Construction of “small-intelligent” focused mutagenesis libraries using well-designed combinatorial degenerate primers

Lixia Tang¹, Hui Gao², Xuechen Zhu¹, Xiong Wang¹, Ming Zhou², and Rongxiang Jiang¹
¹School of Life Science and Technology and ²School of Computer Science and Technology, University of Electronic Science and Technology of China, Chengdu, China

BioTechniques 52:149-158 (March 2012) doi 10.2144/000113820
Keywords: randomization; library construction; degenerate codon design; codon redundancy; amino acid bias

*L.T. and H.G. contributed equally to this work.

Site-saturation mutagenesis is a powerful tool for protein optimization due to its efficiency and simplicity. A degenerate codon NNN or NNS (K) is often used to encode the 20 standard amino acids, but this will produce redundant codons and cause uneven distribution of amino acids in the constructed library. Here we present a novel “small-intelligent” strategy to construct mutagenesis libraries that have a minimal gene library size without inherent amino acid biases, stop codons, or rare codons of Escherichia coli by coupling well-designed combinatorial degenerate primers with suitable PCR-based mutagenesis methods. The designed primer mixture contains exactly one codon per amino acid and thus allows the construction of small-intelligent mutagenesis libraries with one gene per protein. In addition, the software tool DC-Analyzer was developed to assist in primer design according to the user-defined randomization scheme for library construction. This small-intelligent strategy was successfully applied to the randomization of halohydrin dehalogenases with one or two randomized sites. With the help of DC-Analyzer, the strategy was proven to be as simple as NNS randomization and could serve as a general tool to efficiently randomize target genes at positions of interest.

In the past few decades, directed evolution has emerged as a powerful approach for creating proteins with desired properties (1,2). Harnessing Darwinian evolution in vitro, this evolution strategy comprises the seamless combination of gene diversification, expression of mutant enzymes, and high-throughput screening or selection. The advantage of the directed evolution approach is that there is no prerequisite for the mechanisms of the desired activity or the structure of enzymes. Most early directed evolution studies followed a “nonrational” route and applied random mutagenesis methods, such as error-prone PCR or DNA shuffling, to generate large and diverse libraries (3,4). Although a plethora of successful examples utilizing such strategies have been demonstrated in numerous scientific and patent publications, these approaches rely on the availability of high-throughput screening methods. It is apparent that high-quality enzyme libraries could reduce the screening effort required and make directed evolution experiments more efficient.

Currently, enzyme optimization has moved from pure “blind” to “semi-rational” approaches in order to accommodate the demands of rapid creation of novel industrial biocatalysts (5–8). Mutagenesis “hot spots” and regions are generally identified based on mechanistic and structural information or by applying computationally guided mutagenesis strategies (9,10). Over the years, many computational methods, such as B-FITTER, PatchFinder, and Rate4Site, have been introduced, and the option of generating small focused mutant libraries at the defined positions has become preferred (11–13). Although the conventional degenerate codon NNN or NNK (S) covers all 20 amino acids and is often used to randomize those identified hot spots, the resulting focused libraries would contain redundant codons that encode amino acids unevenly. Moreover, codon redundancy and amino acid biases will increase exponentially when the number of mutation sites is increased, which undoubtedly decreases the quality of the constructed library. Consequently, it is difficult to screen low frequency positive colonies from a large-size library. However, in most cases utilizing only one degenerate codon may not be ideal.

To circumvent this limitation, several chemical and biochemical strategies have been developed, which include the use of specialized dinucleotide and trinucleotide phosphoramidites during the synthesis of randomized oligonucleotides (14–16). Although these systems could efficiently eliminate codon redundancy and introduce any subset of codons at target positions, their costs are high, and their application is limited to laboratories with expertise in specialized DNA synthesis (14). In 2003, Hine and coworkers presented their MAX randomization methodology for constructing maximally diverse libraries with a minimized gene library size (17). The essence of the procedure is to synthesize sets of 9-mer primers for each codon to be randomized, which are subsequently used to hybridize to a synthesized template strand with NNN randomization at each target site. The hybridized and ligated selection strands are amplified by a subsequent PCR step. In
this way, the resulting randomized DNA cassette contains no redundant codons and no inherent amino acid biases. Obviously, the advantage of the MAX system over the aforementioned phosphoramidate oligonucleotide strategy is that the MAX technology needs only simple primers. However, the MAX system is more complex and expensive than NNN/NNK randomization.

Here, we provide an alternative way to construct small-intelligent mutagenesis libraries without inherent amino acid biases by using well-designed combinatorial degenerate primers. Ideally, the constructed library contains colonies with just one gene for one protein as MAX does. Moreover, the software tool DC-Analyzer was developed to assist in primer design for our small-intelligent mutagenesis library construction. With the assistance of DC-Analyzer, our small-intelligent randomization strategy could serve as a good substitute for NNS randomization.

Materials and methods

Mutagenesis

Site-directed saturation mutagenesis experiments were carried out using Phusion high-fidelity DNA polymerase PCR kit (New England Biolabs, Ipswich, MA, USA) with the recombinant expression vector pBADHheC containing the wild-type hheC gene (accession no: AF397297) as a template. For full randomization of one site using the small-intelligent method, a mixture of four pairs of well-designed complementary primers at a ratio of 12:6:1:1 were used (forward primers: 5′-CTTCAACXXXCCCAGCATA-3′, where XXX represents NDT, VMA, ATG, and TGG, respectively), while for NNS randomization, one pair of complementary primers was used (forward primer: 5′-TTCAACNNSCCGACTACCTT-3′). All primers were supplied by Invitrogen (Shanghai, China). The two randomization reactions were performed in 50-µL reactions under the same conditions. The reactions contain 1 × HF buffer, 200 µM each dNTP, 1 mM Mg2+, 100 ng template, 2 µM each mixed primer, and 0.01 U/µL Phusion polymerase. The temperature program used was 98°C for 3 min followed by 30 cycles of 10 s at 98°C, 45 s at 50°C, and 2 min at 72°C, and finished with a final incubation at 72°C for 10 min.

For full randomization of two contiguous sites, as well as two sites in close proximity, partially overlapped primers were used. In a case study, two active-site residues (V136/L141) of halohydrin dehalogenase from Arthrobacter sp. AD2 (HheC) were randomized. For this case, eight partially overlapping oligonucleotides were synthesized independently (four forward primers carrying mutagenized site 136: 5′-AGAGAGCGXXGCTACAA-3′, where XXX represents NDT, VMA, ATG, and TGG; four reverse primers carrying the mutagenized site 141: 5′-TGCCXXXXAGAGGTTAG-3′, where XXX represents AHN, TKB, CAT, and CCA). In the case study to randomize two contiguous sites, two active-site residues (PI35/F136) of halohydrin dehalogenase from Agrobacterium radiobacter AD1 (HheC) were selected. For this case, 16 mutagenic oligonucleotides (mutagenic oligonucleotides: 5′-CTGCAACGXXXXXGGGCCTTG-3′, where XXX represents NDT, VMA, ATG, and TGG; there are 16 different combinations of the four degenerate codons in total) and one partially overlapping nonmutagenic primer (5′-GGTAGAAGCTCCTTCCAAAGC-3′) were synthesized independently. The recombinant expression vector pBADHheC containing the wild-type hheC gene (accession no: AF397296) was used as a template. These two PCR amplifications were performed under the same conditions as described above.

Small-intelligent library construction

After amplification, the reaction mixture was digested with DpnI (New England Biolabs) at 37°C for 1 h to remove the parental template. For one-site randomization, 2 µL DpnI-digested mixture were used to transform Escherichia coli TOP10 chemical competent cells, while 10 µL were used for transformations when two sites were randomized. The transformed culture was plated on LB agar plates supplemented with 100 µg/mL ampicillin, resulting in libraries with a few hundred to thousands of colonies for the above two cases. Sequence analysis was performed by Invitrogen.

DC-Analyzer

The ultimate goal of the DC-Analyzer software developed in this work is to assist in degenerate primer design according to the user-defined randomization scheme utilized for the construction of small-intelligent libraries. In DC-Analyzer, the user can randomly choose the desired amino acids one by one. For other specific requests regarding the type of amino acids, such as randomizing one site with amino acids bearing a polar side chain (R, N, D, Q, E, C, F, G, and V), VMA (encoding 6 amino acids: E, A, Q, P, K, and T), ATG (M), and TGG (W) were designed using DC-Analyzer according to the above criteria. Four primers with the above degenerate codons at the target site were synthesized independently and mixed at a ratio of 12.6:6:1:1 according to the numbers of encoded amino acids. The resulting primer mixture was termed small-intelligent primer and used for the subsequent PCR process. Although degenerate codons have been used in the construction of some restricted mutagenesis libraries (18), this is, to the best of our knowledge, the first report to use combinatorial degenerate primers for full randomization with just one codon per amino acid.

Small-intelligent primer is, in principle, compatible with most PCR-based NNS randomization strategies. Thus, a suitable PCR-based mutagenesis approach could be easily selected for library construction. For instance, a simple, standard QuickChange site-directed mutagenesis protocol (Stratagene, Agilent Technologies, Santa Clara, CA, USA) could be adopted when one site is to be randomized (Figure 1, Step II, Option A). In conjunction with a double-stranded primer mixture (NDT, VMA, ATG, and TGG), a small-intelligent library could be easily constructed. Overlapping PCR is suitable for the randomization of two nearby sites in the above two randomization strategies (Figure 1, Step II, Option B), while megaprimers-based PCR could be used for the randomization of

The software is available from: http://cobi.uestc.edu.cn/resource/dc_analyzer/view.

Results and discussion

Outline of the small-intelligent strategy

The small-intelligent strategy comprises the two steps of degenerate codon design and PCR-based randomization (Figure 1). To construct a small-intelligent library with the prerequisite of no amino acid biases, the design of an optimal set of degenerate codons for each target site is critical and should obey the following criteria: (i) degenerate codon sets should comprise minimized codon variants that encode all user-desired amino acids with an equal probability of occurrence in the constructed library and (ii) stop codons and the rare codons of E. coli should be eliminated. For this, 53 out of 61 sense codons that make up the genetic codons were used in degenerate codon design to encode 20 amino acids by eliminating eight rare codons of E. coli (CGA, CCG, AGA, AGG for Arg; CUA for Leu; AU for Ile; GGA for Gln; and CCC for Pro). As exemplified in the full randomization of one mutation site with 20 amino acids, a set of four degenerate codons of NDT (encoding 12 amino acids: N, S, I, H, R, L, Y, C, F, G, and V), VMA (encoding 6 amino acids: E, A, Q, P, K, and T), ATG (M), and TGG (W) were designed using DC-Analyzer according to the above criteria. Four primers with the above degenerate codons at the target site were synthesized independently and mixed at a ratio of 12.6:6:1:1 according to the numbers of encoded amino acids. The resulting primer mixture was termed small-intelligent primer and used for the subsequent PCR process. Although degenerate codons have been used in the construction of some restricted mutagenesis libraries (18), this is, to the best of our knowledge, the first report to use combinatorial degenerate primers for full randomization with just one codon per amino acid.
Step I. Design of degenerate codons

Step II. PCR stage

Step III. Digest the methylated templates with DpnI and transform PCR production into E. coli

Figure 1. Illustration of the construction of a small-intelligent saturation mutagenesis library and NNS randomized codons are surrounded by a gray dashed-line box, and small-intelligent codons are shown inside black frames. Step I: Design a set of degenerate codons for small-intelligent saturation mutagenesis library construction based on the criteria that one amino acid is encoded by only one codon and that stop codons and rare codons of E. coli are eliminated. Amino acids with one-letter notation are shown at the left side of the corresponding codons. Stop codons are represented by X; rare codons of E. coli are marked with an asterisk. The degenerate codon set shown in this step is used to fully randomize one site with 20 amino acids. Primers with those degenerate codons at the mutated site are synthesized independently and mixed at a ratio according to the encoded amino acid numbers (12:6:1:1 in this case). The primer mixture is then used for subsequent PCR procedure. Step II: Selection of PCR-based mutagenesis procedures for library construction. The desired mutation sites are shown as “×” in the arrows, indicating the primers in the schematic diagrams. Option A: Classical QuikChange site-directed mutagenesis procedure is suitable for introducing one mutation site. For this, four pairs of primers incorporating one degenerate codon each are synthesized independently. Option B: Overlapping PCR is suitable for randomizing two mutation sites in close proximity (a) or that are contiguous (b). Option C: Megaprimer-based PCR procedure could be applied in the case that two mutation sites are far from each other. With options B and C, except for two contiguous sites, two sets of single-stranded oligonucleotides with one mutation in each set were synthesized to ensure a high mutant diversity for the constructed library. To randomize two contiguous sites, 16 forward mutagenic oligonucleotides (F) and one partially overlapping reverse primer (R) were synthesized. Small-intelligent libraries were constructed by transforming E. coli competent cells with DpnI-digested PCR mixture in Step III.

two sites located far from each other (Figure 1, Step II, Option C) (19,20). When coupled with these procedures, two-site randomized small-intelligent libraries could be constructed using the small-intelligent system with two sets of single-stranded primers (four oligonucleotides per set: NDT, VMA, ATG, and TGG). Although the libraries could also be constructed by using double-stranded primers carrying two randomized sites, 16 oligonucleotides instead of eight would be required in this case. Obviously, using the single-stranded primers greatly reduces primer costs. However, such a primer design recommendation could not be applied in the case of two contiguous randomized sites, for which 16 mutagenic oligonucleotides (forward) and one nonmutagenic oligonucleotide (reverse) are needed. It is worthwhile to note that the megaprimer-based PCR has been proven to be applicable for megaprimer with sizes up to a few kilobases (21). Thus, in principle, the small-intelligent system could be applied in any cases where two sites are to be randomized.

Moreover, Wang et al. have recently reported a novel method for the construction of multiple-site saturation mutagenesis libraries (NNS randomization) by the combination of overlapping PCR and megaprimer-based PCR, and the strategy has been successfully applied in the randomization of an E. coli K12 malic enzyme at three target positions (22). Combined with this PCR strategy, the small-intelligent system could, in principle, randomize more than two sites simultaneously using (2 × N - 2) sets of single-stranded oligonucleotides, where N represents the number of mutagenized sites. In practice, it was found that for a saturation mutagenesis library containing three randomized sites, the minimal library size is 8000, while it is 32,768 for NNS randomization. An extensive screening effort is needed for these two cases. In response to such drawbacks, Reetz et al. reported their iterative CASTing method (23), which has proven to be a valuable means to accelerate enzyme evolution. The essence of the methodology is to construct several small focused libraries (each library containing only two spatially close randomized sites) and to combine the positive hits by repeating the process until the final hits with desired catalytic properties are achieved. For such a small library, about 3100 transformants are required for 95% library coverage using NNS randomization. By applying our small-intelligent system, only 1200 transformants are needed. Apparently, the screening effort is minimized by a factor of 2.6 in this way. Thus, the following case studies were mainly focused on library construction when no more than two sites are randomized.

Theoretical analysis and case studies

To evaluate the feasibility of the small-intelligent strategy, one active-site residue of HheA was fully randomized using both conventional NNS randomization and small-intelligent methods. For each library, about 55 randomly picked colonies were sequenced, and the encoded amino acids were analyzed. The results showed that in the small-intelligent library, all 20 amino acid codons were identified, and no rare codons and stop codons were observed. In the conventional NNS randomized library, only 16 amino acids were...
Figure 2. Experimental and theoretical amino acid distributions of small-intelligent and NNS randomizations. All randomizations were performed at the active site of both HheA and HheC. For each randomization, about 55 colonies were sequenced, and the obtained sequence data were then translated into the encoded amino acids. Colonies containing insertions or deletions were not included in the data analysis. All 20 amino acids are represented by one-letter notation. Stop codons and rare codons of E. coli are represented by X and *, respectively. (A and B) Theoretical amino acid distributions of NNS (A) and small-intelligent (B) randomizations for one mutation site. (C and D) Experimental amino acid distributions of NNS (C) and small-intelligent (D) randomizations for one mutation site. (E and F) Experimental amino acid distributions of small-intelligent randomization for two mutation sites that are in close proximity (E) or are contiguous (F). Stacked bars in panel E represent the number of mutations occurring in the V136/L141 library at positions 136 (bottom bar) and 141 (top bar). Stacked bars in panel F represent the number of mutations occurring in the P136/F136 library at positions 135 (bottom bar) and 136 (top bar).

The utility of the small-intelligent strategy obtained, and two stop codons and three rare codons of E. coli were present.

Moreover, the encoding amino acids biases of the above gene sequences were analyzed. Theoretically, for NNS randomization (Figure 2A) Ser, Arg, and Leu have the highest frequency of occurrence, followed by Ala, Gly, Pro, Thr, and Val, while the remaining amino acids have the lowest frequency of occurrence, and such amino acid biases could be eliminated by applying the small-intelligent strategy (Figure 2B). It was found that there is a good correlation between our results and the theoretical amino acid bias expectations for both NNS and small-intelligent randomizations. In our experimental data from NNS randomization, Ser had the highest redundancy, followed by Arg and Leu, while Gln, Gln, Asn, and Lys, which theoretically have the lowest frequency of occurrence, were not obtained (Figure 2C). In comparison, the observed amino acid biases in the small-intelligent library are much less profound than those observed in the constructed NNS mutagenesis library (Figure 2D), and this might become even smaller when the sequencing sample is large enough. Taken together, these results demonstrated that randomization using the small-intelligent strategy could greatly improve the quality of the constructed library by reducing library size and inherent amino acid biases, which would in turn facilitate library screening efficiency.

The utility of the small-intelligent strategy for constructing two-site small-intelligent libraries was also tested when coupled with...
Table 1. The abbreviations for degenerate bases used in this study

<table>
<thead>
<tr>
<th>Degenerate base designation</th>
<th>Actual bases coded</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>A or C or G or T</td>
</tr>
<tr>
<td>B</td>
<td>C or A or G or T</td>
</tr>
<tr>
<td>D</td>
<td>A or G or T</td>
</tr>
<tr>
<td>H</td>
<td>A or C or T</td>
</tr>
<tr>
<td>V</td>
<td>A or C or G</td>
</tr>
<tr>
<td>K</td>
<td>G or T</td>
</tr>
<tr>
<td>M</td>
<td>A or C</td>
</tr>
<tr>
<td>R</td>
<td>A or G</td>
</tr>
<tr>
<td>S</td>
<td>C or G</td>
</tr>
<tr>
<td>W</td>
<td>A or T</td>
</tr>
<tr>
<td>Y</td>
<td>C or T</td>
</tr>
</tbody>
</table>

For example, in AA-Calculator, codon redundancy, stop codons, and the rare codons of host cells could not be completely eliminated from the resulting library in most cases, which could affect the quality of the constructed library.

In this work, we developed a software tool called DC-Analyzer for the design of small-intelligent primers. For one degenerate codon $XYZ(X, Y, Z, \text{represent one of the 15 degenerate nucleotides; see Table 1}$), there are $15^3$ possible variants. After eliminating the three stop codons and eight rare codons of $E. coli$, the remaining 1279 possible variants can be used for small-intelligent primer design.

DC-Analyzer carries out iterations of exploration over the number of degenerate codons. In each round $k$, DC-Analyzer exhaustively adds one of 1279 variants of $XYZ$ to all the former schemes $(XYZ)_i + \ldots + (XYZ)_k$, produced in the previous round without repetition and generates all possible schemes $(XYZ)_i + \ldots + (XYZ)_k$ for this round. Each scheme $(XYZ)_i + \ldots + (XYZ)_k$ contains $k$ degenerate codons and maps to a tuple of $(out_i, f_{ij}, r_{ij}, n_{ij})$, where $out_i$ is 0 when the scheme encodes some undesired amino acids, and is 1 otherwise; $f_{ij}$ is 0 when the frequency of all the codons being covered by the scheme is not the same, and is 1 otherwise; $r_{ij}$ is the ratio of the number of the distinct amino acids that can be encoded by the scheme; $n_{ij}$ is the ratio of the number of codons being covered by this scheme. DC-Analyzer ranks all the generated schemes according to their corresponding tuples in lexicographical order, such that larger is the better.

To make DC-Analyzer more efficient, after the ranking, DC-Analyzer chooses to preserve at most the top 2000 best current schemes for the next round of exploration. This means that the number of possible schemes being generated is significantly reduced. Moreover, to eliminate the repetition of schemes, DC-Analyzer numerates all of 1279 degenerate codons in some order and requires that the order number of degenerate codons being added to a scheme should be strictly increasing. The order of degenerate codons can be described as follows: the order number of $(XYZ)_i$ is less than that of $(XYZ)_j$, if $(XYZ)_i$ contains a superset of codons contained in $(XYZ)_j$, or contains more codons than $(XYZ)_j$. DC-Analyzer was proven to allow the exploration of the best schemes for encoding desired amino acids with the minimal degeneracy in around $20s$ on a single core CPU (T2300, 1.66 GHz) of the Dell Inspiron 640 m computer.

We have demonstrated that saturation mutagenesis libraries with no codon redundancy and amino acid biases could be easily constructed by our newly introduced small-intelligent strategy. Since the only difference between small-intelligent primer and the conventional NNS primer is the codon composition, it is reasonable to assume that any PCR strategies that work with NNS primer are adoptable with small-intelligent primer. Thus, the small-intelligent strategy is feasible for the construction of high-quality saturation libraries and could serve as a good substitute for NNS randomization.

In comparison with the MAX system, the small-intelligent system offers an alternative for producing libraries with the same quality as MAX does, but are less complex and less expensive in most cases, since there are no requirements for ligation, restriction digestion, as well as template synthesis, and fewer oligonucleotides are needed in the small-intelligent system relative to the MAX system, except for the case where multiple contiguous sites are randomized. Most importantly, our system would allow for randomization of multiple contiguous amino acids, although more oligonucleotides would need to be synthesized. However, this situation could be improved by applying the iterative CASTing method (23).

In addition to full randomization, the small-intelligent system should also allow us to perform any restricted randomizations using a specific primer set, and such primer sets could be easily achieved with the assistance of DC-Analyzer. Taken together, the small-intelligent strategy could serve as a simple and valuable tool to construct high-quality focused mutagenesis libraries.

Acknowledgments
This research was supported by the National Natural Science Foundation of China (no. 20872014).

Competing interests
The authors declare no competing interests.

References

Received 1 November 2011; accepted 13 January 2012.

Address correspondence to Lixia Tang, School of Life Science and Technology, University of Electronic Science and Technology of China, No. 4, Section 2, North Jianshe Road, Chengdu 610054, China. e-mail: lixiatang@uestc.edu.cn

To purchase reprints of this article, contact: biotechniques@fosterprinting.com

www.BioTechniques.com