Chapter 10: References


References


42. Kapust, R. B. and Waugh, D. S. 1999. *Escherichia coli* maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused. Protein Sci. 8: 1668-1674.


96
References


Expression, purification, and biochemical characterization of *Mycobacterium tuberculosis* aspartate decarboxylase, PanD

Sidharth Chopra, a Harish Pai, b and Anand Ranganathan a,•

a Recombinant Gene Products Group, International Centre for Genetic Engineering and Biotechnology, ICGEB Campus, Aruna Asaf Ali Marg, New Delhi 110 067, India  

b National Brain Research Centre, SCO 516&7, Sector 15/1, Gurgaon, Haryana, India

Received 24 January 2002, and in revised form 15 March 2002

Abstract

Like all bacteria, *Mycobacterium tuberculosis* (Mt) possesses the genes necessary for coenzyme A biosynthesis and metabolism. In the present work, the Mt *panD* gene was PCR amplified, overexpressed, and purified by metal affinity chromatography. The recombinant Mt panD was found to exist as a tetramer in solution. Incubation of Mt panD at 37 °C for several hours resulted in a complete cleavage of the inactive (α) form into the two subunits (α and β). The cleavage was confirmed by Western blot analysis as well as by N-terminal sequencing. Cleaved Mt panD was assayed for decarboxylase activity with L-aspartate as substrate. The kinetic parameters *Km* and *kcat* were found to be 219 µM and 0.65 s⁻¹, respectively. These results provide the means for further studies based on the identification of the Mt panD as well as other components of pantothenate metabolism as potential drug targets. © 2002 Elsevier Science (USA). All rights reserved.

Over one-third of the world’s population is infected with Mt,1 the etiological agent of tuberculosis. The high human and monetary costs resulting from this infection necessitate an urgent need for identifying factors that control the growth and metabolism of this pathogen. The complete sequencing of the Mt H37Rv genome has opened avenues for exploring Mt proteins as potential drug targets [1]. Recent findings suggest that the long-chain fatty acids and polyketides produced by Mt play a central role in its pathogenesis and growth [2,3]. The enzymatic systems that synthesize fatty acids and polyketides involve the use of the acyl-carrier protein (ACP). This enzyme shuttles nascent fatty acid and polyketide chains through the use of a coenzyme A-derived 4′phosphopantetheinyl moiety [4].

The immediate precursor of coenzyme A is pantothenate, a product of the condensation between β-alanine and pantoate. The formation of β-alanine is one of the most crucial steps in the pantothenate synthesis pathway and bacteria obtain it through decarboxylation of L-aspartate using the enzyme L-aspartate-β-decarboxylase or panD [5]. PanD is an unusual enzyme in that it contains a covalently bound pyruvoyl group that is necessary for carrying out catalysis. The protein is initially translated as an inactive proenzyme called as the π-protein [6]. The latter is then cleaved at a specific Gly-Ser bond. Upon cleavage, the π-protein breaks down into an α subunit with a pyruvoyl group at its N terminus and a β subunit with X-OH at its C terminus (Fig. 1). In solution as well as in crystal-form, the *Escherichia coli* panD has been found to exist as a tetramer [7,8]. The Mt homologue (15950 MW), however, does not share significant amino acid sequence identity with the rest of the panD family and has an extended C-terminus with an extra 13 amino acids as compared with the *E. coli* enzyme. In view of the pivotal role played by this class of enzymes in the growth and survival of the host, the Mt panD would represent a viable drug target, especially if the panD gene inactivation in Mt is able to prove the importance of this

1046-5928/02/$ - see front matter © 2002 Elsevier Science (USA). All rights reserved.
Pii: S1046-5928(02)00039-6
enzyme with regard to the growth and virulence of this pathogen.

Materials and methods

Materials

Chemicals and reagents were of the highest grade available. Ni-NTA resin and anti-His monoclonal antibody were purchased from Qiagen. The S-200 gel filtration column was from Pharmacia. Materials for SDS-PAGE were from Bio-Rad. Restriction enzymes were purchased from MBI Fermentas. pGEMT-Easy cloning kit and T4 DNA ligase were purchased from Promega.

Bacterial strains, plasmids, and growth conditions

E. coli BL21(DE3) strain and pET21c(+) expression vector were from Novagen. Mtb H37Rv genomic DNA was a gift from Dr. J. Tyagi, A.I.I.M.S., New Delhi, India. The Mtb H37Rv BAC library was obtained from Dr. S.T. Cole, Pasteur institute, Paris, France. All cloning steps were performed using E. coli JM109 cells. E.coli BL21 (DE3), that expresses the phage T7 RNA polymerase, was used for the overproduction of Mtb aspartate decarboxylase. Liquid cultures of recombinant E. coli were grown in LB Broth at 37°C with 150 μg/ml ampicillin. Routine cloning and transformation procedures for E. coli were as described earlier [9].

Multiple sequence alignments

Multiple amino acid sequence alignment of various aspartate decarboxylases was carried out using the CLUSTALW program available at http://www.ebi.ac.uk.

PCR and cloning strategies

The Mtb H37Rv panD gene (Rv 3601c, GI:2113975) was PCR amplified from BAC-Rv222 using a modification of the strategy described earlier [10]. Oligonucleotide primers were designed for directional cloning into NdeI and XhoI digested pET 21c(+) vector (Forward primer: 5'-AAGAATTCTATATGTACGGAGCATGC TG-3'. Reverse primer: 5'-AATTCGAGTCCCACACCCAGGCAGGGGAGGT-3'). The PCR protocol consisted of an initial 5-min denaturation step (94°C), followed by 25 cycles of (1) 94°C, 1 min, (2) 58°C, 1 min, and (3) 72°C, 1 min, and then a final elongation step at 72°C for 7 min. The 435 bp PCR product was eluted from a 1.0% agarose gel using DEAE cellulose, purified, and cloned into pGEMT-Easy vector to yield vector panDpGEMT. The DNA fragment containing the panD gene was then excised using the restriction enzymes NdeI and XhoI and ligated with pET21c(+) that had been cut earlier with the same two enzymes to give plasmid panDpET. The panD gene was inserted in-frame with the nucleotide sequence, corresponding to six histidine residues at the C-terminus. The ligation products were used to transform E. coli JM109 cells and the cells selected on ampicillin. Colony-PCR screening method was used to identify positive clones containing the panD gene. The identity of the selected clone was verified by restriction enzyme mapping as well as by DNA sequencing. No mistakes in the panD gene were detected.

DNA sequencing and protein analysis

DNA sequencing was carried out on the panDpET plasmid using T7 primer on an Applied Biosystems ABI Prism model 310 machine. SDS-PAGE analysis was carried out on 15% gels using the method described previously [11]. All protein samples were treated with mercaptoethanol before loading them onto SDS-PAGE.
Expression and purification of aspartate decarboxylase

The panDpET plasmid was transformed into E. coli BL21(DE3) cells by heat shock procedure for overproduction of the recombinant protein. The Mtb panD was expressed in a soluble form by induction of mid-log phase cultures \( (A_{600} = 0.6-0.7) \) with 0.5 mM IPTG at 37°C for 4 h. Cells were harvested by centrifugation, washed twice and resuspended in buffer A (10 mM Tris-HCl, pH 8.0, 1 mM EDTA, and 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°C. The supernatant was collected and applied 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 aspartate decarboxylase was detected by 15% SDS-PAGE. Fractions containing the purified protein were pooled and dialyzed against 10 mM Tris-HCl, pH 8.0.

The Ni\(^{2+}\)-NTA-purified material was concentrated using YM-5 concentrators (Vivascience, UK) and then loaded onto the gel filtration column. The Mtb aspartate decarboxylase was further purified using a Superdex-200 column (Amersham Pharmacia) equilibrated with buffer B (50 mM Tris-HCl, pH 7.5, and 150 mM NaCl). The Mtb panD was eluted as a tetramer around 13 min and its presence was detected by SDS-PAGE. The peaks were collected, pooled, and dialyzed extensively against 10 mM Tris-HCl, pH 8.0, and stored at 4°C.

Formation of active aspartate decarboxylase

The formation of active aspartate decarboxylase was studied by incubating recombinant aspartate decarboxylase (1 µg/µl) in 10 mM Tris-HCl, pH 8.0, at varying temperatures of 37, 50, and 70°C for 48 h. Once the temperature at which cleavage was seen to be optimum was standardized, a time course experiment was done with samples removed after every 12 h until 72 h to determine the optimum conditions for cleavage.

N-terminal sequencing of aspartate decarboxylase

Approximately 10 µg purified and cleaved Mtb panD was transferred onto a PVDF membrane in 100 mM CAPS buffer (pH 11) at 100 V for 2 h. This sample was then given for N-terminal sequencing on an Applied Biosystems Procise Sequencer.

Western blot analysis

Cleaved Mtb panD (10 µg) was 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).

cDNA analysis

Mtb H37Rv mRNA was prepared according to manufacturer's instructions (Gibco BRL, Life Technologies). Mtb H37Rv cDNA was prepared by using random hexamer primers (100 ng) per reaction and Superscript II Reverse transcriptase (gift of Dr. J. Tyagi, A.I.I.M.S., New Delhi, India). Using the cDNA as a template, the panD gene was PCR amplified using the primers described earlier.

Biochemical characterization of recombinant aspartate decarboxylase

Mtb panD activity was measured by derivitization of \( L\)-aspartate and \( \beta\)-alanine by fluorescamine, followed by quantification on HPLC. The activity assays were performed at 37°C in 50 mM potassium phosphate, pH 7.0/1 mM EDTA. All components were mixed and the assay was initiated by the addition of the substrate. At given time intervals of 1, 6, 10, 20, and 30 min, 100 µl aliquots were removed and mixed with 0.1 volume of 1 M NaOH to quench the enzyme reaction. Before HPLC quantification, 290 µl of 500 mM Borate (pH 10) and 100 µl fluorescamine (0.3 mg/ml in acetonitrile) were added to the mix and vortexed. The derivatized samples were quantified on reverse-phase HPLC using a C-18 column (Supelcosil C-18, 4.6 mm x 250 mm). Isocratic column elution was carried out at 0.75 ml/min using 20 mM sodium acetate, pH 5.9 (adjusted with acetic acid) containing 15% (v/v) acetonitrile. All plots were constructed using the GraphPad Prism software.

Results

Sequence comparison

The crystal structure of E. coli panD postulates that the residues, lysine, histidine, arginine, threonine, and tyrosine at position nos. 9, 11, 54, 57, and 58 in the protein chain, respectively, play important roles during catalysis by panD [8]. All the above residues are found to be absolutely conserved amongst the enzyme family (Fig. 2). As expected, the Mtb panD shares maximum identity (83%) with its M. leprae homolog while with E. coli panD, the sequence identity is only 45%. Further, the Mtb panD possesses as many as 13 extra amino acids at the C-terminus as compared with the E. coli panD. The crystal structure of E. coli panD shows that the C-terminus of the protein forms a contact point with the other subunits of the tetramer. The implications of these differences between the Mtb panD and its
non-mycobacterial homologs on the structure–activity relationship of aspartate decarboxylases are not yet clear.

**Cloning and cDNA analysis of the Mtb panD gene**

Using the BacRv222 plasmid containing Mtb H37Rv DNA as a template, the panD gene was PCR amplified and subsequently cloned as an NdeI–XhoI fragment into the pET21c(+) expression vector. The cloned panD gene was sequenced using T7 primer and the nucleotide sequence was found to be identical to the panD sequence available from the database. In the context of studying the Mtb panD as a viable drug candidate in future, it was important to find out whether the panD gene is non-functional in its natural possibility that this gene is non-functional in its natural state.

Fig. 2. Multiple sequence alignment of selected aspartate decarboxylases. Sequences were aligned using the CLUSTALW program. Deduced amino acid sequences were obtained from Escherichia coli (GI:12512846), Streptomyces coelicolor (GI:14970643), Corynebacterium glutamicum (GI:10190283), Mycobacterium leprae (GI:13092575), Mycobacterium tuberculosis (GI:2113975), Ratstia eutropha (GI:3786394), Aquifex aeolicus (GI:2983134), Neisseria meningitides (GI:7380140), Bacillus subtilis (GI:1146242), Staphylococcus aureus (GI:13702554), Campylobacter jejuni (GI:6967770), Helicobacter pylori (GI:2313109), Wolinella succinogenes (GI:2649792), and Mesorhizobium loti (GI:12044308). The asterisks and the dots denote the conserved and similar residues, respectively. The conserved glycine–serine residues that are at the site for cleavage in the panO gene are denoted by an arrow. The residues that are thought to play an important role during catalysis by panO are shown in bold. The asterisks and the dots denote the conserved and similar residues, respectively. The conserved glycine–serine residues that are at the site for cleavage in the panO gene are denoted by an arrow. The residues that are thought to play an important role during catalysis by panO are shown in bold.

**Non-exhaustive Taxonomic List**

- S. coelicolor
- C. glutamicum
- M. leprae
- M. tuberculosis
- E. coli
- R. eutropha
- A. aeolicus
- N. meningitides
- B. subtilis
- C. jejuni
- S. aureus
- H. pylori
- W. succinogenes
- M. loti

<table>
<thead>
<tr>
<th>Species</th>
<th>Accession Number</th>
<th>Accession Number</th>
<th>Accession Number</th>
<th>Accession Number</th>
<th>Accession Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>S. coelicolor</td>
<td>MLRTIKS...</td>
<td>MLRTIKS...</td>
<td>MLRTIKS...</td>
<td>MLRTIKS...</td>
<td>MLRTIKS...</td>
</tr>
<tr>
<td>C. glutamicum</td>
<td>MLRTILG...</td>
<td>MLRTILG...</td>
<td>MLRTILG...</td>
<td>MLRTILG...</td>
<td>MLRTILG...</td>
</tr>
<tr>
<td>M. leprae</td>
<td>MLRTLIG...</td>
<td>MLRTLIG...</td>
<td>MLRTLIG...</td>
<td>MLRTLIG...</td>
<td>MLRTLIG...</td>
</tr>
<tr>
<td>M. tuberculosis</td>
<td>MLRTNLK...</td>
<td>MLRTNLK...</td>
<td>MLRTNLK...</td>
<td>MLRTNLK...</td>
<td>MLRTNLK...</td>
</tr>
<tr>
<td>E. coli</td>
<td>MLRTMLK...</td>
<td>MLRTMLK...</td>
<td>MLRTMLK...</td>
<td>MLRTMLK...</td>
<td>MLRTMLK...</td>
</tr>
<tr>
<td>R. eutropha</td>
<td>MLRTMLG...</td>
<td>MLRTMLG...</td>
<td>MLRTMLG...</td>
<td>MLRTMLG...</td>
<td>MLRTMLG...</td>
</tr>
<tr>
<td>A. aeolicus</td>
<td>MLRTMLK...</td>
<td>MLRTMLK...</td>
<td>MLRTMLK...</td>
<td>MLRTMLK...</td>
<td>MLRTMLK...</td>
</tr>
<tr>
<td>N. meningitides</td>
<td>MLRTMLK...</td>
<td>MLRTMLK...</td>
<td>MLRTMLK...</td>
<td>MLRTMLK...</td>
<td>MLRTMLK...</td>
</tr>
<tr>
<td>B. subtilis</td>
<td>MLRTMLK...</td>
<td>MLRTMLK...</td>
<td>MLRTMLK...</td>
<td>MLRTMLK...</td>
<td>MLRTMLK...</td>
</tr>
<tr>
<td>C. jejuni</td>
<td>MLRTMLK...</td>
<td>MLRTMLK...</td>
<td>MLRTMLK...</td>
<td>MLRTMLK...</td>
<td>MLRTMLK...</td>
</tr>
<tr>
<td>S. aureus</td>
<td>MLRTMLK...</td>
<td>MLRTMLK...</td>
<td>MLRTMLK...</td>
<td>MLRTMLK...</td>
<td>MLRTMLK...</td>
</tr>
<tr>
<td>H. pylori</td>
<td>MLRTMLK...</td>
<td>MLRTMLK...</td>
<td>MLRTMLK...</td>
<td>MLRTMLK...</td>
<td>MLRTMLK...</td>
</tr>
<tr>
<td>W. succinogenes</td>
<td>MLRTMLK...</td>
<td>MLRTMLK...</td>
<td>MLRTMLK...</td>
<td>MLRTMLK...</td>
<td>MLRTMLK...</td>
</tr>
<tr>
<td>M. loti</td>
<td>MLRTMLK...</td>
<td>MLRTMLK...</td>
<td>MLRTMLK...</td>
<td>MLRTMLK...</td>
<td>MLRTMLK...</td>
</tr>
</tbody>
</table>
To detect panD mRNA, total RNA prepared from Mtb H37Rv was first reverse transcribed using random primers and then subjected to PCR using panD-specific primers. The obtained 435 bp product was cloned in pGEMT-Easy vector. The authenticity of the cloned fragment as the panD gene was ascertained by extensive restriction enzyme mapping. The results therefore indicate that the panD gene is transcribed in Mtb.

Overexpression and purification of Mtb panD

Cells containing the plasmid panDpET were grown in LB broth and induced with 0.5 mM IPTG. The overexpressed protein could clearly be visualized as a band corresponding to a molecular weight of 16,000 (Track 3, Fig. 3). The cells were harvested, lysed, and centrifuged and the supernatant containing the soluble protein was loaded onto a Ni-NTA column, following which the protein was eluted (Track 6, Fig. 3). The purified protein was concentrated and loaded onto an S-200 gel permeation column. Fractions corresponding to the tetrameric form of the protein (that appeared as a major peak relating to an approximate MW of 64 kDa) were pooled and concentrated. The protein was found to be greater than 95% pure by SDS–PAGE analysis. Protein concentration as determined by using the calculated extinction coefficient of 0.401 at 280 nm was found to be 22 μg/μl. Approximately 45 mg purified Mtb panD could be routinely recovered from 12 g (wet weight) of cells.

Formation of active aspartate decarboxylase

As mentioned earlier, panD is initially expressed as a pro-protein (α) that is then processed into two subunits (α and β). In the case of overexpression of the Mtb panD, we found that the protein is produced essentially as 16 kDa α subunit (Track 3, Fig. 3). This is in contrast to the case of histidine decarboxylase, where the protein that is produced is completely processed into the corresponding α and β subunits [12]. The Mtb panD protein purified by Ni-NTA chromatography was further purified by gel filtration chromatography. The peak corresponding to panD tetramer was eluted and shown by SDS–PAGE to consist exclusively of the 16 kDa protein (results not shown). These results indicate that protein processing is not essential for tetramer formation. Previous reports have indicated that the cleavage of the α protein is temperature assisted [7]. Incubation of Mtb panD at varying temperatures showed that the protein is cleaved at a temperature of 37°C and not at the higher temperatures of 50 and 70°C (Fig. 4A). This is in contrast to earlier results employing the E. coli panD that showed increased α protein processing at temperatures higher than 37°C [7]. It is, however, reasonable to assume that in the case of Mtb, maximal cleavage is likely to be at the optimal growth tempera-
ture of the pathogen (37°C). The processing of the purified α protein was confirmed by N-terminal sequencing (the N-terminal sequence obtained was VTIDADLMD, which is the N-terminal sequence of the α subunit of Mtb panD). The processing of the α protein was further confirmed by Western blot analysis. Taking advantage of the fact that the histidine tag will also be present at the C-terminus of the α subunit, the cleaved protein was probed with anti-His monoclonal antibodies. Upon processing of the Western blot, two bands corresponding to the molecular weights of the α and the α subunits could clearly be seen, thus, confirming that the protein had been cleaved (Fig. 4B).

To find out the minimum time taken for complete processing, we incubated the purified α protein at 37°C for various time intervals (Fig. 4C). As can be seen from the SDS-PAGE analysis, the protein was found to be completely cleaved after incubations that lasted longer than 24 h. Here again, it is worth noting that the doubling time for Mtb is around 24 h. The in vitro behavior of the Mtb panD enzyme therefore for all practical purposes reflects the natural physiological

---

Fig. 5. (A) HPLC analyses of the recombinant Mtb panD-mediated decarboxylation of L-aspartate as a function of time. The positions at which the fluorescamine derivatized L-aspartate and β-alanine elute are indicated. (B) Rate curves at increasing amounts of L-aspartate. (C) Michaelis-Menten plot of Mtb panD activity as a function of L-aspartate concentration obtained using 'best fit' employing the following equation: 

\[ Y = \frac{V_{\text{max}}}{K_m + X} \]

where \( V_{\text{max}} \) is the maximal velocity at saturation, \( K_m \) is the substrate concentration required to reach half-maximal velocity \( (V_{\text{max}}/2) \), \( Y \) represents the rate, and \( X \) is the concentration of L-aspartate. (D) Lineweaver–Burk 'double reciprocal' plot of \( 1/Y \) versus \( 1/[S] \), where \( V \) represents the rate of formation of β-alanine and \( [S] \) represents the concentration of the substrate, L-aspartate. The plot was obtained by using the equation: 

\[ \frac{1}{Y} = \frac{1}{V_{\text{max}}} \cdot \frac{1}{[S]} + \frac{1}{V_{\text{max}}} \]

The values obtained were as follows: \( X \) intercept (\( -1/K_m \)) = -4.552; \( Y \) intercept (\( 1/V_{\text{max}} \)) = .04067; slope \( (K_m/V_{\text{max}}) = .008933. \)
constraints under which the enzyme may have to
form.

**Termination of the catalytic activity of panD**

The fully processed Mtb panD protein was employed in enzyme assays with L-aspartate as the substrate. For these assays, we decided to use the novel method of Ramjee et al. [7], as it was found to be more sensitive than the methods involving monitoring the formation of the released CO$_2$ [13,14]. Detection of β-alanine was based on the formation of its fluorescamine adduct and its formation, as well as the concomitant depletion of the substrate (L-aspartate) was monitored online as described earlier (Fig. 5A). Using these assay conditions, the easy-state kinetic parameters $K_m$ and $k_{cat}$ for the conversion of L-aspartate into β-alanine by Mtb panD were calculated and found to be 219.6 μM and 0.65 s$^{-1}$, respectively (Fig. 5B, C, and D). The specific activity of the panD protein was found to be 2100 nmol/min/mg. These values correspond well with those of the *E. coli* panD [7].

**Discussion**

We have cloned and sequenced the Mtb *panD* gene. Using the H37Rv cDNA as a template, we have shown that *panD* is transcribed in Mtb. It has been suggested that the formation of β-alanine is the rate-limiting step in the pantothenate biosynthesis pathway [15]. Given the importance of β-alanine synthesis, it is highly likely therefore that the panD gene is expressed in Mtb.

We have also shown that the Mtb panD protein verexpresses in its uncleaved form in *E. coli*. The purified Mtb panD behaved as a tetramer under native conditions analogous to its *E. coli* homolog. We have found that optimal cleavage of Mtb panD results from a 8-h incubation period at 37°C. This is in contrast to the optimal cleavage conditions for the *E. coli* panD protein (48 h at 50°C; [7]).

Much of the information that we have today on the biochemical nature of aspartate decarboxylases is derived from the seminal work of Ramjee et al. [7]. However, the questions relating to the mechanism and rigginger of the processing of the α subunit still remain unanswered. Ramjee et al. have suggested the cleavage of the α protein being an autocatalytic process. The autocatalysis does occur, as has been shown by our emperature and time course experiments with the uncleaved pure protein. Further, the optimal time and temperature for autocatalytic processing more or less mirror the doubling time and growth temperature of Mtb. The rather long time period required to achieve protein processing in the case of Mtb panD may therefore act as a mechanism for regulating β-alanine synthesis in the cell.

We have also shown that the Mtb panD is catalytically active when overexpressed and purified as a recombinant protein from *E. coli*. The kinetic parameters obtained were found to be similar to those that have been obtained for other members of the aspartate decarboxylase family [7,13,14]. As mentioned before, the panD class of enzymes provides a vital link in the pantothenate biosynthesis of an organism. In the event that the panD protein is not produced in a cell, pantothenate may be obtained from outside. For this purpose, the bacteria utilize the properties of pantothenate permease and pantothenate kinase—enzymes that transport and phosphorylate extraneous pantothenate, respectively [16,17]. We therefore believe that the identification of the Mtb panD protein as a viable drug target would be the first step towards a multi-target approach, resulting in pantothenate starvation for the pathogen. Further studies employing the recombinant Mtb panD and L-aspartate analogues would reveal the flexibility that may be inherent in the environs of the Mtb panD active site. In addition, such studies would pave the way for utilizing this enzyme as a potential drug target.

**Acknowledgments**

The authors thank N.M.I.T.I.I, C.S.I.R. and T.C.M.R., New Delhi, India, for financial support. S.C. thanks U.G.C. for providing scholarship. The authors thank Dr. Jaya Tyagi, A.I.I.M.S., New Delhi, India for help in cDNA analysis and for the gift of Mtb H37Rv genomic DNA; Dr. Stewart Cole, Pasteur Institute, Paris, France for the kind gift of Mtb BAC library; Dr. Dinkar Salunke, N.I.I., New Delhi, India for N-terminal sequencing; Ms. Ritu Diwan, I.C.G.E.B., New Delhi, India for DNA sequencing; Dr. Rajesh Gokhale, N.I.I., New Delhi, India for helpful discussions and Drs. Navin Khanna, S. Swaminathan, and Amit Sharma, I.C.G.E.B., New Delhi, India for helpful support during the course of this work.

**References**


Protein Evolution by “Codon Shuffling”: A Novel Method for Generating Highly Variant Mutant Libraries by Assembly of Hexamer DNA Duplexes

Sidharth Chopra and Anand Ranganathan*
Recombinant Gene Products Group
International Centre for Genetic Engineering and Biotechnology
Aruna Asaf Ali Marg
New Delhi 110067
India

Summary

The laboratory evolution of functional proteins holds great fascination as an effective tool for searching and discovering novelty in the protein space. Here, we employ a simple molecular approach for generating a multitude of structurally diverse yet functionally similar proteins that have all “evolved” from the parent enzyme, in this case a β-lactamase. By randomly combining a set of 14 DNA-hexamer duplexes, each corresponding to judiciously chosen amino acid pairs, we were able to generate functional proteins that contained large regions of previously unknown sequences. Some of the generated proteins were not only more active than the parent, they were also significantly smaller in size. Our approach could therefore be effectively used in searching for novel structural as well as functional proteins.

Introduction

Charles Darwin first introduced the concept of “Unconscious Evolution” to describe the sudden improvement of an individual member of a species through generations of selective breeding and natural stress selection, by means of processes beyond the control and understanding of man [1]. The idea of laboratory evolution or directed evolution [2–4] of proteins and nucleic acids, however, is only as recent as a decade old, although much progress has since been made [5–8]. The use of degenerate oligonucleotides for cassette mutagenesis, sometimes referred to as “saturation mutagenesis” has been helpful for directed evolution as well as protein structure and function studies, particularly as it allows one to obtain mutants that may differ in every base position in a chosen stretch of DNA sequence [10–12]. Seminal studies by Loeb and coworkers as well as by Struhl and Ollphant have demonstrated the advantage of saturation mutagenesis over the earlier conventional methods [11, 12]. However, the variance is limited to a “portion” of the total gene length, which rarely exceeds 50 bp (correspondingly ~15–20 amino acids). This may be seen as a possible drawback especially in cases where mutations in the whole gene and not just one particular region are desired. Stemmer and coworkers were the first to address this problem by expanding upon the combinatorial potential of earlier techniques such as site-directed mutagenesis, error-prone PCR, and degenerate-oligonucleotide (or cassette) mutagenesis in order to create libraries of protein and nucleic acid mutants, with their gene-shuffling (or sexual PCR) technique [5, 13]. Since its inception, the gene-shuffling method has been used extensively to select for protein mutants, and there are now a host of new techniques that are variants of the gene-shuffling method [8–10]. The mutational rate in gene-shuffling ranges from 0.3% to 2% of the total length of the gene, a factor that may be seen as a drawback of the technique. Nonetheless, gene shuffling and saturation mutagenesis are currently the methods of choice in directed evolution experiments. More recently, Szostak and coworkers have developed a molecular evolution technique (mRNA display) by utilizing a large collection of synthetic degenerate DNA molecules in order to produce a library of proteins of equivalent size [14, 15]. However, these advances in molecular evolution of proteins are not entirely devoid of limitations, ranging from low on degeneracy (i.e., error rate) to high on complexity of experimental design and expense [5, 14].

We now report an altogether novel approach toward protein evolution, one that introduces the concept of “codon shuffling” in order to obtain a vast number of shuffled degenerate gene sequences. The approach entails in vitro assembly of hexamer DNA duplexes that represent all 20 amino acids as codon pairs. The assembled “dicodons” are ligated with an inactive gene scaffold and the ligation mixture used to transform a bacterial host. Upon stress, the cell selects the appropriate shuffled gene whose length has been increased by the dicodon addition, for eventual translation into a catalytic protein that is fit to combat the provided stress (it may be noted that the term “codon shuffling” has been used here to represent the events that occur during the in vitro assembly process, i.e., the dicodons are shuffled in a random manner; the earlier usage of the term “shuffling,” for example in gene shuffling, was to denote shuffling of the wild-type DNA sequences). We have applied this technique to an Inactive β-lactamase parent gene scaffold, and upon codon shuffling, have selected progenies that are vastly different from the parent with some differing in their primary sequence by as much as 94%, while still retaining the β-lactamase activity.

Results and Discussion

Protein Evolution through Codon Shuffling
In order to design a concept of laboratory evolution of proteins that would either address or circumvent the limitations of earlier methods, we began by asking a simple question: can the degeneracy provided for by the genetic code at the nucleotide level be exploited to yield a library of proteins? After all, the degeneracy of a protein at its primary sequence level is the direct fallout of the degeneracy of its gene that is being translated (e.g., the assembly of a 300 base pair DNA fragment,
made up of 100 codon positions that are to be filled up by a set of 20 provided codons having complete freedom to combine with each other, will result in a degeneracy of 20^{100} different fragments, each of which, when translated, will produce a unique protein). In theory, if a set of codons were to be ligated with each other, and the assembled DNA efficiently translated, one could end up with a collection of proteins whose degeneracy is a factor of the number of codons chosen in the particular set, as well as the manner in which the codons themselves have assembled. In practice, however, the ligation of a 3 base pair DNA (i.e., a codon) duplex with another codon duplex using T4 DNA ligase has not yet been reported, although it is common to ligate 6 base pair DNA duplexes with other DNA fragments of varying lengths [16].

Therefore, in an effort to develop over-and-above this theme of so-called "codon-shuffling," we carried out the following experiment: we randomly assembled in vitro, a collection of 8 base pair DNA duplexes (called dicodons) that represented all 20 amino acids, to form degenerate DNA fragments. The collection of dicodons was devoid of the three stop codon sequences so that all degenerate DNA fragments may be read always in a single desired open reading frame. This collection, comprising 14 dicodons, was based on the following "design rules": any given codon was paired up with a codon that is the exact complement of the former and in addition is a "neutral" codon; high-usage codons were used as representative of amino acids, with E. coli usage as the guide (the codon usage cable can be accessed at http://www.kazusa.or.jp/codon/); the collection was a reflection of the percentage of the type (e.g., neutral, polar, hydrophobic, etc.) of amino acids present in the parent protein that was to be evolved, as well as a reflection of the distribution of the various types of amino acids used in E. coli. The comparison of the percentage of amino acids present in an equimolar dicodon pool with E. coli total codon usage is shown in Supplemental Figure S1 at http://www.chembiol.com/cgi/content/full/10/ 10/917/DC1. As can be seen from the figure, the percentage values are similar to a large extent, thereby indicating that any progeny developed from the use of the equimolar dicodon pool would be as well represented in the various protein attributes (hydrophobic, acidic, basic nature, etc.), as are most proteins synthesized by E. coli. The collection of dicodons chosen for the experiment is shown in Table 1. Wherever possible, care was taken to associate any given codon with one that would be representative of a hydrophobic, small, or a nonpolar amino acid (thus termed as a "neutral" codon). The reasoning behind this was two fold. First, there is a preponderance of neutral codons in E. coli—the 7 amino acids Leu, Val, Ile, Met, Pro, Ala, and Gly together make up 50% of total codons used. Through such a pairing, therefore, the equimolar pool of 14 dicodons would work as a truer representative of the E. coli codon usage. Second, these so-called neutral codons represent amino acids that, among them, enjoy superior secondary-structure forming propensities (see reference 4 in the Supplemental Data available at Chemistry & Biology/website). Next, every codon was paired with its exact complement, as this would allow such single-stranded palindromic DNA molecules to associate spontaneously and form a homodimer, or duplex DNA upon cooling. Therefore, only those palindromes were selected wherein the possibility of a mismatched duplex was energetically unfavorable or nonexistent. The complete absence of a mismatched dicodon duplex in the ligation pool would mean that a frameshift in the reading frame was not possible, as only perfectly matched, hence blunt-ended dicodons would be present in the ligation reaction. All 20 natural amino acids were represented by the 14 dicodons (with amino acids Glu, Gly, Ala, Pro, Thr, and Val being represented twice and the amino acid Leu three times); the percentage hydrophobicity, defined as the percentage of amino acids having a positive value on a Kyte-Doolittle hydrophobicity scale in an equimolar dicodon pool, was 39%, compared with a value of 38% for the parent enzyme (see below; [17]).

Codon Shuffling of a Parent β-Lactamase

As a proof-of-concept experiment, we chose the TEM-1 enzyme, a 286 amino acid long class A β-lactamase as the parent protein [18]. TEM-1 is the best studied of all β-lactam hydrolyzing enzymes, with its active site comprising of a catalytic serine residue (S°, Figure 1) to which the open form of the β-lactam is covalently linked [19]. Numerous structural and functional studies have indicated active involvement of other 3 amino acid residues in catalysis, along with a string of 15 residues lying almost 100 residues downstream of the active site serine—the so-called "loop"—thought to play a crucial role during the binding and hydrolysis of the lactam moiety (Figure 2A) [20-22]. In fact, all of the residues that are involved in catalysis lie downstream of S°, including a lysine (involved in the acylation step) that is just 2 residues away. In our belief, TEM-1 represents an attractive target for laboratory evolution using codon shuffling for many reasons. First, it is a reasonably large, monomeric, and soluble protein that can be assayed for its activity both in vivo and in vitro with ease using a variety of substrates. Second, it possesses an active site that is composed of residues that are well spread out in the primary sequence of the protein. Finally, as the results of codon shuffling would be such as to overwhelmingly change the primary sequence of the parent protein, any new evolved protein would yield important structural and functional insights into the functioning of β-lactamases in general.

As a first step, we severely truncated the parent TEM-1 enzyme, so that it was now only 86 amino acids long (including the 23 amino acid signal peptide at the N terminus and a His×6 tag at the C terminus, respectively). The plasmid containing the gene for this protein, pSC2, was unable to confer resistance against any of the β-lactams tested (Table 2). This was not at all surprising considering that the truncated protein was all of only the first 53 amino acids of the wild-type TEM-1, with S° and K° being the only remnants left from the active site of the parent (Figure 2B). In order to build on this inactive protein scaffold, we engineered a dicodon insertion point in its gene, at a position 15 bases downstream of the K° codon. The insertion point was in the form of the sequence for the restriction enzyme SnaBI (5' TAGGTA3')
such that cleavage of the gene with this enzyme would result in blunt-ended products on either side of the excision—a TACO and a GTA in codon, respectively. Plasmid pSC2 was then excised with SnaBI and introduced to a soup of randomly assembled dicodons, followed by transformation of E. coli DH5α and plating on media

**Figure 1. Primary Sequence Analyses of Obtained Mutant β-Lactamasess**

(A) Primary sequence of wild-type and mutant β-lactamasess. For the proteins other than TEM-1, the regions in red depict the evolved sequences, while those in black represent the wild-type sequence. The catalytic serine (S) and lysine (K) residues are shown in bold. The conserved Tyr residue (boxed) in the mutant sequences is a result of the SnaBI dicodon insertion site design. %MR, percentage mutagenic rate, defined as the percentage of primary sequence that is different from the wild-type TEM-1; GRAVY, grand average of hydropathicity (17). (B) A register of the number and type of dicodons found in the mutant β-lactamasess. The numbers for the second generation mutants BlaSC7-9 do not include the first generation dicodon numbers. (C) Cloning of mutant lactamases according to the amino acid attribute patterning developed by Hecht and coworkers. Polar amino acids, namely S, T, Y, W, O, N, C, H, R, K, D, and E, are depicted as open circles while nonpolar amino acids, namely, G, A, I, V, L, P, F, and M, are shown as filled circles. The dicodon sequences are in red. The sequences are numbered according to (9). As a reference, sequence no. 7 shows the patterning of an insoluble protein that contains β sheet repeats and forms amyloid structures. This sequence was designed by Hecht and coworkers to test their hypothesis of obtaining amyloid type structures when alternating polar-nonpolar repeats (shown boxed) are used (24).
containing antibiotics (see Experimental Procedures). Out of the initial library size of approximately 10^6 transformants, we isolated three cfu that were highly resistant to ampicillin and amoxycillin (Table 2). It may be noted that the expression vector pET28a (+) was chosen as a starting vector for all our experiments as it possesses a kanamycin-resistant gene and not a β-lactamase gene as the selection marker. However, in order to discount for the theoretical possibility that our observations were a result of an external contaminating β-lactamase gene, we isolated the plasmid DNA from the resistant strains and PCR amplified the causative agent for the resistance by using universal primers (the so-called universal primers—T7 forward and reverse primers—do not bind to the regions flanking the TEM-1 gene in any of the commonly used cloning vectors, like for example pGEMT, pBluescript, pUC, or pET-based vectors, all of them being possible sources of a contaminating lactamase gene. On the other hand, the universal primers bind perfectly to the T7 promoter and terminator regions of a pET28-based vector. These regions thus flank any “evolving” gene and are therefore ideal primers for amplification of such a gene). The amplified genes were cloned first in a different vector (pGEMT-Easy) and then inserted back into the pET28a expression vector, followed by introduction of these new plasmids into E. coli DH5α. The resulting strains too were resistant to β-lactams, thus confirming our initial observations that the resis-

---

**Table 2. Minimum Inhibitory Concentration Values for Various β-Lactams in Units of μg/ml**

<table>
<thead>
<tr>
<th>Strain/Plasmid (Protein)</th>
<th>AMP</th>
<th>AMX</th>
<th>PEN</th>
<th>CTX</th>
</tr>
</thead>
<tbody>
<tr>
<td>DH5α/pSC1 (TEM-1)</td>
<td>20,000</td>
<td>&gt;20,000</td>
<td>10,000</td>
<td>0.03</td>
</tr>
<tr>
<td>DH5α/pSC3 (BlaSC3)</td>
<td>40,000</td>
<td>&gt;20,000</td>
<td>5,000</td>
<td>0.03</td>
</tr>
<tr>
<td>DH5α/pSC4 (BlaSC4)</td>
<td>10,000</td>
<td>10,000</td>
<td>500</td>
<td>0.03</td>
</tr>
<tr>
<td>DH5α/pSC5 (BlaSC5)</td>
<td>10,000</td>
<td>10,000</td>
<td>1,000</td>
<td>0.03</td>
</tr>
<tr>
<td>DH5α/pSC7 (BlaSC7)</td>
<td>20,000</td>
<td>&gt;20,000</td>
<td>6,000</td>
<td>0.40</td>
</tr>
<tr>
<td>DH5α/pSC8 (BlaSC8)</td>
<td>10,000</td>
<td>&gt;20,000</td>
<td>2,000</td>
<td>0.03</td>
</tr>
<tr>
<td>DH5α/pSC9 (BlaSC9)</td>
<td>10,000</td>
<td>&gt;20,000</td>
<td>2,000</td>
<td>0.03</td>
</tr>
<tr>
<td>DH5α/pSC10 (BlaSC10)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DH5α/pSC11 (TEM-1')</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DH5α/pSC24 (BlaSC24')</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DH5α/pSC26 (BlaSC26')</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DH5α/pET28a (+)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

MIC, minimum inhibitory concentration.

* MIC values of less than 10 μg/ml.
* MIC values of less than 0.02 μg/ml.
Protein Evolution from Hexamer DNA Duplex Assembly

The presence of the contaminant would always show colony growth on ampicillin, even at the point where an inactive BlaSC3 was being tested for its resistance to the antibiotic.

Library Size
Our estimate of the library size was based on the observation that greater than 10^6 cfu were obtained when the vector-locked codon fragments (i.e., the ligation mixture) were plated on media containing only kanamycin, while plating of just the self-ligated SmalI-cut and dephosphorylated vector on media containing kanamycin resulted in not more than 20 cfu. This number, when subtracted from the first value, gives the library size. It can now be configured that 3 cfu out of this library were able to overcome the ampicillin as well as kanamycin stress in the dicon experiment. This is at best a crude estimate of the library size as extraneous factors like the efficiency of the ligation process as well as of the competent cells will affect the size estimate.

Analysis of the Progeny β-Lactamase
The progeny sequence from the three evolved genes is shown in Figure 1A. Indeed, all three genes, BlaSC3-5, were diconon incorporations (in red), sharing little sequence similarity between them or for that matter with any of the proteins in the database. The longest gene among the three, contained in pSC3, yields a protein that is less than half the size of the parent TEM-1 but with a similar GRAVVY value (Figure 1A). It is worth noting that the evolved proteins from previous directed-evolution techniques are always of the same size as their parent and differ only in their primary sequences [5]. Growing cultures of E. coli strains harboring pSC3 were found to be consistently more resistant than the other two TEM-1 derivative genes, and displayed an MIC value double that of the TEM-1 gene-containg strains (Table 2). The BlaSC3 protein was also found to be more soluble than the other BlaSC3 proteins and could routinely be purified for assaying (see below). As is clear from the sequence of BlaSC3, there is an overwhelming deviation from the wild-type TEM-1 sequence with no hint as to the nature of the new active site residues, apart from S^- and K^- . We confirmed the active role of these 2 residues during catalysis by BlaSC3 using site-directed mutagenesis (see Experimental Procedures). The S48A and K48A BlaSC3 mutants showed no catalytic turnover (Table 3). While the entire nature of the BlaSC3 active site can only be made clear through a more detailed structural analysis, it is evident that altogether new active site structural motifs have taken the place of their counterparts that were present in the TEM-1 enzyme (like the w loop). Furthermore, primary sequence analysis of BlaSC3 and its comparison with wild-type TEM-1 show that although BlaSC3 is 79% different from TEM-1, it is strikingly similar in the overall attributes that pertain to hydrophobicity, charge, and secondary structural propensities (Figure 3). This finding does indeed point tantalizingly to the suggestion that the natural course for selecting a progeny from the parent during protein evolution is in fact an emphasis on the overall "functional" similarity between the two proteins even though the progeny is different from the parent by as much as 79% (or for that matter 94% as in the case of the second generation mutant described later). In other words, the diconon-lead protein evolution mimics natural evolution that generates any given protein family. A relevant example of this is the protein family of β-lactamases itself, where we find tens of proteins among the four different classes of lactamases, some differing dramatically in primary protein sequence as well as size, yet all acting as β-lactam hydrolyzing enzymes. The profile of obtained progeny sequences in accordance with the elegant hypothesis of Hecht and coworkers [23, 24], whereby particular secondary structures can be predicted to occur on the basis of patterning of polar and nonpolar amino acids in the primary sequence, also does not conclusively point toward any particular structural repeat. Encouragingly, comparison of progeny profiles with that of an insoluble amyloid-forming β sheet repeat structure (no. 7, Figure 1C) [24] indicates that such structural elements may not be present in the obtained sequences. One reason why such repeats are not obtained may be because of the patterning of the 14 dicodons itself (Figure 1C)—not all dicodons are patterned as polar-nonpolar pairs. Indeed, it may be that this facet is in part responsible for the progenies being water soluble.

Expression and Purification of the Evolved β-Lactamase
As a last check, we truncated the wild-type TEM-1 enzyme, to bring it to a length similar to that of BlaSC3 (see Supplemental Data at http://www.chemblol.com/ cgi/content/full/10/10/917/DC1). The truncated protein BlaSC10 was found to be catalytically inactive, and E. coli cultures bearing the plasmid pSC10 were not able to grow on ampicillin plates. This shows that the activity of the evolved β-lactamase is because of the nature of the primary sequence of the protein and not its length.
sion bodies. The β-lactamases are therefore commonly purified through uninduced expression of the gene, from the resultant periplasmic fractions. We therefore expressed the BlaSC3 gene in uninduced growing cultures of *E. coli* DH5α and isolated soluble BlaSC3 as C-terminally His-tagged protein (MW 14.2 kDa) using Ni-NTA chromatography. We found, for reasons as yet unexplained, the expression levels of BlaSC3 to be around 30–40 times less than TEM-1 levels under similar conditions. The isolated protein was adjudged to be >90% pure by densitometric analysis using SDS-PAGE. The identity of this protein was also confirmed by Western blot using anti-His×6 antibodies (Figure 4; see Supplemental Data at http://www.chembiol.com/cgi/content/full/10/10/917/DC1). Purified periplasmic BlaSC3 was assayed for its activity using ampicillin, nitrocefin, and cefotaxime as substrates. The $k_{\text{cat}}/K_m$ values were in the micromolar range, similar to those obtained for the parent TEM-1 enzyme in our hands, as well as those previously reported for class A β-lactamases (Table 3) [25].

Further Codon Shuffling of a Progeny β-Lactamase

We then decided to further minimize the presence of TEM-1 structural remnants in BlaSC3. We shifted the
dicodon insertion point to the N terminus of the S46 region while additionally reducing the size of the region to just 13 residues. This time the codon shuffling resulted in the generation of three new genes with their predicted proteins, BlaSC7-9, showing little sequence similarity between them (Figure 1A). However, the primary sequence analysis of one of the second generation mutants, BlaSC7, suggests good similarity at the “sequence attribute” level with the wild-type TEM-1 as well as the first generation mutant BlaSC3 (Figure 3). BlaSC7 is the smallest (91 aa) among the three proteins and it is 94% different from TEM-1, yet it shares a similar hydrophobicity and charge profile. In retrospect, these similarities are quite possibly responsible for the export of these proteins into the periplasmic space as well as their presence in the periplasm as soluble proteins. The strains harboring the shuffled genes displayed similar albeit lower MIC values compared with the TEM-1 or the earlier BlaSC series, with one notable exception. We found that the strain containing the gene for the BlaSC7 protein showed an MIC value for cefotaxime that was more than a fold higher than the TEM-1 enzyme itself (Table 2). Cefotaxime is a third generation β-lactam that displays high potency against the class A β-lactamases (MIC of around 0.02 μg/mL), but is more readily hy-

---

**Table 3. Kinetic Parameters of TEM-1 and Other β-Lactamases**

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>AMP</th>
<th>NIT</th>
<th>CTX</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$k_m$ (μM)</td>
<td>$k_{cat}$ (s⁻¹)</td>
<td>$k_{cat}/K_m$ (μM⁻¹·s⁻¹)</td>
</tr>
<tr>
<td>TEM-1</td>
<td>48</td>
<td>992</td>
<td>18.4</td>
</tr>
<tr>
<td>BlaSC3</td>
<td>23</td>
<td>1011</td>
<td>43.5</td>
</tr>
<tr>
<td>BlaSC7</td>
<td>76</td>
<td>400</td>
<td>5.24</td>
</tr>
<tr>
<td>TEM-114</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>BlaSC7114</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>BlaSC76</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

AMP, ampicillin; NIT, nitrocefin; CTX, cefotaxime; ND, not determinable ($K_m > 1000$ μM).

*No measurable activity.*

---

**Figure 3. Primary Sequence Analysis of Evolved β-Lactamases Using the SAPS Program**

Comparison of the two representative first and second generation evolved mutants (BlaSC3 and BlaSC7, respectively) with the wild-type TEM-1 enzyme is shown. The y axis represents the percentage amino acids of the total number of amino acids in the proteins. Hydrophobic: defined in SAPS (Statistical Analysis of Protein Sequences) [33] as the L, V, I, F, M set; polar: S, T, C; acidic: D, E; basic: K, R; small: A, G; net charge: DE +, KF, α helix and β sheet propensity: % total of amino acids values defined in the modified version of the Chou-Fasman scale (see reference 4 in the Supplemental Data available at https://www.chembiol.com/cgi/content/full/10/10/917/DC1) that have a tendency to partake in the formation of α helices or β sheets; % mutational rate: defined as the percentage of primary sequence that is different from the wild-type TEM-1.
Protein Evolution with the plasmic, Figln stress, Uke shows [5].

malE found a and Expreslon obtain sufficiently large gene devoid to be previously made with the maltose binding protein. Previous work has indicated that N-terminal MBP fusion results in increased amounts of soluble proteins that were previously difficult to obtain as single entities [26]. In the first instance, the fusion was made with the malE gene devoid of its 25 amino acid long periplasmic secretion signal so that the fusion protein would exclusively be present in the cytoplasmic fraction. We were able to obtain sufficiently large amounts of the fusion protein in soluble form (see Supplemental Figure S2 at http://www.chembiol.com/cgi/content/full/10/10/917/DC1). Unfortunately, upon cleavage from MBP, we found BlaSC3 to be preferentially in the insoluble fractions. However, treatment with urea, followed by refolding of the protein resulted in BlaSC3 regaining its catalytic activity (see Supplemental Data at Chemistry & Biology’s website). This was also found to be the case for signal-less TEM-1 and BlaSC3 proteins that were found to be insoluble when isolated from the cytoplasm but regained hydrolytic activity upon urea treatment and refolding (see Supplemental Data at Chemistry & Biology’s website).

Indeed, the regain of activity upon refolding of TEM-1 has previously been reported [27]. We then fused the mutant as well as the wild-type TEM-1 genes with the periplasmic signal-containing malE gene. All three fusion proteins are overexpressed in the cell and are generously visible at their predicted lengths on SDS-PAGE (see Supplemental Figure S3 at Chemistry & Biology’s website). Preliminary results indicate that the MBP fusions are found predominantly in the soluble fraction. In addition to removing any doubts regarding the absence of a contaminating lactamase, this finding also means that progenies can now be constructed on inactive “MBP-parent protein” scaffolds, thus yielding greater amounts of soluble protein, a prerequisite for conducting 3D structural studies. Such studies are currently underway in our laboratory.

Comparison with Other Notable Directed Evolution Methods

Cassette Mutagenesis

One of the earliest examples of the utility of cassette mutagenesis involved the “reformatting” of the 17 amino acid long active site region of TEM-1 using a 51 bp degenerate oligonucleotide [12]. From an initial library of 10^9, 2000 cul were able to confer moderate to high ampicillin resistance. Out of the 50 clones that were sequenced, it was found that any given clone contained not more than three amino acid changes (mutation rate around 10%-15% over the 17 residue stretch). The comparison can now be made on many counts. First, while the library size is two orders of magnitude larger than what is obtained from our method, the “whole-gene” mutation rate is 1% compared to values of up to 94% obtained from our method. Second, the lengths of the progenies are all identical to the parent, as this parameter is dictated by the length of the degenerate oligonucleotide itself. Here, there is a striking difference between the two methods—the lengths of the active lactamases obtained through our method range from being 40% to just 25% of the total TEM-1 length.

Gene Shuffling

The very first application of gene shuffling was to the TEM-1 lactamase [5]. Although the initial library size was not determined, it is likely that the number would have been as high as the transformation efficiency of the host, i.e., around 10^9. This number is much higher than what is obtained using our method. There is also a clear advantage when using gene shuffling, in that all areas of the parent gene are prone to mutagenesis, unlike in our method or in cassette mutagenesis. However, the mutation rate per cycle was in the range of 1%, albeit variant methods like family shuffling, ICHY, etc. yield much higher mutagenesis rates. Finally, the progeny length is always identical to that of the parent, unlike in our method. As of the present, we believe that gene shuffling, in its simplicity and control, remains as the method of choice for a truly combinatorial approach toward improvement of any given protein.

Methods for De Novo Protein Synthesis

Perhaps a more pertinent comparison would be with regard to the prospect of generating totally de novo proteins or chimeras wherein large regions of the se-
quence are created de novo. Here, attention is drawn toward the seminal works of Szostak and coworkers [14], Riechmann and Winter [28], and finally of Hecht and coworkers [23]. In a pioneering effort, Szostak and coworkers created totally de novo proteins that were able to counter stress, which in their case amounted to the binding of ATP moiety by the proteins [14]. The successful proteins were obtained from a maximum possible set of 10^8 sequences, which in their estimate is tantamount to obtaining a fully folded functional protein from a collection of 10^10 totally random sequences. In another scenario, Riechmann and Winter demonstrated the possibility of obtaining protein chimeras, starting from an inactive protein scaffold, in their case a cold shock protein, capA [26]. The scaffold was made active by “filling up” the missing regions of the protein sequence through the use of a library of small segments comprising the total E. coli genome. From a library of 10^8 sequences, as many as 600 chimeras were able to counter the provided stress, given that nearly half of each progeny sequence was always that of the wild-type capA. We on the other hand have obtained functional proteins from a library that is several degrees of magnitude lower than the starting libraries in the abovementioned experiments. Why then are functional proteins obtained from libraries in one case and not the other? It is believed because of the following two reasons. First, although the chimeras BlaSci-7 differ greatly from their TEM-1 parent in the primary sequence, all of them without exception possess the two most critical residues required for β-lactamase activity, namely the active site serine and tyrosine. In the context of a directed evolution experiment, this may be construed as the inactive scaffold being “sufficiently close” to becoming an active protein. Thus, all that was required was a “little push” in the right direction that came from the codon-assembled amino acid addenda. It therefore remains to be seen whether our approach can be extended to generate totally de novo functional proteins. Second, for such amino acid attachments to result in functional proteins, given that the library is very large, it perhaps is helpful if the selected codons themselves have an inherent preference for forming favorable structures. As has been mentioned earlier, the 14 codons do not all conform to one single identical pattern in terms of their polar-nonpolar nature. Therefore, it may very well be that the obtained amino acid attachments are from the beginning itself able to form foldable structures, thereby reducing the need for a large primary library. This is presently the best a conjecture that nonetheless can be tested by employing a range of different codon sets. For such future experiments, the hypothesis of Hecht and coworkers mentioned earlier would be particularly helpful for codon design.

Significance

We have reported here a novel and simple method for protein evolution capable of generating variants that differ vastly from their parent protein. We have also shown that the variance can be applied at multiple regions on the parent sequence depending upon the position of the codon insertion point. The method can easily be programmed to skew the type of amino acids one wishes to be present in the evolved proteins, depending upon the nature, amount, and number of codons used. Ours is a method that yields proteins of vastly varying lengths and primary sequence. In many cases, therefore, the “mutation rate” is as high as 94%. Indeed, this variance in progeny length makes for an attractive prospect for generating “minimized” enzymes that contain a bare minimum catalytic center, yet are able to perform as well or even better than their large parents. Our method can also be utilized in the context of elegant, wholly in vitro systems of Szostak and coworkers, thereby further extending the amount of degeneracy accessible at the DNA level. There is also the attractive possibility of applying our method for the construction of combinatorial protein libraries, based on the binary patterning developed by Hecht and coworkers [23]. For example, alternating patterns of polar (C)-nonpolar (Ω) amino acid sequences (e.g., OOOOOOOOOOOOOO) that give rise to βsheet repeats can be obtained by listing only those codons that are themselves polar-nonpolar pairs (e.g., DI, KL, EL, NV, etc). On the other hand, patterns leading to α helices would have to obtained by assembly of modules like OOOOO, so that the assembled sequence corresponds to a nonpolar residue every three or four residues (e.g., OOOOOOOOOOOOOO). Success in such a strategy promises exciting implications for de novo protein design.

Experimental Procedures

All antibiotics, with the exception of Nitrocefin (Merck) were purchased from Sigma. Ni-NTA resin and Anti-Flag monoclonal antibody were purchased from Clagen. PGEML-Easy cloning kit were purchased from Promega. T4 DNA Ligase was purchased from New England Biolabs. E. coll BL21(DE3) strain, pET21a (+) and pET28a (+) expression vectors were from Novagen. QuickChange-XL site-directed mutagenesis kit was purchased from Stratagene.

DNA Cloning

Routine cloning and transformation procedures for E. coli were as described earlier [25]. All new plasmids generated during this work were sequenced using the dideoxy method in order to validate their authenticity.

Construction of Plasmid pSC1

The TEM-1 β-lactamase gene was PCR amplified (25 cycles) from plasmid pET21c using taq DNA polymerase and the oligonucleotides 5'-AACCATGGGTATTCAACAATCCGTGTCGCCCTT-3' and 5' - AAC TCAAGTTACCATCATTTCACTTGGTGCGGCCC-3' as forward and reverse primers respectively. The resulting 875 bp long PCR product was cloned in pGEMT-Easy vector. The TEM-1 gene was excised using NcoI and XhoI restriction enzymes and cloned in pET28a vector previously cut with the same two enzymes. The resulting expression plasmid was designated pSC1.

Construction of Plasmid pSC2

Using plasmid pSC1 as template, a fragment from TEM-1 gene was amplified using the oligonucleotides 5'-AACCATGAAGAGATTGGCSCCCCAACAGTCCCC-3' and 5' - AATAGGTAAGCAGAGAAGAC TTCAAAGGCACCAGAGCGCGAAC TTCAAAGGT-3'. The resulting 975 bp long PCR product was cloned in pGEMT-Easy vector. The desired fragment of the TEM-1 gene was excised with Sphl and SmaI and cloned in pSC1 previously cut with the same two enzymes. The resulting expression plasmid was designated pSC2. This vector served as the inactive TEM-1 scaffold for the first-generation dicond experiments.
Construction of Libraries

100 ng of each of the 14 5'-phosphorylated DNA hexamers (referred to as dodecans or DC in text) in a 20 μl reaction mixture containing 2 μl of 10× ligase buffer and 14 μl of double-distilled water was gently heated to 56°C. The temperature was then slowly brought down to 4°C. 14 μl of each DC was then mixed and distributed equally into seven tubes. To each of these tubes were added 2 μl of 10× ligase buffer and 14 DNA ligase respectively and the tubes incubated at 4°C. This was considered as time t = 0 hr. 2 μl (~250 ng) of SnaBI-cut and dephosphorylated plasmid pSC2 was added to each tube after defined time intervals of 0.15, 0.5, 1, 4, 8, 16, and 24 hr. All tubes were then incubated at 4°C for a further 8 hr. The contents of each tube were then independently used to transform E. coli DH5α competent cells. The transformation efficiency of these competent cells had earlier been standardized to 2 × 10^8 cfu per 100 μl of cells per μg of DNA. The transformation mixture was plated on LB agar plates containing 50 μg/ml kanamycin, 0.1 mM IPTG (see reference 1 in the Supplemental Data available at http://www.chembiol.com/cgi/content/full/1/10/971/D1C1), and increasing amounts of ampicillin (50–200 μg/ml). Initially 3 cfu, exhibiting growth at ampicillin concentrations of 20,000; 5,000; and 2,000 μg/ml respectively were isolated, plated again on ampicillin/kanamycin plate, and plasmid DNA recovered from them. The plasmids were designated as pSC2, pSC4, and pSC5. Using these plasmids as DNA templates, the three genes of unknown sequence were independently PCR amplified by employing the universal forward and reverse primers 5'-AATTAATGACACTAGTATAGGAAGTTG-3' and 5'-AGGACGACTACCTTCTTGCGGTCTT-3' and the PCR products cloned in pGEM-T-Easy vector. The complete genes were then excised using the restriction enzymes NcoI and XhoI and cloned in pET28a expression vector previously cut with the same two enzymes. The newly derived plasmids were sequenced using the T7 forward and reverse primers and were designated pSC3, pSC4, and pSC5. The sequence of the genes harbored in these plasmids—designated BlaSC3, BlaSC4, and BlaSC5—was found to be identical to the gene sequence contained in plasmids pSC3', pSC4', and pSC5'. As a final check, we digested plasmids pSC3-5 with restriction enzymes in order to destroy the newly formed β-lactamase genes. The digested plasmids were self-ligated and used to transform DH5α. The transformants were able to grow on kanamycin but not on ampicillin.

Construction of Plasmid pSC6

Using plasmid pSC7 as template, a fragment from TEM-1 gene was amplified using the oligonucleotides 5'-AGGACGACTACCTTCTTGCGGTCTT-3' and 3'-CTGCTATGTGGC-3'. The resulting 392 bp long PCR product was cloned in pGEM-T-Easy vector and the new vector designated pSC6. A 387 bp Xhol fragment from plasmid pSC3 was ligated with pSC6 previously cut with EcoRV, and the new plasmid was designated as pSC6B. The desired fragment of the altered BlaSC3 gene was excised using NcoI and XhoI and cloned in pET28a previously cut with the same two enzymes. The resulting expression plasmid was designated pSC6B. This vector served as the inactive BlaSC3 scaffold for the second generation dodecan experiments.

Construction of Plasmids pSC7, pSC8, and pSC9

The generation of these plasmids was carried out using the DC-ligation method described previously except that 2 μl (~250 ng) of SnaBI-cut and dephosphorylated plasmid pSC2 was added to each tube containing the DC-ligated products. The E. coli DH5α transformants were plated on increasing amounts of cefotaxime, in addition to their selection on ampicillin and kanamycin. Following on, the three plasmids, pSC7, pSC8, and pSC9 were generated and sequenced. The genes harbored in these plasmids were designated as BlaSC7, BlaSC8, and BlaSC9.

Site-Directed Mutagenesis

The synthetically synthesized and lysine residues of TEM-1, BlaSC3, and BlaSC7 were changed to alanine using the primer pairs 5'-GTTTCC AATATGATGAACTTTTAAAGTTCTGCG-3' and 5'-GCAAGAATTTAA AATGTCGATCATGGAAAAC-3', and 5'-ATGATGACACTTTTTG GGTCTTGCGATTGCGG-3' and 5'-GGGCTATAAGGCGAAGGCAA-3' respectively. The altered genes were sequenced and the changes validated.

Determination of MIC Values

MICs were determined by microdilution method as well as by plating cells on LB agar plates containing varying amounts of antibiotic. In the former method, a bacterial inoculum of ~10^6 cfu per tube was grown in the presence of 2-fold increases of each antibiotic used. The results were interpreted according to the guidelines provided [30]. In the latter method, procedures were followed as published previously, especially for determining the MIC for cefotaxime [5]. The plates or growing cultures were examined after 18 hr incubation at 37°C. The lowest concentration of antibiotic that inhibited growth was adjudged as the MIC.

Purification of TEM-1, BlaSC3, and BlaSC7 Proteins

6 liters of LB media containing 200 μg/ml of ampicillin and 50 μg/ml of kanamycin were inoculated with overnight cultures of DH5α strains harboring the plasmids pSC1, pSC3, and pSC7. All cultures were grown for 16 hr at 37°C under shaking. Cells were harvested by centrifugation, washed twice, and resuspended in buffer A (50 mM Tris-HCl, pH 8.0, 1 mM EDTA, 20% w/v sucrose) and purified using the osmotic-shock procedure [31]. The supernatant was collected and applied onto a 5 ml Ni²⁺-NTA column that had been equilibrated with buffer B (200 mM Tris-HCl, pH 8.0, 1 mM EDTA, 150 mM NaCl). The column was extensively washed with buffer B and eluted with two bed volumes of 250 mM imidazole in buffer B. Fractions displaying β-lactamase activity against nitrocefin were collected and pooled. The presence of β-lactamases was also detected by 15% SDS-PAGE. The N114-NTA-purified material was concentrated using YM-5 concentrators (Vivascience, UK). The purified proteins were pooled and dialyzed against buffer C (50 mM phosphate, pH 7.0). All proteins were stored at 4°C until further use.

Determination of Kinetic Parameters

Kinetic measurements were performed on a Shimadzu UV-1601 double-beam spectrophotometer using a 1.0 cm path-length cuvette at room temperature in 50 mM phosphate buffer (pH 7.0) using enzyme concentrations of either 1 or 5 nM. The rate constants were obtained for the following antibiotics: ampicillin (232 nm, kcat = 820 M⁻¹ s⁻¹), nitrocefin (486 nm, kcat = 20,500 M⁻¹ s⁻¹), and cefotaxime (264 nm, kcat = 6,900 M⁻¹ s⁻¹). All experiments were recorded in triplicate using six to seven substrate concentrations bracketing the Km. Kinetic parameters were determined by fitting obtained values to the differential form of the michaelis-Menten equation using unweighted nonlinear least squares. The data was analyzed by Graphpad-PRISM and by E-Z-FIT software programs.

Acknowledgments

We thank Dr. Marcel Lubér and his team at Microsynth, Switzerland for DNA sequencing and oligonucleotide synthesis and Dr. S. Pascal, University of Rochester, Rochester, NY for the kind gift of MBP plasmids HMBP-3C and HMBP-3C-Pro. The help and support of RGP Group members during the course of this work is gratefully acknowledged. S.C. would like to thank UGC for financial support.

Received: June 26, 2003
Revised: July 29, 2003
Accepted: July 30, 2003
Published: October 17, 2003

References