JBC Papers in Press. Published on April 20, 2005 as Manuscript M503056200

APPLICATION OF THE "CODON-SHUFFLING" METHOD: SYNTHESIS AND SELECTION OF *DE NOVO* PROTEINS AS ANTIBACTERIALS

Alka Rao[1], Sidharth Chopra[1,2], Geeta Ram, Ankit Gupta and Anand Ranganathan*
From the Recombinant Gene Products Group
International Centre for Genetic Engineering and Biotechnology
Aruna Asaf Ali Marg, New Delhi –110067, India

Running title: Synthesis and selection of bioactive *de novo* proteins

*Corresponding author: Anand Ranganathan, Recombinant Gene Products Group, International Centre for Genetic Engineering and Biotechnology, Aruna Asaf Ali Marg, New Delhi - 110067, India, Tel. +91-11-26195007; Fax. +91-11-26162316; E-mail. anand@icgeb.res.in
[1]Both authors contributed equally to this work; [2]Present address: School of Medicine, Stanford University, Palo Alto, California - 94305, USA

Library-based methods of non-rational and part-rational designed *de novo* peptides are worthy beacons in the search for bioactive peptides and proteins of medicinal importance. In the present report, we have used a recently developed directed evolution method called 'codon-shuffling', for the synthesis and selection of bioactive proteins. The selection of such proteins was based on the creation of an inducible library of 'codon-shuffled' genes that are constructed from the ligation-based assembly of judiciously designed hexamer DNA-duplexes, called dicodons. Upon induction with IPTG, some library members were found to express dicodon-incorporated proteins, because of which the host cells, in our case *E. coli*, were unable to grow any further. The bactereostatic/lytic nature of the dicodon proteins was monitored by growth curves as well as by zone clearance studies. Transmission Electron Microscopy of the affected cells illustrated the extent of cell damage. The proteins themselves were over-expressed as fusion partners and subsequently purified to homogeneity. One such purified protein was found to strongly bind heparin, an indication that the interaction of the *de novo* proteins may be with the nucleic acids of the host cell, much like many of the naturally occurring antibacterial peptides, e.g. Buforin. Our approach may therefore help in generating a multitude of finely tuned anti-bacterial proteins that can potentially be regarded as lead compounds, once the method is extended to pathogenic hosts, like for example Mycobacteria.

The objective of recasting or constructing proteins from scratch in order to achieve altogether new entities that are then able to carry out a preordained catalytic or structural function is currently an area of intense scientific research. That small proteins or peptides can be constructed based on the first principles of secondary structure design has been known for sometime, largely through some seminal contributions of DeGrado & coworkers and others (1-3). This so-called *de novo* protein design is the rational approach towards the creation of functional proteins and it is perhaps only now, with realization of the hitherto unreachable requirements of computational power, that one is witness to an increasing number of functional peptides and proteins being successfully generated *de novo* (4,5). On the other hand, non-rational and part-rational approaches towards creating *de novo* proteins designed for specific purposes seem to be much more powerful methods at generating functional proteins, especially for *in vivo* biological requirements (6,7). Methods like the 'phage-display' (8) and 'mRNA display' (9) increasingly allow one to effectively generate and then subsequently pan non-rational peptide and protein libraries, in order to select for the right protein. In situations where an effective assay-based selection system is available, part-rational and non-rational approaches may be more effective than a rational approach in the isolation of a functional protein, largely because the former mimic evolutionary pathways of selection (of the fittest). In this context, two such studies are worthy of special attention.

A little more than a decade ago, Hecht and coworkers (10) put forth a most novel hypothesis, that of binary patterning of proteins. According to this simple concept, it is the binary patterning of polar (○) and non-polar (●) amino acids in the designed protein, and not the exact identity of those amino acids, that eventually dictates the secondary structure of the designed protein. The binary-patterning concept may be viewed as a

part-rational approach, whereby one can predetermine the library folds but not the exact folds of its individual members. Indeed, *de novo* protein libraries based on binary-patterning have furnished many functional as well as structural proteins (11,12), although problems of a suitable assay system, as well as the limitation of a fixed length of library members, have yet to be addressed satisfactorily. The second seminal study, by Mekalanos and coworkers (13), involved the generation of non-rational peptide libraries, based on degenerate oligonucleotide design. Their work elegantly demonstrated that an inducible non-rational DNA library, once translated *in vivo*, could lead to the selection of a functional peptide. This so-called ABBIS approach was successfully used for isolation of a number of peptides that, upon overexpression in the host cell, led to severe growth attenuation of the host. The ABBIS method too has its limitations that are primarily because of its use of fixed-length degenerate oligonucleotides as a starting point for making a peptide library. Nonetheless, it remains a powerful example of a non-rational approach towards the *de novo* synthesis of useful proteins, in their case anti-microbial peptides (AMP).

Of late, there is an urgent need for generating newer, more potent AMPs, given their newfound importance in the context of widespread antibiotic resistance amongst human pathogens (14-16). AMPs have been isolated from a variety of species, including around forty odd from humans (15,17). Thus far, around 700 AMPs have been isolated and characterized (18) and many among them are in advanced clinical trials as drug candidates against bacterial and fungal infections (15,19). A general characteristic of AMPs is their net positive charge that is believed to be crucial for gaining entry into negatively charged bacterial outer membranes, although lately some anionic AMPs have also been isolated (20,21). Consequently, many models for bacterial entry have been postulated (17,22). However, recent DNA-microarray studies on bacteria affected by AMPs have suggested that cell wall lysis may not be the sole mode of AMP action, given that the expression of a host of cytoplasmic genes is severely affected upon AMP cell entry (23,24). Notwithstanding the uncertainty surrounding their action, it is clear that AMPs represent a class of molecules that may in the near future provide an effective alternative to the currently used antibiotics (14). Therefore, an approach that leads to the synthesis of a diverse pool of AMPs, and one that addresses the earlier problems associated with AMP synthesis, would we believe, be of considerable help.

In this report, we describe one such approach that is based on the application of a recently developed directed-evolution technique by us, called "codon-shuffling" (25). We also describe how, through codon-shuffling, libraries are generated that: (a) can be both non-rational as well as part-rational, (b) have members large enough to be classified as proteins and not just peptides, (c) are not restricted by limits of length of their corresponding genes, and (d) can be preferentially skewed in amino acid attributes like charge and hydrophobicity, thereby narrowing down the search for a successful AMP.

EXPERIMENTAL PROCEDURES

Construction of expression plasmid pDNp28 - An unrelated DNA fragment of length 875 bp from the plasmid *pET21c* was PCR-amplified using *pfu* polymerase and the oligonucleotides 5'-aaccatggg-ttacgtacatttccgtgtcgccctt-3' and 5'-aactcgagtacgta-ccaatgcttaatcagtgaggc-3' as forward and reverse primers respectively. The PCR product was cloned in *pBluescriptSRF* vector, the fragment excised using *Nco*I and *Xho*I restriction enzymes and cloned in *pET28a* vector previously cut with the same two enzymes. The resulting plasmid was isolated and digested with *Sna*BI to finally yield the expression plasmid *pDNp28*. Plasmid *pDNp28* carries the *Sna*BI site into which DC fragment libraries can be inserted. The site is immediately downstream of the ATG start codon, and just upstream of the hisX6 tag.

Construction of ds hairpin-encapsulated DC fragment (HPDN) library - For the initial part of library construction, the protocol followed was as described earlier (25), with one variation. Immediately after the mixing of 14 DCs, the solution was made to 7.5 % PEG and the DC mix incubated at 4 ^{o}C for 24 h. To this mix, was added 100 pmoles of 5'-phosphorylated and PAGE-purified ds hairpin: 5'-tttaaacacgtggcggccgctctaga-ggcccgcgcgggcctctagagcggccgccacgtgtttaaa-3' that had been self-annealed earlier. The ligation temperature was increased to 16 ^{o}C and the incubation prolonged for another 12 h. The ligation mixture DNA was precipitated and extracted once with phenol/chloroform. The resuspended DNA was digested with *Xba*I for 4 h at 37 ^{o}C, after which 1 µl of the digested DNA was

used as a template for PCR using the 5'-phosphorylated oligonucleotide: 5'-agcggccgccac-gtgtttaaa-3', that served both as a forward and reverse primer. The PCR products were eluted using DEAE membrane and fractionated based on their lengths (50-400 bp). The purified fragments were used directly as inserts for creation of *de novo* libraries.

Construction of expression plasmid pTEMDNp28 - A 420 bp long DNA fragment from the *TEM-1* gene of plasmid *pSC1* (25) was PCR-amplified using *pfu* polymerase and the oligonucleotides 5'-aagcatgcaaggagatggcgcccaacagtccc-3' and 5'-cctacgtatgcacccaactgatcttcagcatc-3' as forward and reverse primers respectively. The PCR product was cloned in *pBluescriptSRF* vector, the fragment excised using *Nco*I and *Sna*BI restriction enzymes and cloned in *pDNp28* vector previously cut with the same two enzymes. The resulting plasmid was designated *pTEMDNp28*. Plasmid *pTEMDNp28* carries the *Sna*BI site into which DC fragment libraries can be inserted. The site is immediately downstream of the *TEM-1* secretion signal sequence, and just upstream of the hisX6 tag.

Western Blot analysis - Whole-cell lysates of EQAMP1-3 cultures were run on 15 % SDS-PAGE and transferred onto supported nitrocellulose membrane (Gibco BRL, Life Technologies). The membrane was blocked with 1 % PVP and then probed with anti-His monoclonal antibody (1:7500 dilution). The blot was developed by using BCIP/NBT (Promega).

Heparin binding studies - Procedures relating to the construction of GST-SKAMP1 gene and subsequent purification of the corresponding protein are described in Supplemental Experimental Procedures. 20 µg of pure GST-SKAMP1 protein was loaded on to a column containing heparin sepharose. The column was washed extensively with buffer C. Elution buffer (Buffer C with an NaCl gradient 250 mM to 1 M) was then poured over the column and fractions collected.

Western blot: The eluted fractions, as well as the heparin resin were probed with anti-GST antibodies (1:25000 dilution) for 1 h. Nitrocellulose membrane was then washed four times with 1X PBST (each wash of 5 min duration). Secondary probing was done with anti-rabbit IgG-heavy & light chain, AP conjugated antibodies (1:5000 dilution), for 1 h. After washing the membrane with 1X PBST, the blot was developed with BCIP/NBT substrate.

RESULTS AND DISCUSSION

Exploring the possibility of using codon-shuffling for constructing de novo protein libraries

The codon-shuffling method was initially conceptualized by us as a directed-evolution technique. However, because the method is based not on selection of a progeny through inducement of random mutations in the parent gene, like in gene-shuffling (26), but rather on extending a truncated parent gene with new DNA fragments that are constructed *de novo*, codon-shuffling can also be viewed as a method for creating stand-alone new DNA fragments. This is because codon-shuffling fundamentally entails randomly assembling 6 bp DNA duplexes (called dicodons or DCs) from a set containing 14 such dicodons (Table 1), each of which are judiciously chosen to encode amino acid pairs, such that in an equimolar dicodon pool, all 20 natural amino acids are represented and in proportions that mirror their usage in *E. coli*. Furthermore, there is no restriction on the lengths of the resulting DNA fragments. Consequently, the amount of degeneracy achievable from ligation of the dicodons in making DNA fragments of an average length of say 250-300 bp is astronomically high and much beyond what may be required to exhaust all possible protein folds. At first glance therefore, codon-shuffling represents a useful method for generating *de novo* protein libraries.

Here, it is pertinent to mention one disadvantageous feature of peptide libraries made from degenerate oligonucleotides. In addition to the problems of length restriction (the upper limit being 100 nt on average, resulting in 30-40 aa long peptides) as well as random appearances of stop codons, such peptide libraries yield few folded peptides, thereby severely limiting their usefulness (27,28). One reason why few peptides from such libraries are able to fold properly may be the completely non-rational nature of their parent DNA fragments, that may result in a peptide that contains a string of polar (○○○○○○…) or nonpolar (●●●●●●…) residues, rather than an arrangement of such residues that promotes a secondary structure feature (○●○●○●○●○●○ or ○●○○●●○○●○○). Indeed, by predetermining the degeneracy of parent oligonucleotides such that the resulting peptides may possess polar/nonpolar

arrangements, Hecht and coworkers were able to generate a library of peptides wherein most members displayed a well-folded behaviour (10,29). Consequently, the library size in such situations may not be necessarily large to begin with, much like in 'codon shuffling'. One possible explanation in the 'codon shuffling' context may be that the 14 dicodons themselves possess a range of inherent secondary structure forming attributes, because of the manner in which they are paired – as polar/nonpolar, nonpolar/polar, nonpolar/turn, nonpolar/nonpolar and other such combinations (Table 1). For example, 'EL' or 'DI' repeats yield a binary pattern (○●○●○●○●○●○●) that represents beta-sheet structure, while the string MH-DI-MH-EL-ST yields a binary pattern (●○○●●○○●○○) that would preferentially form an alpha-helix. Additionally, appearance of the dicodons 'WP', 'PG' or 'GA' may induce turns or breaks, while sputtering of 'CA' within the DC protein sequence may create intra or inter-molecular disulfide bridges.

To further investigate this hypothesis, we undertook a search of all proteins in the PDB crystal structure database, to find the location of dicodon-pairs within protein structures. The search string was established as a dicodon-dicodon pair (DC-DC) amounting to searching of a total of 196 DC-DC motifs (14DC × 14DC). The DC preferences for particular secondary structure elements are listed in Table 1. Predictably, the proline-containing DCs (WP and PG) were preferentially found in turn elements, while DCs that were formed of amino acid pairs of high alpha-helix forming propensity, like for example FE, were found to be embedded in alpha-helices. Note worthily, only 11% of DC-DC pairs were found in unstructured and random coil elements. This may indicate that a *de novo* DC-protein library would contain proteins that possess secondary structure elements and are therefore well folded, as opposed to proteins from a degenerate oligonucleotide library. In order to test our hypothesis, we cloned codon-shuffled genes in specially designed vectors (see Experimental Procedures), using the equimolar DC-protocol described previously (25). Two representative clones were sequenced and shown to be wholly DC incorporates (Fig. S1A, Supplemental Figures). When over-expressed in *E. coli*, the clones yielded DC-proteins that were preferentially found in the insoluble fractions. Nonetheless, because of reasonably good expression, the proteins could be routinely purified by metal-affinity chromatography in 8 M urea buffer. Post-dialysis, one of the proteins could be solublised in as little as 300 mM urea (Fig. S1B).

These results indicated to us that *de novo* DC-protein libraries were viable at generating reasonably large quantities of proteins of expected sizes, and that subsequent application of a selection system could fish out desirable proteins. However, one caveat needed to be addressed before codon-shuffling could be used efficiently for constructing *de novo* protein libraries. Because codon-shuffled DNA fragments were ligated directly to expression vectors for their eventual translation, the numbers of such fragments available for cloning were not all exhaustibly represented in the DC library, since there was only one ds DNA fragment (i.e. insert fragment) of each type available. In other words, the impossibility of experimentally cloning each & every ds DNA fragment meant that the available complexity of the DC library was being compromised.

Use of ds DNA hairpins for amplification of codon-shuffled fragments

In order to address the above-mentioned problem, we envisaged the use of ds DNA hairpin that would firstly cap the two exposed ends of DC assembled DNA fragment and then be used as a template for fragment amplification using oligonucleotides that are complementary to the hairpin sequence (30). Moreover, the ds hairpin, in addition to its primary role of amplifying DC fragments, could also be used as a scaffold for encapsulating DC-proteins. Protein scaffolds have previously been employed to 'tighten' randomly generated *de novo* peptides, in the hope that the latter may fold better as their N- and C-termini are no longer free (31,32). One such scaffold, the thioredoxin protein (trx) has been so employed on numerous occasions (13,33,34). However, whether the peptides are as effective once they are removed from their scaffolds, or more worryingly, whether the *de novo* peptide activity is partly due to the scaffold itself, are some issues that remain largely unaddressed. One could however be prudent and reduce the deleterious effects of the scaffold, simply be keeping its size to a minimum, as using a large protein scaffold may stimulate the *de novo* protein to take a shape that, without the scaffold would be altogether different. Therefore, in designing the ds hairpin, we focused on what was our initial goal – to isolate proteins that act as antibacterial.

A significant number of isolated AMPs are alpha-helical in nature and many that are not, become so once they interact with bacterial membranes (15,35). It would therefore be helpful to construct AMPs over an alpha-helical scaffold. As a starting point, we envisaged that the ds hairpin should code for a string of 6-7 amino acid residues. After a preliminary study of exploring binary patterns for the scaffold, we settled on the pattern shown in Figure 1. The binary patterns, ○●○○●●○, along with its 'anti-sense' pattern, ●○○●●○●, are both predicted to form an alpha-helix, based on the Hecht-hypothesis. We then searched the PDB crystal structure database in order to document the presence of these two patterns in the various fold elements. Interestingly, both patterns show a generous preference for alpha-helices, thereby confirming the practicability of using Hecht-binary patterning for *de novo* peptide design (Fig. 1). As a next step, we embedded some unique restriction enzyme (RE) sites (none of which appear in any given DC fragment) within the ds hairpin sequence such that they would not disturb the binary pattern of the translated hairpin (Fig. 1B). The final sequence of the designed hairpin (shown in Figure 1) coded for the following two sequences, 'SGRHVFK' and 'FKHVAAA', depending upon which hairpin strand was translated. Importantly, both these sequences conform to an alpha-helix forming Hecht-pattern.

Construction of de novo DC-protein libraries with the hairpin scaffold

Our next objective was to experimentally investigate the merit of using ds hairpin for amplifying DC fragments, as well as for ably providing a scaffold for the DC proteins. As a starting point, we added a molar excess of ds hairpin to a mixture of equimolar DC pool (see Experimental procedures for details). This was followed by amplification of encapsulated DC fragments with a 5'-phosphorylated oligonucleotide that was complimentary to the ds hairpin, and therefore could act as both forward and reverse primer for the purposes of amplification. The PCR products were separated based on their length (50 to 400 bp) and ligated directly with *Sna*BI-cut expression vector *pDNp28*. Upon transformation of *E. coli* BL21(DE3) followed by selection on kanamycin, we obtained colonies roughly to the order of 10^4. We selected and sequenced some representative clones and found all of them to indeed contain DC incorporates that were flanked on either end by the ds hairpin. The corresponding DC protein library showed little sequence similarity amongst its members and was vastly variant in terms of length and primary sequence attributes (Fig. 2A). To test the protein expression levels of our DC library, we induced growing cultures of a representative clone HPDN5 with IPTG. We observed overexpression of HPDN5 protein and were subsequently able to purify the protein using Ni-NTA chromatography (Fig. 2B). These results firmly indicated to us that generation of DC proteins from ds hairpin-encapsulated DC fragment library represents a viable route towards isolation of *de novo* proteins that are expressed from their corresponding genes at good levels and therefore can be easily purified.

Finally, we wanted to explore a case where the DC protein is encapsulated by the traditional, often employed thioredoxin (trx) scaffold. We therefore ligated the hairpin-encapsulated DC fragment (HPDC) library with an expression vector (*pTRXDNp21*) that carried the modified *trx* gene wherein a unique *Sna*BI site had been positioned within the *trx* sequence that corresponded to the region between the two cystines (see Supplemental Experimental Procedures for details). The DNA of many of the obtained clones was sequenced and found to contain HPDC sequences expectedly positioned within the *SnaB*I site (Fig. S2A; Supplemental Figures). One representative member of the trx-DC library was further analysed to check for protein expression. Upon induction with IPTG, the protein THPDN3 was found to be generously overexpressed. However, most of the protein was found to be in the insoluble fractions, albeit the protein could be purified to homogeneity under denaturing conditions (8 M urea) using Ni-NTA chromatography (Fig. S2B). Upon dialysis, we found the protein to be soluble in a minimum of 150 mM urea. Circular Dichroism studies of THPDN3 along with native trx protein showed that the chimera displayed a predominantly alpha-helical content (helix – 24.1 %, sheet – 10.0 %, turn – 34.4 %, random – 31.4 %), much like the parent trx protein (helix – 21.5 %, sheet – 14.3 %, turn – 31.3 %, random – 32.8 %; Fig. S2C). Without the use more incisive techniques like x-ray crystallography or NMR, it is not possible to speculate whether the DC protein has folded outside the trx scaffold, or whether the obtained CD data is for a single protein entity. Further work in this direction is currently being carried out in our laboratory.

In summary, we have investigated the usefulness of DC proteins as candidates for *de novo* protein libraries. We have also found that it is worthwhile to encapsulate the DC proteins using scaffolds. Of the two scaffolds that we used for the purpose, we found the trx scaffold to yield chimeras that were insoluble and thus of little use *in vivo*.

Design and synthesis of de novo Antibacterial Proteins

Having put in place a method for the design, synthesis and selection of *de novo* proteins with two different scaffolds, we now focused our efforts on the synthesis of *de novo* proteins that would act as antibacterial. Our strategy was to select for AMPs *in vivo*, by identifying those transformed bacterial colonies that were unable to grow in the presence of IPTG, in contrast to robust growth in absence of the inducer. Additionally, we decided to tether the DC proteins with an N-terminal signal sequence, for the following reason. Because the AMP site of action could be the cytoplasm, the periplasm, or indeed the extracellular environment of the host, in our case *E. coli*, we felt that it was important not to exclude, by a method of selection, DC proteins that were acting at any of the above-mentioned three sites. For example, by not choosing to tether the DC library with an N-terminal signal, all DC proteins would be restricted to the cytoplasm of the cell. On the other hand, choosing the ompA signal sequence as the primary tether for the AMPs, all DC proteins would be directed towards the extracellular environment of the host cell. As a result, AMPs that would have had a cytoplasmic or a periplasmic site of action would not be selected.

Keeping this viewpoint in mind, we decided to employ the N-terminal signal sequence of TEM-1 beta-lactamase as the primary AMP tether. This was because we and others have shown that, under induction conditions, the TEM signal sequence directs transport of the beta-lactamase protein to all three regions of the bacterial cell - the cytoplasm, the periplasm and the extracellular environment, in the approximate ratio of 70:25:5 (36,25). By choosing the TEM-1 signal, we would therefore not be restricting the site of action of *de novo* AMPs to a particular cellular environment. The expression vector that would accept the in coming hairpin-encapsulated DC fragments was redesigned to include the segment of TEM-1 gene that codes for the 23 amino acid-long N-terminal signal (see Experimental Procedures). This vector (*pTEMDNp28*; Fig. S3, Supplemental Figures) was digested with *Sna*BI and introduced to an equimolar hairpin-encapsulated DC mix. The ligation mixture was used to transform *E. coli* BL21(DE3) competent cells and the latter plated on to media containing kanamycin (the vector marker). We obtained a library size of approximately 10^3 c.f.u; the size varied by a factor of a fold when the experiments were repeated, presumably because of differences in ligation and transformation efficiencies. The c.f.u were then replica-plated on to plates that, in addition to kanamycin, also contained 0.5 mM IPTG. While the majority of colonies were able to display robust growth, we were able to isolate four plate I (kan) colonies that showed no growth on plate II (kan + IPTG). The sequencing of DNA isolated from the colonies showed that all of them were hairpin-DC incorporates. Predicted translation products as well as physical properties of the three unique sequences are shown in Figure 3. There is wide variance seen in the protein length and pI among the three proteins, named EQAMP1, 2 and 3. As a control, the DNA from a colony that grew well on both plates I and II was also sequenced (cEQ1). Consensus secondary structure predictions on all four proteins displayed a dramatic difference between the three proteins and the control protein (Fig. 3). While EQAMP1-3 were predicted to be rich in alpha-helical content, the control protein, cEQ1 was predicted to be rich in random coils. This incongruity between the three EQAMP proteins and the control made us revisit the initial HPDN DC protein library (Fig. 2). Indeed, all HPDN1-8 proteins were predicted to have a random coil content of greater than 55%. It is tempting therefore to suggest that the EQAMP proteins, because of their *in vivo* AMP behavior, display pronounced helical fold characteristics that stand in contrast to the predicted fold characteristics of non-selected DC proteins.

Confirmation of AMP behavior of EQAMP proteins

In order to confirm beyond reasonable doubt that the *in vivo* AMP activity witnessed was because of DC proteins and nothing else, we carried out the following experimental checks: (a). Plasmid DNA from the strains exhibiting growth inhibition upon addition of IPTG were isolated, purified and used to transform a fresh lot of *E. coli* competent cells. Addition of IPTG to growing cultures of the

freshly transformed cells also displayed growth inhibition, thereby pointing to the fact that the latter was a property of the plasmid DNA used for transformation. (b). The DC fragment within each plasmid was amplified using universal primers that bound to only the vector regions flanking the DC fragments and nowhere else. The amplified products were digested with restriction enzymes that cut within the vector sequences and ligated to freshly cut *pET28a* vector. The resulting plasmids (that should be identical to the starting plasmids) were used to investigate growth inhibition and were all found to exhibit the same. This ruled out growth inhibition properties by any means (contamination etc) other than the DC fragments themselves. (c). The supernatant of IPTG-induced growth cultures of EQAMPs were isolated and spread on agarose plates containing *E. coli* DH5α lawns, and the plates incubated for a period of 14-18 h at 37 ^{o}C. Absence of any plaque formation ruled out phage-mediated inhibition being the reason for growth inhibition of EQAMP cultures. (d). As a final check, we wanted to confirm that the growth inhibition was due to DC proteins and not because of their DC-mRNA. For this purpose, the unique *Nco*I site in each EQAMP plasmid (that also contained the ATG start codon) was cut, filled-in using Klenow polymerase, and the plasmids self-ligated. This resulted in a frame shift, starting from the *Nco*I ATG codon, which no longer yielded the DC proteins as the translated products. However, the mRNA in each case remained identical to the wild-type EQAMP mRNA. When the resulting plasmids EQAMPnco1-3 were used, no growth inhibition was seen (Fig. S4; Supplemental Figures), thereby confirming that it was the DC protein and not the DC-mRNA that caused inhibition.

Monitoring of growth inhibition by EQAMPs

The extent of inhibition by EQAMP1-3 was directly followed by 'growth-inhibition' curves. Again, the experiments were carried out in combination with extensive controls that were as described: *pET28a* vector; the scaffold vector *pDNp28*; *E. coli* BL21(DE3) strain; the scaffold vector containing a dummy gene (in our case *trx*), the translated product of which does not cause growth inhibition. The control curves as well as those for EQAMP1-3 are shown in Figure S5 (Supplemental Figures) and Figure 4 respectively. Also shown in Figure 4 is the growth curve for the positive control cEQ1. As can be seen, there is a near complete growth inhibition upon IPTG addition in EQAMP samples, while there is little or no difference between the induced and uninduced curves for the controls, including cEQ1. Interestingly, the EQAMP samples seemed to display a bacteriostatic rather than a bacteriolytic behavior, as small aliquots of induced cultures taken at the end-points, when used as seed samples to inoculate IPTG-less media resulted in robust growth (results not shown). Identical growth inhibition results were also obtained when we employed 'zone-clearance' studies. Small filter paper discs that were resting on growing lawns of EQAMP1-2 cultures were gently soaked with IPTG and the plates incubated for 14-18 h. Clearance zones appeared in EQAMP1-3 samples, that were not present in the control plates (Fig. S6; Supplemental Figures), thereby confirming that the inhibition is inducer-driven and is because of DC-protein expression.

Identification of DC proteins responsible for growth inhibition

It is well recorded that gratuitous overexpression of a protein can also lead to severe growth inhibition or bacteriostasis of the host (37). To rule out the possibility that the growth inhibition we were witnessing was because the EQAMP proteins were being hugely over-expressed upon IPTG induction, we decided to look for the proteins in lysed cultures of EQAMPs as well as of cEQ1. Collection of reasonable amounts of biomass for the above-mentioned purpose proved to be a tough challenge because of the bactereostatic nature of the expressed DC proteins. Nonetheless, pellets of 1 litre cultures could routinely be lysed and an adequate amount of total protein extracted for SDS-PAGE analysis. We could not detect EQAMP1-3 proteins at their expected sizes through coomassie-staining. On the other hand, the control cEQ1 protein was readily visible under same conditions (with and without the TEM-1 signal; Fig. 5A). However, because all DC proteins are expressed as C-terminal His-tag fusions, proteins EQAMP1-3 could be detected in western blots using monoclonal anti-his antibodies (Fig. 5B). The presence of minute amounts of EQAMP proteins at their expected sizes, and large amounts of the control cEQ1 protein, provided strong evidence that the growth inhibition of the host was not because of gratuitous overexpression of DC proteins, but rather a consequence of the antibacterial nature of the latter. All efforts to isolate purified EQAMP proteins have as yet proven to be unsuccessful. *In vitro*

transcription/translation of EQAMP genes may however yield sufficient amounts of the proteins for us to carry out further investigations on their antibacterial nature as well as their structure, and such efforts are currently underway in our laboratory.

As was stated earlier, two major factors, namely, protein charge and its tertiary structure, are responsible for AMP activity of peptide and proteins. Total charge profiles of EQAMP proteins provided an inconclusive picture – while EQAMP1 was predicted to be positively charged at physiological pH, EQAMP2 and 3 possessed a pI of 6.6 and 6.5 respectively. However, an inspection of charge distribution of the latter two proteins provided for a possibility that these proteins may possess a domain structure wherein one domain could be positively charged (results not shown). Also, the charge state (whether negative or positive) at physiological pH, of multiple histidine residues that are present in all EQAMPs could affect the over all charge of the proteins. On the matter of tertiary structure, consensus secondary structure predictions of the three EQAMP proteins label them as rich in alpha-helical elements (Fig. 3). Indeed, if one were to assume this prediction as correct, helical-wheel projection drawing of EQAMP2, for example, nicely illustrates the demarcation of its residues - non-polar in the interior and polar at the exterior of the assumed all-helical bundle (Fig. S7; Supplemental Figures). However, presently these are at best conjectures; isolation of purified EQAMP proteins and their subsequent structural characterization would lead us to an accurate representation. As a next best alternative, we isolated representative AMPs (EQAMP2 and SKAMP1) as trx fusion proteins and obtained their CD spectra (Fig. 6). The AMP fusion proteins displayed a well folded structure as indicated by the CD data: for EQAMP2 - helix – 31 %, sheet – 0 %, turn – 38 %, random – 31 %); for SKAMP1 - helix – 29 %, sheet – 4 %, turn – 34 %, random – 33 %; Fig. 6B).

Direct visualization of the in vivo affect of EQAMPs

Although growth inhibition experiments had earlier indicated to us that the EQAMP1-3 proteins were most probably bacteriostatic rather than bacteriolytic in nature, we decided to investigate by Transmission Electron Microscopy (TEM) the state of the host cell post-IPTG induction with a non-induced cell culture as reference control. TEM results (Fig. S8A-D; Supplemental Figures) illustrated to us that induced samples displayed a morphology that was markedly different from the non-induced samples. In addition to the disintegration of cell wall, widespread cytoplasmic contraction was also clearly visible. The latter is generally caused when a membrane protein of a cell is affected or sequestered, or when the cell's nucleic acids are targeted (17). We decided to investigate this phenomenon further by studying the possible association of EQAMPs with heparin (heparan-sulphate). It is well known that proteins that bind heparin also bind nucleic acids (38,39); In addition, many AMPs have been shown to bind heparin directly (40). While obtaining pure EQAMPs had earlier proved to be unsuccessful, we were able to obtain good quantities of soluble EQAMPs as fusion partners of the Glutathione S-transferase protein (GST). As a representative example of such fusion proteins, we decided to investigate the SKAMP1-GST fusion protein (see below for the synthesis and isolation of SKAMP1 protein). Purified SKAMP1-GST (MW 36.7 KDa) was able to bind to heparin at varying pH conditions and the fusion protein could partly be eluted with buffer containing 1 M NaCl (Fig. 7); some protein was seen to be bound to heparin resin even after elution with the above-mentioned buffer, indicating very strong binding. It has previously been shown that GST on its own does not bind to heparin (41), a finding also confirmed by us using purified GST protein as a control (Fig. 7). These results therefore indicate association of AMPs with a negatively-charged moiety (in this case heparin) that could be either a nucleic acid or indeed a negatively charged membrane protein. Further studies in this direction are ongoing.

Generating AMPs using skewed DC libraries

As mentioned earlier, one exclusive advantage of 'codon-shuffling' over other methods of library generation is that skewed libraries can be generated wherein proteins can possess desired physical or chemical attributes, on account of the excess presence or complete absence of one or many dicodons. We decided to take advantage of this facet of the 'codon-shuffling' method by constructing an HPDN protein library from a DC mix wherein the positively charged dicodons KL and RT were increased three times in proportion to the other 12 dicodons. Following the same method for synthesis and selection of AMPs as described previously, we obtained (from an initial library size of 10^2) one colony that was not able to grow

on plate II (kan + IPTG). The predicted sequence of this protein, designated SKAMP1, is shown in Figure 3. The protein is extremely positively charged (pI 10.1) and is predicted to preferentially possess alpha-helical elements. Growth inhibition and zone clearance studies with SKAMP1 yielded similar results to those with EQAMPs, indicating that expression of SKAMP1 was also severely deleterious to the host cell (Figs. 4 and S6). Thus, the isolation of an AMP from a skewed DC library is a step in the direction of tailoring protein libraries to meet specific needs; we are currently studying the prospect of creating 'severely skewed' libraries wherein the negatively charged dicodons have been altogether removed from the DC mix.

Conclusions

In this report, we have described an application of the 'codon-shuffling' method, towards the creation of *de novo* protein libraries of use. From the initial idea of using the DC set for making *de novo* proteins, we have streamlined the method for library creation, by designing tailored hairpin scaffolds that encapsulate the DC proteins favorably. The method that we have described results in *de novo* proteins and not just peptides, because of which we believe the library is much less amenable to proteolysis by cellular proteases. The longer average length of the library members also presumably enhances protein folding. Secondly, we have also explored an exclusive property of the codon-shuffling' method, which is the allowance for skewing library properties by inclusion, exclusion or predominance of some DCs. The ability to skew a protein library, we believe, further narrows the protein space that needs to be explored. One can therefore design skewed libraries to isolate proteins for specific needs. For example, a protein library of arginine-rich sequences may be generated by an overwhelming use of the 'RT' dicodon; poly-arginine peptides have lately gained much interest for their ability to act as cell-penetrating peptides that act by binding to the cell-surface of a pathogen (42). Furthermore, the recent sequencing of organisms such as *M. tuberculosis* and *P. facliparum* has unearthed many virulence-determining proteins that are exclusively composed of short sequence repeats (43). It is not known what selective advantage this may provide the pathogen. Proteins rich in DC-repeats, constructed using our method could serve as a model for studying this hitherto unexplored natural phenomena. Finally, our proof-of-concept experiments relating to the generation of AMPs against *E. coli* can be conveniently extended to pathogenic microorganisms like *M. tuberculosis*. The molecular targets of AMPs can then be identified using DNA microarrays. Such studies, we believe would provide a novel direction towards discovering new molecules against common pathogens, and importantly, would further extend the scope of *de novo* protein design as an important chemical tool for tackling generic problems in medicine.

ACKNOWLEDGEMENTS

The authors wish to thank Prof. Michael Hecht, Princeton University, for critically reading the manuscript. We also thank RGP lab members for help and support, and Malaria group members for their help in generating CD data. Expert analysis of TEM data by Dr. Nirupam Roy Chaudhary is gratefully acknowledged. This work was funded by internal grants of ICGEB. Alak Rao and Geeta Ram wish to thank CSIR, India for financial support.

REFERENCES

1. Regan, L., and DeGrado, W.F. (1988) *Science* **241**, 976-978
2. Cordes, M.H., Walsh, N.P., McKnight, C.J., and Sauer, R.T. (1999) *Science* **284**, 325-328
3. Hecht, M.H., Richardson, J.S., Richardson, D.C., and Ogden, R.C. (1990) *Science* **249**, 884-891
4. Schneider, G., Schrodl, W., Wallukat, G., Muller, J., Nissen, E., Ronspeck, W., Wrede, P., and Kunze, R. (1998) *Proc. Natl. Acad. Sci. USA* **95**, 12179-12184
5. Dahiyat, B.I., and Mayo, S.L. (1997) *Science* **278**, 82-87
6. Taylor, S.V., Walter, K.U., Kast, P., and Hilvert, D. (2001) *Proc. Natl. Acad. Sci. USA* **98**, 10596-10601
7. Roberts, R.W. (1999) *Curr. Opin. Chem. Biol.* **3**, 268-273
8. Rodi, D.J., and Makowski, L. (1999) *Curr. Opin. Biotechnol.* **10**, 87-893

9. Keefe, A.D., and Szostak, J.W. (2001) *Nature* **410**, 715-718
10. Kamtekar, S., Schiffer, J.M., Xiong, H., Babik, J.M., and Hecht, M.H. (1993) *Science* **262**, 1680-1685
11. Moffet, D.A., and Hecht, M.H. (2001) *Chem. Rev.* **101**, 3191-3203
12. Rojas, N.R., Kamtekar, S., Simons, C.T., McLean, J.E., Vogel, K.M., Spiro, T.G., Farid R.S., and Hecht, M.H. (1997) *Protein Sci.* **6**, 2512-2524
13. Blum, J.H., Dove, S.L., Hochschild, A., and Mekalanos, J.J. (2000) *Proc. Natl. Acad. Sci. USA* **97**, 2241-2246
14. Boman, H.G. (2003) *J. Intern. Med.* **254**, 197-215
15. Ganz, T. (2003) *Nat. Rev. Immunol.* **3**, 710-720
16. Lehrer, R.I. (2004) in '*Advances in molecular and cellular microbiology: Mammalian host defense peptides*' (Devine, D.A. and Hancock, R.E.W., ed), pp. 5-8, Cambridge University Press
17. Brogden, K.A. (2005) *Nat. Rev. Microbiol.* **3**, 238-250
18. Wang, Z., and Wang, G. (2004) *Nucleic Acids Res.* **32**, 590-592
19. Hancock, R.E., and Lehrer, R.I. (1998) *Trends Biotechnol.* **16**, 82-88
20. Lai, R., Liu, H., Hui Lee, W., and Zhang, Y. (2002) *Biochem. Biophys. Res. Commun.* **295**, 796-799
21. Brogden, K.A., Ackermann, M., and Huttner, K.M. (1998) *Infect. Immun.* **66**, 5948-5954
22. Powers, J.P., and Hancock, R.E. (2003) *Peptides* **24**, 1681-1691
23. Tomasinsig, L., Scocchi, M., Mettulio, R., and Zanetti, M. (2004) *Antimicrob. Agents Chemother.* **48**, 3260-3267
24. Hong, R.W., Shchepetov, M., Weiser, J.N., and Axelsen, P.H. (2003) *Antimicrob. Agents Chemother.* **47**, 1-6
25. Chopra, S., and Ranganathan, A. (2003) *Chem. Biol.* **10**, 917-926
26. Stemmer, W.P. (1994) *Nature* **370**, 389-391
27. Corey, M.J., and Corey, E. (1996) *Proc. Natl. Acad. Sci. USA* **93**, 11428-11434
28. Prijambada, I.D., Yomo, T., Tanaka, F., Kawama, T., Yamamoto, K., Hasegawa, A., Shima, Y., Negoro, S., and Urabe, I. (1996) *FEBS Lett.* **382**, 21-25
29. Wei, Y., and Hecht, M.H. (2004) *Protein Eng. Des. Sel.* **17**, 67-75
30. Bittker, J.A., Le, B.V., and Liu, D.R. (2002) *Nat. Biotechnol.* **20**, 1024-1129
31. Klevenz, B., Butz, K., and Hoppe-Seyler, F. (2002) *Cell. Mol. Life. Sci.* **59**, 1993-1998
32. Walker, J.R., Altman, R.K., Warren, J.W., and Altman, E. (2003) *J. Pept. Res.* **62**, 214-226
33. Butz, K., Denk, C., Fitscher, B., Crnkovic-Mertens, I., Ullmann, A., Schroder, C.H., and Hoppe-Seyler, F. (2001) *Oncogene* **20**, 6579-6586
34. Colas, P., Cohen, B., Jessen, T., Grishina, I., McCoy, J., and Brent, R. (1996) *Nature* **380**, 548-550
35. Giangaspero, A., Sandri, L., and Tossi, A. (2001) *Eur. J. Biochem.* **268**, 5589-5600
36. Bowden, G.A., and Georgiou, G. (1990) *J. Biol. Chem.* **265**, 16760-16766
37. Dong, H., Nilsson, L., and Kurland, C.G. (1995) *J. Bacteriol.* **177**, 1497-1504
38. Capila, I., and Linhardt, R.J. (2002) *Angew. Chem. Int. Ed. Engl.* **41**, 391-412
39. Aoki, K., Matsumoto, S., Hirayama, Y., Wada, T., Ozeki, Y., Niki, M., Domenech, P., Umemori, K., Yamamoto, S., Mineda, A., Matsumoto, M., and Kobayashi, K. (2004) *J. Biol. Chem.* **279**, 39798-39806
40. Andersson, E., Rydengard, V., Sonesson, A., Morgelin, M., Bjorck, L., and Schmidtchen, A. (2004) *Eur. J. Biochem.* **271**, 1219-1226
41. Rusnati, M., Coltrini, D., Oreste., P, Zoppetti, G., Albini, A., Noonan, D., d'Adda di Fagagna, F., Giacca, M., and Presta, M. (1997) *J. Biol. Chem.* **272**, 11313-11320
42. Goncalves, E., Kitas, E., and Seelig, J. (2005) *Biochemistry* **44**, 2692-2702
43. Ravi Chandra, B., Gowthaman, R., Raj Akhouri, R., Gupta, D., and Sharma, A. (2004) *Protein Eng. Des. Sel.* **17**, 175-182
44. Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual*, 2nd Ed., Cold Spring HarborLaboratory, Cold Spring Harbor, NY
45. Laemmli, U.K., (1970) *Nature* **227**, 680-685

46. Combet, C., Blanchet, C., Geourjon, C. and Deléage, G. (2000) *Trends Biochem. Sci.* **25**, 147-150

FOOTNOTES

The abbreviations used are: NTA- nitrolotriacetic acid; SDS- sodium dodecyl sulfate; PAGE- polyacrylamide gel electrophoresis; EDTA- ethylenediaminetetraacetic acid; PVDF- polyvinylidine difluoride; LB-luria broth; DEAE- diethylaminoethyl; IPTG – isopropyl β-D-thiogalactoside; Mab- monoclonal antibody; BCIP/NBT- 5-bromo-4-chloro-3-indolyl phosphate/Nitro Blue Tetrazolium.

LEGEND TO TABLE

Table 1. The dicodon (DC) set containing the 14 DCs with the listing of search-string occurrences within protein sequences in the PDB crystal structure database. A total of 8356 protein structures were inspected visually using the VIEWERLITE software in order to determine the fold elements of each DC search string. DCs that have high preferences within each fold element have their percentages depicted in bold. The database includes 15 % redundancies.

LEGENDS TO FIGURES

Fig. 1. A. Search-string sequences denoted by the drawn Hecht patterns that are predicted to form alpha-helical secondary structure elements. The PDB database was searched using 'regular expressions' wherein 'n' represents any of the following non-polar residues: L, V, I, F, M, G and A, and 'p' represents polar residues D, E, K, H, R, S and C. The search was carried out from the MRC server (http://csc-fserve.hh.med.ic.ac.uk/pattern.html). Bold numbers within brackets depict the percentages of particular search strings found within helices. **B.** Illustration of ds hairpin design and its use as a scaffold for DC proteins. Unique restriction enzymes sites that cannot appear within any possible DC fragment were incorporated within the hairpin sequence and are as shown. The two strands of the annealed hairpin are shown in red and black, as is the corresponding translated amino acid hairpin sequence.

Fig. 2. A. Amino acid sequences of some ds hairpin-encapsulated DC protein library members, showing variance in protein length, pI and molecular weight (given in kDa). The hairpin sequences are displayed in bold letters. Segements derived from the expression vector are underlined. **B.** 15 % denaturing SDS-PAGE showing expression and purification of one library member – HPDN5. lane 1 - Pharmacia low molecular weight marker (MW shown in KDa); lane 2 - total protein obtained from an IPTG-induced culture of *E. coli* BL21(DE3)/HPDN5; lane 3 – soluble protein loaded on to Ni-NTA column; lanes 4 and 5 – eluted samples of HPDN5 (MW 19 kDa); lanes 6 and 7 – concentrated HPDN5 protein.

Fig. 3. Listing of primary sequences and physical characteristics of anitmicrobial proteins (AMP) described in this study. Predominant secondary structure elements are depicted in bold and are boxed.

Fig. 4. A-E. Growth Inhibition curves of isolated AMPs. Generally, growing *E. coli* cultures were induced with IPTG after 2-3 hours of growth, or after an OD_{600} of 0.4 had been reached. Curves for induced cultures are shown in black. A – EQAMP1; B – EQAMP2; C – EQAMP3; D – SKAMP1; E – cEQ1. **F.** Plating of the culture containing the scaffold plasmid *pDNp28* as well as those containing the AMPs. The plate contained kanamycin and IPTG.

Fig. 5. A. Coomassie-stained 15 % SDS-PAGE showing the whole-cell expression of the control cEQ1. Arrows depict the protein with (top arrow – 19.6 KDa) and without the N-terminal TEM-1 signal (bottom arrow – 17.0 KDa). Lane 1 – Pharmacia low MW marker; lanes 2 and 3 – whole cell lysate of *E. coli* BL21(DE3)/cEQ1. **B.** Western blot of EQAMP1-3 proteins detected using monoclonal anti-his antibodies. Lane 1 – MW markers with sizes as indicated (in KDa); lane 2 – EQAMP3 (MW 20.2 KDa); lane 3 – EQAMP2 (MW 13.3 KDa); lane 4 – EQAMP1 (MW 10.5 KDa).

Fig. 6. **A.** Coomassie-stained 15 % SDS-PAGE of purified SKAMP1trx (lane 2) and EQAMP2trx (lane 3) proteins (expected MW 23.8 KDa and 21 KDa respectively). **B.** CD spectra of purified control as well as AMP fusion proteins: a – trx (thioredoxin); b – SKAMP1trx; c – EQAMP2trx.

Fig. 7. Analysis of Heparin-binding by GST-SKAMP1 fusion protein. **A.** Coomassie-stained 15 % SDS-PAGE illustrating the heparin-binding ability of GST-SKAMP1 fusion protein (expected MW 37.6 KDa). Lane 1 - Pharmacia low MW marker (MW in KDa); lane 2 – protein eluted with 750 mM NaCl buffer; lane 3 – protein eluted with 1 M NaCl buffer; lane 4 – heparin sepharose resin; **B.** Western blot showing heparin-binding by GST-SKAMP1 protein, developed using anti-GST antibodies. Lane 1 – prestained MW markers with sizes as indicated (in KDa); lane 2 – protein eluted with 750 mM NaCl buffer; lane 3 – protein eluted with 1 M NaCl buffer; lane 4 – heparin sepharose resin; lane 5 – purified GST protein.

Table 1

DC	BINARY PATTERN	SEARCH STRING (SS)	NUMBER OF SS	TOTAL EMBEDDED	FOUND EMBEDDED IN:							
					HELIX	(%)	SHEET	(%)	TURN	(%)	RANDOM	(%)
EL	○●	XX-EL-DC-XX	14	1052	549	(**52**)	168	(16)	239	(23)	96	(9)
DI	○●	XX-DI-DC-XX	14	778	193	(25)	333	(**43**)	141	(18)	111	(14)
KL	○●	XX-KL-DC-XX	14	971	270	(28)	260	(27)	303	(31)	138	(14)
NV	○●	XX-NV-DC-XX	14	387	92	(24)	160	(**41**)	84	(22)	51	(13)
GA	●●	XX-GA-DC-XX	14	1262	339	(27)	205	(16)	487	(39)	231	(**18**)
ST	○○	XX-ST-DC-XX	14	882	161	(18)	151	(17)	435	(**49**)	135	(**16**)
CA	○●	XX-CA-DC-XX	14	188	53	(28)	69	(**37**)	60	(32)	6	(3)
PG	●●	XX-PG-DC-XX	14	727	119	(16)	19	(3)	533	(**73**)	56	(8)
MH	●○	XX-MH-DC-XX	14	89	36	(**40**)	20	(23)	26	(29)	7	(8)
WP	○●	XX-WP-DC-XX	14	116	11	(9)	0	(0)	99	(**86**)	6	(5)
QL	○●	XX-QL-DC-XX	14	393	115	(29)	134	(34)	97	(25)	47	(12)
YV	○●	XX-YV-DC-XX	14	362	142	(39)	136	(38)	32	(9)	52	(14)
FE	●○	XX-FE-DC-XX	14	726	522	(**72**)	84	(12)	111	(15)	9	(1)
RT	○○	XX-RT-DC-XX	14	423	73	(17)	65	(15)	231	(55)	54	(13)
			196	**8356**	**2675**	**(30)**	**1804**	**(23)**	**2878**	**(36)**	**999**	**(11)**

Figure 1

A.

BINARY PATTERN	SEARCH STRINGS	TOTAL EMBEDDED	FOUND EMBEDDED IN: HELIX (%)	SHEET (%)	TURN (%)	RANDOM (%)
○●○○●●○	[p][n][p][p][n][n][p]	544	341 (**63**)	52 (10)	136 (24)	15 (3)
●○○●●○●	[n][p][p][n][n][p][n]	1253	800 (**64**)	267 (21)	114 (9)	72 (6)

B.

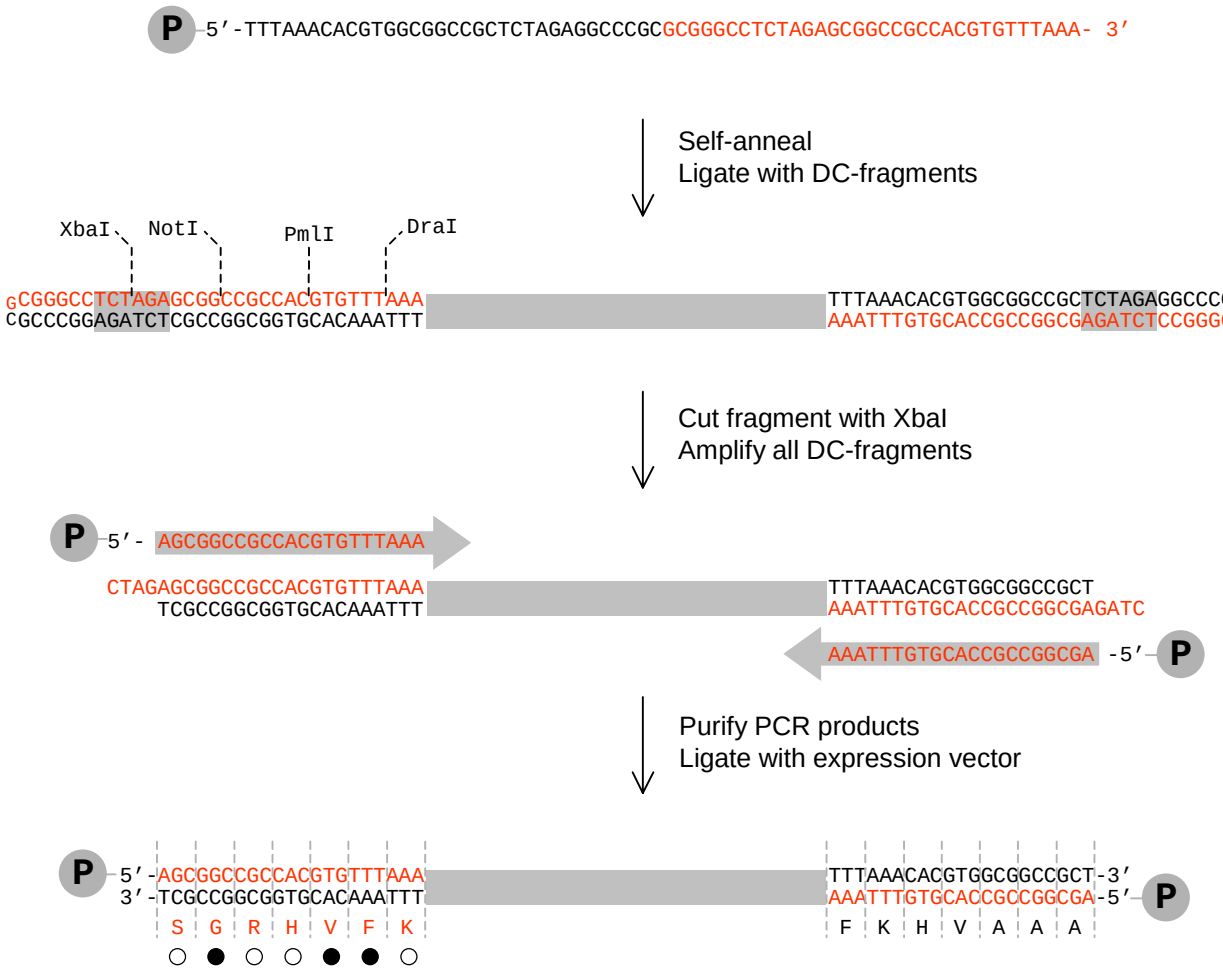


Figure 2

A.

NAME	SEQUENCE	aa	pI	MW
HPDN1	MGY**SGRHVFK**GAMHMHKLDIPGMHDIKLMH ELELDIELDIRTMHELRTDIGAMHDIGADI DIDIPGSTMHELMHMHELRTMHELSTMHDI MHDIKLPGDIQLDIELSTMHNVDIRTSTEL **FKHVAAA**VLEHHHHHH	136	5.7	15.7
HPDN2	MGY**SGRHVFK**ELELSTRTMHRTMHGAMHMH STPGMHDIMHSTCAELDIRTELGA MHNVMH DIDISTDISTMHRTELDIPGKLPGMHSTRT STMH**FKHVAAA**VLEHHHHHH	110	6.6	12.5
HPDN3	MGY**SGRHVFK**ELQLMHELGAMHRTMHNVDI RTSTSTMHELRTPGKLMHGAKLSTRTELGA STYVSTMHMHPGDIRTKLSTNVDIRT **FKHV AAA**VLEHHHHHH	102	9.8	11.6
HPDN4	MGY**SGRHVFK**PGSTDIMHMHNVSTGAELMH MHDISTRTMHNVSTMHRTSTKLCAKLQLST STWPELRTELSTMHMHDISTNVDISTQLEL RTDIRTEL**FKHVAAA**VLEHHHHHH	114	6.9	13.0
HPDN5	MGY**SGRHVFK**KLQLDIMHPGMHDIELDINV ELMHELMHMHMHQLSTSTDIMHDIMHFEMH STNVELMHPGGAKLPGSTPGDIRLDIPGMH DIMHMHMHKLQLMHELMHSTSTRTMHMHKL MHSTSTNVSTRTSTDIQLSTSTRTDI **FKHV AAA**VLEHHHHHH	162	6.5	18.7
HPDN6	MGY**SGRHVFK**STNVQLDIMHPGSTDIELMH ELRTMHPGELDIKLMHDIMHKLELGASTEL STMHELRTDIDIRTMHSTMHEL **FKHVAAA**V LEHHHHHH	98	6.2	11.3
HPDN7	MGY**SGRHVFK**MHKLMHELQLDIMHMHQLGA STRTSTKLPGMHDIMHELKLQLMHMHKLMH STDIRTELRTKLPGDIRTMHSTELRTDIEL STDIELGA**FKHVAAA**VLEHHHHHH	114	7.3	13.2
HPDN8	MGY**SGRHVFK**PGSTDIMHDIELQLDIRTDI MHSTMHDIPGKLMHPGELQLEL **FKHVAAA**V LEHHHHHH	68	6.2	7.8

B.

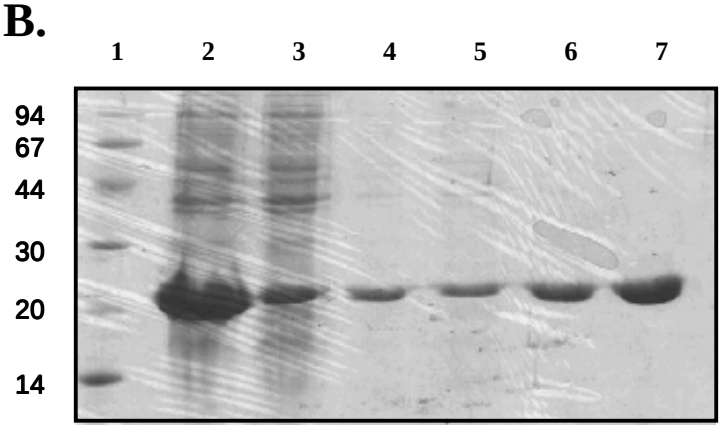


Figure 3

NAME	SEQUENCE	aa	pI	MW	SECONDARY STRUCTURE (%) HELIX	SHEET	RANDOM
EQAMP1	MGIQHFRVALIPFFAAFCLPVFAHPETLVKVKDAED QLGAYSGRHVFKMHCAKLELRTDINVMHKLMHMHCA MHPGFKHVAAAVLEHHHHHH	92	8.6	10.5	**63**	9	28
EQAMP2	MGIQHFRVALIPFFAAFCLPVFAHPETLVKVKDAED QLGAYSGRHVFKGASTKLMHPGELMHSTMHMHMHCA MHDIRTELDISTMHKLMHKLDIDIELDIFKHVAAAV LEHHHHHH	116	6.6	13.3	**55**	10	35
EQAMP3	MGIQHFRVALIPFFAAFCLPVFAHPETLVKVKDAED QLGAYSGRHVFKDIELMHRTDIRTELKLELELKLST ELMHMHQLDIDIRTELELQLDIELMHELELMHSTRT DIMHELKLSTSTSTSTELMHQLMHKLNVKLELMHKL KLMHKLQLMHKLFKHVAAAVLEHHHHHH	172	6.5	20.2	**79**	3	18
cEQ1	MGIQHFRVALIPFFAAFCLPVFAHPETLVKVKDAED QLGAYSGRHVFKDIRTSTMHELSTSTMHKLDIELPG MHMHSTSTMHKLNVMHKLMHELSTSTSTPGSTSTRT STMHPGMHDIGARTKLMHMHSTDIHTDIRTDIELKL PGMHDIMHPGSTELFKHVAAAVLEHHHHHH	174	6.9	19.6	26	6	**68**
SKAMP1	MGIQHFRVALIPFFAAFCLPVFAHPETLVKVKDAED QLGAYSGRHVFKELRTKLKLSTRTDIRTSTDIRTST RTELRTSTDIRTDIDIDIRTMHDIKLMHMHMHRTDI RTRTDIELRTKLRTFKHVAAAVLEHHHHHH	138	10.0	16.2	**54**	4	42

Figure 4

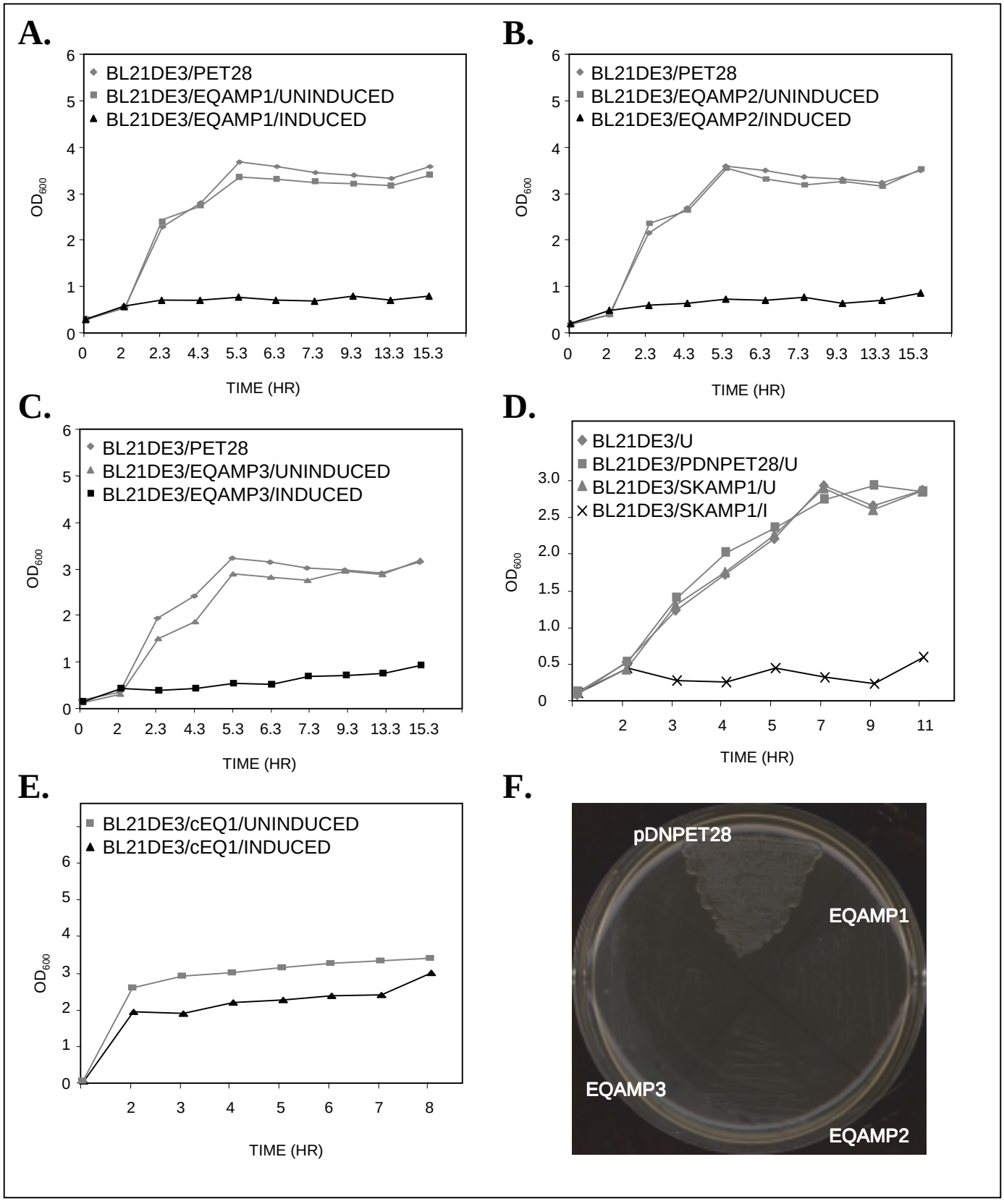


Figure 5

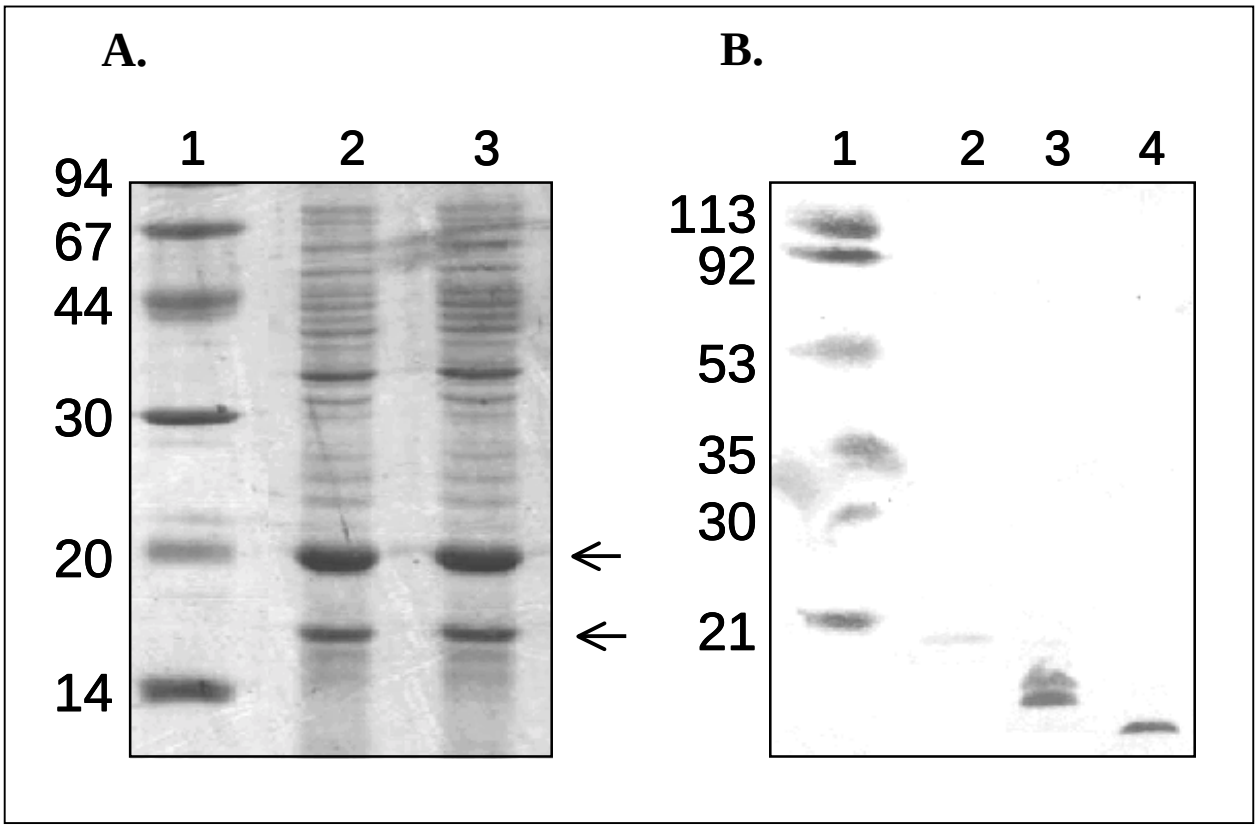


Figure 6

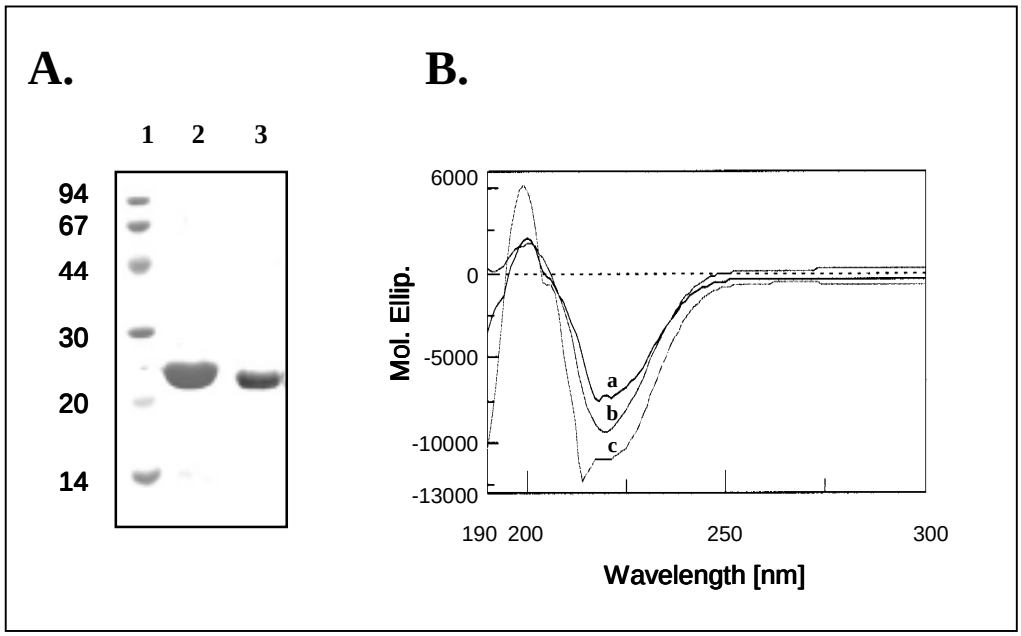
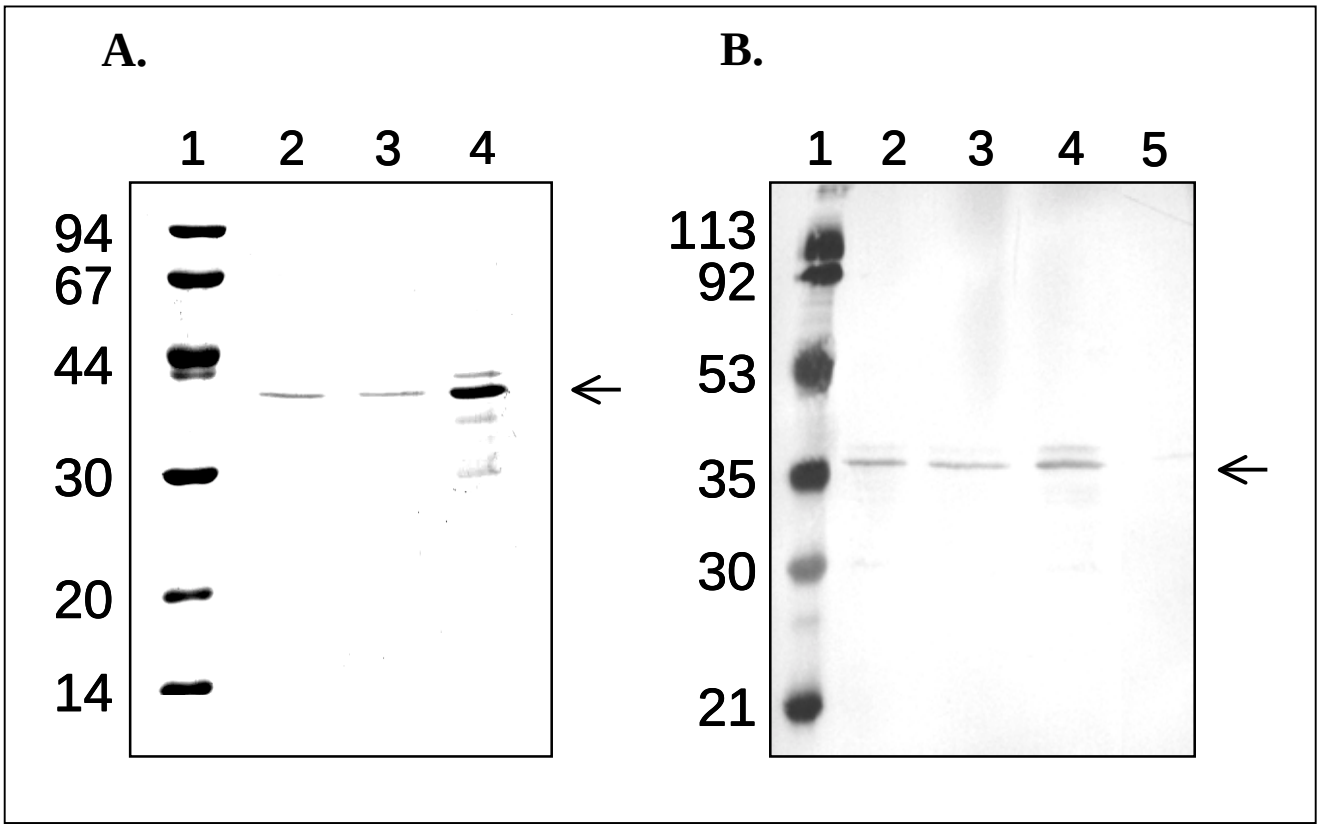


Figure 7



SUPPLEMENTARY EXPERIMENTAL PROCEDURES AND LEGENDS TO SUPPLEMENTARY FIGURES

SUPPLEMENTARY EXPERIMENTAL PROCEDURES

All antibiotics were purchased from Sigma. Ni-NTA resin and Anti-His monoclonal antibody were purchased from Qiagen. *pGEMT–Easy* and *pBluescriptSRF* cloning kits were purchased from Promega and Stratagene respectively. T4 DNA Ligase was purchased from New England Biolabs. *E. coli* BL21(DE3) strain, *pET21c(+)* and *pET28a(+)* expression vectors were from Novagen. Routine cloning and transformation procedures for *E. coli* were as described earlier (44). All new plasmids generated during this work were sequenced using the di-deoxy method in order to validate their authenticity. SDS-PAGE analysis was carried out on 15 % gels using the method described previously (45). All protein samples were treated with mercaptoethanol before loading them on to SDS-PAGE.

Multiple sequence alignments - Multiple amino acid sequence alignment was carried out using the CLUSTALW program available at http://www.ebi.ac.uk .

Expression and purification of DN1 protein - The plasmid *pDNp28/DN1* was transformed into *E. coli* BL21(DE3) cells by heat shock procedure for overproduction of the recombinant protein. Induction of mid log phase cultures (A_{600} = 0.4-0.6) with 0.5 mM IPTG at 37 ^{o}C for 4 h led to overexpression of the protein. The protein was found mostly in the insoluble inclusion bodies. The inclusion bodies were solubilized in Denaturing Buffer A (100 mM NaH_2PO_4, 10 mM Tris.HCl, pH 8.0, 8 M urea) followed by application onto a 5 ml Ni^{+2}-NTA column that had been equilibrated with buffer A. The column was extensively washed with buffer A and eluted with two bed volumes of 250 mM imidazole in buffer A. The presence of DN1 protein was detected by 15 % SDS-PAGE. Fractions containing the purified protein were pooled and dialyzed against buffer B (10 mM PBS, pH 7.4, 300 mM urea).

Expression and purification of HPDN5 protein - *E. coli* BL21(DE3) was transformed with the plasmid *pDNp28/HPDN5* and colonies selected on kanamycin. Overexpression of the protein was induced by addition of IPTG (0.5 mM) in the mid log phase (OD_{600} ~ 0.4-0.6). Cells were harvested by centrifugation, resuspended in buffer A (10mM Tris-HCl, pH 8.0, 1 mM EDTA, 0.8 % NaCl) and lysed by sonication (Sonic Dismembrator, Fisher Scientific, USA). The lysate was centrifuged at 17,000 rpm for 30 min at 4 ^{o}C. The supernatant was collected and applied onto a 5 ml Ni^{+2}-NTA column previously equilibrated with buffer A. After extensive washing of the column with buffer A, the protein was eluted with 250 mM imidazole in buffer A. Fractions containing the purified HPDN5 protein were pooled after analysis be SDS-PAGE and dialyzed against buffer B (10 mM Tris-HCl, pH 8.0). Finally, the purified protein was concentrated using YM-5 concentrators (Vivascience, UK).

Construction of expression plasmid pTRXDNp21 - The first and second halves (123 and 255 bp respectively) of the 378 bp *E. coli* thioredoxin gene (*trx*) were amplified from genomic DNA with *pfu* polymerase, using the following pair of oligonucleotides: first half – 5'-ccgcggccgcacatatggatatcaaaatt-attcacctgactg-3' and 5'-cctacgtaaccgcaccactctgccca-3'; second half - 5'-cctacgtaccgtgcaaaatgatcgcc-3' and 5'-ggtgcggccgcctcgaggatatccgccaggttagcgtcgagg-3'. The two halves were independently cloned in *pBluescriptSRF* and then combined to yield the complete *trx* gene. This altered gene was identical to the wild-type *trx* gene, except for the presence of a unique in-frame *Sna*BI site at a position in the gene that corresponds to the DC protein being flanked by the two trx cystines (---**C**G-'*Sna*BI'-P**C**---). The altered *trx* gene was then moved to the expression plasmid *pET21c* using the restriction enzymes *Nde*I and *Xho*I. The final plasmid was designated as *pTRXDNp21*.

Construction of the dummy vector pDUMDNp28 - The expression plasmid *pDUMDNp28* that was used as a control in growth inhibition experiments was constructed as follows: a 330 bp long *Eco*RV fragment from the vector *pTRXDNp21*, containing the full-length *trx* gene was ligated with *Sna*BI-cut and dephosphorylated *pDNp28*. The resulting plasmid, named *pDUMDNp28*, contains the *trx* gene in the correct orientation, for eventual translation into the full-length thioredoxin protein.

Expression and purification of THPDN3 protein - THPDN3 and the control protein TRX (thioredoxin) were overexpressed by induction of mid-log phase *E. coli* BL21(DE3) cultures with 0.5 mM IPTG. Both proteins were found in insoluble inclusion bodies. Consequently, the protocol for their purification was similar to that employed for the purification of DN1, except for a few minor changes. After solubilisation of the inclusion bodies with Denaturing Buffer A (see above), and application onto a 5 ml Ni^{+2}-NTA column previously equilibrated with the same buffer, the proteins were eluted with 500 mM Imidazole in buffer A. The purified proteins were refolded by dialyzing sequentially against buffer B (10 mM PBS, pH 7.4) that contained concentrations of urea ranging from 0-8 M. Dialyzed samples were analysed on 15% SDS-PAGE and their CD spectrum was recorded.

Construction of expression plasmid pSKAMP1GEX4T3: Plasmid *pSKAMP1DNp28* was digested with *Not*I and the 262 bp long *SKAMP1* gene excised. The *SKAMP1* gene was then ligated with *Not*I-cut and dephosphorylated GST-fusion vector *pGEX4T3* (Amersham Biosciences, USA). The resulting plasmid was designated *pSKAMP1GEX4T3*. Expression of this plasmid resulted in a 322 aa long GST-SKAMP1 fusion protein (MW 37.6 KDa).

Construction of expression plasmids pEQAMP2TRXDNp21 and pSKAMP1TRXDNp21: Plasmid *pTRXDNp21* was digested with *Sna*BI and the full-length genes EQAMP2 and SKAMP1 cloned as 5'-phosphorylated amplified products using the single oligonucleotide: 5'-agcggccgccac-gtgtttaaa-3', that served both as a forward and reverse primer. The resulting plasmids were designated *pEQAMP2TRXDNp21* and *pSKAMP1TRXDNp21* respectively. Expression of these plasmids resulted in trx-fusion proteins EQAMP2trx and SKAMP1trx, of MW 21 kDa and 24 kDa respectively. These proteins were purified using the procedure described earlier for trx fusion proteins.

Expression and purification of GST-SKAMP1 fusion protein: *E. coli* DH5α was transformed with plasmid *pSKAMP1GEX4T3* and the culture (1 l) induced with 0.5 mM IPTG for 12 h at 28 ^{o}C. The cells were subsequently lysed and soluble GST-SKAMP1 was loaded onto a column containing GST-sepharose 4B resin that had earlier been equilibrated with buffer A (1X PBS, pH 7.4). The column was extensively washed with buffer A and the fusion protein eluted with four bed volumes of Buffer B (10 mM reduced glutathione in 50 mM Tris-HCl, pH 8.0). Fractions containing pure protein were pooled and dialysed against Buffer C (50 mM Sodium Phosphate, pH 5.5) and visualized on 15 % SDS-PAGE.

Circular dichroism studies - CD spectra were recorded on a JASCO model J-810 spectropolarimeter. Data was collected at room temperature with a pathlength of 1 mm. The concentrations of proteins were as follows: THPDN3 - 7.5 μM; TRX – 8.2 μM; EQAMP2trx - 12 μM; SKAMP1trx – 5.2 μM. The proteins were suspended in 1X PBS, 150 mM urea, pH 8.0.

Transmission Electron Microscopy studies - Fresh colonies of *E.coli* BL21(DE3) transformed with EQAMP expression plasmids were cultured till log phase followed by induction with 0.5 mM IPTG at 37ºC for 4h. 1 ml of each induced culture was centrifuged and the pellet washed once with Buffer A (10 mM PBS pH 7.4). The cell pellet was fixed using 2.5 % glutaraldehyde and 2 % paraformaldehyde on ice. After 45 min of incubation, all pellet samples were washed three times with buffer A. Pellet samples were post fixed with 1 % buffered Osmium tetroxide for 1 h at room temperature followed by dehydration against a graded series of acetone (30-100 %). The post-fixed samples were infiltrated with resin + toluene and finally embedded in LR white resin (TAAB, UK). Polymerization of resin was done at 60 ºC for 12 h. For Transmission Electron Microscopic analysis, ultra thin sections (~70 nm) of each embedded sample were cut and lifted on copper grids followed by staining with 4 % uranyl acetate and 0.4 % lead acetate for 5 min. The sections were examined and photographed using FEI-PHILIPS MORGAGMI

268D digital Transmission electron Microscope. Control samples were also processed and analysed in parallel.

LEGENDS TO SUPPLEMENTARY FIGURES

Fig. S1. **A.** Amino acid sequences and primary sequence characteristics of *de novo* proteins DN1 and DN2. Residues underlined are those derived from the expression vector. **B.** Consensus secondary structure element estimates of DN1 and DN2 obtained using NPS@: Network Protein Sequence Analysis program (reference 46; http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_seccons.html). **C.** Purification of DC protein DN1. All samples were run on denaturing 15 % SDS-PAGE. Left panel, lane 1 – Pharmacia low molecular weight Marker; lane 2 – induced sample of *E. coli* BL21(DE3)/DN1 gene. Middle panel, lane 1 – MW marker; lane 2 – soluble fraction; lane 3 – insoluble fraction. Right panel, lane 1 – MW marker; lane 2 – 11.6 kDa DN1 protein eluted from Ni-NTA column in 8M urea.

Fig. S2. **A.** Amino acid sequences of some thioredoxin-encapsulated hairpin-DC protein library members, showing variance in protein length, pI and molecular weight (given in KDa). The hairpin sequences are displayed in bold letters. Segments derived from the thioredoxin scaffold protein are underlined. **B.** 15 % denaturing SDS-PAGE showing expression and purification of one library member – THPDN3. Lane 1 – uninduced *E. coli* BL21(DE3)/THPDN3; lane 2 – IPTG-induced *E. coli* BL21(DE3)/THPDN3 in 8 M urea and loaded on to Ni-NTA column; lanes 3 & 4 – wash; lane 5 - eluted sample of THPDN5 (MW 29 KDa); lane 6 - Pharmacia low molecular weight marker. **C.** CD spectra of purified THPDN5 protein.

Fig. S3. **A.** Map of the expression vector *pTEMDNp28* showing the *Sna*BI restriction enzyme location. **B.** Translated product of *pTEMDNp28.* The 23 amino acid N-terminal TEM1 signal is shown as a string of white circles.

Fig. S4. **A.** Schematic drawing illustrating the DC-protein frame shift induced by *Nco*I fill-in and self-ligation of the EQAMP expression vectors. The ribosome-binding site and the ATG start codon are underlined. **B.** Growth of EQAMPnco mutants on kanamycin / IPTG plate, in contrast to EQAMP1.

Fig. S5. Growth Inhibition curves of controls employed for AMP experiments. Generally, growing *E. coli* cultures were induced with IPTG after 2-3 hours of growth, or after an OD_{600} of 0.4 had been reached. Curves for induced cultures are shown in black. SCAFFOLD – the expression plasmid *pDNp28* used for expression of hairpin-encapsulated DC fragments; PET28 – expression plasmid *pET28a* that was used as the parent vector for creating the scaffold plasmid; STRAIN – *E. coli* BL21(DE3) that was used as a transformation host; DUMMY – scaffold expression plasmid containing the *trx* gene, and serving as a dummy control.

Fig. S6. Zone clearance studies with AMPs. Filter paper discs were soaked with 0.5 mM IPTG and placed on growing cultures of: **A.** *E. coli* BL21(DE3)/cEQ1; **B.** *E. coli* BL21(DE3)/EQAMP1; **C.** *E. coli* BL21(DE3)/EQAMP2; **D.** *E. coli* BL21(DE3)/EQAMP3 (here, IPTG was poured into a small cavity of 4 mm diameter); **E.** *E. coli* BL21(DE3)/SKEQ1.

Fig. S7. Hypothetical helical-wheel projection of EQAMP2. Nonpolar residues are shown as black circles while polar residues are depicted as white circles. Arrows depict the N- and C-termini of the protein. The protein sequence proceeds in the order of wheels A-E.

Fig. S8. TEM pictures of AMP-expressing *E. coli* BL21(DE3) strains. **A.** – Uninduced *E. coli* BL21(DE3)/EQAMP1; **B.** - IPTG-induced *E. coli* BL21(DE3)/EQAMP1; **C.** - IPTG-induced *E. coli* BL21(DE3)/EQAMP2; **D.** - IPTG-induced *E. coli* BL21(DE3)/EQAMP2.

Supplementary Figure 1

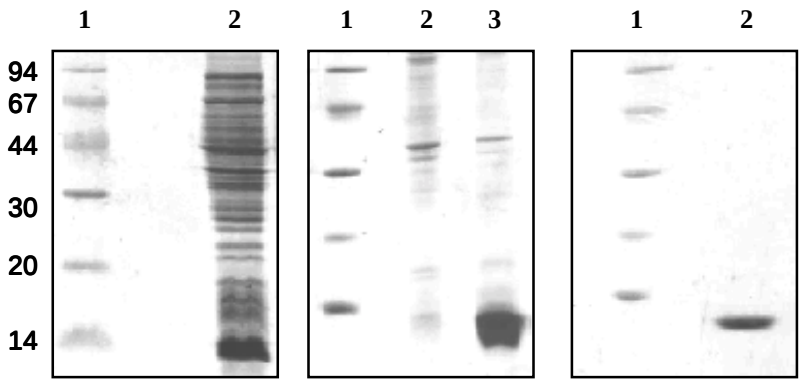
A.

NAME	SEQUENCE	aa	MW	pI
DN1	MGYELELDIRTDISTGAMHRTDIMHNVM HPGDIRTMHKLMHRTDISTDIELMHGAM HDISTPGMHRTMHMHMHSTPGDIMHRTE LELRTGAVLEHHHHHH	100	11.6	6.3
DN2	MGYMHKLMHSTRTMHDINVSTMHSTELR TELRTMHELRTELMHDICADIDINVSTD IMHGASTMHPGMHGAMHNVMHSTSTSTR TDINVSTPGELMHRTDIRTDISTKLSTD IPGSTMHRTELDIRTELMHNVRTPGDID IKLMHELVLEHHHHHH	156	17.9	6.2

B.

NAME	% SECONDARY STRUCTURE		
	HELIX	SHEET	RANDOM
DN1	24	9	67
DN2	29	6	65

C.

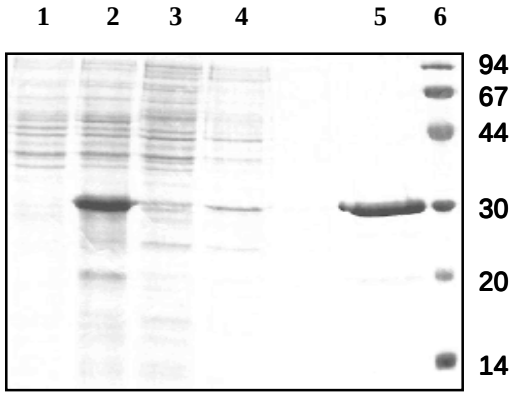


Supplementary Figure 2

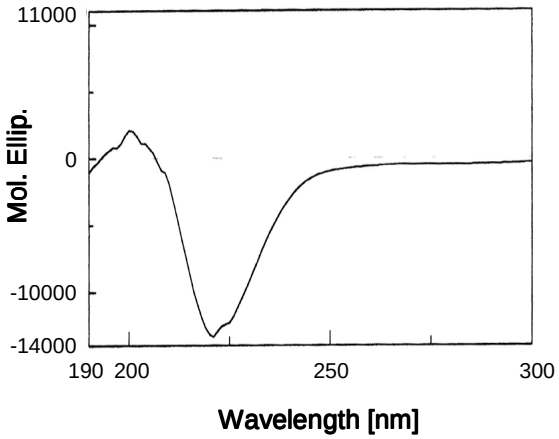
A.

NAME	SEQUENCE	aa	pI	MW
THPDN1	MDIKIIHLTDDSFDTDVLKADGAILVDFWAEWCGY**SGRHVFK**DIRTSTSTDIELRTSTDIDIMHELDIPGM HKLSTKLMHSTQLKLRTMHDISTKLMHPGELSTRTSTELQLDIQLDINVMHSTQLELMHMHSTRTMHGAMH PGKLRTDIMHSTDIMHPGMHMHMHMH**FKHVAAA**VPCKMIAPILDEIADEYQGKLTVAKLNIDQNPGTAPKY GIRGIPTLLLFKNGEVAATKVGALSKGQLKEFLDANLADILEHHHHHH	261	6.3	29.4
THPDN2	MDIKIIHLTDDSFDTDVLKADGAILVDFWAEWCGY**SGRHVFK**GAMHPGMHMHKLSTMHNVMHFEELQLKLD IPGMHGASTNVSTPGELMHSTELPGDIMHDIMHELMHRTKLELQLMHELDIRTSTPGNVMH**FKHVAAA**VPC KMIAPILDEIADEYQGKLTVAKLNIDQNPGTAPKYGIRGIPTLLLFKNGEVAATKVGALSKGQLKEFLDAN LADILEHHHHHH	225	6.0	25.1
THPDN3	MDIKIIHLTDDSFDTDVLKADGAILVDFWAEWCGY**SGRHVFK**DIMHPGELMHPGMHRTGASTPGMHELRTM HSTRTMHDIMHPGMHSTDIMHSTELDIRTRTSTELDIMHKLNVMHRTMHRTNVRTDIDIGASTQLSTMHMH PGELPGMHSTKLMHDIRTELEL**FKHVAAA**VPCKMIAPILDEIADEYQGKLTVAKLNIDQNPGTAPKYGIRG IPTLLLFKNGEVAATKVGALSKGQLKEFLDANLADILEHHHHHH	257	6.3	28.8
THPDN4	MDIKIIHLTDDSFDTDVLKADGAILVDFWAEWCGY**SGRHVFK**DISTMHSTRTKLELDIDIPGSTPGMHSTP GMHSTDIQLMHDIELKLMHMHMHNVMHELMHSTKLPGSTMHMHGADISTMHMHDIMHSTPGMHDIQLDIRT KLDIGAKLKLMHMHPG**FKHVAAA**VPCKMIAPILDEIADEYQGKLTVAKLNIDQNPGTAPKYGIRGIPTLLL FKNGEVAATKVGALSKGQLKEFLDANLADILEHHHHHH	251	6.2	28.0
THPDN5	MDIKIIHLTDDSFDTDVLKADGAILVDFWAEWCGY**SGRHVFK**DIQLMHRTELNVDIRTMHELMHDIRTDIR TDIELMHRTMHDIQLDIMHRTDIRTGAELMHSTQLELDIRTMHSTMHPGMHDIMHKLDIDIPGKLDIQLDI MHKL**FKHVAAA**VPCKMIAPILDEIADEYQGKLTVAKLNIDQNPGTAPKYGIRGIPTLLLFKNGEVAATKVG ALSKGQLKEFLDANLADILEHHHHHH	239	5.9	27.3
THPDN6	MDIKIIHLTDDSFDTDVLKADGAILVDFWAEWCGY**SGRHVFK**DIRTSTMHELSTSTMHKLDIELPGMHMHS TSTMHKLNVMHKLMHELSTSTSTPGSTSTRTSTMHPGMHDIGARTKLMHMHSTDIMHDIRTDIELKLPGMH DIMHPGSTEL**FKHVAAA**VPCKMIAPILDEIADEYQGKLTVAKLNIDQNPGTAPKYGIRGIPTLLLFKNGEV AATKVGALSKGQLKEFLDANLADILEHHHHHH	245	6.4	27.3

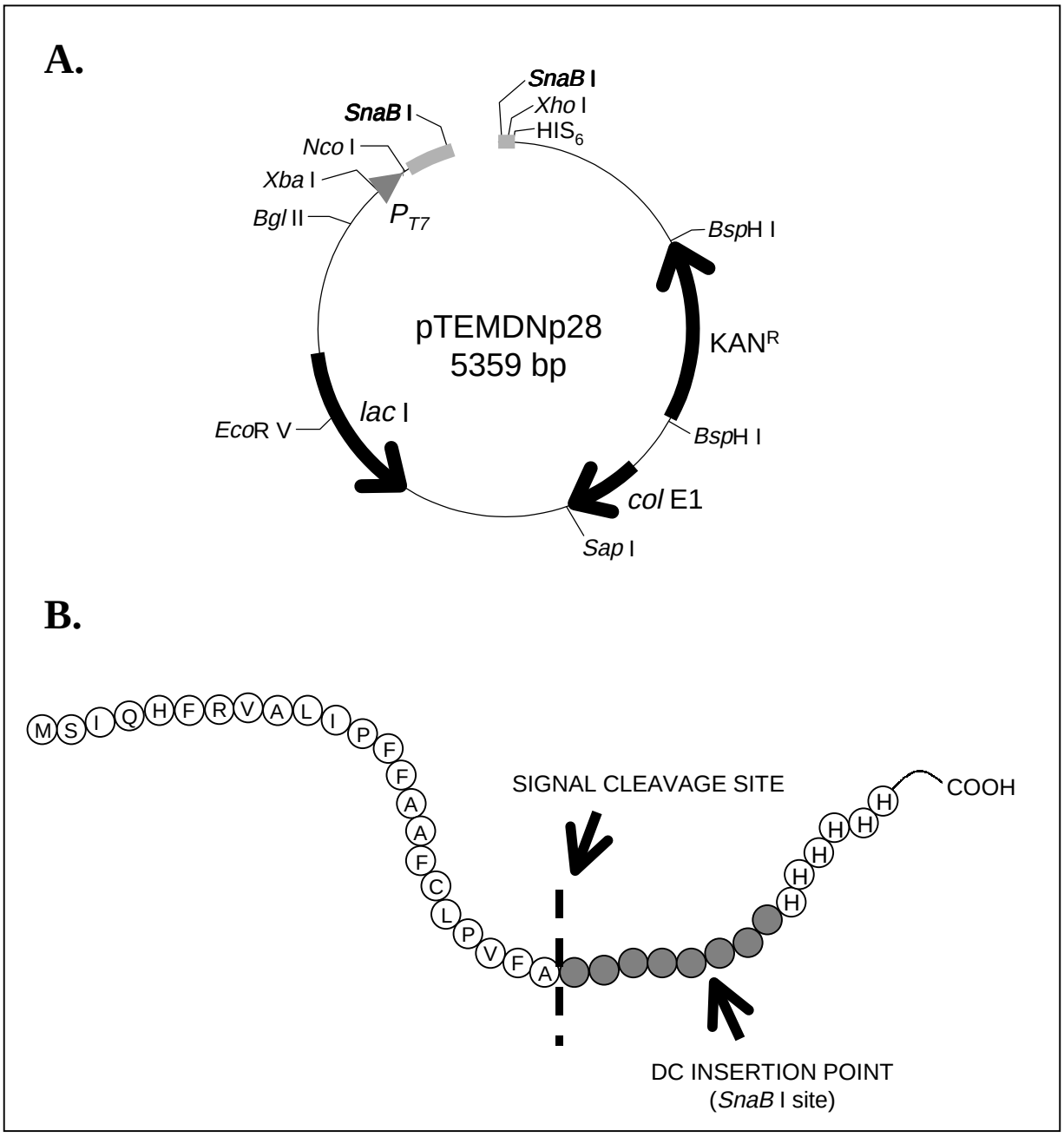
B.



C.

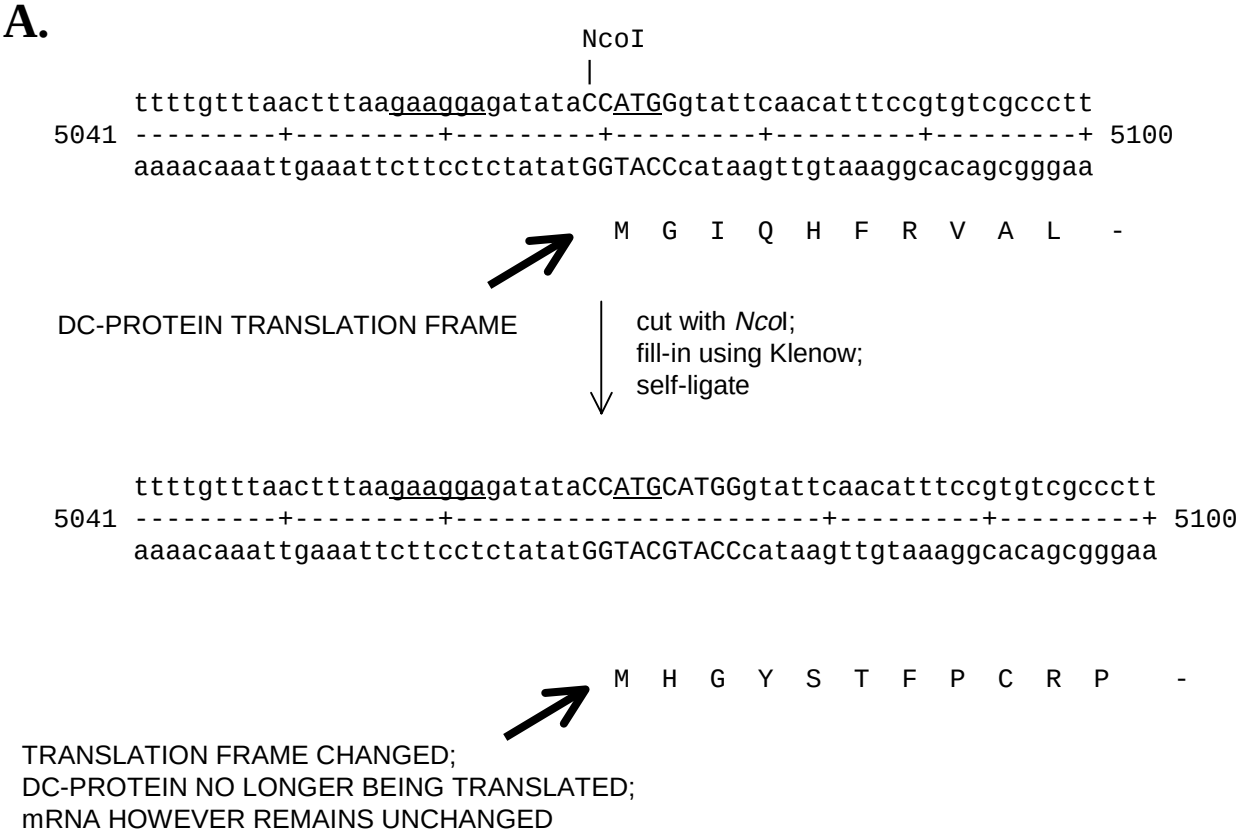


Supplementary Figure 3

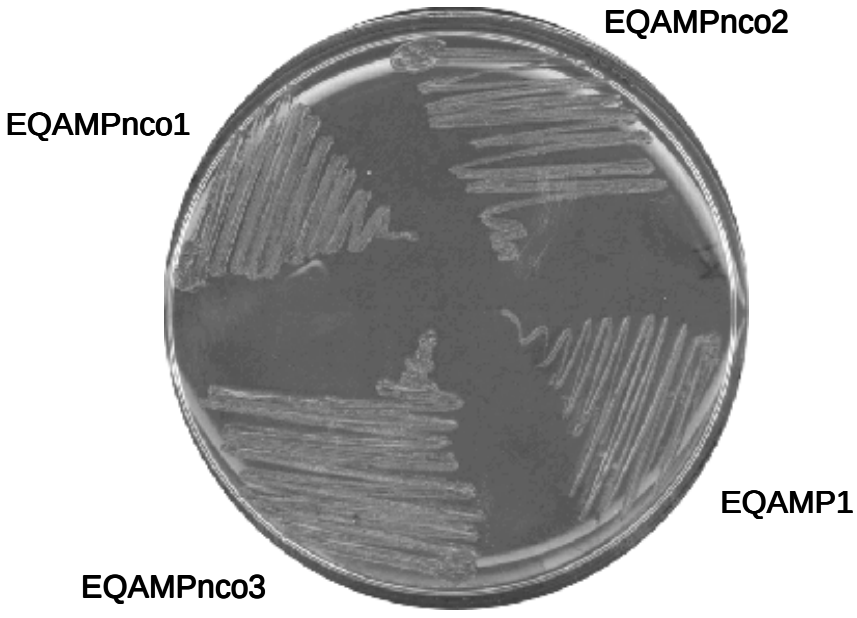


Supplementary Figure 4

A.

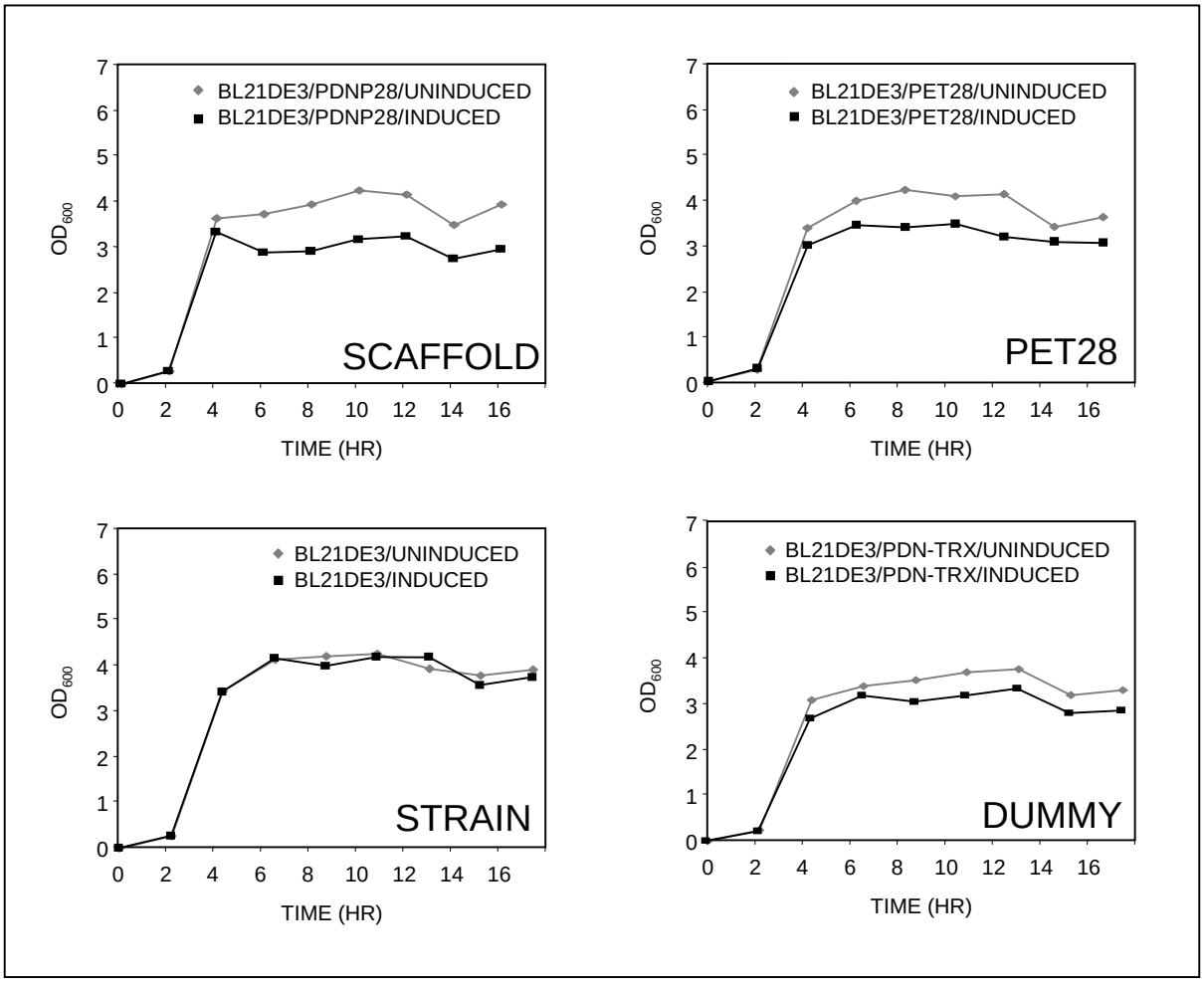


B.

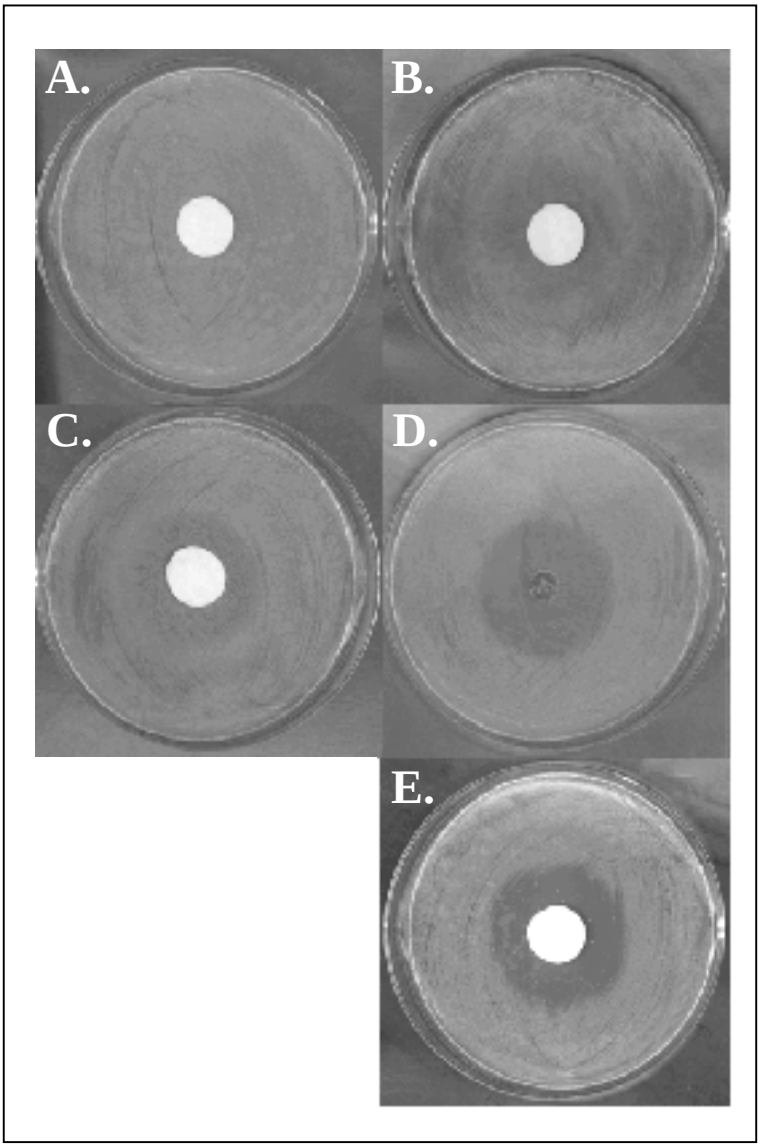


E. coli BL21DE3 / KAN / IPTG

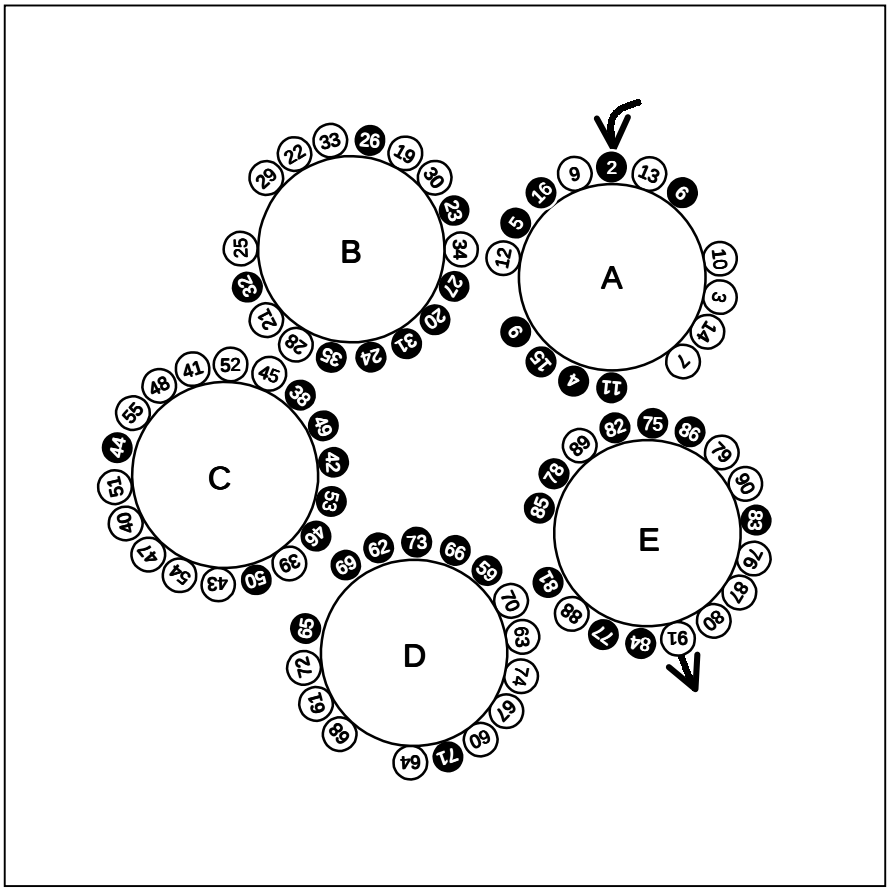
Supplementary Figure 5



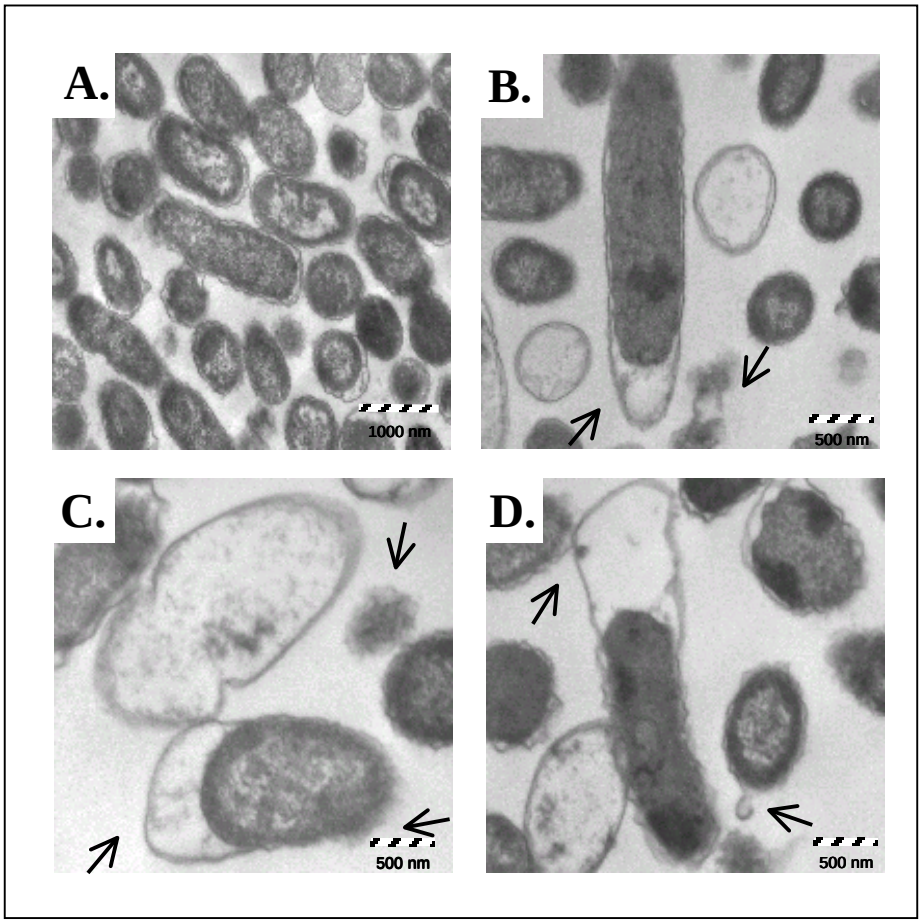
Supplementary Figure 6



Supplementary Figure 7



Supplementary Figure 8



Application of the "CODON-shuffling" method: Synthesis and selection of de novo protein as antibacterials

Alka Rao, Sidharth Chopra, Geeta Ram, Ankit Gupta and Anand Ranganathan

J. Biol. Chem. published online April 20, 2005
