microbiol*., **3**, 238–250.

Chapple,D.S. *et al*. (2004) Structure and association of human lactoferrin peptides with Escherichia coli lipopolysaccharide. *Antimicrob. Agents Chemother*., **48**, 2190–2198.

Durbin,R. *et al*. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University press, Cambridge, UK.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Finlay,B.B. and Hancock,R.E.W. (2004) Can innate immunity be enhance to treat microbial infections? *Nat. Rev. Microbiol*., **2**, 497–504.

Hamilton-Miller,J.M.T. (2004) Antibiotic resistance from two perspectives: man and microbe. *Int. J. Antimicrobial Agents*, **23**, 209–212.

Hancock,R.E.W. (2001) Cationic peptides: effectors in innate immunity and novel antimicrobials. *The Lancet Infectious Diseases*, **1**, 156–164.

Hancock,R.E.W. and Rozek,A. (2002) Role of membranes in the activities of antimicrobial cationic peptides. *FEMS Microbiol. Lett*., **206**, 143–149.

Hwang,P.M. and Vogel,H.J. (1998) Structure-function relationships of antimicrobial peptides. *Biochem. Cell Biol*., **76**, 235–246.

Jack,R.W. *et al*. (1995) Bacteriocins of gram-positive bacteria. *Microbiol. Rev*., **59**, 171–200.

Koczulla,A.R. and Bals,R. (2003) Antimicrobial peptides: current status and therapeutic potential. *Drugs*, **63**, 389–407.

Levy,S.B. and Marshall,B. (2004) Antibacterial resistance worldwide: causes, challenges and responses. *Nature Medicine*, **10**, S122–S129.

Madera,M. and Gough,J. (2002) A comparison of profile hidden Markov model procedures for remote homology detection. *Nucleic Acids Res*., **30**, 4321–4328.

Park,J. *et al*. (1998) Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol*., **284**, 1201–1210.

Patrzykat,A. *et al*. (2002) Sublethal concentrations of pleurocidin-derived antimicrobial peptides inhibit macromolecular synthesis in Escherichia coli. *Antimicrob. Agents Chemother*., **46**, 605–614.

Powers,J.P.S. and Hancock,R.E.W. (2003) The relationship between peptide structure and antibacterial activity. *Peptides*, **24**, 1681–1691.

Schutte,B.C. *et al*. (2002) Discovery of five conserved beta-defensin gene clusters using a computational search strategy. *Proc. Natl. Acad. Sci. USA*, **99**, 2129–2133.

Scheetz,T. *et al*. (2002) Genomics-based approaches to gene discovery in innate immunity. *Immunol Rev*.

Sima,P. *et al*. (2003a) Mammalian antibiotic peptides. *Folia Microbiol*., **48**, 123–137.

Sima,P. *et al*. (2003b) Non-mammalian vertebrate antibiotic peptides. *Folia Microbiol*., **48**, 709–724.

Sing,T. *et al*. (2005) ROCR: visualizing classifier performance in R. *Bioinformatics*, **21**, 3940–3941.

Thompson,J.D. *et al*. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*., **22**, 4673–4680.

Yeaman,M.R. and Yount,N.Y. (2003) Mechanisms of antimicrobial peptide action and resistance. *Pharmacol Rev*., **55**, 27–55.

Yount,N.Y. and Yeaman,M.R. (2004) Multidimensional signatures in antimicrobial peptides. *PNAS*, **101**, 7363–7368.