doi:10.1016/j.jmb.2009.02.022 *J. Mol. Biol.* (2009) **387**, 949–964

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Protein Design through Systematic Catalytic Loop Exchange in the $(\beta/\alpha)_8$ Fold

Adrián Ochoa-Leyva, Xavier Soberón*, Filiberto Sánchez, Martha Argüello, Gabriela Montero-Morán and Gloria Saab-Rincón*

Departamento de Ingeniería Celular y Biocatálisis, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Apartado Postal 510-3, Cuernavaca, Morelos 62271, México

Received 18 October 2008; received in revised form 2 February 2009; accepted 10 February 2009
Available online 20 February 2009

Protein engineering by directed evolution has proven effective in achieving various functional modifications, but the well-established protocols for the introduction of variability, typically limited to random point mutations, seriously restrict the scope of the approach. In an attempt to overcome this limitation, we sought to explore variant libraries with richer diversity at regions recognized as functionally important through an exchange of natural components, thus combining design with combinatorial diversity. With this approach, we expected to maintain interactions important for protein stability while directing the introduction of variability to areas important for catalysis.

Our strategy consisted in loop exchange over a $(\beta/\alpha)_8$ fold. Phosphoribosylanthranilate isomerase was chosen as scaffold, and we investigated its tolerance to loop exchange by fusing variant libraries to the chloramphenicol acetyl transferase coding gene as an *in vivo* folding reporter. We replaced loops 2, 4, and 6 of phosphoribosylanthranilate isomerase with loops of varied types and sizes from enzymes sharing the same fold.

To allow for a better structural fit, saturation mutagenesis was adopted at two amino acid positions preceding the exchanged loop. Our results showed that 30% to 90% of the generated mutants in the different libraries were folded. Some variants were selected for further characterization after removal of chloramphenicol acetyl transferase gene, and their stability was studied by circular dichroism and fluorescence spectroscopy. The sequences of 545 clones show that the introduction of variability at "hinges" connecting the loops with the scaffold exhibited a noticeable effect on the appearance of folded proteins. Also, we observed that each position accepted foreign loops of different sizes and sequences.

We believe our work provides the basis of a general method of exchanging variably sized loops within the $(\beta/\alpha)_8$ fold, affording a novel starting point for the screening of novel activities as well as modest diversions from an original activity.

© 2009 Elsevier Ltd. All rights reserved.

Edited by C. R. Matthews

Keywords: loop grafting; PRAI; directed evolution; CAT; folding reporters

**Corresponding authors.* E-mail addresses: soberon@ibt.unam.mx; gsaab@ibt.unam.mx.

Present address: G. Montero-Morán, Department of Nutritional Sciences, Rutgers, The State University of New Jersey, New Brunswick, NJ 08901, USA.

Abbreviations used: PRAI, phosphoribosylanthranilate isomerase; ePRAI, monofunctional version of E. coli PRAI; ADA, adenosine deaminase; FBPA, fructose-bisphosphate aldolase; PBGS, porphobilinogen synthase; DHDPS, dihydrodipicolinate synthase; αTS, tryptophan synthase α chain; TPS, thiamine phosphate synthase; Ure, urease; MR, mandelate racemase; dsDNA, double-stranded DNA; CAT, chloramphenicol acetyl transferase; Cm, chloramphenicol; WT, wild type; EDTA, ethylenediaminetetraacetic acid; BME, β-mercaptoethanol.

0022-2836/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.

Petitioner Merck, Ex. 1144, p. 949
Merck Sharp & Dohme LLC v. Halozyme Inc.
PGR2025-00039

Introduction

It is recognized that the evolution of proteins has proceeded through processes where gene duplication and point mutations have played a major role.[1] This is especially true for the relatively recent evolutionary times, but this simple process is incapable of explaining the first appearance of primordial functional proteins and many of the more drastic modifications apparent in extant protein families and superfamilies. These additional modifications suggest processes involving modular exchange between segments of unrelated proteins.[2–4]

Understanding these natural processes is a very valuable asset to the engineering of proteins, in particular when directed evolution techniques are employed.[5] For the study of the most fundamental questions, it is advantageous to rely on suitable protein architectures, several of which have become model systems due to their versatility and/or potentially wider relevance. Proteins such as those possessing the immunoglobulin fold[6] or the $(\beta/\alpha)_8$ fold[7] can be viewed as "scaffolds" for engineering due to their apparently modular architecture and evolutionarily variable loops. Combining protein design and directed evolution with scaffold-based protein libraries provides an excellent route to engineering new protein functions. The utility of such libraries depends on how tolerant the scaffolds are to randomization, because the selected variants must remain folded and soluble. The $(\beta/\alpha)_8$ or TIM barrel scaffold should be an ideal starting point for engineering novel enzymatic activities by rational design and directed evolution.

TIM barrel proteins are the most common fold among protein catalysts, constituting approximately 10% of all known enzyme structures. They are catalytically versatile: five of the six enzyme classes contain members with this fold.[3] The barrel structure is composed of eight parallel β-strands in the interior of the protein, surrounded by eight α-helices. The pseudo symmetry of the fold is simultaneously an attractive feature for engineering and a suggestion of the modular construction and evolution of this protein architecture.[8,9] The residues at the active sites of all known TIM barrel enzymes are located on the catalytic face of the barrel, which is composed of the C-terminal end of the β-strands and residues coming from the β/α loops (the loops that link β-strands with α-helices). In contrast, the remainder of the fold, including the opposite face of the barrel, is important for conformational stability.[10] It has long been hypothesized that varying the residues of the active site, including the loops, might alter enzymatic function without affecting the stability of the fold.[11–14] Indeed, it has been shown that engineering of protein loops can afford functional changes, notably substrate specificity. This has been demonstrated for proteins with TIM barrel[15] and other folds.[16–18] Furthermore, recent work demonstrated that more profound changes in the enzymatic function are possible through loop swapping as a component of the directed evolution strategy.[19–21]

On the other hand, the design of new enzymes by generating variability through point mutation of an existing gene has shown modest progress.[22] As stated above, for the natural evolution of proteins, this is likely because more drastic changes are needed. Following a related idea, it has been suggested that new enzymes can be generated by recombining different segments of modern proteins, and the concept has been shown to be successful.[4,23]

Here we describe Systematic Catalytic Loop Exchange (SCLE), a new strategy for exchanging loops on a TIM barrel fold. As a scaffold, we chose a variant of phosphoribosylanthranilate isomerase (PRAI-LoxP) from *Escherichia coli*, a monomeric enzyme where we have previously done significant work.[24,25] This gene has an insertion coding for a cre-lox recognition site in the loop linking α-helix 4 and β-strand 5, as previously described.[24] The purpose of using this gene is to carry out *in vivo* recombination of generated variants in future experiments. This gene will be referred to as *trpF*-loxP and the respective protein PRAI-LoxP throughout the rest of the article. We set out to recombine loops compatible with this scaffold that have already been explored and optimized (albeit in different contexts) by natural evolution, hoping that the resulting libraries preserve a high enough percentage of folded proteins to be screened for function. In order to design specific modification sites, we investigated the tolerance of PRAI-LoxP to loop exchange, as estimated by stable folding *in vivo*. We decided to initially generate variability toward 3 of the 8 (β/α) loops, where there is higher participation of their residues in substrate binding and catalysis.[3] In this communication, we describe the exchange of 14 different (β/α) loops within the PRAI-LoxP structure. The exchanged loops belong to eight different proteins sharing the same fold but have diverse functions. Functional diversity will be explored in the future after *in vivo* recombination of the different loop libraries combined with random mutagenesis and additional designed mutations.

Results

Design of the loop exchange in PRAI

As an ultimate goal, the present work aims at contributing to the generation of *de novo* protein activities. The scaffold with the cre-lox insertion is derived from a modified monofunctional version of *E. coli* PRAI (WT ePRAI), in which part of the gene coding for the bifunctional IGPS-PRAI protein has been excised to express the PRAI gene separately.[26] The three-dimensional structure of the *E. coli* enzyme [Protein Data Bank (PDB) entry 1PII] was used to identify the sites for loop exchange. Our exchange strategy focused specifically on loops 2, 4, and 6, where residues identified as binding and catalytic sites are more abundant in a set of TIM barrel proteins analyzed.[3] The loop-donor proteins were selected under the following criteria: they should

Table 1. Characteristics of the exchanged loops in the PRAI-LoxP scaffold

Enzyme (EC number)	Position of the exchanged loop[a]	Sequence of the loop	Lengths
MR (EC 5.1.2.2)	β/α loop 2	GYPAL	5
FBPA (EC 4.1.2.13)	β/α loop 2	SNGGASFIAGKGVKSDVPQ	19
Ure (EC 3.5.1.5)	β/α loop 2	GGTGPAAGTHATTCTPG	17
PRAI WT (5.3.1.24)	β/α loop 2	VATSPRCVN[b]	9
DHDPS (EC 4.2.1.52)	β/α loop 4	PYYNRPS	7
TPS (EC 2.5.1.3)	β/α loop 4	LGQEDLH	7
MR (EC 5.1.2.2)	β/α loop 4	EPTLEHD	7
FBPA (EC 4.1.2.13)	β/α loop 4	DLSEES	6
αTS (EC 4.2.1.20)	β/α loop 4	DVPVQQS	7
Ure (EC 3.5.1.5)	β/α loop 4	EDWGAT	6
ADA (EC 3.5.4.4)	β/α loop 4	GDELGFPGSLF	11
PBGS (EC 4.2.1.24)	β/α loop 4	AAMDG	5
PRAI WT (5.3.1.24)	β/α loop 4	GNEE[b]	4
DHDPS (EC 4.2.1.52)	β/α loop 6	TGNL	4
PBGS (EC 4.2.1.24)	β/α loop 6	PAGAY	5
αTS (EC 4.2.1.20)	β/α loop 6	SRAGVTGAENRAALP	15
PRAI WT (5.3.1.24)	β/α loop 6	NGQGGSGQRFD[b]	11

[a] Structural position in which the WT loop was replaced for the new loop in the PRAI-LoxP scaffold.
[b] Sequence corresponding to the WT loops from PRAI-LoxP.

be TIM barrels, they should comprise diverse functionalities and substrate kinds, and they should preferably have a monomeric state; if this last condition was not met, their oligomerization interface should not involve the catalytic site (i.e., the top of the barrel). Based on these criteria, the selected loop donors were the proteins adenosine deaminase (ADA; PDB code 2ada), fructose-bisphosphate aldolase (FBPA; PDB code 1dos), porphobilinogen synthase (PBGS; PDB code 1l6s), dihydrodipicolinate synthase (DHDPS; PDB code 1dhp), tryptophan synthase α chain (αTS; PDB code 1xc4), thiamine phosphate synthase (TPS; PDB code 1xi3), urease (Ure), and mandelate racemase (MR) from *E. coli*. The structures from Ure and MR were modeled from the homologous proteins from *Klebsiella aerogenes* (PDB code 2kau) and *Pseudomonas putida* (PDB code 1mdr), respectively. The amino acid sequences of loops 2, 4, and 6 of these proteins are shown in Table 1. The loops were inserted at the zone delimited by the last position of the previous β-strand and the first residue of the next α-helix, as shown in Fig. 1a. For loop 2, the exchanged fragment is located between residues Val31 and Asn39; for loop 4, between residues Gly82 and Glu85; and finally, for loop 6, between Asn127 and Arg135 (numbering according to gene reported by Kirschner *et al.*[26]). In order to have a general strategy to exchange loops occupying topological positions different from their original site in the donor protein, we designed "general" double-

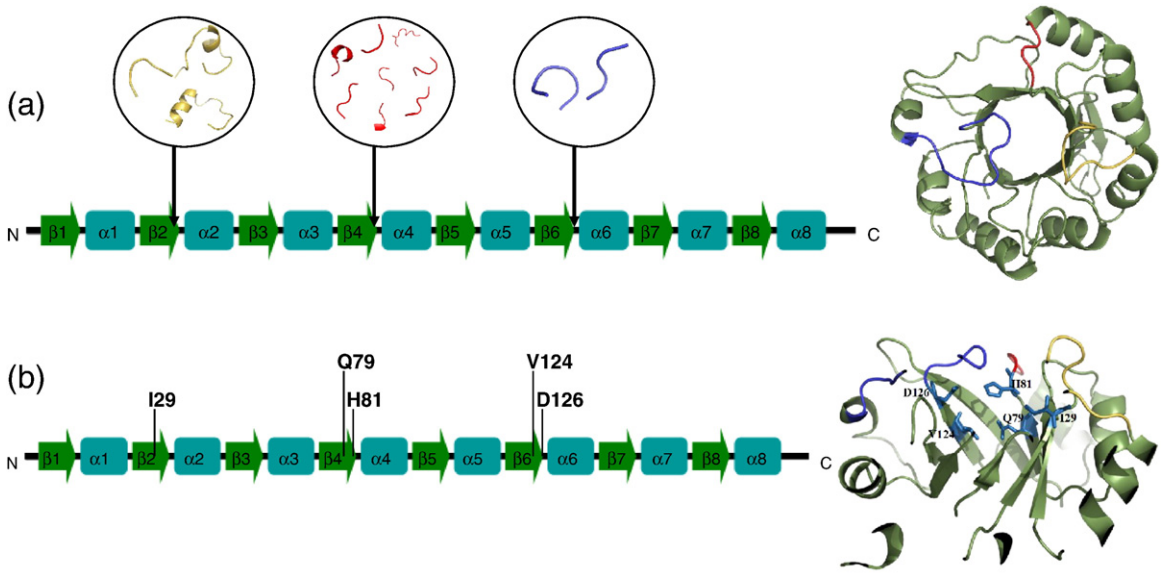

Fig. 1. Secondary-structure elements and tertiary structure (PDB code 1pii) of WT PRAI (ribbon diagram showing a view from the top of the central β-barrel). (a) The loops 2, 4, and 6 are shown in yellow, red, and blue, respectively. The circular insets show the loops to exchange in the PRAI-LoxP scaffold at the three structural positions (β/α 2, β/α 4, and β/α 6). (b) The mutated positions are indicated and a side-view slice of the TIM barrel structure of the PRAI scaffold is shown on the right.

stranded DNA (dsDNA) connectors that would direct the insertion to any of the particular locations explored in PRAI-LoxP (i.e., loop 2, loop 4, or loop 6). There were two sets of dsDNA connectors, the NH2 connector that coded for β2, β4, and β6 (on the side 5′ from the loop) and the CO connector that coded for α2, α4, and α6 (on the side 3′ from the loop). These connectors had different single base overhangs at both ends to allow directional ligation of the incoming dsDNA coding for the loop with the resulting replacement of the original loop present in PRAI, as shown in Fig. 2a. A schematic representation of the strategy used to exchange the loops in the gene of PRAI-LoxP is shown in Fig. 2a and b.

A change of function would likely require, in addition to the loop transplant, the substitution of the last residues in the β-strands, as these are also typically part of the binding and catalytic sites.[27] Therefore, another component in the design of the NH2 connectors was the introduction of variability in the final positions of the β-strand preceding the loop to be exchanged and pointing inward into the barrel (where catalysis takes place), as shown in Fig. 1b. In the case of the NH2 connector for loop 2, the last residue in the β-strand points toward the external shell; therefore, variability was introduced in the previous position, Ile29. In the cases of loops 4 and 6, the residues preceding the loops, His81 and Asp126, respectively, point toward the interior of the barrel; therefore, we decided to introduce variability in these positions as well as in the previous residues also pointing to the interior of the barrel, that is,

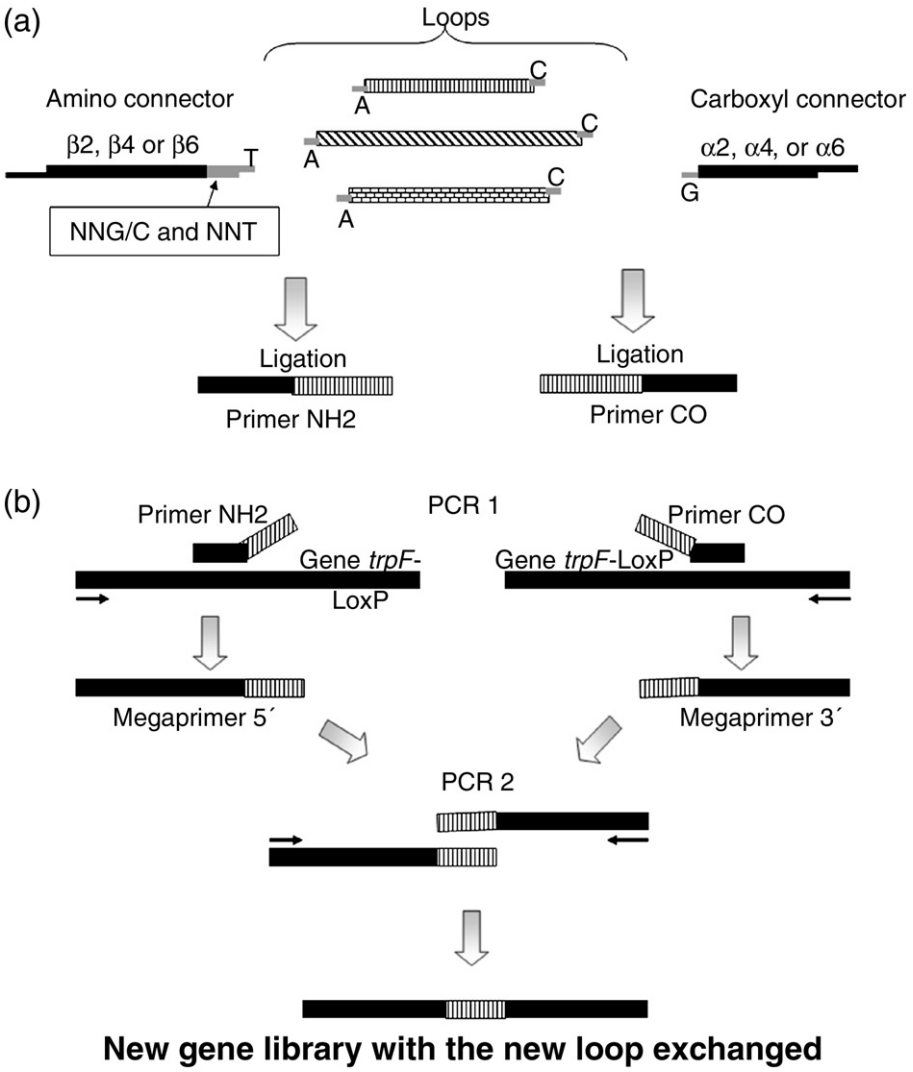

Fig. 2. Strategy for systematic loop exchange in the trpF-loxP gene. (a) dsDNA coding for each loop is ligated with the amino and the carboxyl connector independently, yielding two primers. The amino connectors (that prime to either β2, β4, or β6 coding sequences), the carboxyl connectors (prime to either α2, α4, or α6 coding sequences) and the "loops" (dsDNA coding for the diverse loops) all had different single base overhang at both ends to allow directional ligation. The mutated sites shown in the box are encoded in the oligonucleotides that constitute the amino connectors. (b) The resulting primers from the ligation described in (a) are used in combination with the corresponding end primers (arrowed lines) to amplify both the 5′ and the 3′part of the gene, respectively, by PCR using the *trpF*-loxP gene as template. The only common sequence between the two obtained megaprimers is the coding sequence for the exchanged loop; these are used in a second-round PCR to obtain the final gene with exchanged loop sequence by PCR primer extension.

Gln79 and Val124, respectively. The degenerate codon NNG/C was used when no other restrictions were present, allowing exploration of all 20 amino acids; however, the requirement of fixing the last base to T in the last codon of the NH2 connector restricts the variability of this position, preventing the exploration of residues Met, Gln, Lys, Glu, and Trp. Similarly, the CO connectors had a G base overhanging in the noncoding strand, designed to anneal with the overhanging C in the coding strand of the guest loop. This made unavoidable the introduction of a size conservative mutation from Gln to His in one of the exchanged loops, specifically loop 2 from FBPA from *E. coli*.

Folding selection method

The selection of folded proteins through the use of folding reporter genes has been previously documented.[28,29] Briefly, in the system we chose, the gene or family of genes to be screened for folding is fused to a reporter gene, in this case the gene that codes for chloramphenicol acetyl transferase (CAT) and provides resistance to the antibiotic chloramphenicol (Cm). A protein capable of folding will allow the correct folding of CAT and, therefore, the resultant clones will be resistant to Cm, while no growth will be observed for those clones bearing a folding-deficient fusion gene product.

To validate the selection method, we made a construction of the wild-type (WT) PRAI gene fused to CAT gene as a positive control and, as negative controls, two proteins proven to have poor folding: a circular permutation of PRAI previously described[30] and an engineered antibody provided by Dr. J. Osuna (unpublished result). Additionally, a construction of WT PRAI with two stop codons between PRAI and CAT was also engineered. The resultant plasmids were transformed in four *E. coli* strains, XL1-Blue, JM101Δ*trpF*, JMB9, and MC1061Δ*thiE* and plated on Luria broth (LB) medium plates containing ampicillin (200 μg/ml) and Cm at concentrations ranging from 0 to 100 μg/ml. We found that JM101, XL1-Blue, and JMB9 strains show lower resistance to Cm. Additionally, as also noticed by others,[31] the JM101 and XL1-Blue gave many false-positive results. In contrast, selection with the strain MC1061Δ*thiE* gave consistent results. None of the negative controls, including the fusion of WT PRAI with two stop codons, showed growth at Cm concentrations above 15 μg/ml (Table 2). The positive control, on the other hand, showed a higher proportion of colonies under selective conditions. In this case, we observed that Cm concentration as low as 20 μg/ml is enough to discriminate unfolded variants while maintaining high viability of cells bearing viable genes.

Assessment of folding in selected clones

To further confirm the accuracy of positive clone scorings, Western blot analysis of the cell extract of 8 clones selected in Cm plates from each of nine libraries was carried out. As observed in Fig. 3, of 72 clones selected, all but one expressed variable but clearly detectable amounts of fusion protein in the soluble fraction. This demonstrates that the method is reliable for the identification of true positive folded proteins.

Ratio of folded proteins in the generated libraries

Using the conditions described in the folding selection method section, we evaluated the fraction of folded proteins in each loop-exchanged library. Figure 4 shows the percentage of fusion proteins conferring Cm resistance in libraries generated on loop 2, loop 4, and loop 6, respectively. We then took the survival ratio of a population of cells transformed with the PRAI-LoxP-CAT fusion as our higher limit and normalized the survival rate observed for the different libraries with this value. The percentage of folded proteins for the three different libraries generated in loop 2 had a very similar value, around 30–40%, and in fact they were the lowest values found in the whole set of constructions. Exchanges at loop 4, where more libraries were generated, afforded a higher viability, from 40% to approximately 70% of folded peptides, depending on the incoming guest loop. Finally, in the case of loop 6, for which only three libraries were constructed, we observed the highest viability, ranging from 50% to 90%.

In order to further investigate the properties of these clones, 10 colonies were isolated from the plate

Table 2. Resistance toward Cm conferred by different CAT fusions when expressed in MC1061Δ*thiE*

		Concentration of Cm (μg/ml)[b]				
Variant	Description[a]	0	5	10	15	20
pDan5	No insert	+++	—	—	—	—
Vh-CAT	Antibody fused to CAT	+++	+++	+	—	—
Per3-β4/α3-CAT	Permutation of PRAI fused to CAT	+++	+++	+	—	—
PRAI-Stop-CAT	PRAI WT with the insertion of two stop codons between their and CAT fusion.	+++	—	—	—	—
PRAI-CAT	PRAI WT fused to CAT	+++	+++	+++	+++	+++

Resistance was estimated by the growth rate.
[a] Description of the genes from the different variants.
[b] Cells were grown in LB medium plates supplemented with ampicillin and with different concentrations of Cm. The number of colonies appearing in the presence of Cm was counted after 18 h and compared with that without Cm. The percentages of colonies that survived to different concentrations of Cm relative to the control without it are as follows: +++, 100%; ++, 60%; +, ≤10%.

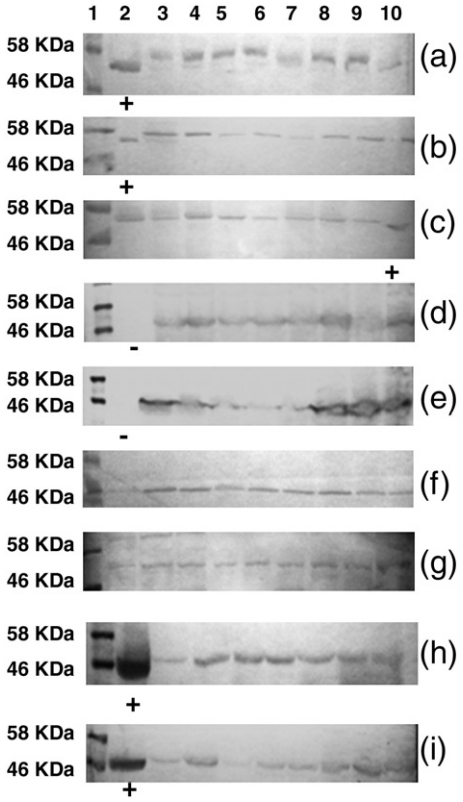

Fig. 3. Western blot analysis from soluble extract with an antibody rose against CAT protein. The expected molecular mass of the proteins fused to CAT is approximately 50 kDa. Lane 1 shows a molecular mass marker, lane 2 or lane 10 shows either a positive or negative control indicated by a + or − symbol under the lane, according to the case. Lanes 2–10 show the soluble fraction of eight independent colonies chosen from LB/ampicillin/Cm plates from (a) loop 2 FBPA, (b) loop 6 αTS, (c) loop 4 Ure, (d) loop 4 MR, (e) loop 6 DHDPS, (f) loop 2 MR, (g) loop 4 αTS, (h) loop 4 DHDPS, and (i) loop 4 FBPA libraries.

without selection of representative libraries of each loop comprising the broad range of survival rates observed, that is, loop 2 from MR, loop 4 from Ure, and loop 6 from PBGS. Figure 5a–c shows the Western blot of the cell extract soluble fraction of the selected clones with an antibody rose against CAT. It is important to note that the results of percentage of clones showing the presence of fusion protein product in the gel summarized in Table 3 are in good

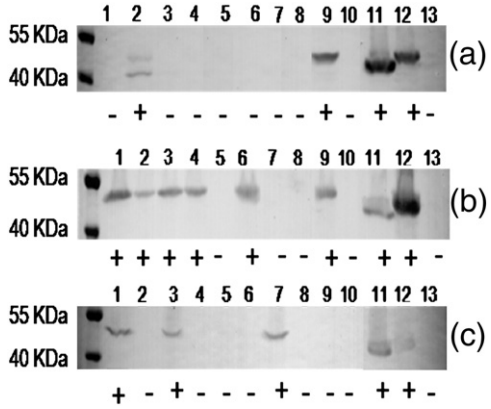

Fig. 5. Western blot analysis from soluble extract from colonies isolated without selective pressure using an antibody rose against CAT protein. The expected molecular mass of the proteins fused to CAT is approximately 50 kDa. 1–10, Western blots of the soluble fractions of 10 clones randomly chosen from the LB/ampicillin plates from each loop library. 11, WT PRAI; 12, clone selected in presence of Cm (20 μg/ml); 13, plasmid pDAN5. The isolated plasmid DNAs from analyzed clones were retransformed in the strain MC1061Δ*thiE* and plated on LB supplemented with ampicillin and Cm. Those that conferred Cm resistance are shown as (+) and the ones that did not are shown as (−). (a) Loop 2 library from MR. (b) Loop 4 library from Ure. (c) Loop 6 library from PBGS.

agreement with those found by Cm resistance *in vivo*. Plasmid DNA was retransformed in MC1061Δ*thiE E. coli* competent cells. There is an invariable correspondence between growth in Cm and the presence of fusion protein in the Western experiment, indicating that Cm resistance is an adequate criterion to evaluate the ratio of folded protein variants (Fig. 5a–c). All these plasmids were sequenced and, as expected, none of the sequences found in clones reported as nonfolded by Western blot analysis are present in the clones reported as folded by the genetic selection method.

Sequence analysis

Sequence analysis of a few clones (about 20 from each selective and nonselective condition) was carried out from each library to verify if the insertions were introduced as planned, as well as to find out if

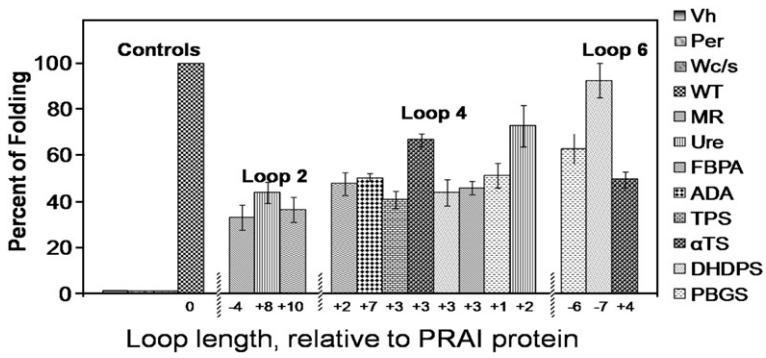

Fig. 4. Percentage of folded proteins fused to CAT was calculated as described in Materials and Methods. Values and error bars represent the average and standard deviation for each library analyzed. Controls are designated as follows: Vh, antibody; Per, circular permutation of PRAI; Wc/s, PRAI-LoxP with stop codons before the fusion to CAT; WT, PRAI-LoxP. The differences in length of loops of the libraries, relative to the PRAI WT protein, are shown in the *x*-axis.

Table 3. Percentage of folding for loop libraries by Western blot analysis

Library[a]	Percentage of mutants positive to fusion to CAT[b]
Loop 2 of MR	28
Loop 4 of Ure	60
Loop 6 of PBGS	43

[a] Name of the library analyzed by Western blot.
[b] The percentage of folding was calculated by dividing the number of clones that present fusion with CAT by the total of clones analyzed. This value was normalized by the sequence found as WT or with an aberrant construction in the 10 clones.

there were amino acid preferences at the end of the β-strand preceding the inserted loops in the folded proteins. From the sequence analysis of 545 clones selected and nonselected (354,250 pb), we conclude that between 70% and 100% of the generated mutants were well constructed, demonstrating the correct loop replacement and the presence of variability in mutagenized positions in the different libraries.

Sequence analysis of clones under selective pressure

There is a true biased representation of some amino acids as a result of design. After correction by normalizing the observed frequencies of each amino acid in selected *versus* nonselected clones for each library, we still observe overrepresentation of some amino acids in selected clones. However, the size of the sample is very small to draw conclusions.

Libraries of loop 2

Some amino acids at the variable positions had either a significantly higher or lower frequency than expected from chance. Some loops show a strong bias for a specific amino acid, as is the case for the insertion of loop 2 from Ure, which shows a positive selection for both serine and alanine at position 29, as well as a negative selection for threonine. Loop 2 from FBPA, on the other hand, showed a strong positive selection for threonine and an equally strong negative selection for leucine at that position. Finally, the exchange of loop 2 with the library representing the loop 2 from MR had its highest preference for arginine and a negative selection against tryptophan.

Libraries of loop 4

The same analysis was performed for the two randomized residues upon replacement of loop 4 in PRAI-LoxP. The insertion of loop 4 from TPS shows a strong preference for arginine at position 79, while position 81 shows negative selection against proline. The replacements of the rest of loops 4 do not display significant preference for any particular residue at position 79, with the exception of loops from MR and DHDPS, which show negative selection against isoleucine and glycine, respectively. Position 81 on the other hand, shows a small but significant preference for serine and threonine when the loop from PBGS is introduced and for serine when the guest loop is from MR. Interestingly, serine is counterselected when the inserted loops 4 derive from the αTS and from DHDPS. Negative selection was also observed against glycine when loop 4 was replaced by that from ADA. These results indicate that some residues preceding the exchanged sequence are more compatible with the new loop than with others. Perhaps more important was the observation of some correlated preferences. With the size of the sample analyzed, the probability of finding the same pair of residues twice is quite small. In five of the eight libraries with loop 4 exchanged, clones with repeated pairs of residues were found among the approximately 20 selected clones that were sequenced. For instance, the pairs S–T, D–T, and N–P at positions 79–81 were found twice when loop 4 from Ure replaced the PRAI loop. In the library with replacement of loop 4 from PBGS, the combination H–V was selected twice, while in that from ADA, the pair S–R was found. The pair R–V was selected in the library from the DHDPS loop and the loop from MR showed two repeated pairs, T–V and P–N. These combinations were not present in the sample of randomly chosen clones, so the possibility of overrepresentation of these amino acids in the library is unlikely. When loop 4 of αTS, of TPS, and of FBPA are exchanged, there is no significant preference for specific combinations at positions 79 and 81.

Libraries of loop 6

In the three libraries constructed, there were some significant selections at position 124. In the case of the PBGS loop, leucine was selected, while lysine was counterselected; in the case of the guest loop from DHDPS, there was a small but significant preference for cysteine, while for the guest loop from the α subunit of αTS, there was a preference for threonine. In both DHDPS and αTS guest loops, there was a negative selection for serine at position 124. On the other hand, no preference was evident for any residue at position 126, but counterselection was observed for serine in the case of the DHDPS guest loop and for threonine in the case of the αTS guest loop. Just as observed in the libraries from loop 4, there were combinations at the two positions that were selected twice. The combination T–C in both PBGS and DHDPS libraries and the combination T–R in the αTS library were observed twice in folded clones.

Structural CD and fluorescence spectroscopic analysis of some selected clones

Considering the drastic changes introduced in the PRAI-LoxP sequence, the relatively high percentages of folded proteins that were obtained according to the Cm resistance test are somewhat surprising. To find out if some selected mutants were folded and stable in the absence of CAT, we further investigated

their structural properties. Figure 6 shows the far UV–CD spectra of the purified proteins overlapped with the spectrum of the PRAI-LoxP (the parental gene product used for the constructions).[24] Some subtle changes in the spectra could be observed, suggesting some structural differences. The secondary-structure predictions based on the CD spectra were carried out. Albeit these results have to be taken with reservation, it is interesting to note that loop 4 from the αTS showed a calculated content of secondary-structure elements closest to that of PRAI-LoxP protein (14% α-helix, 38% β-strand), while variants with loop 2 and loop 6 exchanged show an increase in the α-helical content (28%). Loop 2 from FBPA forms a helical structure in the context of its parental protein. The increase in signals at 208 and 222 nm suggests that this loop may also be helical in the context of the PRAI-LoxP scaffold. The increase in helical structure upon substitution of loop 6 by the corresponding loop from PBGS is harder to explain, since in this case it is a small loop of only five residues compared to the 11 residues in loop 6 from PRAI-LoxP. However, the shortening of this loop may have contributed to the propagation or better formation of α-helix 6, which shows only one turn in the crystal structure of the WT protein[32] and/or in the configuration of the adjacent α-helix 5, which is not a well-structured helix in the native protein. Nonetheless, it is important to highlight that all the spectra are characteristic of folded proteins. Figure 7 shows the fluorescence emission spectra of the selected variants under native and denaturing conditions. The blue shift observed under native conditions for all the proteins is characteristic of folded proteins, and their red shift in urea indicates the exposure of the fluorophores to the aqueous environment upon loss of tertiary structure. The fluorescence spectrum of WT PRAI was recorded for comparison (Fig. 7a). The center of mass of the spectra under denaturing conditions is the same for all the proteins (around 366), while under native conditions all the PRAI-LoxP mutants show a blue shift higher than that of the WT ePRAI protein

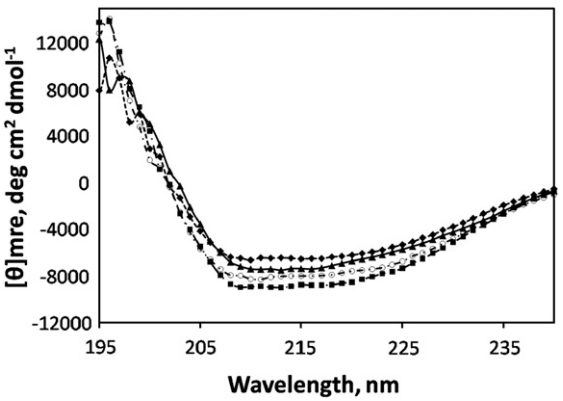

Fig. 6. Far-UV CD spectra of representative constructs engineered in this study. Triangles, PRAI-LoxP; open circles, Mut_L2_FBPA; diamonds, Mut_L4_αTS; squares, Mut_L6_PBGS.

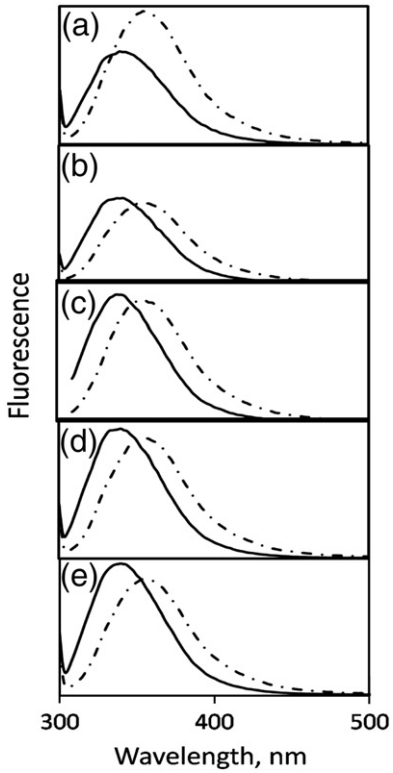

Fig. 7. Fluorescence emission spectra of representative constructs engineered in this study under native conditions (continuous line) and at 9 M urea (dashed line). (a) WT ePRAI, (b) PRAI-LoxP, (c) Mutant_L2_FBPA, (d) Mutant_L4_αTS, and (e) Mutant_L6_PBGS.

(around 349 nm *versus* 352 nm for the WT enzyme). Interestingly, the WT PRAI spectrum shows an increment of fluorescence upon unfolding, while the rest of the variants show a slight decrease in fluorescence. This is indicative of less fluorescence quenching among the fluorophores in the folded state of the variants, but the same behavior was observed for the parental PRAI-LoxP protein, indicating that this difference is related to the introduction of the lox-P loop at the bottom of the barrel and not to the exchanged loops at the catalytic face.

Figure 8 shows the loss of secondary structure of the variants in 8 M urea. It is noteworthy that there is some residual secondary structure in the proteins after incubation for 2 h at denaturing conditions. This is in agreement with the high stability of the secondary structure observed for the proteins subjected to thermal unfolding (data not shown). This behavior was shared by all the variants, including the parental PRAI-LoxP protein. Therefore, we attributed it to the presence of the insertion coded by the cre-lox recognition sequence.

Oligomeric state of the mutants

The oligomeric state of the mutants was investigated by size-exclusion chromatography using three different initial protein concentrations ranging

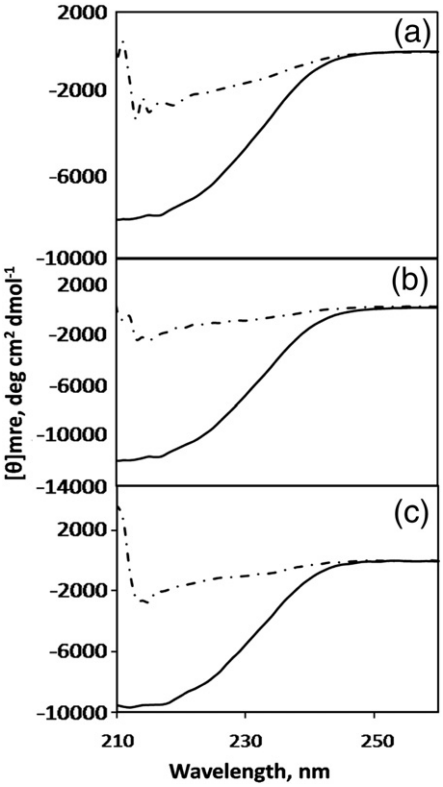

Fig. 8. Far-UV CD spectra of representative constructs engineered in this study under native conditions (continuous line) and at 9 M urea (dashed line). (a) Mutant_L2_FBPA, (b) Mutant_L4_ αTS, and (c) Mutant L6 PBGS.

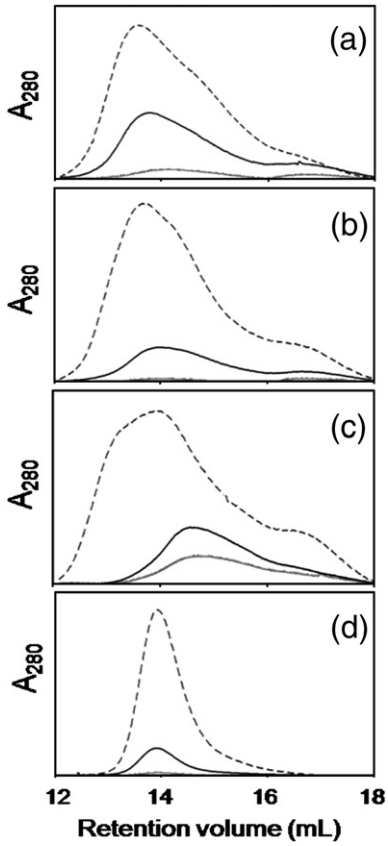

Fig. 9. Analytical gel-filtration chromatograms of PRAI-LoxP and its mutants using a Superose HR12 column. The initial protein concentrations were 2 μM (thin lines), 20 μM (thick lines), and 100 μM (broken lines). (a) PRAI-LoxP, (b) Mutant_L2_FBPA, (c) Mutant_L4_αTS, and (d) Mutant_L6_PBGS.

from 2 to 100 μM. Table 4 shows the apparent molecular masses of the variants determined by their elution profile. As shown in Fig. 9, all the proteins, including the PRAI-LoxP parental protein, are present as a mixture of monomer–dimer as judged by their elution profile. This behavior has also been observed for the WT ePRAI.[33] The variant with loop 6 from PBGS shows dimer as the dominant species (Fig. 9d). In the other cases, the dimer seems to be in a fast equilibrium with higher oligomers when the protein concentration increases to 100 μM, as suggested by the increase in apparent molecular mass and the deformation of the peaks, especially for the variant with loop 4 from αTS (Fig. 9c).

Discussion

In this study, the PRAI-LoxP protein was chosen as a scaffold and its tolerance to loop exchange was investigated by fusing the libraries to the CAT coding gene as an *in vivo* folding reporter. When such fusion proteins are soluble, they provide

Table 4. Apparent molecular masses and deduced association states

Protein[a]	Calculated molecular mass (kDa)[b]	Apparent molecular mass (kDa)[c]			Deduced association state
		2.0 μM	20 μM	100 μM	
PRAI-LoxP	24.1	22.9 + 43.3	22.9 + 47.6	49.8	Monomer − dimer
Mut_L2_FBPA	25.0	23.8 + 46.5	23.9 + 46.7	24.3 + 49.7	Monomer − dimer
Mut_L4_αTS	24.3	40	40	23.9 + 46.4	Dimer − monomer
Mut_L6_PBGS	23.7	46.2	46.6	46.4	Dimer

[a] The proteins were dissolved in 20 mM potassium phosphate buffer (pH 7.6) and 300 mM KCl. Each protein carries a His-tag at its N-terminus and the protein concentrations were 2, 20, and 100 μM.
[b] The calculated molecular mass was theoretically calculated from the sequences of the respective proteins.
[c] Apparent molecular masses were calculated using the protein elution volumes and a calibration curve generated using elution volumes for proteins of known molecular mass.

resistance to Cm to their host cells. Our design allows the exchange of the loops in the same position or in other topological positions from the original protein in PRAI. As an ultimate goal, the present work aims at contributing to the generation of *de novo* protein activities by applying selection regimes to the folded proteins from the loop libraries. This approach allows the exploration of a more divergent sequence space and is not limited to point mutations. Also, it enables the introduction of variability to areas important for catalysis while maintaining a higher probability for a folded protein.

The behavior of the controls for folding selection allowed us to evaluate different *E. coli* strains for the selection of true positive clones and the identification of the proper Cm concentration to discriminate the unfolded variants while maintaining the viability of cell growth. As mentioned, JM101 and XL1-Blue strains gave many false-positive results.[31] On the other hand, the MC1061Δ*thiE* was the only strain tested with zero false positives. The MC1061 strain is a Str^r mutant that has a point mutation (K43R) within the *rpsL* gene (encoding ribosomal protein S12).[34] Mutation K43R has been suggested to stabilize the *ram* state by increasing the tRNA affinity in the A site of the ribosome.[35–37] The influence of translational processes on protein folding has been amply debated, and a recent study suggests its influence on the relative abundance of some proteins, including CAT.[35,38] These peculiarities may be related to the different expression behavior of these proteins in the strains we have tested. We propose that the change in the stability of the ribosome is probably the main difference between the MC1061 and the other strains that permit the proper expression of our fused proteins, but more experiments are needed to corroborate this hypothesis.

An important finding of this study is that each position accepted foreign loops of different sizes and sequences (most likely due to the allowance of hinge variability). Our data suggest that the sequence of the loop inserted has much less impact on structural adaptations than does the site of the insertion (Fig. 3). This has also been suggested for ubiquitin, where the structural effects of a loop insertion depend primarily on the site of the insertion and much less on the sequence of the insert.[39] A larger data set will be needed to draw more drastic conclusions regarding any significant tolerance difference between other alternative loop positions in the barrel scaffold. The relatively high tolerance to the diversity of sequences tested and the lengths of the inserted loops is in agreement with the general notion that the active-site end of the TIM barrel is inherently tolerant to variability.[10] The high proportion of chimeric proteins (30–90%) that remained foldable and relatively stable after the introduction of foreign loops also indicates the feasibility of our approach to afford useful libraries for subsequent functional screening.

It is clear from the data that the introduction of variability at the hinges connecting the loops with the scaffold has a noticeable effect for obtaining folded proteins. The distribution of amino acids found in randomized positions without selective pressure of the analyzed samples shows only the normal bias inherent to the combinations NNG/C and NNT in the oligonucleotides used for mutagenesis. Thus, the amino acid preferences observed with the introduction of each particular loop are indicative of the importance of the position adjacent to the loop for conformational fitness. The constraints imposed on these positions are likely to derive both from the set of interactions with the guest loop and from those corresponding to tightly packed environments within the core of the protein, as suggested by the lower viability observed in libraries from loop 2, where only one of the barrel-core positions was randomized. It is interesting to note that the amino acids with higher propensity at the hinge positions are dependent on the guest loop and always different from the original residue at the host protein. However, these mutated positions may have been selected for function rather than stability in the host protein because it is well documented that amino acids constituting the catalytic site may be in unstable conformations to favor catalysis.[40] A sequence alignment of all nonredundant PRAI proteins shows that the residues mutated are well conserved (I29 by 74%, Q79 by 96%, H81 by 91%, V124 by 80%, and D126 by 98%). Furthermore, computational analysis of the protein stability with ΔΔG Rosetta and Rosetta Design algorithms[41] shows that these mutated residues are destabilizing in the PRAI structure. These observations suggest a role that is more functional than structural for these sites. Since our selection method is based on stability, the appearance of different amino acids is to be expected. The small size of the sample and the different patterns found with various loops precludes a systematic analysis of the sequence preferences, but the fact that many sequences did not produce folded proteins highlights the importance of introducing variability in the libraries. It will be interesting to compare the preferences observed for folding with those observed when searching for activity.

Overall, the far UV–CD and fluorescence emission spectra of the purified chimeras are remarkably similar to the parental protein spectra, demonstrating that the global interactions that maintain the structure are preserved in the chimera proteins tested (Figs. 6 and 7). The rather small differences observed may reflect some local structural changes. Far UV–CD spectra are more sensitive to the formation of secondary-structure elements; however, the purification of fair amounts of the different proteins from the soluble cell extract fraction and the blue shift in the fluorescence spectra suggest that the proteins are properly folded and not molten globule intermediates.[42,43] Interestingly, two of the tested constructs seem to have a higher α-helical content. The guest loop of one of them, loop 2 from 1,6-biphosphate aldolase (Mut_L2_FBPA), forms indeed an α-helix in its original protein. This suggests that this loop probably retains its secondary structure in the PRAI-LoxP scaffold—a likely fact, since interactions that maintain α-helical structure depend on

local contacts.[44] The other, loop 6 from PBGS is a shorter loop than the host loop (Mut_L6-PBGS). Loop 6 in PRAI forms part of a mobile lid that opens and closes the active site of the enzyme. PRAI from *E. coli* has a canonical TIM barrel structure, although helix 6 is of only one-turn length and the segment corresponding to helix 5 is not recognized as such in the crystal structure. These two helices are in the vicinity of the exchanged loop 6. It is possible that a change in this loop promotes the formation or stabilization of α-helical structure. Also, it is important to note that the percentages of secondary structure predicted between PRAI-LoxP and Mut_L4 αTS are the same. The original loop 4 from αTS is only three amino acids longer in length than the same loop from PRAI and the tertiary structures of both are very similar. This probably explains the minimal structural changes in this mutant with respect to those in the other mutants and PRAI-LoxP. The selection of a robust protein as starting point as well as the limitation of mutagenesis to areas relevant to function but not so committed to structural integrity is likely to be important for finding a high proportion of folded chimeras from which to search for activity. The proposed strategy allows exploration of distant mutants not reachable by any of the conventional directed evolution strategies, traversing through extensive valley regions to new peaks in the folding-sequence landscape. Thus, search of novel function is restricted to screening the small population of folded proteins instead of a vast amount of nonviable proteins.

Several observations on natural protein evolution were inspiring for our work. First, it is clear that different members of enzyme superfamilies, which have common ancestors but sometimes rather different catalytic activities,[45] incorporate indels or insertions in the loops besides point mutations. Second, the mechanism of antibody generation, which can be viewed as a natural "directed evolution" machinery, relies on a few scaffolds with variability concentrated at the loops.[46,47] Furthermore, in the antibody system, the loop repertoire is itself limited, suggesting the efficiency of employing modular exchange as opposed to randomization.[48] Third, in enzymes with the TIM barrel fold, the active site is always composed of loops and residues in the same face of the protein architecture.

It is tempting then to speculate that even modest diversion from an original activity and specificity would benefit from the use of loops as modular components of variability. Our work suggests the possibility of "structural evolvability" of the TIM barrel fold by loop grafting. A combinatorial approach would lead to the evolution of new proteins with the same scaffold hopefully generating novel binding and catalytic activities. Nature has probably used this mechanism to evolve different functions in the same fold, as hypothesized in the transition from microbial phosphotriesterase-like lactonase into modern bacterial phosphotriesterase, an enzyme that degrades a synthetic insecticide introduced in the 20th century, by insertions in loop 7.[12] Certainly, insertions and deletions into loops, or loop grafting, are believed to be a primary mechanism for creating enzyme diversity.[19,20] Here we demonstrate that in a protein with the TIM barrel architecture, a general method for exchanging variable-sized loops is possible and efficient enough to provide a reasonable starting point for the selection of new enzymatic activities. The mutability of these loops in isolation offers a broad scope for further engineering of multiple loops simultaneously.

Materials and Methods

Construction of the loop insertions in the PRAI gene scaffold

We designed two oligonucleotides, one that corresponds to coding DNA strand and the other to noncoding DNA strand, for assembling the loops and connectors. The coding oligonucleotide of the different loops had a hanging cytosine at the 3′ end, and the noncoding oligonucleotide includes a hanging adenine at the 3′end. On the other hand, the coding oligonucleotide of the NH2 connectors includes a dangling thymine at the 3′ end and the noncoding oligonucleotide for the CO connector had a dangling guanine at the 3′ end (Fig. 2a). Each coding oligonucleotide was hybridized with its respective noncoding oligonucleotide to allow the production of 20 dsDNA—14 for loops and 6 for connectors.

In the NH2 connector of loop 2, an NNS codon was introduced in the coding position for Ile29 to generate variability. In the case of loops 4 and 6, two NNS and NNT codons were introduced to generate variability at positions Gln79, His81 (loop 4), Val124, and Asp126 (loop 6), respectively. The dsDNA NH2 and CO connectors were independently ligated with dsDNA of each loop to produce two primers. These primers are named NH2 and CO, respectively. The 14 different libraries of PRAI-LoxP were constructed in three steps (Fig. 2b). First, the amino halves of the *trpF*-loxP genes were constructed by PCR using a trpF-loxP containing pDAN5 vector (pDAN5B1A4LinkerB5A8PRAI)[24] as template and the oligonucleotide 5′-ATGACAGTCCGAAGCTTCAGGAGGGGTGTTGATG-3′ with a HindIII restriction site (underlined) as 5′-primer and the NH2 megaprimers as 3′-primer. Second, the carboxyl halves were constructed by PCR using the same vector as template and the CO megaprimer as 5′-primer and the oligonucleotide 5′-ATTGGTTTGCCGCTAGCT-CATTAATATGCGCG-3′ with NheI restriction site (underlined) as 3′-primer. Third, each library was amplified by overlapping extension PCR, using the two PCR products previously obtained and the oligonucleotides 5′-ATGACAGTCCGAAGCTTCAGGAGGGGTGTTGATG-3′ with a HindIII restriction site (underlined) as 5′-primer and 5′-ATTGGTTTGCCGCTAGCTCATTAATATGCGCG-3′ with an NheI restriction site (underlined) as 3′-primer. The amplified loop libraries were cloned into pDAN5 vector (pDAN5B1A4LinkerB5A8PRAI) using HindIII and NheI restriction sites, yielding 14 different libraries of plasmids with the new loop exchanged in the corresponding position in the *trpF*-loxP gene.

The plasmids obtained were used to transform *E. coli* XL1-BlueMrf′ cells (Stratagene) by electroporation. Transformed cells were plated on LB medium containing ampicillin to select for plasmid uptake. The individual loop libraries contained from 2.4×10^4 to 4.5×10^5 different variants as estimated from the numbers of grown colonies.

Grown colonies were picked from the plates and amplified in liquid medium to purify plasmid DNA. Plasmid DNA of 283 clones (approximately 20 clones from each loop library) grown under such nonselective conditions were sequenced.

Construction of vector for fusion to CAT gene

The CAT fusion vector pDAN5-*trpF*-loxP-CAT was derived from pDAN5B1A4LinkerB5A8PRAI. pDAN5-*trpF*-LoxP-CAT was constructed in three steps. First a BamHI site was deleted in the original vector pDAN5-*trpF*-loxP by using PCR of whole plasmid (MEGAWHOP)[49] with the complementary oligonucleotides 5′-TGGCTTTAATGAGGGTCCATTCGTTTGTGA-3′ and 5′-TCACAAACGAATGGACCCTCATTAAAGGCA-3′ as overlapping primers, yielding a vector with the deletion of BamHI restriction site by the change in one nucleotide (underlined). This vector was named pDAN5-*trpF*2. Second, the *trpF*-CAT fusion was amplified by PCR using a *trpF*-CAT containing pT4 vector (pT4-*trpF*-CAT) kindly provided by Dr. Lorenzo Segovia (unpublished results) as template and the oligonucleotide 5′-CAGTCGAAGCTTCAGGAGGGGTGTTGATGCTGGAGAAT-AAGGTAT-3′ with HindIII site (underlined) as 5′-primer, and 5′-TGCCGCTAGCTTAGTGGTGGTGGTGGTGGTGAGATCTTTA-3′ with NheI site (underlined) as 3′-primer. The amplified fragment was cloned into pDAN5-*trpF*2 using HindIII and NheI restriction sites, introducing a new BamHI site between *trpF* and CAT genes and yielding pDAN5-*trpF*-CAT vector. Sequences inserted into the multiple cloning site of pDAN5-*trpF*-CAT between the restriction sites HindIII and BamHI led to the expression of proteins fused to CAT at their C-terminus.

Construction of positive and negative controls for folding selection

The positive control to select folded proteins was the WT PRAI protein fused to CAT (pDAN5-*trpF*-CAT vector). Two proteins proven to have poor folding, a circular permutation of PRAI previously described[30] and an engineered antibody (kindly provided by Dr. J. Osuna, unpublished result) fused to CAT, were used as negative controls. Additionally, a construction of WT PRAI with two stop codons between PRAI and CAT fusion was used as negative control.

The circular permutation constructed for this study was the β4/α3 variant previously reported.[30] For its construction, new amino and carboxyl termini were created between α-helix 3 and β-strand 4 using the oligonucleotide 5′-AGTCCGAAGCTTCAGGAGGGGTGTTGATGCGCGATCTGGCAGTCCAACTGCATGG-3′ with HindIII site (underlined) as 5′-primer and the oligonucleotide 5′-CCCGCCAGAGGGATCGTAATATGCGCGCAGCGTC-3′as 3′-primer to generate the new amino termini at β-strand 4. The oligonucleotide 5′-TACGATCCCTCTGGCGGG-CTCGAGAATAAAGTATGT-3′ as 5′-primer and the oligonucleotide 5′-TGCC-GGATCCGTGATCCGATAACACCTTAGCTTTG-3′ with BamH1 site (underlined) as 3′-primer were used to amplify the new C-terminal segment. Both amplifications used a WT PRAI containing vector (pUC18-*trpF*-WT) as template. The two amplified fragments were used to amplify the permuted gene by overlapping extension PCR using oligonucleotide 5′-AGTCCGAAGCTTCAGGAGGGGTGTTGATGCGCGATCTGGC-AGTCCAACTGCATGG-3′ with HindIII site (underlined) as 5′-primer and the oligonucleotide 5′-TGCCGGATCCGTGATCCGATAACACCTTAGCTTTG-3′ with BamH1 site (underlined) as 3′-primer. The amplified fragment was cloned into pDAN5-*trpF*-CAT using HindIII and BamHI restriction sites, yielding the variant Per3-β4/α3-CAT in fusion with CAT.

The fusion of the antibody gene to CAT gene was assembled by PCR using the previously described antibody (Vh) containing pT4 vector as template and the oligonucleotides 5′-AGTCCGAAGCTTCAGGAGGGGTGTTGATGGACGTCCAGCTCCAGCAGTCTGGA-3′ with HindIII site (underlined) as 5′-primer and the oligonucleotide 5′-TGCC-GGATCCACTAGTGACAGTGACCAGAGTACCTTG-3′ with BamH1 site (underlined) as 3′-primer. The amplified fragment was cloned into pDAN5-*trpF*-CAT using HindIII and BamHI restriction sites, yielding the Vh-CAT variant.

The construction of WT PRAI with two stop codons between PRAI and CAT fusion was assembled by PCR using a WT PRAI containing vector (pUC18-*trpF*-WT) as template and the oligonucleotide 5′-ATGACAGTCCGAAGCTTCAGGAGGGGTGT-TGATG-3′ with HindIII site (underlined) as 5′-primer and the oligonucleotide 5′-TGCCGGATCC*TCATTA*ATATGCGCGCAGCGTC-3′with BamH1 site (underlined) and with the two stop codons (italic) as 3′-primer. The amplified fragment was cloned into pDAN5-*trpF*-CAT using HindIII and BamHI restriction sites, yielding the PRAI variant with two stop codons before the fusion to CAT gene.

Chloramphenicol resistance analysis

E. coli MC1061Δ*thiE* cells were transformed with the different variants. Dilutions of the transformed cells were spread on LB medium plates containing Cm (0–100 μg/ml) and ampicillin (200 μg/ml) and incubated for 18 h at 30 °C and the number of colonies under each condition was quantified. The vector pDAN5-*trpF*-CAT, containing a fusion of the WT *trpF* gene fused to CAT gene was used as a control to normalize the survival rate under each condition. A concentration of Cm that discriminated between the negative controls and the WT PRAI, with the highest survival rate for the last one, was picked as the selection condition. In this case, we observe that a Cm concentration as low as 20 μg/ml was adequate to discriminate unfolded variants while maintaining high viability of cells bearing viable genes.

Assessment of expression of soluble fusion protein in selected clones

The *E. coli* MC1061Δ*thiE* strain was transformed by electroporation with each of the constructed loop libraries fused to CAT. Following shaking of the transformants for 1 h at 37 °C, the cells were streaked on LB agar plates containing Cm (20 μg/ml) and ampicillin (200 μg/ml). After incubation at 30 °C for 18 h, eight colonies from each library were selected and precultured in TB liquid medium supplemented with ampicillin (200 μg/ml, for maintenance of plasmid) and incubated for 12 h at 37 °C in the shaker. These precultures (200 μl) were used to inoculate 5 ml of fresh TB liquid medium supplemented with ampicillin (200 μg/ml) and incubated for 18 h at 18 °C. Cells were harvested by centrifugation (Eppendorf/5804-R5, FA-45-30-11 rotor, 5 min, 4000 rpm at 4 °C). The cells were resuspended in 200 μl of 10 mM potassium phosphate buffer at pH 7.6 supplemented with protease inhibitor cocktail [Complete, Free EDTA (ethylenediaminetetraacetic acid), Roche]. The cells were lysed by sonication

(Branson Sonifier 450; 10 s six times at 30-s intervals, 50% pulse, 0 °C) and centrifuged again (Eppendorf/5804-R5, FA-45-30-11 rotor, 20 min, 11,000 rpm at 4 °C) to separate the soluble and insoluble fractions of the cell extract. After separation of 10 μL of soluble extract on a 13% SDS-PAGE and transfer to nitrocellulose membranes (Amersham Pharmacia Bioscience) for 90 min at 80 mA in a semidry transfer unit (Hoefer SemiPhor-Amersham Pharmacia Biotech), Western blotting was performed according to standard protocols,[50] using affinity-purified anti-CAT-digoxigenin (Roche) at a 1:3000 dilution as a primary antibody. The second antibody used was the anti-digoxigenin-AP (Roche) at a 1:5000 dilution and visualized with BCIP/NBT alkaline phosphatase substrate solution (Sigma) for the detection by colorimetric reaction.

Determination of the percentage of folding of the loop libraries

Each loop library fused to CAT gene (pDAN5-Loop Library-CAT), as well as the controls pDAN5-*trpF*-CAT containing the WT PRAI, pDAN5-*trpF*C/S-CAT containing the WT PRAI with two stop codons between *trpF* and CAT gene fusion, pDAN5-Vh-CAT containing the antibody fused to CAT, and pDAN5-per3-CAT including the circular permutation of PRAI, were independently transformed in *E. coli* MC1061Δ*thiE* electrocompetent cells in triplicate. Dilutions of transformed cells were spread in two separate LB medium agar plates, one supplemented with Cm (20 μg/ml) and ampicillin (200 μg/ml), and the other only with ampicillin (200 μg/ml). These plates were incubated for 18 h at 30 °C and the numbers of colonies grown between these two conditions for each plasmid were compared. The folding percentage of the libraries was calculated as the ratio of number of colonies grown under the selective pressure [Cm (20 μg/ml) and ampicillin (200 μg/ml)] to the number of colonies grown without this selective pressure [only ampicillin (200 μg/ml)]. This value was corrected by the number of WT contaminant colonies found under selective pressure for each loop library.

In vivo selection for soluble variants

The *E. coli* MC1061Δ*thiE* strain was transformed by electroporation with the corresponding loop library fused to CAT. Following shaking of the transformants for 1 h at 37 °C, the cells were streaked onto LB agar plates containing Cm (20 μg/ml) and ampicillin (200 μg/ml) and incubated at 30 °C for 18 h. Grown colonies were selected from the plates and grown in liquid LB medium to purify the plasmid DNA. A total of 262 clones grown under selective pressure for folding were sequenced (approximately 18 clones from each loop library).

Statistical analysis of sequences

The sequences found in the mutagenized positions under selective pressure (ampicillin and Cm) were compared with the sequences found without selective pressure (ampicillin). The amino acids observed in these positions were converted to frequencies and the difference between these indicates the discrepancy of occurrence for each residue between these two conditions. The average and the standard deviation of the frequencies were used to determine with a 95% confidence which amino acids were negatively or positively selected in proteins folded for each loop library.

Western blot analysis

The *E. coli* strain MC1061Δ*thiE* was transformed by electroporation with three loop libraries and pDAN5-*trpF*-CAT containing the WT *trpF*. The libraries used for this analysis were loop2 from MR, loop 4 from Ure, and loop 6 from PBGS. After 1-h incubation at 37 °C under shaking, the cells were streaked onto LB plates containing ampicillin (200 μg/ml) and incubated at 37 °C for 12 h. Ten grown colonies were scratched from the plates and resuspended in 5 ml of TB liquid medium supplemented with ampicillin (200 μg/ml, for maintenance of plasmid) and incubated for 12 h at 37 °C in the shaker. Precultures (200 μl) were used to inoculate 5 ml of fresh TB liquid medium supplemented with ampicillin (200 μg/ml) and incubated for 18 h at 18 °C. Cells were harvested and treated as described above. Total protein concentration was measured by Bradford reagent (Bio-Rad). After separation of 11 μg of soluble extract on a 13% SDS-PAGE, the protein was transferred to nitrocellulose membranes (Amersham Pharmacia Bioscience) and analyzed as previously described.

Cloning of selected variants into pET28b(+)

In order to produce the selected variants in the absence of CAT as well as to introduce a C-terminal His6 tag, the *trpF*-loxP and the genes from four true positive colonies grown in the presence of Cm (20 μg/ml) were subcloned from pDAN5-*trpF*-CAT into pET28b(+) vector. To this end, the PRAI-LoxP and the PRAI variant genes (Mut_L4_Ure, Mut_L4_αTS, Mut_L6_PBGS, and Mut_L2_FBPA) were amplified by PCR using the oligonucleotides 5′-GCCA-TACCATGGGGGAGAATAAGGTATG-TGGC-3′ with a NcoI site (underlined) as 5′-primer and 5′-GTCCGAAA-GCTTTCATT-AGTGGTGGTGGTGGTGGTGGGATCCA-TATGCGCGCA-3′with a HindIII site (underlined) as 3′-primer. The amplified product was ligated with pET28b (+), yielding the PRAI-LoxP and the PRAI variants cloned in the pET28b(+) vector. All constructs were sequenced entirely to exclude inadvertent mutations.

Analysis of protein solubility

The mutants Mut_L2_FBPA, Mut_L4_ αTS, Mut_L6_PBGS, and PRAI-LoxP were produced in 5-ml cultures of *E. coli* Rosetta2 cells (Novagen) as described above. Each culture was centrifuged (Eppendorf/5804-R5, FA-45-30-11 rotor, 5 min, 4000 rpm at 4 °C). The cells were resuspended in 0.4 ml of 10 mM potassium phosphate, 0.5 mM EDTA (pH 7.6), 50 mM NaCl, 5% glycerol, 0.1 mM DTT, 0.1 mM PMSF, and 2.5 mg lysozyme and lysed by sonication (Branson Sonifier 450; 10 s six times at 30-s intervals, 50% pulse, 0 °C). The soluble and insoluble fractions of the cell extract were separated. The soluble fractions were applied to SDS-PAGE [13% (w/v) acrylamide].

Expression of PRAI-LoxP variants and protein purification

The expression of the PRAI-LoxP variants was performed in *E. coli* Rosetta2 cells (Novagen) transformed with various pET28b(+) plasmids containing the encoding sequences. To this end, 1 l of LB medium supplemented with kanamycin [50 μg/ml, for maintenance of pET28b(+)] and Cm (25 μg/ml, for maintenance of pRARE) was inoculated with a preculture and incubated at 37 °C. After

an OD_{600} of 0.7 was reached, expression was induced by addition of 1 mM IPTG, and growth was continued for another 15 h at 20 °C. The cells were harvested by centrifugation (Eppendorf/5804-R5, F34-6-38 rotor, 5 min, 4000 rpm at 4 °C) and resuspended in 25 ml of 10 mM potassium phosphate buffer at pH 7.6, 0.5 mM, 50 mM NaCl, 5% glycerol, 0.1 mM DTT, 0.1 mM PMSF, and 2.5 mg of lysozyme, lysed by sonication (Branson Sonifier 450; 20 s six times at 30-s intervals, 50% pulse, 0 °C), and centrifuged again (Eppendorf/5804-R5, F34-6-38 rotor, 20 min, 13000 rpm at 4 °C) to separate the soluble from the insoluble fraction of the cell extract.

The variants Mut_L2_FBPA (mutant with loop 2 from the FBPA library), Mut_L4_αTS (mutant with loop 4 from the αTS library), Mut_L6_PBGS (mutant with loop 4 from the PBGS library), and PRAI-LoxP were purified from the soluble cell fraction. To this end, the extract was loaded onto a nickel Sepharose column (HisTrap FF crude 5 ml, GE Healthcare) previously equilibrated with 50 mM potassium phosphate and 300 mM NaCl buffer at pH 7.6. The column was equilibrated with the same buffer, and the bound His6-tagged protein was eluted by applying a linear gradient from 1 to 500 mM imidazole. Fractions with pure protein were pooled and these were concentrated in the Amicon Ultra-15 system (Millipore) until a final volume of 300 μl was reached. This volume was loaded onto a gel-permeation column (Sephacryl S200, GE Healthcare) that had been previously equilibrated with 50 mM potassium phosphate, 1 mM EDTA, and 0.4 mM DTT buffer at pH 7.6. The different proteins were eluted with the same buffer, and the fractions with pure protein were pooled.

Analytical methods

The different proteins were concentrated using 10-kDa cutoff Amicon® membranes and dialyzed against 4× 1-l degassed 10 mM sodium phosphate, 1 mM EDTA, and 1 mM 2-mercaptoethenol, pH 7.6 buffer. Protein concentration was estimated using an extinction coefficient of 22,900 M^{-1} cm^{-1} at 280 nm. Protein samples were prepared in the same buffer and in 9 M urea, 10 mM sodium phosphate, 1 mM EDTA, and 1 mM 2-mercaptoethanol, pH 7.6 buffer, with a final protein concentration of 16.5 μM for CD studies and 2.0 μM for fluorescence studies.

CD spectra were recorded with a JASCO model J-715 spectropolarimeter equipped with a Peltier temperature control supplied by Jasco. Spectra were collected from 260 to 190 nm. Buffer conditions were 10 mM potassium phosphate, 1 mM EDTA, and 1 mM β-mercaptoethanol (BME) at pH 7.6 and 25.0 °C. Eight replicate spectra were collected from each sample to improve signal-to-noise ratio. The final protein concentration was 16.5 μM, and spectra were collected in a 0.01-cm path-length cell. The secondary-structure prediction was performed using the CDSSTR algorithm, which requires data from 190 to 240 nm.[51–55] Loss of secondary structure was measured in samples of the different proteins in 8.0 M urea incubated at 25 °C for at least 2 h before the spectrum was recorded.

Fluorescence emission spectra were recorded on an LS50B spectrofluorimeter (Perkin Elmer, Norwalk, CT) equipped with a thermostated cell compartment. Spectra were collected from 300 to 540 nm using a 295-nm excitation wavelength at 25 °C (excitation and emission slit widths were 4.0 nm). Protein samples were prepared at 50 μg/ml concentration either in 10 mM phosphate, 1 mM EDTA, and 1 mM BME buffer at pH 7.6 or in 9.0 M urea in the same buffer. Spectra were recorded after 2 h incubation at 25 °C.

FPL size-exclusion chromatography data analysis

The purified enzymes were analyzed by size-exclusion chromatography to determine their nature of oligomerization in an Äkta FPLC system equipped with a UV detector and a size-exclusion Superose HR12 column from Amersham Biosciences (Uppsala, Sweden). The samples were eluted with a 20 mM phosphate, 300 mM KCl, and 1 mM BME buffer at pH 7.6 at a flow rate of 0.5 ml/min. The column was calibrated with lysozyme, lectin, and bovine serum albumin. The proteins were injected at initial concentrations of 2, 20, and 100 μM.

Acknowledgements

This work was supported by grant IN220802 from PAPIIT and CONACyT 49590-Q to G.S.R. We thank Dr. Helena Wright and Dr. Vilmos Fülop for experimental help in the overexpression and purification of the PRAI-LoxP and PRAI variants. We thank Azucena Carrillo Hernández, Yara Sánchez Corrales, and Paulina Tapia Quiroz for their technical assistance in the construction of some of the libraries.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2009.02.022

References

1. Grauer, D. a. & Li, W. -H. (2000). *Fundamentals of Molecular Evolution*. Sinauer Associates, Inc., Sunderland, MA.
2. Graziano, J. J., Liu, W., Perera, R., Geierstanger, B. H., Lesley, S. A. & Schultz, P. G. (2008). Selecting folded proteins from a library of secondary structural elements. *J. Am. Chem. Soc.* **130**, 176–185.
3. Nagano, N., Orengo, C. A. & Thornton, J. M. (2002). One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *J. Mol. Biol.* **321**, 741–765.
4. Voigt, C. A., Martinez, C., Wang, Z. G., Mayo, S. L. & Arnold, F. H. (2002). Protein building blocks preserved by recombination. *Nat. Struct. Biol.* **9**, 553–558.
5. Shao, Z. & Arnold, F. H. (1996). Engineering new functions and altering existing functions. *Curr. Opin. Struct. Biol.* **6**, 513–518.
6. Nilsson, B. (1995). Antibody engineering. *Curr. Opin. Struct. Biol.* **5**, 450–456.
7. Wierenga, R. K. (2001). The TIM-barrel fold: a versatile framework for efficient enzymes. *FEBS Lett.* **492**, 193–198.
8. Hocker, B., Beismann-Driemeyer, S., Hettwer, S., Lustig, A. & Sterner, R. (2001). Dissection of a (betaalpha)8-barrel enzyme into two folded halves. *Nat. Struct. Biol.* **8**, 32–36.
9. Lang, D., Thoma, R., Henn-Sax, M., Sterner, R. & Wilmanns, M. (2000). Structural evidence for

evolution of the beta/alpha barrel scaffold by gene duplication and fusion. *Science*, **289**, 1546–1550.
10. Urfer, R. & Kirschner, K. (1992). The importance of surface loops for stabilizing an eightfold beta alpha barrel protein. *Protein Sci.* **1**, 31–45.
11. Sterner, R. & Hocker, B. (2005). Catalytic versatility, stability, and evolution of the (betaalpha)8-barrel enzyme fold. *Chem. Rev.* **105**, 4038–4055.
12. Afriat, L., Roodveldt, C., Manco, G. & Tawfik, D. S. (2006). The latent promiscuity of newly identified microbial lactonases is linked to a recently diverged phosphotriesterase. *Biochemistry*, **45**, 13677–13686.
13. Wright, H., Noda-Garcia, L., Ochoa-Leyva, A., Hodgson, D. A., Fulop, V. & Barona-Gomez, F. (2008). The structure/function relationship of a dual-substrate (betaalpha)8-isomerase. *Biochem. Biophys. Res. Commun.* **365**, 16–21.
14. Patrick, W. M. & Blackburn, J. M. (2005). In vitro selection and characterization of a stable subdomain of phosphoribosylanthranilate isomerase. *FEBS J.* **272**, 3684–3697.
15. Cheon, Y. H., Park, H. S., Kim, J. H., Kim, Y. & Kim, H. S. (2004). Manipulation of the active site loops of D-hydantoinase, a (beta/alpha)8-barrel protein, for modulation of the substrate specificity. *Biochemistry*, **43**, 7413–7420.
16. Forrer, P., Stumpp, M. T., Binz, H. K. & Pluckthun, A. (2003). A novel strategy to design binding molecules harnessing the modular nature of repeat proteins. *FEBS Lett.* **539**, 2–6.
17. Binz, H. K., Stumpp, M. T., Forrer, P., Amstutz, P. & Pluckthun, A. (2003). Designing repeat proteins: well-expressed, soluble and stable proteins from combinatorial libraries of consensus ankyrin repeat proteins. *J. Mol. Biol.* **332**, 489–503.
18. DeLano, W. L., Ultsch, M. H., de Vos, A. M. & Wells, J. A. (2000). Convergent solutions to binding at a protein–protein interface. *Science*, **287**, 1279–1283.
19. Park, H. S., Nam, S. H., Lee, J. K., Yoon, C. N., Mannervik, B., Benkovic, S. J. & Kim, H. S. (2006). Design and evolution of new catalytic activity with an existing protein scaffold. *Science*, **311**, 535–538.
20. Tawfik, D. S. (2006). Biochemistry. Loop grafting and the origins of enzyme species. *Science*, **311**, 475–476.
21. Li, S., Li, B., Fei, Y., Jiang, D., Sheng, Y., Sun, Y. & Zhang, J. (2007). Exon grafting yields a "two active-site" lysozyme. *Biochem. Biophys. Res. Commun.* **358**, 997–1001.
22. Barona-Gómez, F., Ochoa-Leyva, A. & Soberón, X. (2008). Advances and perspectives in protein engineering: From natural history to directed evolution of enzymes. In *Advances in Protein Physical Chemistry* (García-Hernández, E. & Fernández-Velasco, D. A., eds), pp. 407–438, Transworld Research Network, Kerala, India.
23. Tsuji, T., Onimaru, M. & Yanagawa, H. (2006). Towards the creation of novel proteins by block shuffling. *Comb. Chem. High Throughput Screen.* **9**, 259–269.
24. Saab-Rincon, G., Mancera, E., Montero-Moran, G., Sanchez, F. & Soberon, X. (2005). Generation of variability by in vivo recombination of halves of a (beta/alpha)8 barrel protein. *Biomol. Eng.* **22**, 113–120.
25. Soberon, X., Fuentes-Gallego, P. & Saab-Rincon, G. (2004). In vivo fragment complementation of a (beta/alpha)(8) barrel protein: generation of variability by recombination. *FEBS Lett.* **560**, 167–172.
26. Eberhard, M., Tsai-Pflugfelder, M., Bolewska, K., Hommel, U. & Kirschner, K. (1995). Indoleglycerol phosphate synthase-phosphoribosyl anthranilate isomerase: comparison of the bifunctional enzyme from *Escherichia coli* with engineered monofunctional domains. *Biochemistry*, **34**, 5419–5428.
27. Farber, G. K. & Petsko, G. A. (1990). The evolution of alpha/beta barrel enzymes. *Trends Biochem. Sci.* **15**, 228–234.
28. Maxwell, K. L., Mittermaier, A. K., Forman-Kay, J. D. & Davidson, A. R. (1999). A simple in vivo assay for increased protein solubility. *Protein Sci.* **8**, 1908–1911.
29. Sieber, V., Martinez, C. A. & Arnold, F. H. (2001). Libraries of hybrid proteins from distantly related sequences. *Nat. Biotechnol.* **19**, 456–460.
30. Akanuma, S. & Yamagishi, A. (2005). Identification and characterization of key substructures involved in the early folding events of a (beta/alpha)8-barrel protein as studied by experimental and computational methods. *J. Mol. Biol.* **353**, 1161–1170.
31. Seitz, T., Bocola, M., Claren, J. & Sterner, R. (2007). Stabilisation of a (betaalpha)8-barrel protein designed from identical half barrels. *J. Mol. Biol.* **372**, 114–129.
32. Priestle, J. P., Grutter, M. G., White, J. L., Vincent, M. G., Kania, M., Wilson, E. *et al.* (1987). Three-dimensional structure of the bifunctional enzyme *N*-(5′-phosphoribosyl)anthranilate isomerase-indole-3-glycerol-phosphate synthase from *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **84**, 5690–5694.
33. Akanuma, S. & Yamagishi, A. (2008). Experimental evidence for the existence of a stable half-barrel subdomain in the (beta/alpha)8-barrel fold. *J. Mol. Biol.* **382**, 458–466.
34. Casadaban, M. J. & Cohen, S. N. (1980). Analysis of gene control signals by DNA fusion and cloning in *Escherichia coli*. *J. Mol. Biol.* **138**, 179–207.
35. Chumpolkulwong, N., Hori-Takemoto, C., Hosaka, T., Inaoka, T., Kigawa, T., Shirouzu, M. *et al.* (2004). Effects of *Escherichia coli* ribosomal protein S12 mutations on cell-free protein synthesis. *Eur. J. Biochem.* **271**, 1127–1134.
36. Hosaka, T., Tamehiro, N., Chumpolkulwong, N., Hori-Takemoto, C., Shirouzu, M., Yokoyama, S. & Ochi, K. (2004). The novel mutation K87E in ribosomal protein S12 enhances protein synthesis activity during the late growth phase in *Escherichia coli*. *Mol. Genet. Genomics*, **271**, 317–324.
37. Carter, A. P., Clemons, W. M., Brodersen, D. E., Morgan-Warren, R. J., Wimberly, B. T. & Ramakrishnan, V. (2000). Functional insights from the structure of the 30S ribosomal subunit and its interactions with antibiotics. *Nature*, **407**, 340–348.
38. Horjales, S., Cota, G., Senorale-Pose, M., Rovira, C., Roman, E., Artagaveytia, N. *et al.* (2007). Translational machinery and protein folding: evidence of conformational variants of the estrogen receptor alpha. *Arch. Biochem. Biophys.* **467**, 139–143.
39. Ferraro, D. M., Hope, E. K. & Robertson, A. D. (2005). Site-specific reflex response of ubiquitin to loop insertions. *J. Mol. Biol.* **352**, 575–584.
40. Shoichet, B. K., Baase, W. A., Kuroki, R. & Matthews, B. W. (1995). A relationship between protein stability and protein function. *Proc. Natl Acad. Sci. USA*, **92**, 452–456.
41. Cheng, G., Qian, B., Samudrala, R. & Baker, D. (2005). Improvement in protein functional site prediction by distinguishing structural and functional constraints on protein family evolution using computational design. *Nucleic Acids Res.* **33**, 5861–5867.
42. Kuwajima, K. (1989). The molten globule state as a clue for understanding the folding and cooperativity of globular-protein structure. *Proteins*, **6**, 87–103.

43. Baum, J., Dobson, C. M., Evans, P. A. & Hanley, C. (1989). Characterization of a partly folded protein by NMR methods: studies on the molten globule state of guinea pig alpha-lactalbumin. *Biochemistry*, **28**, 7–13.
44. Creighton, T. E. (1992). *Proteins: Structures and Molecular Properties*, 2nd edit. W. H. Freeman Company, New York.
45. Babbitt, P. C. & Gerlt, J. A. (1997). Understanding enzyme superfamilies. Chemistry as the fundamental determinant in the evolution of new catalytic activities. *J. Biol. Chem.* **272**, 30591–30594.
46. Tomlinson, I. M., Walter, G., Marks, J. D., Llewelyn, M. B. & Winter, G. (1992). The repertoire of human germline VH sequences reveals about fifty groups of VH segments with different hypervariable loops. *J. Mol. Biol.* **227**, 776–798.
47. Chothia, C., Lesk, A. M., Gherardi, E., Tomlinson, I. M., Walter, G., Marks, J. D. *et al.* (1992). Structural repertoire of the human VH segments. *J. Mol. Biol.* **227**, 799–817.
48. Almagro, J. C., Quintero-Hernandez, V., Ortiz-Leon, M., Velandia, A., Smith, S. L. & Becerril, B. (2006). Design and validation of a synthetic VH repertoire with tailored diversity for protein recognition. *J. Mol. Recognit.* **19**, 413–422.
49. Miyazaki, K. (2003). Creating random mutagenesis libraries by megaprimer PCR of whole plasmid (MEGAWHOP). *Methods Mol. Biol.* **231**, 23–28.
50. Sambrook, J. F. E. & Maniatis, T. (1989). *Molecular Cloning: A Laboratory Manual*, 2nd edit. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
51. Whitmore, L. & Wallace, B. A. (2008). Protein secondary structure analyses from circular dichroism spectroscopy: methods and reference databases. *Biopolymers*, **89**, 392–400.
52. Whitmore, L. & Wallace, B. A. (2004). DICHROWEB, an online server for protein secondary structure analyses from circular dichroism spectroscopic data. *Nucleic Acids Res.* **32**, W668–W673.
53. Sreerama, N. & Woody, R. W. (2000). Estimation of protein secondary structure from circular dichroism spectra: comparison of CONTIN, SELCON, and CDSSTR methods with an expanded reference set. *Anal. Biochem.* **287**, 252–260.
54. Compton, L. A. & Johnson, W. C., Jr (1986). Analysis of protein circular dichroism spectra for secondary structure using a simple matrix multiplication. *Anal. Biochem.* **155**, 155–167.
55. Manavalan, P. & Johnson, W. C., Jr (1987). Variable selection method improves the prediction of protein secondary structure from circular dichroism spectra. *Anal. Biochem.* **167**, 76–85.

Supplementary material

Table S1. Oligonucleotides used to construct the connectors and the loops.

Name of the oligonucleotide	Sequence (from 5´- to -3´)	Description
L2NHCod	GCGCGATTTACGGTG GGTTGNNSTTT	Coding sDNA of the NH2-connector of loop 2
L2NHNoCod	AASNNCAACCCACCGT AAATCGCGCCCGCGTCATA	Non-coding sDNA of the NH2 connector of loop 2
L2COCod	GTTGAACAGGCGCAG GAAGTGATGGCTGCGGC	Coding sDNA of the CO-connector of loop 2
L2CONOCod	CCTGCGCCTGTTCAACG	Non-coding sDNA of the CO-connector of loop 2
L2MarCod	GGCTATCCGGCACTC	Coding sDNA of the loop 2 from mandelate racemase
L2MarNoCod	AGTGCCGGATAGCCA	Non-coding sDNA of the loop 2 from mandelate racemase
L2AldCod	TCCAACGGTGGTGCTTCCTTTATCGCTGGTAAAGGCGTGAAATCTGAC GTTCCGCAC	Coding sDNA of the loop 2 from fructose-bisphosphate aldolase
L2AldNoCod	TGCGGAACGTCAGATTTCACGCCTTTACCAGCGATAAAGGAAGCACC ACCGTTGGAA	Non-coding sDNA of the loop 2 from fructose-bisphosphate aldolase
L2UreCod	GGCGGTACTGGCCCGACAGCGGGTTCTAACGCCACAACCTGTACCCC AGGC	Coding sDNA of the loop 2 from urease
L2UreNoCod	CCTGGGGTACAGGTTGTGGCGTTAGAACCCGCTGTCGGGCCAGTACC GCCA	Non-coding sDNA of the loop 2 from urease
L4NHCod2	TTATCGCTGGCGGCAGTGNNSCTGNNT	Coding sDNA of the NH2-connector of loop 4
L4NHNoCod	NNCAGSNNCACTGCCGCCAGCGATAACACCTTAGCTTTG	Non-coding sDNA of the NH2 connector of loop 4
L4COCodbis	CAGCTGTATATCGATACGCTGCGTGAAGCTCTGCCAGCA	Coding sDNA of the CO-

		connector of loop 4
L4CONoCod	CAGCGTATCGATATACAGCTGG	Non-coding sDNA of the CO-connector of loop 4
L4DiPCod	CCTTACTACAATCGTCCGTCC	Coding sDNA of the loop 4 from dihydrodipicolinate synthase
L4DiPNoCod	GACGGACGATTGTAGTAAGGA	Non-coding sDNA of the loop 4 from dihydrodipicolinate synthase
L4ThiCod	TTGGGGCAGGAAGATTTGCAC	Coding sDNA of the loop 4 from thiamin phosphate synthase
L4ThiNoCod	TGCAAATCTTCCTGCCCCAAA	Non-coding sDNA of the loop 4 from thiamin phosphate synthase
L4MaRCod	GAGCCGACGCTTCAACACGAC	Coding sDNA of the loop 4 from mandelate racemase
L4MaRNoCod	TCGTGTTGAAGCGTCGGCTCA	Non-coding sDNA of the loop 4 from mandelate racemase
L4AldCod	GACCTGTCTGAAGAATCC	Coding sDNA of the loop 4 from fructose-bisphosphate aldolase
L4AldNoCod	GATTCTTCAGACAGGTCA	Non-coding sDNA of the loop 4 from fructose-bisphosphate aldolase
L4TrpSCod	GATGTGCCAGTTGAAGAGTCC	Coding sDNA of the loop 4 from tryptophan synthase alpha chain
L4TrpSNoCo	GACTCTTCAACTGGCACATCA	Non-coding sDNA of the loop 4 from tryptophan synthase alpha chain
L4UreCod	GAAGACTGGGGAGCTACC	Coding sDNA of the loop 4

		from urease
L4UreNoCod	GTAGCTCCCCAGTCTTCA	Non-coding sDNA of the loop 4 from urease
L4ADACod	GGTGATGAACTTGGTTTCCCGGGAAGTCTGTTC	Coding sDNA of the loop 4 from adenosine deaminase
L4ADANoCod	AACAGACTTCCCGGGAAACCAAGTTCATCACCA	Non-coding sDNA of the loop 4 from adenosine deaminase
L4AlevCod	GCCGCGATGGACGGC	Coding sDNA of the loop 4 from porphobilinogen synthase
L4AlevNoCo	CCGTCCATCGCGGCA	Non-coding sDNA of the loop 4 from porphobilinogen synthase
L6NHCod	CAGCACGTTGATAAATATNNSTTANNT	Coding sDNA of the NH2-connector of loop 6
L6NHNoCod	NNTAASNNATATTTATCAACGTGCTGAAACTCGCG	Non-coding sDNA of the NH2 connector of loop 6
L6COCod	TTTGACTGGTCACTATTAAATGGTCAATCGCTTGCC	Coding sDNA of the CO-connector of loop 6
L6CONoCod	CGATTGACCATTTAATAGTGACCAGTCAAAG	Non-coding sDNA of the CO-connector of loop 6
L6DiPCod	ACAGGGAACCTC	Coding sDNA of the loop 6 from dihydrodipicolinate synthase
L6DiPNoCod	AGGTTCCCTGTA	Non-coding sDNA of the loop 6 from dihydrodipicolinate synthase
L6AlevCod	CCTGCTGGAGCGTAC	Coding sDNA of the loop 6 from porphobilinogen synthase
L6AlevNoCod	TACGCTCCAGCAGGA	Non-coding sDNA of the loop 6 from porphobilinogen synthase
L6TrpSCod	TCACGAGCAGGCGTGACCGGCGCAGAAAACCGCGCCGCGTTACCC	Coding sDNA of the loop 6

		from tryptophan synthase alpha chain
L6TrpSNoCod	GGTAACGCGGCGCGGTTTTCTGCGCCGGTCACGCCTGCTCGTGAA	Non-coding sDNA of the loop 6 from tryptophan synthase alpha chain

Table S2. Sequences obtained from clones that show fusion to CAT in the western blot analysis.

Variants	Library of loop 2 from mandelate racemase	Library of loop 4 from urease		Library of loop 6 from porphobilinogen synthase	
	129 [a]	**Q79** [a]	**H81** [a]	**V124** [a]	**D126** [a]
1	W	H	H	A	L
2	R	R	S	A	Y
3		L	S	A	H
4		S	N		
5		G	A		
6		N	H		

[a]Sites mutated in the loop libraries.

Table S3. Sequences obtained from clones that did not show fusion to CAT in the western blot analysis

Variants	Library of loop 2 from mandelate racemase	Library of loop 4 from urease		Library of loop 6 from porphobilinogen synthase	
	129 [a]	**Q79** [a]	**H81** [a]	**V124** [a]	**D126** [a]
1	N	H[b]	H[b]	H	L
2	F	V	G	N	L
3	F	P	N	X[c]	H
4		H	V	T	H

[a]Sites mutated in the loop libraries.
[b]The clone has a stop codon in other region of the gene.
[c]The clone has a stop codon in the position V124.

Table S4. Sequences found without selective pressure in the variable position I29 for libraries of loop 2 .

Variants	Urease	Mandelate racemase	Fructose-bisphosphate aldolase
	I29	**I29**	**I29**
1	L	L	T
2	V	A	S
3	T	W	L
4	T	W	L
5	F	W	T
6	K	W	S
7	S	D	H
8	D	D	L
9	L	G	L
10	T	L	L
11	I	W	L
12	I	N	N
13	I	R	G
14	W	R	R
15	G	G	G
16	L	D	I
17	S	D	G
18	R	P	A
19		F	
20		W	

Table S5. Sequences found without selective pressure in variable positions for libraries of loop 4.

Variants	Urease		Porphobilinogen synthase		Adenosine deaminase		Thiamin phosphate synthase		Mandelate racemase		Dihydrodipicolinate synthase		Tryptophan synthase alpha chain		Fructose-bisphosphate aldolase	
	Q79	**H81**	**Q79**	**H81**	**Q79**	**H81**	**Q79**	**H81**	**Q79**	**H81**	**Q79**	**H81**	**Q79**	**H81**	**Q79**	**H81**
1	L	T	S	R	R	G	D	P	D	N	V	Y	T	P	R	P
2	L	S	S	C	C	I	D	C	D	D	K	T	G	I	R	D
3	L	P	V	Y	R	A	I	P	P	P	E	R	X[a]	Y	L	V
4	I	S	V	V	G	Y	I	V	P	C	P	S	K	T	C	N
5	I	T	V	A	V	A	P	P	I	P	S	L	I	S	V	D
6	R	P	C	P	H	V	P	V	I	V	G	V	V	P	V	F
7	R	S	C	H	W	G	P	C	I	N	R	V	S	V	S	Y
8	R	L	T	V	M	R	H	F	I	I	D	A	K	P	S	I
9	R	T	T	I	Y	G	H	S	I	L	V	S	Y	S	H	P
10	P	H	T	D	F	R	M	N	E	H	F	S	A	V	T	N
11	G	N	L	N	F	Y	M	C	H	N	T	C	V	S	T	I
12	F	P	L	R	I	H	R	C	H	F	Y	T	I	F	T	F
13	V	F	L	S	L	G	R	H	M	C	L	C	D	D	P	S
14	V	R	I	I	K	C	L	F	Q	F	W	S	H	D	N	S
15	Y	I	I	S	L	N	L	A	Q	A	P	V	H	S	N	F
16	Y	R	H	N	A	D	V	R	K	P	P	C	F	S	I	N
17	A	Q	H	R	H	P	G	D	K	H	K	N	W	I	I	A
18	A	I	P	V	S	I	G	R	G	H	G	I	R	D	I	L
19	S	A	P	Y			A	P	V	R					K	A
20	S	V	K	F			S	N	S	N					Y	A
21	S	S	R	H			S	P	W	D					G	T
22	K	R	Y	I			S	I							G	V
23	K	P	Y	V			S	S							G	L
24							W	N								

[a]A stop codon was found in this site.

Table S6. Sequences found without selective pressure in variable positions for libraries of loop 6 .

Variants	Porphobilinogen synthase		Dihydrodipicolinate synthase		Tryptophan synthase alpha chain	
	V124	**D126**	**V124**	**D126**	**V124**	**D126**
1	V	Y	T	S	H	L
2	N	L	M	S	V	L
3	P	L	I	C	N	T
4	A	Y	V	C	C	P
5	P	H	T	G	F	N
6	R	C	Q	I	S	T
7	V	D	R	I	T	A
8	R	A	L	C	N	L
9	P	T	I	R	C	M
10	T	F	S	T	W	L
11	Y	C	L	L	L	S
12	D	S	P	V	S	L
13	A	D	V	L	F	C
14	N	P	S	S	G	L
15	W	G	S	L	S	T
16	G	G	G	T	L	C
17	S	A	A	S	N	N
18	N	T	E	F	C	F
19			E	V		
20			L	R		
21			T	H		
22			N	T		
23			S	F		

Table S7. Sequences found under selective pressure in the variable position for libraries of loop 2.

Variants	Urease	Mandelate racemase	Fructose-bisphosphate aldolase
	I29	**I29**	**129**
1	S	R	T
2	S	R	T
3	S	R	T
4	S	R	T
5	S	R	T
6	I	R	T
7	I	R	T
8	F	A	T
9	A	A	T
10	A	A	T
11	A	A	T
12	F	G	S
13	F	D	N
14	V	D	G
15	L	D	N
16	L	D	M
17	Q	W	C
18	K	V	S

Table S8. Sequences found under selective pressure in the variable positions for libraries of loop 4.

Variants	Urease		Porphobilinogen synthase		Adenosine deaminase		Thiamin phosphate synthase		Mandelate racemase		Dihydrodipicolinate synthase		Tryptophan synthase alpha chain		Fructose-bisphosphate aldolase	
	Q79	**H81**	**Q79**	**H81**	**Q79**	**H81**	**Q79**	**H81**	**Q79**	**H81**	**Q79**	**H81**	**Q79**	**H81**	**Q79**	**H81**
1	G	R	P	S	K	S	K	G	I	H	V	G	S	A	G	H
2	Q	H	I	I	R	P	R	T	F	S	K	N	K	G	N	P
3	P	H	P	V	I	R	S	F	M	N	E	I	I	R	Q	N
4	**S**[a]	**T**[a]	V	P	S	C	N	D	V	L	V	T	F	G	H	L
5	**D**[a]	**T**[a]	I	N	V	H	S	I	H	Y	F	D	I	D	S	A
6	L	Y	**H**[a]	**V**[a]	L	H	R	L	**T**[a]	**V**[a]	T	Y	H	L	I	R
7	C	P	L	S	M	R	L	S	W	I	L	H	N	P	K	G
8	K	P	T	N	**S**[a]	**R**[a]	R	N	C	P	D	Y	Q	N	L	G
9	**S**[a]	**T**[a]	T	G	N	R	S	L	R	S	S	S	G	H	R	D
10	A	T	F	R	R	Y	S	N	**P**[a]	**N**[a]	**R**[a]	**V**[a]	G	D	T	N
11	**D**[a]	**T**[a]	C	S	K	H	V	V	W	A	P	R	T	Y	G	T
12	R	D	S	H	G	N	P	H	N	D	Y	D	V	H	R	A
13	**N**[a]	**P**[a]	Y	S	C	H	L	N	**P**[a]	**N**[a]	T	P	H	N	Y	D
14	M	A	E	C	R	C	C	N	P	T	R	P	L	L	H	V
15	T	Y	T	Y	A	V	C	D	W	S	A	N	L	D	W	T
16	**N**[a]	**P**[a]	**H**[a]	**V**[a]	K	D	M	C	**T**[a]	**V**[a]	A	C	Y	C	R	H
17	M	H	G	A	Q	T	R	S	N	A	V	I	G	S	A	P
18	P	T	S	T	**S**[a]	**R**[a]	R	G	P	S	**R**[a]	**V**[a]	I	H	G	L
19	L	H	N	T	P	D			Y	I	P	I	E	H	K	T
20			E	R							S	A	Y	P		

[a]These clones were found as duplicate in each library.

Table S9. Sequences found under selective pressure for libraries of loop 6.

Variants	Porphobilinogen synthase		Dihydrodipicolinate synthase		Tryptophan synthase alpha chain	
	V124	**D126**	**V124**	**D126**	**V124**	**D126**
1	V	G	**T**[a]	**C**[a]	F	F
2	D	H	T	I	L	P
3	T	T	L	H	T	M
4	F	T	R	T	A	N
5	K	I	C	C	C	L
6	H	C	C	Y	L	F
7	Y	L	A	L	E	L
8	L	R	G	T	M	L
9	E	C	P	T	V	V
10	Y	W	**T**[a]	**C**[a]	A	I
11	L	P	K	F	**T**[a]	**R**[a]
12	P	F	S	V	M	L
13	**T**[a]	**C**[a]	V	L	**T**[a]	**R**[a]
14	L	A	R	P	W	L
15	**T**[a]	**C**[a]	P	I	N	S
16	R	V	L	P	W	P
17	P	H	F	L	N	C
18	P	R	S	T	T	M

[a]These clones were found as duplicate in each library.

Figure 1S: Sequence analysis of clones under selective pressure in the positions mutated by the NNG/C or NNT combination codons. Amino acids are represented by the one code letter. The histograms represent the normalized frequency of the selected versus non-selected amino acids. The red colored histograms (*) represent those amino acids at the variable positions that had a frequency either significantly higher (2 standard deviations) or lower than expected by chance. The codes for loop libraries are as follows: (Ure), urease; (FBPA), fructose-bisphosphate aldolase; (MR), mandelate racemase; (TPS), thiamin phosphate synthase; (PBGS), porphobilinogen synthase; (αTS), tryptophan synthase alpha chain; (ADA), adenosine deaminase; (DHDPS), dihydrodipicolinate synthase. (a) Frequencies found at position I29 (mutated to NNG/C combination) in libraries from loop 2. (b) Frequencies found at position Q79 (NNG/C) in libraries from loop 4. (c) Frequencies found in the position H81 (NNT) in libraries from loop 4. (d) Frequencies found at position V124 (NNG/C) in libraries from loop 6. (e) Frequencies found at position D126 (NNT) in libraries from loop 6.

Figure S1a

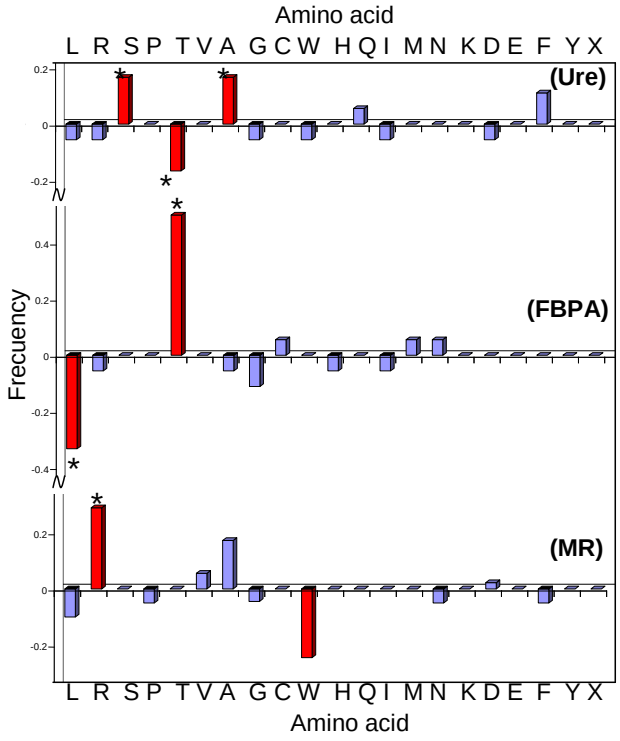
Amino acid
L R S P T V A G C W H Q I M N K D E F Y X
(Ure)
(FBPA)
(MR)
Frecuency
0.2
0
-0.2
0.4
-0.4
Amino acid

Figure S1b

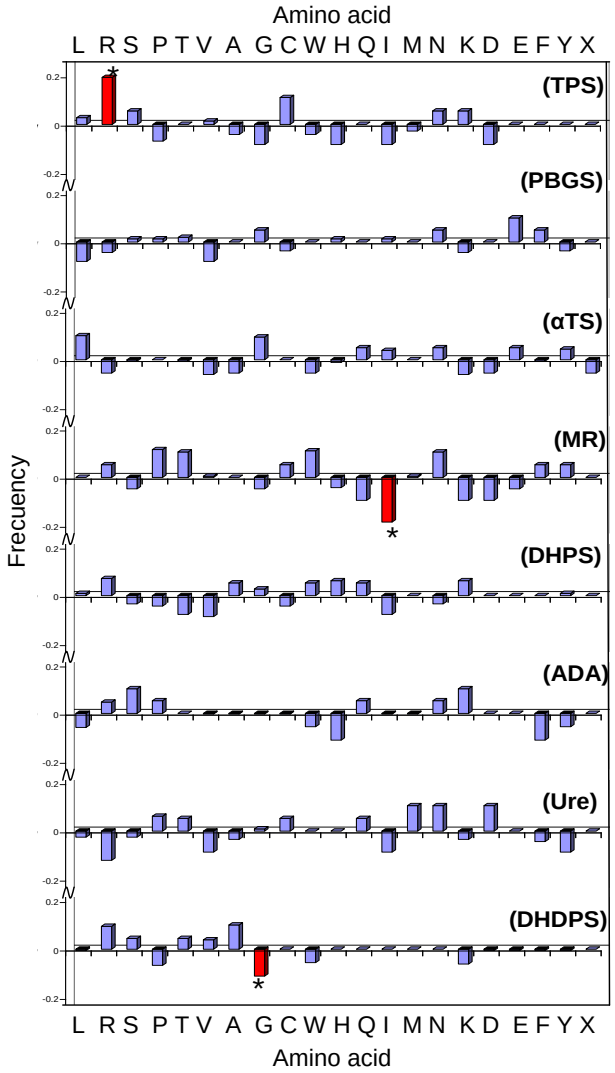
Amino acid
L R S P T V A G C W H Q I M N K D E F Y X
(TPS)
(PBGS)
(αTS)
(MR)
(DHPS)
(ADA)
(Ure)
(DHDPS)
Frecuency
0.2
0
-0.2
*
L R S P T V A G C W H Q I M N K D E F Y X
Amino acid

Figure S1c

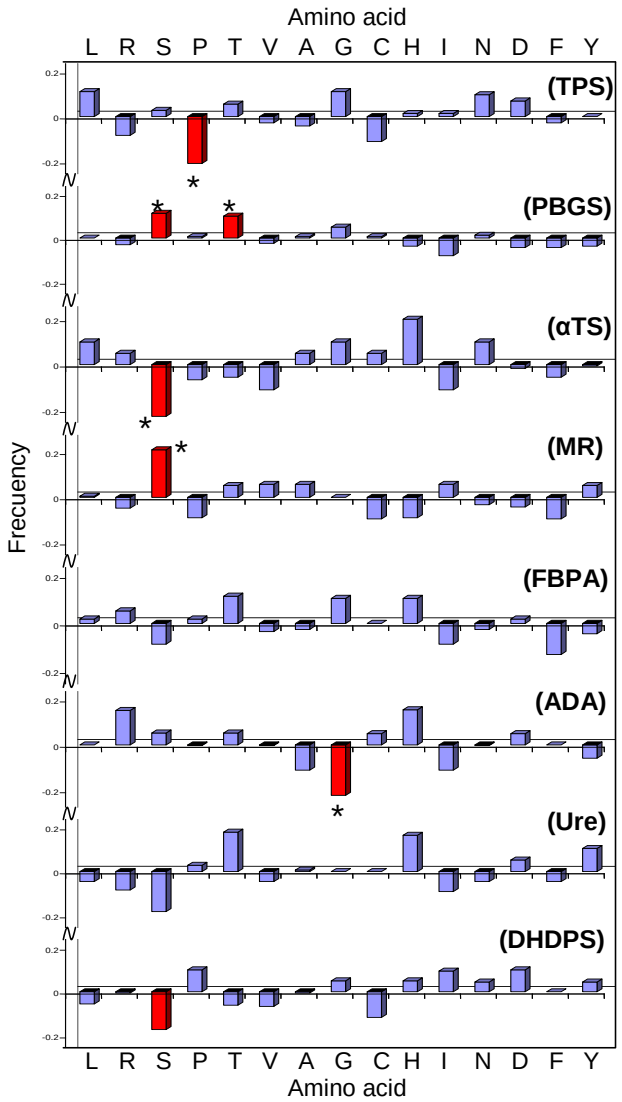
Amino acid
L R S P T V A G C H I N D F Y
(TPS)
(PBGS)
(αTS)
(MR)
(FBPA)
(ADA)
(Ure)
(DHDPS)
Frecuency
0.2
0
-0.2
*
L R S P T V A G C H I N D F Y
Amino acid

Figure S1d

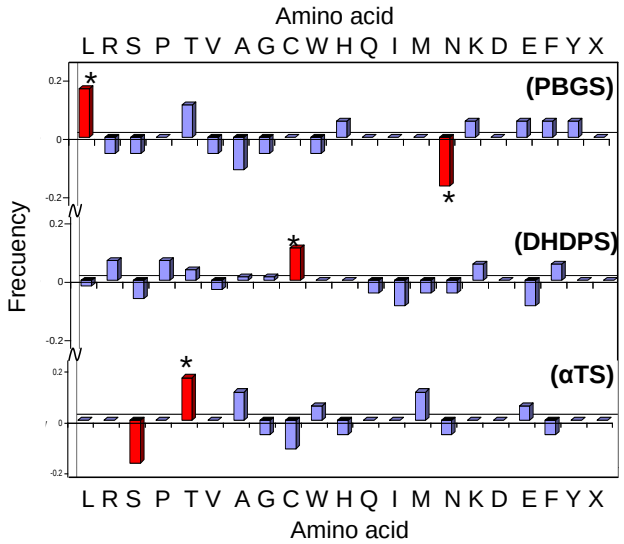
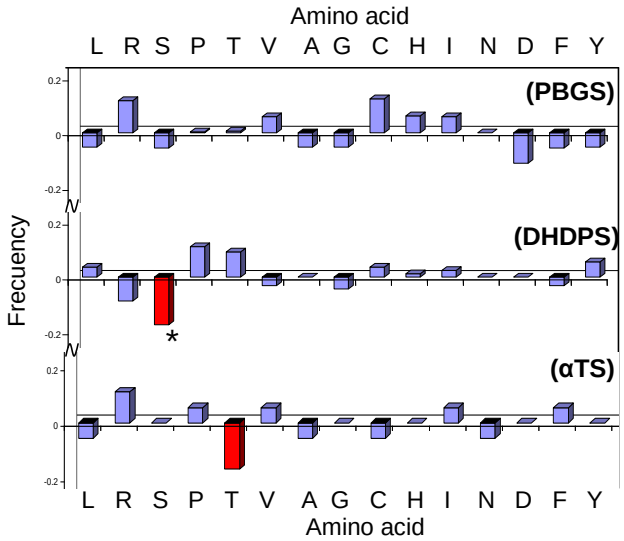
Figure S1e