Physicochemical Determinants of Antimicrobial Activity

Daniel Veltri[1] Amarda Shehu[2,3,*]
[1]School of Systems Biology,
[2]Department of Computer Science,
[3]Department of Bioengineering,
George Mason University,
Fairfax, VA 22030
*amarda@gmu.edu

Abstract

Antimicrobial peptides (AMPs) are lately receiving significant attention as targets for antibacterial drug research. While many machine learning techniques are shown effective for AMP recognition, their utility for the rational design of novel AMP-based drugs in the wet laboratory is questionable. In this paper we seek to elucidate determinants of antimicrobial activity in a well-studied class of AMPs, cathelicidins. We do so by considering an extensive set of physicochemical properties at the residue level as features in the context of SVM-based classification, employing a carefully-constructed decoy dataset. A detailed statistical analysis of feature profiles reveals interesting physicochemical properties to preserve when modifying or designing novel AMPs in the wet laboratory. The method presented here is a first step towards assisting *de novo* design of AMPs in the wet laboratory.

1 Introduction

As bacterial resistance to known antibiotics grows, new targets are sought for antibacterial drug research. Antimicrobial peptides (AMPs) constitute a promising class of potential targets due to diverse killing mechanisms and high activity against a broad spectrum of bacteria. Understanding what confers activity to AMPs is central to wet-lab efforts on de novo design of AMP-based antibacterial drugs [19].

Many machine learning methods are devised to address AMP recognition. Table 1 summarizes the state of the art. Direct comparisons are hard to draw, however, due to great diversity amongst methods in terms of the algorithm employed, the features constructed, and the positive and negative datasets used to demonstrate AMP recognition. Moreover, although such methods are touted as black boxes for automatic recognition, their value for drawing rules for wet-lab design of novel AMP-based drugs has not been demonstrated. It is yet unclear how one can modify the sequence of a peptide for antimicrobial activity [19].

Table 1 shows that great recognition accuracy can currently be obtained. For instance, recent work in [4] achieves a high MCC value of 0.94. Interestingly, the high recognition accuracy of AMPs is obtained with only two features calculated over the peptide sequence, length (number of amino acids) and propensity for aggregation. Other interesting physicochemical properties proven useful in the wet lab for modifying antimicrobial activity of various AMPs (such as, hydrophobicity, propensity for certain secondary structures, and more) were shown not to be important for automatic recognition [4]. Differences between what is shown important in the dry-lab versus the wet-lab highlight the question as to what the features employed for the recognition are really capturing, and what role, if any, the employed negative dataset plays. A hastily constructed negative dataset can potentially bias the model towards features that capture differences in characteristics other than activity.

In this paper we argue that better progress can be made by focusing on a specific well-studied class of AMPs that allows for constructing a high-quality decoy dataset. The first contribution of this paper is a carefully-constructed decoy dataset. Controlling the extent to which the negative dataset resembles the positive dataset is important to extract features that relate to antimicrobial activity rather than other potentially trivial differences. We focus here on cathelicidins and construct a high-quality decoy dataset that resembles cathelicidins in properties other than antimicrobial activity. The second contribution of this paper is in the construction of features. Rather than relying on global features computed over the entire sequence of a peptide, we employ local features. We consider at each residue an extensive list of known physicochemical properties documented in the AAIndex [11] database. The third contribution of this paper is a detailed statistical analysis that highlights top features related to antimicrobial activity on a per-residue level.

The AAIndex-based features we employ here are shown useful in the context of SVM-based classification. The validation proceeds in two steps. We first show the general relevance of such features beyond cathelicidins in the context of SVM-based comparison with other work. Our comparison is limited to work where the method can be re-implemented and the results reproduced. In many of the methods listed in Table 1, it is difficult or not possible to reproduce the datasets due to the presence of peptides of variable length or of synthetic AMPs of unusual sequence com-

position. In some cases, negative datasets do not explicitly have member peptides listed [12, 13, 17]. After establishing the general relevance of the features, we then take a detailed look at the top features reported from our analysis. Many of these features are shown to be in strong agreement with properties demonstrated as important for antimicrobial activity in the wet lab. Our analysis also elucidates novel features that may be interesting to biological researchers. We believe the feature profiles presented in this paper are a first step towards obtaining a better understanding of what confers cathelicidins their activity and aiding the wet-lab modification or design of novel AMP-based drugs.

We now proceed with a description of methods and results, followed by a summary and discussion of future work. All the datasets employed in this paper and the top features reported by our analysis are made available for download at: `http://binf.gmu.edu/dveltri/bicob2013`.

Algorithm	MCC			Database
	Training Dataset	Validation Dataset	Testing Dataset	
HMM [5]		0.98		AMPer
HMM [6]		0.88		RANDOM
ANN [2]		0.60		CAMEL QSAR
DA [16]	0.75		0.74	CAMP
RF [16]	0.86		0.86	CAMP
SVM [16]	0.88		0.82	CAMP
SVM [13]			0.84	AntiBP2
ANFIS [4]		0.94		APD2
ANN [4]		0.85		APD2

Table 1: Summary of algorithms and their performance on datasets drawn from different databases.

2 Methods

2.1 Dataset Generation

Two positive datasets are constructed, each consisting of 18-residue long N- and C-terminal cathelicidin subsequences extracted from from various databases. The reason for considering the termini individually in two different datasets (thus building two SVM models) is due to the fact that a different subset of features may be relevant for each of the termini. The individual role and contribution to activity of the termini is not yet conclusive [21]. Since the first four N-terminal residues in a cathelicidin are involved in enzymatic cleavage, it is important to differentiate top features that may be important for activity rather than cleavage. A third dataset is employed for this purpose, which consists of neutrophil elastase-cleaved substrates. A statistical test detailed below identifies top N-termini features reported by the SVM that are not statistically different from those found in this third dataset. This information is used to essentially lower confidence in such features, which is of particular use in a wet-lab design study aiming to modify features for activity rather than cleavage.

2.1.1 Positive Datasets Extracted from Cathelicidins

A total of 45 mature (activated post-cleavage) cathelicidin sequences with no more than 90% sequence identity were collected; 10 were extracted from the Antimicrobial Peptide Database (APD2) [20], a repository of AMPs extracted from literature, and the rest from UniProt [14] (details provided in Supplementary Material).

2.1.2 Negative Datasets of Decoy Sequences

HeliQuest [7] was used to generate two negative sets of 18-residue long sequences to be paired with the positive N- and C-termini datasets. The length limit is due to the maximum scan-length allowed by the server. The negative sequences are designed to be helical through HeliQuest, so they can share this structural characteristic with cathelicidins. The profile was constructed based on consensus patterns. For the N-terminus, a consensus pattern of KRR[RL]GLF[RL][KR]KAR[KE]KIKKG was determined (amino acids in brackets represent an equal number of observations at that position), resulting in 16 different sequences based on ties at positions 4, 8, 9, and 13. The sequences were passed on to the screening module of HeliQuest to obtain peptides extracted from UniProt. Entries were limited to the human proteome, and the set was further reduced to a sequence identity of less than 50%. Additionally, UniProt annotations were used to excise entries with antimicrobial, anti-fungal, anti-viral or cytotoxic activity and those with cellular location other than the cytoplasm. 180 sequences were then drawn at random from the remaining entries (resulting in a $1:4$ positive-to-negative sample ratio). The process was repeated to obtain 180 C-terminal decoys using the consensus pattern KIGQKIKDFLGI[LP]VPRTG.

Similarity between the negative and positive datasets was checked through dimensionality reduction techniques. Results for Principal Component Analysis (first 5 principal components) and Local Linear Embedding ($k = 10$) are provided in Supplementary Materials. Neither technique was able to adequately separate the two classes at either termini, suggesting an expected high level of classification difficulty with the constructed datasets.

The third negative dataset employed for the statistical analysis consists of neutrophil elastase substrates. 45 non-AMP substrates were extracted from the PMAP-CutDB Proteolytic Event Database [8], provided as 8-mers centered about the cleavage site. The 4 residues upstream of cleavage were discarded. The analysis below compares features present in this (substrate) dataset with those over the first 4 N-termini residues for the 45 peptides in the positive dataset. The goal is to mark or discard features identified as important by the SVM but found present in the substrate dataset.

2.2 Feature Construction

Each sequence was converted into a numeric vector by expanding each position into a list of all known physicochemical properties for the amino acid at that position. We employ the AAIndex (AAIndex1,Vr.9) [11], which is a collection of 544 quantified amino-acid physicochemical properties obtained from literature. Removing 13 entries which contain "NA" values leaves 531 properties per amino acid. This set is comprehensive but presents problems for long sequences. While all 531 features are employed for the neutrophil elastase dataset that contains only 4-residue long sequences (essentially converting each sequence into a numeric vector of $2124 = 531 \times 4$ elements), the feature list is reduced for the datasets with 18 residue-long sequences. Removing entries found to share $\pm 80\%$ or greater correlation, as defined in [11], reduces this set down to 299 features. This allows for mapping each 18-residue long sequence into a vector of $5382 = 299 \times 18$ elements. We include some more information into the vectors, by arranging them as: $\{C, (R_1, X_1), \ldots, (R_n, X_1), (R_1, X_2), \ldots, (R_n, X_{299})\}$, where C is a class label, R_i is a residue over n positions, and X_j is an AAIndex entry over the entries considered. This format allows any feature to be traced back to a specific physicochemical property at a particular residue position.

2.3 SVM Classification

Two SVM models are trained separately on the N- and C-termini datasets using LibSVM [3]. Both the RBF and Linear kernels are used and found to yield similar performance. Kernel parameters and the SVM cost functlion are tuned through the standard grid search mechanism. Features are scaled from -1 to 1. Results are obtained after 3-fold cross-validation on each of the training datasets.

2.4 Feature Selection Based on F-score Ranking

The F-scores that SVM models associate with support vectors provide an estimate of the discriminatory power of features. The F-score measures the discrimination of two sets of real numbers, with a higher score denoting a feature with better discriminatory power. We employ the F-score to elucidate the top ranking features after the SVM training. Essentially, features with the highest F-scores are added iteratively, and the classification performance is evaluated until a minimum set of features with the lowest validation error are detected. Further details are available in Supplementary Material, and the protocol can be found in [1].

2.5 Cleavage Site Analysis

A statistical approach is used to evaluate if features of cleavage site amino acids (N-termini residues $1-4$) in cathelicidins are different from those in a set of natural, yet non-AMP, neutrophil elastase substrates. The dataset of 45 substrates was prepared as described above. The dataset of cathelidicin cleavage sites consists of the first 4 amino acids of the same N-termini subsequences in the N-termini positive dataset employed for SVM classification.

Each feature is treated separately, and most are not normally distributed (data not shown). The Brown-Forsythe test is conducted [15] to assess the quality of variance between feature populations of the two datasets. Features with differing variance ($p \leq 0.05, \alpha = 0.95$), shown to be statistically independent, are removed. Remaining features are passed on to a second round of assessment with the Mann-Whitney-Wilcoxon Test. Features shown to be statistically independent from the test ($p \leq 0.05, \alpha = 0.95$) are removed. Remaining features represent those that cannot be confidently associated with antimicrobial activity over protease specificity. These features can now be removed, or marked, in the list of top features reported by the SVM-based feature selection technique described above. This annotation allows biologists to focus on features according to the confidence with which they consider them relevant for antimicrobial activity. For analysis of additional AMP substrates not cleaved by neutrophil elastase, this process can be repeated given sufficient non-AMP cleavage examples. Full results of the tests described here can be found in the Supplementary Material.

3 Results

3.1 Summary of SVM Performance

SVM analysis was conducted using 3-fold validation and the RBF kernel. Performance on the N-termini dataset demonstrated an average ACC of 94.67% and an MCC of 0.83. ACC and MCC values

on the C-termini dataset were 93.33% and 0.78, respectively, as summarized in Table 2. Average ROC curves shown in Figure 1 further make the case that the employed features allow an SVM to achieve high classification accuracy.

Dataset	Sen.(%)	Spec.(%)	ACC(%)	MCC
N-Term.	95.21	94.80	94.67	0.8277
C-Term.	90.56	93.75	93.33	0.7761

Table 2: Performance on N- and C-termini datasets.

The rest of the analysis highlights features relevant for antimicrobial activity. The analysis employs ranking based on SVM-based F-scores and explicitly discounts features potentially important for cleavage rather than activity.

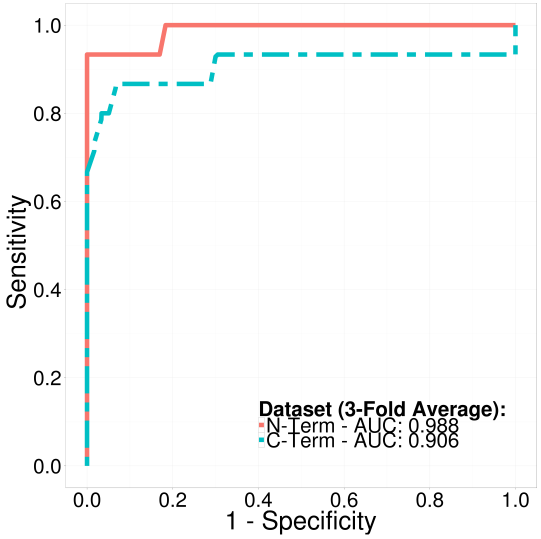

Figure 1: ROC results on N- and C-termini datasets. Average ACC values in Table 2 correspond to the area under the ROC curves.

3.2 Cleavage Site Analysis

A total of 2124 (4×531) features were independently tested. The Brown-Forsythe test removed 510 features due to differing group variance ($p \leq 0.05$). Remaining features were fed to the Mann-Whitney-Wilcoxon test (two-tailed), and 77% were found not significantly different ($p > 0.05$). The 1243 features (aggregate over the 4 residues) potentially encode signals related to positive selection for protease specificity (available in Supplemental Materials). Inspection of ranked features reveals that cleavage-associated ones can be present starting at rank 122.

3.3 F-score based Feature Reduction

The 299 non-redundant physicochemical properties were used for each residue position to construct feature vectors for SVM training. F-scores obtained by the SVM were analyzed, and the selection procedure in Methods was employed to elucidate top features with high discriminatory power. Training on the entire N-termini dataset, the F-score based selection procedure reports a maximum ACC of 96% with 936 features (out of 299×18). On the full C-termini dataset, a maximum ACC of 96.4% was obtained with 520 features. The procedure was also implemented separately on the 3 folds reported by the SVM classification to obtain average ranks; a high ACC of 96.4% was obtained using an average of 529 features on N-termini, and a high ACC of 96.3% with an average of 647 features was obtained on C-termini.

R	P	F	ID	Description
1	7	0.217	WILM950102	Hydrophobicity coeff. in RP-HPLC ... (Wilce et al. '95)
2	3	0.184	SNEP660103	Principal comp. III (Sneath, '66)
3	12	0.177	FAUJ880111	Pos. charge (Fauchere et al., '88)
4	7	0.176	MEEJ800101	Retention coeff. in HPLC, pH7.4 (Meek, '80)
5	15	0.176	WILM950104	Hydrophobicity coeff. in RP-HPLC ... (Wilce et al. '95)
1	-4	0.334	BUNA790101	α-NH chem. shifts (Bundi-Wuthrich, '79)
2	-4	0.309	FINA910102	Helix initiation param. at i,i+1,i+2 (Finkelstein et al., '91)
3	-4	0.291	GEOR030109	Linker propensity from non-helical DSSP dataset (George-Heringa, 2003)
4	-4	0.287	GEOR030101	Linker propensity from all dataset (George-Heringa, 2003)
5	-12	0.250	JANJ790102	Transfer free energy (Janin, '79)

Table 3: Top 5 features for N-termini (top) and C-termini (bottom) datasets.

The top 5 features are shown in Table 3 with the remaining list available in Supplemental Materials. Column 2 shows the residue position of a mature peptide corresponding to a reported top feature. Rank R and positions P are shown in columns 1 and 2, respectively. Negative positions are reported in the C-termini dataset and count backwards from the C-terminus (-1 refers to final C-terminal residue). F-scores F are shown in column 3. Column 4 shows the AAIndex [11] ID corresponding to the physicochemical

property represented in each feature. Column 5 provides a brief explanation of each AAIndex entry, using source descriptions from [11].

3.4 Detailed Analysis of Biological Relevance of Top Features

A number of the reported top features have been found biologically important for activity in literature [18]. Notably, both hydrophobicity and charge are found important for attraction towards bacterial membranes [18]. These same features have also been used successfully for AMP recognition [4, 17]. Figure 2 shows a C-termini profile for a novel feature, BUNA790101, ranked $1st$ at position $i = -4$ and $19th$ at position $i = -6$ in Table 3. The AAIndex describes this feature as "alpha-NH chemical shifts (Bundi-Wuthrich, 1979)." Chemical shifts describe the electronic environment surrounding a peptide [9] and this may have further relevance to membrane interactions. Table 4 outlines correlated entries, and how they may help direct literature searches.

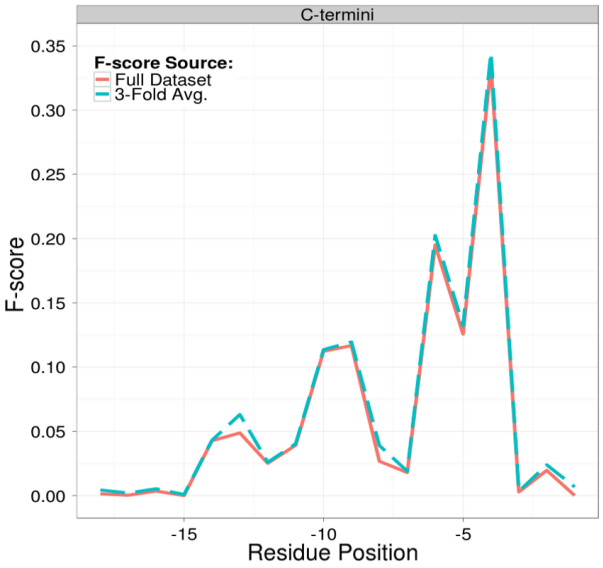

Figure 2: Profile for top-ranked feature BUNA790101.

To show that the features employed here are relevant beyond cathelicidins, a generalized AMP dataset from [4] was considered. A direct comparison of local-to-global features was not possible on this dataset as peptides were of variable length (as required for the SVM). A compromise was found by averaging features across all amino acids to create fixed-length vectors (details in the Supplemental Material). Comparable accuracies in the [80.0−85.9%] range were obtained (details in the Supplementary Material). Interestingly, when the 8 global features used in [4] were incorporated into the feature space, accuracy increased to 93%. Of these, *in vitro* peptide aggregation ranked as the top feature for activity and is in agreement with previous findings [4, 17]. AAIndex features made up the remainder of the top ten features in this set, including WEBA780101 ("RF value in high salt chromatography") which may have relevance to the salt sensitivity of helical AMPs. Experimental studies have shown that peptides particularly sensitive to salt concentrations may assume helical structures too early and reduce performance in disrupting bacterial membranes [18].

ID	Corr.	Description
BLAM930101	0.95	**Alpha helix prop.** of posit. 44 in T4 lysozyme (Blaber et al., '93) **Structural basis of amino acid α helix propensity**
ONEK900101	0.91	Delta G values for the pep. extrapolated to 0 M urea (O'Neil-DeGrado, '90) A thermodynamic scale for **helix-forming tendencies**
ROBB760104	0.82	Information measure for **C-terminal helix** (Robson-Suzuki, '76) Conformational properties of amino acid residues in globular proteins
FAUJ880113	0.82	pK-a(RCOOH) (Fauchere et al., '88) **side chain parameters for correlation studies in bio and pharma.**

Table 4: Other features positively correlated to BUNA790101 (correlations shown in column 2). Descriptions in column 3 contain key words relevant to AMPs (in bold for emphasis).

Additional files, including the full list of F-scores, cleavage analysis results and physicochemical profiles for all features, are available online (`http://binf.gmu.edu/dveltri/bicob2013`).

4 Summary

This paper has presented a supervised learning method for elucidating activity-related physicochemical features at the local amino acid level. While global (whole-peptide) classification schemes have proven useful for screening for new AMPs, features effectively become a "black box" in regards to activity. It is hoped a shift towards applying features which can identify specific motifs related to antimicrobial activity will be of more use to aid wet-labs in AMP modification or directed design of novel AMPs. Ongoing work to this aim focuses on incorporating correlations through

spectrum features. To maintain a manageable sized feature space, an obvious direction is to build on the work presented here and employ only a few top features reported for each amino-acid [10].

Acknowledgements

Work is supported in part by NSF Grant No. DGE-1007911 (GK-12 Award No. 0638680). We thank Uday Kamath for help with scripting interfaces.

References

[1] Yi-Wei Chen and Chih-Jen Lin. Combining SVMs with various feature selection strategies. In Isabelle Guyon, Masoud Nikravesh, Steve Gunn, and Lotfi Zadeh, editors, *Feature Extraction*, volume 207 of *Studies in Fuzziness and Soft Computing*, pages 315–324. Springer Berlin, Heidelberg, 2006.

[2] A. Cherkasov and B. Jankovic. Application of 'inductive' QSAR descriptors for quantification of antibacterial activity of cationic polypeptides. *Moleculess*, 9(12):10341052, 2004.

[3] R.-E. Fan, P.-H. Chen, and C.-J. Lin. Working set selection using the second order information for training SVM. *J. Mach. Learn. Res.*, 6(1532-4435):1889–1918, 2005.

[4] F. C. Fernandes, D. J. Rigden, and O. L. Franco. Prediction of antimicrobial peptides based on the adaptive neuro-fuzzy inference system application. *Peptide Science*, 98(4):280–287, 2012.

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

[6] C. D. Fjell, H. Jenssen, K. Hilpert, W. A. Cheung, N. Pante, R. E. Hancock, and A Cherkasov. Identification of novel antibacterial peptides by chemoinformatics and machine learning. *J. Med. Chem.*, 52(7):2006–2015, 2009.

[7] R. Gautier, D. Douguet, B. Antonny, and Drin G. HELIQUEST: a web server to screen sequences with specific α-helical properties. *Bioinformatics*, 24(18):2101–2102, 2008.

[8] Y. Igarashi, A. Eroshkin, S. Gramatikova, G. Gramatikoff, Y. Zhang, J. W. Smith, A. L. Osterman, and Godzik A. CutDB: a proteolytic event database. *Nucl. Acids Res.*, 35:D546–D549, 2007.

[9] C.J. Jameson. Understanding NMR chemical shifts. *Annual review of phys chem*, 47(1):135–169, 1996.

[10] U. Kamath, J. Compton, R. Islamaj-Dogan, De Jong K. A., and A. Shehu. n evolutionary algorithm approach for feature generation from sequence data and its application to dna splice-site prediction. *Trans Comp Biol and Bioinf*, 2012. in press.

[11] S. Kawashima and M. Kanehisa. AAindex: amino acid index database. *Nucl. Acids Res.*, 28(1):374, 2000.

[12] S. Lata, N. K. Mishra, and G. P. Raghava. AntiBP2: improved version of antibacterial peptide prediction. *BMC Bioinformatics*, 11(Suppl 1):S1–S19, 2010.

[13] S. Lata, B. K. Sharma, and G. P. Raghava. Analysis and prediction of antibacterial peptides. *BMC Bioinformatics*, 23(8):263–272, 2007.

[14] M. Magrane and the UniProt consortium. UniProt knowledgebase: a hub of integrated protein data. *Database*, 2011(bar009):1–13, 2011.

[15] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. ISBN 3-900051-07-0.

[16] S. Thomas, S. Karnik, R. S. Barai, V. K. Jayaraman, and S. I. Thomas. CAMP: a useful resource for research on antimicrobial peptides. *Nucl. Acids Res.*, 38(Suppl 1):D774–D780, 2009.

[17] M. Torrent, P. Di Tommaso, D. Pulido, M. V. Nogues, Notredame. C., E. Boix, and D. Andreu. AMPA: An automated web server for prediction of protein antimicrobial regions. *Bioinformatics*, 28(1):130–1, 2011.

[18] A. Tossi, L. Sandri, and A. Giangaspero. Amphipathic, α-helical antimicrobial peptides. *Peptide Science*, 55(1):4–30, 2000.

[19] G. Wang. *Antimicrobial Peptides: Discovery, Design and Novel Therapeutic Strategies*. CABI Bookshop, Wallingford, England, 2010.

[20] G. Wang, X. Li, and Z. Wang. APD2: the updated antimicrobial peptide database and its application in peptide design. *Nucl. Acids Res.*, 37(Suppl1):D933–D937, 2009.

[21] Igor Zelezetsky, Alessandra Pontillo, Luca Puzzi, Nikolinka Antcheva, Ludovica Segat, Sabrina Pacor, Sergio Crovella, and Alessandro Tossi. Evolution of the primate cathelicidin. *J. Biol. Chem.*, 281(29):19861–19871, 2006.