Antimicrobial Potency of Cationic Antimicrobial Peptides can be Predicted from their Amino Acid Composition: Application to the Detection of "Cryptic" Antimicrobial Peptides.

Katia Pane[a¶], Lorenzo Durante[a¶], Orlando Crescenzi[b], Valeria Cafaro[a], Elio Pizzo[a], Mario Varcamonti[a], Anna Zanfardino[a], Viviana Izzo[c], Alberto Di Donato[a], Eugenio Notomista[a]*

[a]Department of Biology, Università degli Studi di Napoli Federico II, Napoli, Italy

[b]Department of Chemical Sciences, Università degli Studi di Napoli Federico II, Napoli, Italy

[c]Department of Medicine, Surgery and Dentistry, Università degli Studi di Salerno, Baronissi, Italy

*Corresponding author

E-mail: notomist@unina.it

[¶]These authors contributed equally to this work.

Supplemental Section 1: Screening of large pools of proteins

In order to screen large numbers of sequences we implemented in a GNU Awk script a procedure that reads in a list of protein sequences, generates all possible sub-peptides (within a range of lengths), and computes a score for each one of them. Depending on the number of proteins in the list, and on their size distribution, the number of possible peptides can be very high: however, the calculation of AS value is not demanding from the computational point of view, and the approach we adopted proved perfectly adequate for the cases of interest. Moreover, the analysis of a pool of proteins represents a typical "embarrassingly parallel" problem, which can be subdivided among different machines by simply splitting the pool of proteins into several lists, with no particular programming efforts nor infrastructural requirements. In principle, therefore, the procedure is feasible even for very large pools of proteins.

The main difficulty in the analysis of large amounts of sequences is to provide a simple output allowing selecting very interesting candidates to be analyzed more thoroughly. It should be noted that, especially in large proteins, several CAMP-like sequences could be present and that these fragments could be distant, adjacent or even partially overlapped. We have developed a simple iterative procedure to manage these different cases using the following steps: (i) the first analysis cycle, for each sequence, provides the best scoring peptide; (ii) the second cycle provides the best scoring peptide whose central residue is outside the edges of the best scoring peptide identified in the previous cycle; (iii) the procedure is iterated until no peptide with a AS higher than a user-defined threshold is found. For each protein the procedure provides a list of different or at most partially overlapped potential CAMPs. Auxiliary Awk scripts have been prepared to automate the data analysis procedure. The results can be used to plot a simple bar chart where each bar shows the position of a high scoring peptide.

We tested this method by analyzing two pools of proteins of different complexity, α-defensins (136 sequences from pfam database; pfam family PF00323), and cathelicidins (193 sequences; pfam family PF00666).

α-defensins, like cathelicidins, are secreted as precursors with a pro-peptide at the N-terminus. Differently from cathelicidin, α-defensins are homogeneous and well conserved CAMPs characterized by the presence of three disulfides, so that homology can be recognized directly in the region with antimicrobial activity [1]. All the precursors showed a high scoring peptide at the C-terminus as expected (Dataset S1, sheets A-C), furthermore the position of all the experimentally characterized α-defensins was predicted with very good precision (Dataset S1, sheet A). By choosing as threshold AS = 6.2 (corresponding to MIC = 300 μM) only 15 precursors were not detected (Dataset S1, sheet C). Very interestingly our analysis would suggest that at least some α-defensins (in particular those with scores higher than 8, corresponding to MIC values lower than 10 μM) could posses significant antimicrobial activity also in the denatured and reduced state as observed in the case of β-defensin 1. Some experimental data are in agreement with this hypothesis: human α-defensin 5 (score = 6.9, predicted MIC = 100 μM) in the reduced form is significantly less active than in the native form [2]; a mutant of mouse α-defensin 4 with the six cysteine residues mutated to alanine (score = 11.4, predicted MIC < 1 μM) is more active than the folded protein [3]. For five precursors of experimentally characterized α-defensins (Dataset S1; sheet A, chart 4) the window analysis indicated the presence of a high scoring peptide at the N-terminus. These peptides correspond almost exactly to known signal peptides (Dataset S1; sheet A, chart 4). Given the functional role of signal peptide, a partial overlap between the molecular features of CAMPs and signal peptides is not surprising. Moreover, Kim and co-workers demonstrated that the signal peptide of human methionine sulfoxide reductase B3 has antimicrobial activity [4].

Cathelicidins, usually, contain an antimicrobial peptide at the C-terminus of a cathelin domain [5]. Even if cathelin domains are homologous and easily identifiable at the sequence level, the antimicrobial peptides at the C-termini are not homologous and very different from species to species, spanning from thirteen residues –in the case of bovine indolicidin– to almost forty residues. As a consequence, *pfam* family PF00666 contains proteins which share a cathelin domain but not necessarily an antimicrobial peptide sequence, in general, or a CAMP in particular. All the experimentally studied cathelicidins were correctly identified at the C-terminus (Dataset S1; sheet D). Potential CAMPs were also identified at the C-terminus of several uncharacterized precursors (Dataset S1; sheet E, charts 1-7). However, 42 sequences included in the *pfam* family PF00666, according to our analysis, does not contain at the C-terminus a peptide with an AS value higher than the threshold (Dataset S1; sheet E, charts 8-9). Of these precursors, 21 are incomplete sequences and were not further analyzed (Dataset S1; sheet E, chart 9). The remaining 21 entries (Dataset S1; sheet E, chart 8) show several differences from the canonical cathelicidins. For example UniProt entry H9GBP9, a cathelicidin from american chameleon (*Anolis carolinensis*), contains at the C-terminus a peptide composed prevalently by glutamate and lysine residues with negative net charge and few hydrophobic residues (three valine residues). Even if we cannot exclude that it is an antimicrobial peptide, obviously, it is not a CAMP. On the other hand, five entries (P15175, G3HKQ4, F7DVN7, G1Q4G3 and F7ARW0) contain only six residues or less, downstream the last cysteine residue of the cathelin domain. Therefore, it is likely that these proteins do not contain any type of cryptic peptide downstream the cathelin domain. It is also interesting to note that in the cathelicidin family the frequency of precursors with high scoring peptides overlapped to known or potential signal peptides is much higher than among α-defensins (64% and 6%, respectively). The meaning of this difference, if any, remains to be determined.

Computational details of the analysis of large sequence pools

The window extraction and peptide scoring procedure was implemented in a self-contained GNU Awk script. The script reads in from user-specified input streams the control data for the run, the sets of exponents and the corresponding maximum scores (pre-computed over a range of lengths), and the lists of residue-specific hydrophobicity and charge values. Another input stream provides the protein sequences in FASTA format. As soon as a protein sequence has been read in, all possible sub-peptides are generated by means of two nested loops (over peptide length, and over starting position), and the corresponding scores are computed by a user-defined function. The output contains the FASTA protein identifiers and, for each sub-peptide generated, position, length, and scores on a single line.

The computational performance of the script was not characterized in detail. However, for the protein sets of interest, the execution times were quite reasonable. Thus, for example, the complete analysis of a set of 192 proteins (ranging in length from 37 to 276 residues, for a total of over 30,000 residues), took just over 30 seconds on single Linux workstation (equipped with 2 Intel Xeon quad-core X5550 processors cadenced at 2.67 GHz).

Auxiliary scripts have also been employed to filter out irrelevant data from the (often sizeable) complete output file.

References

[1] R.I. Lehrer and W. Lu, alpha-Defensins in human innate immunity, Immunol Rev 245 (2012) 84-112.

[2] H. Tanabe, T. Ayabe, A. Maemoto, C. Ishikawa, Y. Inaba, R. Sato, K. Moriichi, K. Okamoto, J. Watari, T. Kono, T. Ashida and Y. Kohgo, Denatured human alpha-defensin attenuates the bactericidal activity and the stability against enzymatic digestion, Biochem Biophys Res Commun 358 (2007) 349-55.

[3] A. Maemoto, X. Qu, K.J. Rosengren, H. Tanabe, A. Henschen-Edman, D.J. Craik and A.J. Ouellette, Functional analysis of the alpha-defensin disulfide array in mouse cryptdin-4, J Biol Chem 279 (2004) 44188-96.

[4] Y. Kim, G.H. Kwak, C. Lee and H.Y. Kim, Identification of an antimicrobial peptide from human methionine sulfoxide reductase B3, BMB Rep 44 (2011) 669-73.

[5] M. Pazgier, B. Ericksen, M. Ling, E. Toth, J. Shi, X. Li, A. Galliher-Beckley, L. Lan, G. Zou, C. Zhan, W. Yuan, E. Pozharski and W. Lu, Structural and functional analysis of the pro-domain of human cathelicidin, LL-37, Biochemistry 52 (2013) 1547-58.

Dataset S1. Window analysis of α-defensins (pfam family PF00323; sheets A, B and C) and cathelicidins (pfam family PF00666; sheets D and E) precursors. Each bar in the bar charts indicates the position of a high scoring peptide inside the precursor (AS is shown below the bar), the position of an experimental antimicrobial peptide reported in the UniProt database (acronym UPD is reported below the bar) or the position of a signal peptide reported in the UniProt database (acronym SP is reported below the bar).

Sheets A and B report the analysis of α-defensins precursors with at least a peptide with AS higher than the chosen threshold (AS = 6.2, corresponding to a MIC = 300 μM). Sheet A reports the analysis of α-defensins for which the position of the antimicrobial peptide is described in the UniProt database.

Sheet C reports the analysis of α-defensins precursors showing only peptides with AS < 6.2.

Some sequences, shorter then the average of α-defensin precursors, correspond to entries of the database containing only mature peptides instead of the prepropeptide sequences (sheet A, chart 1; sheet B, chart 1; sheet C, chart 1).

Sheet D reports the analysis of cathelicidins for which the position of the antimicrobial peptide is described in the UniProt database. For clarity sequences are arranged in the charts according to their source or type: FALL-39 homologues, charts 1 and 2; mammalian cathelicidins, charts 3 and 4; chicken and snake cathelicidins, chart 5; precursors of short amidated peptides (e.g. pig protegrins), chart 6; precursors of long amidated peptides (e.g. pig protegrins), chart 7. AS of peptides in chart 6 and 7 were calculated without taking into account amidation of C-termini. In chart 7 protein PF11_PIG (P51524) is the precursor of two pig CAMPs, tritrpticin and prophenin (13 and 79 residues, respectively, separated only by a cationic tetrapeptide containing the cleavage signal). Tritrpticin is a typical CAMP, whereas prophenin is a proline/phenylalanine/arginine-rich peptide. The highest scoring peptide corresponds to tripticin plus the cationic tetrapeptide and the N-terminus of prophenin. Precursor PF12_PIG (P51525) is an isoform of PF11_PIG (P51524).

Sheet E reports the analysis of cathelicidins for which the position(s) of the antimicrobial peptide(s) is(are) not described in the UniProt database. For clarity sequences are arranged in the charts according to results of the analysis: charts 1 and 2 show precursors containing a single high scoring peptide at or near the C-terminus (excluding, when present, the high scoring peptide at the N-terminus overlapped to the signal peptide region); charts 3 to 7 show precursors containing a high scoring peptide at or near the C-terminus and one or more additional high scoring peptides (peptides at or near the C-terminus are always those with the highest score with few exceptions shown in the right part of chart 7); chart 8 shows precursors without a high scoring peptide at or near the C-terminus; chart 9 shows the analysis of several incomplete entries (at least at the moment family PF00666 was downloaded from pfam database).