LETTERS

A linguistic model for the rational design of antimicrobial peptides

Christopher Loose[1]*, Kyle Jensen[1,2,3]*, Isidore Rigoutsos[1,4] & Gregory Stephanopoulos[1]

Antimicrobial peptides (AmPs) are small proteins that are used by the innate immune system to combat bacterial infection in multicellular eukaryotes[1]. There is mounting evidence that these peptides are less susceptible to bacterial resistance than traditional antibiotics and could form the basis for a new class of therapeutic agents[2]. Here we report the rational design of new AmPs that show limited homology to naturally occurring proteins but have strong bacteriostatic activity against several species of bacteria, including *Staphylococcus aureus* and *Bacillus anthracis*. These peptides were designed using a linguistic model of natural AmPs: we treated the amino-acid sequences of natural AmPs as a formal language and built a set of regular grammars to describe this language. We used this set of grammars to create new, unnatural AmP sequences. Our peptides conform to the formal syntax of natural antimicrobial peptides but populate a previously unexplored region of protein sequence space.

AmPs are ubiquitous among multicellular eukaryotes and are found in diverse contexts including frog skin, scorpion venom and human sweat. Recent studies show that some AmPs are active against pathogens that are resistant to traditional antibiotics such as penicillin, tetracycline and vancomycin[3]. In addition to their antibiotic uses, AmPs might have other interesting clinical applications: they act as adjuvants for the adaptive immune system[4] and might be useful in treating certain cancers[5].

The many disease-relevant actions of AmPs are a consequence of their broad ability to distinguish eukaryotic cells from pathogenic invaders. In general, AmPs have a net positive charge and an amphipathic three-dimensional structure that gives the peptides an electrostatic affinity to the outer leaflet of the microbial membrane[6]. This affinity leads to binding, disruption of the membrane and, ultimately, microbial cell death[7].

Our preliminary studies of natural AmPs indicated that their amphipathic structure gives rise to a modularity among the different AmP amino-acid sequences. The repeated usage of sequence modules—which might be a relic of evolutionary divergence and radiation—is reminiscent of phrases in a natural language, such as English. For example, the pattern QxEAGxLxKxxK (where 'x' is any amino acid) is found in more than 90% of the insect AmPs known as cecropins. On the basis of this observation, we modelled the AmP sequences as a formal language—a set of sentences using words from a fixed vocabulary. In this case, the vocabulary is the set of naturally occurring amino acids, represented by their one-letter symbols[8].

We conjectured that the 'language of AmPs' could be described by a set of regular grammars. Regular grammars are, in essence, simple rules that describe the allowed arrangements of words. These grammars, such as the cecropin pattern mentioned previously, are commonly written as regular expressions and are widely used to describe patterns in nucleotide and amino-acid sequences[9,10].

To find a set of regular grammars to describe AmPs we used the Teiresias pattern discovery tool[11]. With Teiresias, we derived a set of 684 regular grammars that occur commonly in 526 well-characterized eukaryotic AmP sequences from the Antimicrobial Peptide Database (APD)[12] (see Methods). Together, these ~700 grammars describe the 'language' of the AmP sequences. In this linguistic metaphor, the peptide sequences are analogous to sentences and the individual amino acids are analogous to the words in a sentence. Each grammar describes a common arrangement of amino acids, similar to popular phrases in English. For example, the frog AmP brevinin-1E contains the amino-acid sequence fragment PKIFCKITRK, which matches the grammar P[KAYS][ILN][FGI]C[KPSA][IV][TS][RKC][KR] from our database (the bracketed expression [KAYS] indicates that, at the second position in the grammar, lysine, alanine, tyrosine or serine is equally acceptable). On the basis of this match, we would say that the brevinin-1E fragment is 'grammatical'.

By design, each grammar in this set of ~700 grammars is ten amino-acids long and is specific to AmPs—at least 80% of the matches for each grammar in Swiss-Prot/TrEMBL[13] (the APD is a subset of Swiss-Prot/TrEMBL) are found in peptides annotated as AmPs.

To design unnatural AmPs, we combinatorially enumerated all grammatical sequences of length twenty (see Methods) in which each window of size ten in the 20-mers was matched by one of the ~700 grammars. We chose this length because we could easily chemically synthesize 20-mers and this length is close to the median length of AmPs in the APD. From this set, we removed any 20-mers that had six or more amino acids in a row in common with a naturally occurring AmP; we then clustered the remaining sequences on the basis of similarity, choosing 42 to synthesize and assay for antimicrobial activity. This process is illustrated in Fig. 1 and the designed sequences are shown as Supplementary Table 1.

For each of the 42 synthetic peptides, we also designed a shuffled sequence, in which the order of the amino acids was rearranged randomly so that the sequence did not match any grammars. These shuffled peptides had the same amino-acid composition as their synthetic counterparts and thus, the same molecular weight, charge and isoelectric point (bulk physiochemical factors that are often correlated with antimicrobial activity). We hypothesized that because the shuffled sequences were 'ungrammatical' they would have no antimicrobial activity, despite having the same bulk physiochemical characteristics. In addition, we selected eight peptides from the APD as positive controls and six 20-mers from non-antimicrobial proteins as negative controls. A complete list of designed and shuffled sequences, along with controls, is provided in the Supplementary Information.

[1]Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [2]Harvard–MIT Health Sciences and Technology, Cambridge, Massachusetts 02139, USA. [3]Agrivida, 411 Massachusetts Ave B1, Cambridge, Massachusetts 02139, USA. [4]IBM Research Division, Thomas J.Watson Research Center, Yorktown Heights, New York 10598, USA.
*These authors contributed equally to this work.

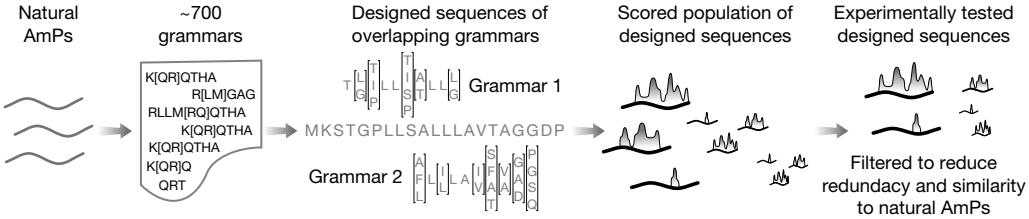

Figure 1 | A schematic of the *in silico* peptide design strategy. Grammars are induced from the set of natural AmP sequences using Teiresias. Overlapping grammars are stitched together to create new 20-amino-acid sequences that correspond to the antimicrobial syntax. Designed sequences are scored on the basis of the prevalence of the underlying grammar in natural peptides. After removing designs with six or more consecutive amino acids in common with a natural peptide, a representative set of sequences is selected for experimental validation.

We characterized the activity of each synthetic AmP using a broth microdilution assay described elsewhere[14]. This assay measures the minimum inhibitory concentration (MIC) at which the peptide inhibits growth of the target organism. Table 1 shows the MICs of synthetic peptides against *Bacillus cereus* and *Escherichia coli*, as representative gram-positive and gram-negative bacteria. Of the 40 soluble (two of the designed and four of the shuffled peptides were insoluble) designed peptides, 18 had an MIC of 256 μg ml^{-1} or less against at least one of the bacterial targets. Only two of the soluble, shuffled peptides showed antibacterial activity. Thus, the activity is not an artefact of molecular weight, charge or isoelectric point.

Of the negative controls—six peptides randomly selected from the middle of non-antimicrobial proteins from Swiss-Prot/TrEMBL—none had antimicrobial activity, whereas six of the eight naturally-occurring AmPs in the positive control group did.

To validate MIC determinations, we measured optical density at varying concentrations of a representative set of designed, shuffled, and natural peptides. Supplementary Fig. 1 shows that peptides inhibit growth at the MICs determined by visual inspection. In addition, plating of these samples confirmed that these peptide-treated bacterial populations were generally not viable at concentrations of peptides above the MIC.

Two of the designed peptides, D28 (FLGVVFKLASKVFPAVFGKV) and D51 (FLFRVASKVFPALIGKFKKK), inhibited *B. cereus* growth at 16 μg ml^{-1}, which is close to the MICs of the strong positive controls melittin and cecropin-melittin hybrid (8 μg ml^{-1}). Peptides with activity against gram-positive bacteria are particularly exciting because of the prevalence of drug-resistant nosocomial *S. aureus* and the threat of bioterror agents such as *B. anthracis*, or anthrax. Therefore, we assayed seven designed peptides that had such activity, including the highly active D28 and D51 peptides, against the Smith diffuse strain of *S. aureus* and the Sterne strain of *B. anthracis*. As shown in Table 2, all seven peptides were active against both bacteria at 256 μg ml^{-1} or less, whereas only one of the seven shuffled controls was active (the only active shuffled peptide had MICs of 128 μg ml^{-1} against *S. aureus* and 256 μg ml^{-1} against *B. anthracis*). Moreover, two designed peptides, D28 and D51, had MICs of 16 μg ml^{-1} against *Bacillus anthracis*, which is equivalent to the activity of Cecropin-melittin hybrid, a strong natural AmP. D28 also had an MIC of 8 μg ml^{-1} against *S. aureus*.

In an attempt to improve the antimicrobial activity of our designed peptides, we optimized our best candidate, peptide D28, using a heuristic approach. We created 44 variants of D28 by introducing mutations that were selected to increase positive charge, increase hydrophobicity, remove an interior proline residue, or improve segregation of positive and hydrophobic residues based on a helical projection. As shown in Supplementary Table 2, 18 of the 44 D28 variants showed improved activity against *E. coli*, *B. cereus* or *S. aureus*. Many of the D28 variants with improved activity against *B. cereus* included a mutation at an internal proline, either to lysine or glycine. D28 and six of its variants were assayed for bactericidal activity, and all had activity within a twofold dilution of their MIC. One variant (R8) had MICs of 16 μg ml^{-1} against *E. coli*, 8 μg ml^{-1} against *B. cereus* and 4 μg ml^{-1} against *S. aureus* (relative to 64, 16 and 8 μg ml^{-1}, respectively, for D28).

Our linguistic approach to designing synthetic AmPs might be successful because of the pronounced modular nature of natural AmP amino-acid sequences. As we have shown, this approach can be used to expand the AmP sequence space rationally without using structure–activity information or complex simulations of the interactions of a peptide with a membrane. The peptides that we designed are different from previously designed synthetic AmPs[15] in that they bear limited homology to any known protein, which might be desirable for AmPs used in clinical settings. Some researchers argue that widespread clinical use of AmPs that are too similar to human AmPs will inevitably elicit bacterial resistance, compromising our own natural defences and posing a threat to public health[16]. In addition, using our approach to develop an arsenal of diverse antimicrobial agents would further reduce concerns about the development of antimicrobial resistance. We hope that this approach will help to expand the diversity of known AmPs well beyond those found in nature, possibly leading to new candidates for AmP-based antibiotic therapeutics.

Our designed AmPs show some degree of homology with natural AmPs because the grammars are based on native sequences. Peptide D28, for example, was matched by grammars derived from 11 natural AmPs, including brevinin, temporin and ponericin. However, Smith-Waterman alignments of our designed peptides against all natural AmPs in the Swiss-Prot/TrEMBL database reveal that the degree of homology is, by design (see Methods), limited. In particular, our two most active peptides, D51 and D28, have only 50 and 60% sequence identity, respectively, with the nearest natural AmP.

Our linguistic design approach might be most valuable as a method for rationally constraining a sequence-based search for new AmPs. Diverse leads generated by our algorithms could be optimized using approaches described in the literature[17]. But the

Table 1 | MICs of peptides against bacterial targets

Class	*E. coli* (gram-negative)		*B. cereus* (gram-positive)		*E. coli* or *B. cereus*
	MIC ≤256 μg ml^{-1}	MIC ≤64 μg ml^{-1}	MIC ≤256 μg ml^{-1}	MIC ≤64 μg ml^{-1}	MIC ≤256 μg ml^{-1}
Designed	16/40	4/40	8/40	4/40	18/40
Shuffled	1/38	0/38	2/38	1/38	2/38
Natural AmPs	6/8	6/8	4/8	3/8	6/8

Table 2 | MICs of peptides against *S. aureus* and *B. anthracis*

MIC (μg ml^{-1})	*S. aureus*	*B. anthracis*
8	D28	
16	D51	D28, D51
32		
64	D22	D22
128	D63, S51	D5, D35, D43, D63
256	D5, D35, D43	S51
>256	S5, S22, S28 S35, S43, S63	S5, S22, S28, S35, S43, S63

D, designed on the basis of motifs; S, shuffled control.

linguistic approach described here has a number of limitations. First, sequence families that are poorly conserved on an amino-acid level would not benefit from this approach. Second, we suspect that the small size of AmPs is helpful. Owing to the simple nature of regular grammars, they would be less useful for designing larger proteins and, in particular, proteins with complex tertiary or quaternary structures.

METHODS

Computational design of unnatural AmPs. To design unnatural AmPs, we combinatorially enumerated all grammatical sequences based on the set of ~700 grammars (see Supplementary Materials, Methods, Derivation of grammars). First, for each grammar, we wrote out all possible grammatical amino-acid sequences. So, for example, the grammar [IVL]K[TEGDK]V[GA]K[AELNH][VA][GA]K produced 600 sequences, where $3 \times 5 \times 2 \times 5 \times 2 \times 2 = 600$, owing to the option of choosing one of many amino acids at each bracketed position. There are roughly 3 million such 10-mers that correspond to antimicrobial patterns. Then we wrote out all possible 20-amino-acid sequences for which each window of 10 amino acids is found in the set of 3 million 10-mers. From this set, we removed any 20-mers that had six or more amino acids in a row in common with a naturally occurring AmP. There are roughly 12 million such 20-mers, each of which is a 'tiling' of ten 10-mers.

We clustered these 12 million sequences using the Mcd-hit software[18] at 70% identity. From these clusters, we chose 42 high scoring sequences to test experimentally (see Supplementary Materials, Methods, Scoring of sequences)[19]. These sequences have varying degrees of similarity to naturally occurring AmPs, as determined by sequence alignment and discussed in the text.

Peptide synthesis and MIC determination. Fmoc (fluorenylmethoxycarbonyl) chemistry was used to synthesize peptides on the Intavis Multipep Synthesizer (Intavis LLC) at the MIT Biopolymers Lab, and confirmed using MALDI-TOF Mass Spectrometry and recombinantly produced AmP standards. The MIC was measured with an assay based on the NCCLS M26A and the Hancock assay for cationic peptides[17]. See Methods in Supplementary Materials for details on synthesis and antimicrobial assay.

Received 1 May; accepted 4 September 2006.

1. Zasloff, M. Antimicrobial peptides of multicellular organisms. *Nature* **415**, 389–395 (2002).
2. Hancock, R. E. & Patrzykat, A. Clinical development of cationic antimicrobial peptides: from natural to novel antibiotics. *Curr. Drug Targets Infect. Disord.* **2**, 79–83 (2002).
3. Tiozzo, E., Rocco, G., Tossi, A. & Romeo, D. Wide-spectrum antibiotic activity of synthetic, amphipathic peptides. *Biochem. Biophys. Res. Commun.* **249**, 202–206 (1998).
4. Biragyn, A. *et al.* Toll-like receptor 4-dependent activation of dendritic cells by β-defensin 2. *Science* **298**, 1025–1029 (2002).
5. Ellerby, H. M. *et al.* Anti-cancer activity of targeted pro-apoptotic peptides. *Nature Med.* **5**, 1032–1038 (1999).
6. Giangaspero, A., Sandri, L. & Tossi, A. Amphipathic α helical antimicrobial peptides. *Eur. J. Biochem.* **268**, 5589–5600 (2001).
7. Shai, Y. Mode of action of membrane active antimicrobial peptides. *Biopolymers* **66**, 236–248 (2002).
8. Jurafsky, D. & Martin, J. H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (Prentice Hall, Upper Saddle River, New Jersey, 2000).
9. Searls, D. B. The language of genes. *Nature* **420**, 211–217 (2002).
10. Hofmann, K., Bucher, P., Falquet, L. & Bairoch, A. The PROSITE database, its status in 1999. *Nucleic Acids Res.* **27**, 215–219 (1999).
11. Rigoutsos, I. & Floratos, A. Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. *Bioinformatics* **14**, 55–67 (1998).
12. Wang, Z. & Wang, G. APD: the Antimicrobial Peptide Database. *Nucleic Acids Res.* **32**, D590–D592 (2004).
13. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**, 45–48 (2000).
14. Wu, M. & Hancock, R. E. W. Interaction of the cyclic antimicrobial cationic peptide bactenecin with the outer and cytoplasmic membrane. *J. Biol. Chem.* **274**, 29–35 (1999).
15. Tossi, A., Sandri, L. & Giangaspero, A. Amphipathic, α-helical antimicrobial peptides. *Biopolymers* **55**, 4–30 (2000).
16. Bell, G. & Gouyon, P-H. Arming the enemy: the evolution of resistance to self-proteins. *Microbiology* **149**, 1367–1375 (2003).
17. Hilpert, K., Volkmer-Engert, R., Walter, J. & Hancock, R. E. W. High-throughput generation of small antibacterial peptides with improved activity. *Nature Biotechnol.* **23**, 1008–1012 (2005).
18. Li, W., Jaroszewski, L. & Godzik, A. Clustering of highly homologous sequences to reduce the size of large protein database. *Bioinformatics* **17**, 282–283 (2001).
19. Maizel, J. V. & Lenk, R. P. Enhanced graphic matrix analysis of nucleic acid and protein sequences. *Proc. Natl Acad. Sci. USA* **78**, 7665–7669 (1981).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The authors would like to thank M. Zasloff, K. D. Wittrup, R. Berwick, and G. Georgiou for valuable input on the draft manuscript, and J. Moxley for figure preparation. The authors gratefully acknowledge the support of the Singapore-MIT Alliance, the NIH, and the Fannie and John Hertz Foundation.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to G.S. (gregstep@mit.edu).