ACS SyntheticBiology

Research Article

pubs.acs.org/synthbio

Reducing Codon Redundancy and Screening Effort of Combinatorial Protein Libraries Created by Saturation Mutagenesis

Sabrina Kille,[†,‡] Carlos G. Acevedo-Rocha,[†,‡] Loreto P. Parra,[†,‡] Zhi-Gang Zhang,[†,‡] Diederik J. Opperman,[†] Manfred T. Reetz,[†,‡] and Juan Pablo Acevedo*[,§]

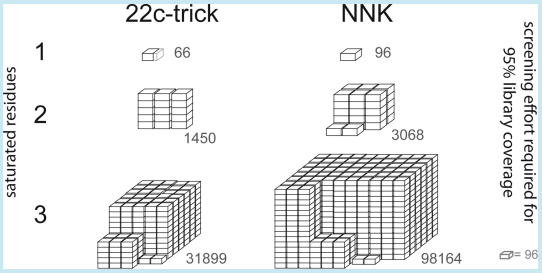
[†]Max-Planck-Institut für Kohlenforschung, Kaiser-Wilhelm-Platz 1, 45470 Mülheim an der Ruhr, Germany

[‡]Fachbereich Chemie, Philipps-Universität Marburg, Hans-Meerwein-Straße, 35043 Marburg, Germany

[§]Facultad de Medicina y Facultad de Ingeniería de la Universidad de los Andes, Santiago, Chile

Supporting Information

ABSTRACT: Saturation mutagenesis probes define sections of the vast protein sequence space. However, even if randomization is limited this way, the combinatorial numbers problem is severe. Because diversity is created at the codon level, codon redundancy is a crucial factor determining the necessary effort for library screening. Additionally, due to the probabilistic nature of the sampling process, oversampling is required to ensure library completeness as well as a high probability to encounter all unique variants. Our trick employs a special mixture of three primers, creating a degeneracy of 22 unique codons coding for the 20 canonical amino acids. Therefore, codon redundancy and subsequent screening effort is significantly reduced, and a balanced distribution of codon per amino acid is achieved, as demonstrated exemplarily for a library of cyclohexanone monooxygenase. We show that this strategy is suitable for any saturation mutagenesis methodology to generate less-redundant libraries.



KEYWORDS: *codon redundancy, primer degeneracy, screening effort, library quality, directed evolution, combinatorial mutagenesis*

Saturation or cassette mutagenesis[1−3] is a powerful tool in protein−protein interaction studies,[4] protein engineering,[5] and specially directed evolution[6−11] because it allows the focused exploration of defined segments of the vast protein sequence space. Saturation mutagenesis can be performed at one or multiple amino acid residues simultaneously, leading to libraries with diverse protein variants.[12−14] Importantly, cooperative (non-additive) effects occur only when saturating simultaneously more than one amino acid residue.[15,16] However, if many positions are targeted simultaneously, the diversity of a combinatorial library can have a staggering vastness.[17] For example, the saturation of 10 amino acids using the 20 canonical ones has $20^{10} = 1.024 \times 10^{13}$ possible combinations! There are several strategies to overcome this "numbers problem".[18] Obviously the most attractive is the application of positive or negative selection strategies during screening,[19] but these are often restricted to case-specific enzymes (e.g., aminoacyl-tRNA synthetases) and are not generally applicable. Other strategies include display technologies (e.g., ribosome, phage, bacteria, and yeast) where huge libraries can be generated and screened in a single experiment.[20] However, these require sophisticated equipment not accessible to many researchers. Other more general strategies for reducing the overall number of variants to a more reasonable number are based on eliminating the genetic code redundancy by using limited amino acid sets or alphabets.[4,18,21−23] Reduced semirationally selected amino acid alphabets avoid the generation with libraries of infinite size, making this approach certainly more efficient when compared to random ones.[24−28] Efficient experimental semirational approaches include Gene Site Saturation Mutagenesis (GSSM),[29] Structure-based COmbinatorial Protein Engineering (SCOPE),[30] and Combinatorial Active-site Saturation Test (CAST),[31] which can be systematized in the form of Iterative Saturation Mutagenesis (ISM).[32] Computer-based semirational approaches are also useful to create combinatorial libraries.[24,33−37] Yet library design is often suboptimal,[38] which can lead to unnecessary screening, waste of resources, and misinterpretation of results. This is partially caused by the redundancy of the genetic code because the 20 canonical amino acids are disproportionally coded by 61 sense codons, i.e., some amino acids are encoded only by a single codon (e.g., methionine and tryptophan), while others are encoded by up to six different codons (e.g., arginine, serine, and leucine). Thus, the proportion of highly encoded amino acids will be much higher than the lesser (under) represented ones when increasing the sites of randomization. For example, the theoretically ratio of serine to tryptophan is 36:1 and 216:1 for randomization sites comprising two or three amino acids positions, respectively.[39]

Received: April 20, 2012
Published: June 15, 2012

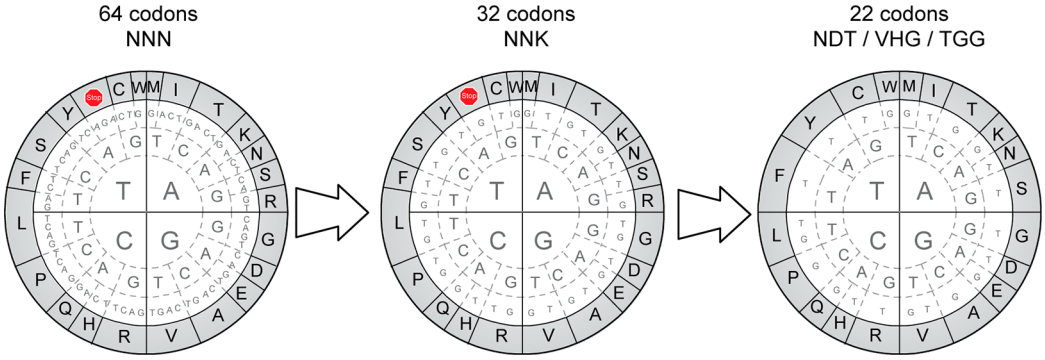


Figure 1. Different redundancies of the genetic code encoding all 20 amino acids represented in sun format. The redundancy of the genetic code can be reduced from 64 codons (left) to 32 codons (center) using a NNK/S degenerate primer or even further to 22 codons (right) using the appropriate combination of degenerate primers NDT (N = A/T/C/G, D = no C) and VHG (V = no T, H = no G) with a non-degenerate TGG (W; tryptophan) primer. In this radial or sun representation of the genetic code, the codons are read from the most inner circle to the outside. Encoded amino acids are presented in one letter code in the outer gray shell.

To reduce the amino acid bias of the genetic code, the most common approach is to generate libraries using degenerate (formerly referred to as contaminated, doped, spiked) oligos that can be produced during their chemical synthesis. To fully explore the probed protein sequence space, the degenerate codons NNK or NNS (N = A/T/G/C; K = T/G and S = G/C) are normally chosen, because they encode all 20 amino acids with the lowest redundancy and price (a single oligo). However, because the NNK/S degeneracy still contains three codons for arginine, leucine, and serine and two codons for the five amino acids alanine, glycine, proline, threonine, and valine, the redundancy is not completely eliminated and a certain amino acid bias is still present.

To overcome these drawbacks some techniques make use of specially prepared phosphoramidite solutions of mono- (known as MAX),[39] di-,[40] or trinucleotides[13,41] during the synthesis of the growing oligonucleotide. However, although these techniques eliminate the redundancy and provide full randomization with 20 codons, none is routinely used for saturation mutagenesis due to practical reasons such as the requirement of a DNA synthesizer or high special handling charges. Furthermore, gene synthesis with a defined set of 20 codons per saturated residue still remains expensive and is therefore not ideal for practical, routinely applications.

As already pointed out, another strategy to obtain redundancy-free libraries is a more stringent restriction of the probed part of the sequence space.[4,18,21−23] This can be achieved by using degenerate codons encoding amino acid alphabets of less than 20 members exhibiting certain properties such as hydrophilicity (VRK, 12:8 codons:amino acids), hydrophobicity (NYC, 8:8 respectively), small size (KST, 4:4 respectively), charge (RRK, 8:7 respectively), or balanced nature (NDT, 12:12 respectively).[22,42,43] Indeed, we have repeatedly employed NDT and other codon degeneracies in the successful quest to enhance enantioselectivity and rate of different enzymes, demonstrating the use of reduced amino acid alphabets for reducing the screening effort drastically.[7,18] Information obtained by bioinformatic techniques such as database analysis can help in identifying the codon degeneracy of choice, an approach that was first demonstrated for a Baeyer−Villiger monooxygenase[23] and later for a hydrolase.[44] However, if the targeted region of a protein is of unknown function, it may be advisable to saturate with all 20 amino acids.[45]

Ideally, the number of codons should be equal to the number of amino acids, as has been shown practically[18] and mathematically.[46] Since all available degenerate bases[47] and triplet combinations thereof are not able to encode the 20 canonical amino acids in a single degenerate primer,[22,42,43] we realized that the combination of more than one degenerate primer could be a simple solution. Previously, we partially eliminated codon redundancy and subsequently screening effort by mixing in equimolar concentrations nine defined primers, from which eight encoded a specific codon and one primer carried the NDT degeneracy (12 codons), thereby generating a codon to amino acid ratio of 20:20.[48]

In the present paper, we have generalized this concept by mixing conventional degenerate oligonucleotides, which allows the researcher to create saturation mutagenesis libraries in which the number of gene sequences is almost equal to the number of protein sequences covering all 20 amino acids. We compare our approach with the commonly used NNK degeneracy not only by applying a statistical analysis for library coverage to explore the effect of codon redundancy removal on screening effort but also by experimentally validating the effectiveness of our approach using a model stereoselective enzyme. Finally, we applied this simple trick when creating randomized libraries at one and two positions in several proteins at both optimal and suboptimal conditions affecting library quality, followed by a quality control analysis, an important factor often neglected in directed evolution studies.

RESULTS AND DISCUSSION

Reducing Codon Redundancy. A common approach to reduce the redundancy in the genetic code, while saturating all 20 amino acids, is the use of NNK/S degenerate primers encoding 32 distinct codons. This 2-fold reduction from the original 64 codons can be further reduced to 22 codons by mixing a total of three oligonucleotides: two degeneracy carrying primers, one with NDT (12 unique codons) and the other with VHG (9 unique codons), and a TGG containing primer (one codon). We call this the "22c-trick", which reduces the genetic code redundancy up to a codon to amino acid ratio of 22:20. This mixture contains no stop codons and only two redundant sets for valine (GTT, GTG) and leucine (CTT, CTG). The stepwise reduction of codon redundancy for all canonical amino acids is schematized in Figure 1.

dx.doi.org/10.1021/sb300037w | ACS Synth. Biol. 2013, 2, 83−92

To generate a 22c-trick library, the total number of synthesized oligonucleotides depends on (i) the technique employed, e.g., QuikChange(QC),[49] MegaPrimer (MP),[50] or overlap-extension PCR (OE-PCR);[51] (ii) the number of residues to be saturated; and (iii) the distance between these residues in the gene sequence. The distance between the individual residues often dictates what technique is more suitable. For a single amino acid residue, 3 forward or reverse primers harboring the NDT, VHG, and TGG combinations could suffice for MP and OE-PCR, but 6 primers are necessary for other techniques such as QC. If two or three residues of close proximity are randomized, a total of 9 or 27 primers, respectively, need to be synthesized for techniques using only a sense or antisense primer, but these numbers double to 18 and 54 when both primers are needed. Supplementary Table S1 lists the necessary ratios for mixing primers containing the NDT, VHG, and TGG combinations to generate libraries targeting one, two, or three residues using the 22c-trick.

We are using our technique in several ongoing directed evolution projects, and during the writing of the present paper Tang et al.[52] reported a similar strategy termed "small-intelligent". They used instead 4 primers with two degeneracies (NDT, VMA) and two coding sequences (ATG, TGG), reducing completely codon redundancy while targeting all 20 amino acids. The attractive feature of this approach lies in a complete theoretical absence of amino acid bias, and both degeneracies are suitable for *E. coli* codon usage. Nevertheless, both approaches have advantages and disadvantages. The ideal case of a codon to amino acid ratio of 20:20 provides the library with the smallest possible set of combinatorial variants. Obviously, the total number of primers should be as low as possible, since this grows with *n*-residues to be saturated, thereby increasing exponentially the costs of synthesis. Certainly, for one site the "small-intelligent" approach requires 4 equimolar mixed sense or antisense primers, but the number can double to 8 depending on the technique. Similarly, a significantly higher number is required when targeting two or three residues: 16 or 64 in the case of sense or antisense and 32 or 128 sense and antisense individually synthesized primers, respectively. In either strategy the reduction of codon redundancy is essential for reducing the screening effort of any designed library.

Reducing Screening Effort. Screening effort refers to the number of samples required to analyze the protein sequence space targeted by saturation mutagenesis. Due to the stochastic nature of the sampling process, the number of samples, i.e., the screening effort, must be properly defined. This can be theoretically calculated with one of the two following concepts. *Library Coverage* deals with the number of colonies obtained after transformation, i.e., the proportion of the probed variant space (total amount of theoretical possible variants) of a generated library and the proportion of the variant space covered by picking a certain number of samples. Another factor, the *full coverage probability*, calculates the likelihood of sampling the complete variant space. These factors, have been mathematically developed[46,53,54] and exemplified.[38] This assumes that all variants have a defined probability to be present (independent of biological factors and practical conditions), which is certainly not the case in library generation (*vide infra*). Whereas a high *full coverage probability* requires between 10- to 25-fold oversampling, only a factor close to 3-fold is required for 95% *library coverage* of the variant space.[53,54] Obviously, the former oversampling numbers are beyond technical[55] and physical (amount of DNA)[38] limitations, and some researchers prefer using instead the oversampling factor of ~3 to determine library size.[18] Using this factor, library size can be calculated with the following formula:[54]

$$L = -V \ln(1 - F) \quad (1)$$

where L = number of samples, library size or screening effort; V = total number of possible variants X^n (where X and n denote the number of codons and saturated residues, respectively); and F = fractional library completeness, e.g., 0.95 for 95%. For example, when one amino acid residue is targeted by NNK/S, 96 colonies [$L = -32^1 \ln(1 - 0.95)$] are necessary to cover 95% of the variants. In contrast, when using the 22c-trick, only 66 (~3 times 22) colonies need to be sampled. These numbers also indicate the minimum number of colonies that a given transformation should yield; otherwise the library would contain less than 95% of the variant space.

Importantly, the redundancy of the genetic code blows up the size of the library, since various codons encoding the same amino acids and junk sequences such as stop codons are not eliminated in NNK/S or NNN.[18] With the removal of almost all redundancy, the screening effort is significantly decreased by 32% using our approach when saturating one residue. Similarly, if a library is diversified to NNK/S at two positions, it will contain 1024 (32^2) sequences but 400 unique variants. Thus, in order to screen a fraction of 95% of these, 3068 (~3 times 1024) clones must be sampled (Figure 2). Here again, if our

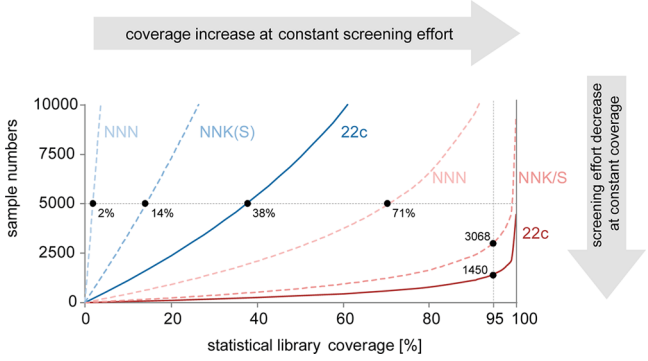


Figure 2. Screening effort required for different randomization schemes regarding sites composed of 2 or 3 amino acid residues. The choice of codon degeneracy dictates the sampling size for a desired statistical coverage of the library. For a 95% *library coverage* targeting two amino acid residues (red lines), 3068 samples have to be screened in the case of NNK/S, whereas only 1450 are necessary when applying the 22c-trick (53% lower screening effort). However, if the assumed capacity of medium-throughput systems is limited to 5000 samples, the *library coverage* drops to 71% when using NNN degeneracy. Similarly, when targeting three amino acid residues (blue lines) and limiting the sample size to 5000 colonies or transformants, the *library coverage* changes drastically to 38%, 14%, and 2% in the case of the 22c-trick, NNK/S, and NNN, respectively.

22c-trick is instead applied for the same coverage, it means that only 1450 [~3 times 484 (22^2)] colonies need to be sampled (Figure 2). Thus, the 22c-trick decreases the screening effort by 53% for a 2-residue site.

Whereas the above numbers can usually be handled by medium-throughput assays such as automated GC or HPLC, any reduction, *ideally having the same library coverage*, would be highly desirable since screening is the bottleneck in most directed evolution studies.[55] If three sites are simultaneously

dx.doi.org/10.1021/sb300037w | *ACS Synth. Biol.* 2013, 2, 83−92

randomized by traditional NNK/S, about 1×10^5 transformants would need to be screened for 95% *library coverage* and 3×10^4 using the 22c-trick degeneracy. One could limit screening to 5000 colonies, for example, but this would mean that only 14% of all variants are present in the library. In the case of the 22c-trick, this number increases to 38%, thus allowing a better exploration of the relevant protein sequence space (Figure 2). Figure 2 convincingly demonstrates how effective our 22c-trick reduces screening effort that otherwise would rise exponentially, making sample sizes impractical to handle. Of course, one can simply ignore such statistical considerations and screen a smaller sample number. However, this would result in libraries of low diversity as only a very small fraction of the variants is present (Supporting Table S2).

Very recently, Nov[46] reported an interesting mathematical analysis that demonstrates that striving to find the best variant (i.e., the sequence with the highest fitness value) is *de facto* not necessary, since it requires having a high *full coverage probability*. He argues that finding one of the two or three best variants is good enough and more advantageous in practical terms due to a lower screening effort. In fact, to ensure 95% probability of discovering at least one of the top two variants when applying the 22c-trick to one single residue, Nov's analysis recommends sampling 30 colonies. This number is reduced marginally to 29 when considering Tang's "small-intelligent" library. This analysis, however, has to be taken cautiously into consideration because it assumes that the probed protein sequence space is smooth or "Fujiyama-type". Finally, although not considered here, a completely different approach to reduce the screening effort is the pooling of mutant libraries,[56] as we have demonstrated elsewhere.[48]

Comparison of 22c-trick and NNK Libraries. Two saturation mutagenesis libraries were created based on cyclohexanone monooxygenase (CHMO) by randomizing residue Leu426, one using the 22c-trick and the other with the traditional NNK codon degeneracy (see Methods). Upon transformation, both libraries yielded several thousand colonies, from which a total of 92 and 144 colonies from the 22c-trick and NNK libraries, respectively, were randomly chosen for sequencing. A total of 3 and 14 samples, respectively, either failed sequencing or did not exhibit the correct gene construct. The respective library completeness is therefore calculated as 98.9% and 98.2%, using the equation for fractional completeness by Patrick et al.[54] The sequencing of individual clones revealed the following diversity in both libraries (Figure 3).

In the case of the 22c-trick library, all 20 expected amino acid variants were found. In contrast, two amino acids (aspartate and isoleucine) were not present in the NNK library. The observed diversity did not provide sufficient evidence to claim whether the observed distribution of codons is normal or whether an experimental parameter has biased it. Since sampling is a random process, a statistical judgment became necessary. The occurrence of a specific codon is a success/failure experiment, i.e., the codon is either found or not found in the library. These experiments are called Bernoulli experiments and are described by a binomial distribution. Accordingly, the statistical hypothesis test χ^2 (chi square) can be used to evaluate if the experimentally observed occurrence of codons (Figure 3) is comparable to the theoretical expected codon distribution (Supplementary Figure S1). If no experimental factors (e.g., annealing bias, suboptimal primer synthesis, or insufficient *Dpn*I digestion) have biased codon diversity in the given library, the χ^2 test will confirm the

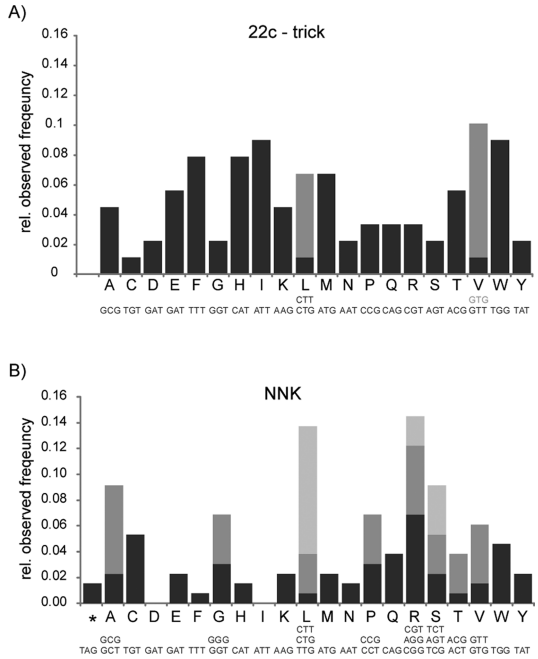


Figure 3. Relative amino acid frequencies in combinatorial libraries of CHMO at position Leu426. Unique nucleotide and amino acid sequences were obtained for (A) 22c-trick and (B) NNK degeneracies from 89 and 130 colonies, respectively. Stacked bars of various gray colors represent redundancy for the particular amino acid. Alphabetically sorted amino acids are given in one letter code (black) with its corresponding codon below. The stop codon is represented by a star.

probability of success for each codon. In the case of the 22c-trick library, the probability of success is $p = 1/22$, whereas it is $p = 1/32$ for each codon present in the NNK library.

The application of the χ^2 test to the 22c-trick library confirms that in this particular sampling experiment the difference between theoretical expected and experimental observed distribution is not quite statistically significant (see Supporting Information). Thus, the observed distribution of codons in the 22c-trick library is within statistical expectations and the library is unbiased on the codon level. Moreover, since there is only little redundancy, the generated 22c-trick library is unbiased at the amino acid level as well.

As can be seen in Supplementary Figures S1 and S2, the event that a codon appears zero times (i.e., absent in the created libraries) occurs with a 2% probability. The application of the χ^2 test to the NNK-sequenced library revealed an experimental bias because the difference between theoretically expected and experimentally observed occurrence of codons was found as very statistically significant (see Supporting Information). A Grubb's test for outlier identification was performed, identifying the appearance of 13 CTT codons as statistical outlier. Conclusively, it can be stated that an experimental parameter has biased the NNK library toward the very unlikely occurrence of 13 WT codons. Nonetheless, repeating the χ^2 test without the CTT data point confirmed that the remaining codon diversity of the NNK library is within the expected statistical distribution and therefore not biased on the codon level. The existence of redundant amino acids with two (A, G, P, T, V, R) or three (R, L, S) codons biases the library at the amino acid level, compromising library quality in terms of diversity (Figure 3B). A less redundant library will always have a higher diversity and a higher probability to find

more unique variants. However, if the *library coverage* would be lowered by sampling fewer colonies, it would become more probable to miss more amino acids. A smaller sampling size will always compromise the diversity and hence the quality of the library (Supplementary Figure S2).

Importance of Quality Control in Library Creation. The individual sequencing of clones from the Leu426 libraries served a second purpose: We wanted to investigate whether our routinely applied "Quick Quality Control" (QQC)[48,57] is useful for evaluating the diversity and hence the quality of a determined library before screening. Briefly, after retransforming cells with the generated libraries, all colonies are scratched from the agar plate with a Drigalski spatula, and plasmid DNA is isolated. Thereafter, the pool of plasmid DNA belonging to all clones is sequenced in a single run and analyzed to determine whether the degeneracy is successfully introduced and whether removal of the WT-sequence is achieved. For a successful library creation, the experimental distribution of bases in the target codon should be very similar to the expected percentages. The distribution of bases for each position obtained from the individual clones of CHMO was calculated and compared to the theoretical (see Supplementary Table S3) and experimental QCC distributions (Figure 4).

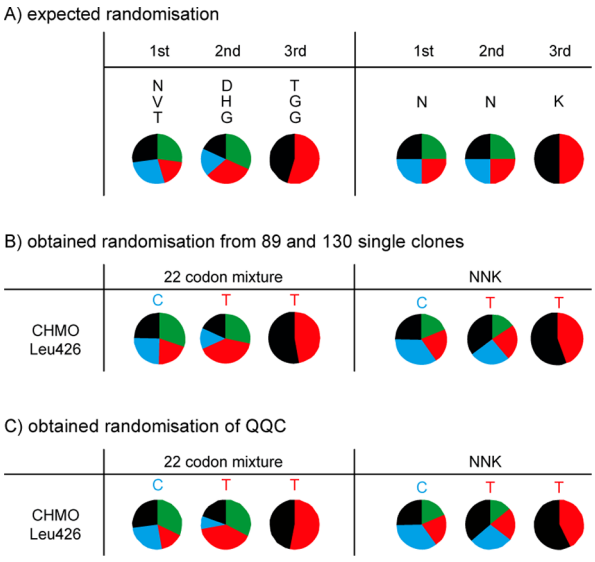


Figure 4. Distribution of nucleotide bases in the randomized residue Leu426 of CHMO. The percentual distribution of nucleotides is shown in pie diagrams for each of the three randomized bases using the 22c-trick (left) and NNK (right) degeneracies. (A) Theoretical expected distribution. (B) Experimental distribution calculated from the sequencing of 89 and 130 individual clones from the 22c-trick and NNK libraries, respectively. (C) Experimental Quick Quality Control from colony pooling. The nucleotide base guanidine (G) is depicted in black, adenosine (A) in green, threonine (T) in red, and cytosine (C) in blue.

Upon comparing the distribution of the encountered bases from both the individual clones (Figure 4B) and the QQC (Figure 4C) with the expected values (Figure 4A), it becomes apparent that our simple QQC is a reliable, quick, and cost-efficient method to assess library quality because the distribution of bases is virtually the same in all cases. This is of particular importance before starting any screening effort, according to the motto "you should not search for something that does not exist".[48] Nevertheless, it should be noted that the QQC requires a minimum amount of transformants for a sample to be representative. About 50−100 colonies are enough for saturation at one residue, but at least 500−1500 colonies is the minimum for combinatorial libraries of two and three residues.

Creation of Other Libraries Using the 22c-trick. We have successfully applied the 22c-trick for the generation of libraries using saturation mutagenesis at one and two residues in ongoing directed evolution studies. Since we did not aim to sequence individual clones in the following examples, we limited ourselves to perform the QQC and to estimate whether the base distribution at the targeted codons using the 22c-trick is comparable to the theoretical values.

The genes coding for CHMO and phenylacetone monooxygenase (PAMO), in appropriate vectors, were used as templates to randomize target positions either by QC (single sites and two consecutive residues) or MP (two distant residues) PCR-based methods. The overview of three single-residue and five double-residue saturation libraries in terms of QQC is summarized in Figure 5. In the case of single residue saturation of CHMO Ala146, a high dominance of WT bases for the first and second nucleotide can be observed, even though the third base is close to perfection (Figure 5, entry 1). Individual randomization of Phe432 and Thr433 of the same protein, though, proved to be more successful (Figure 5, entryies 2 and 3).

Apparent is the poor saturation result of Ser441 in the two-residue containing PAMO library (Figure 5, entry 4), where the second and third bases contain dominantly the WT bases.

The general high frequency for guanidine in all four PAMO residues in the first position is noteworthy (Figure 5C). Another interesting observation can be made for the two-residue libraries of CHMO (Figure 5D): All of the first codons have apparently low or no amount of cytosine in the second base position (Figure 5, entries 6, 7, and 8). From the first codon of entry 7 and second codon of entry 8, it can be concluded that the *Dpn*I digestion of the template was complete, since no A or C was found in the third base. Nevertheless, a high tendency toward incorporation of WT and WT-related bases can be generally observed. The application of the QQC is therefore essential to estimate the generated library diversity in saturation mutagenesis experiments.

Unfortunately, with some exceptions,[45,48,57,58] this or similar tests are often not reported in directed evolution studies, in contrast to other fields such as antibody research.[59−61] Screening of non-created diversity is not only useless, it can also lead to wrong interpretations and conclusions of the results. In fact, we have realized the need to optimize library creation herein because there are many factors that influence the construction of an optimally diverse library. Since other degeneracies[45,48,57,58] result in relatively good or poor QQC, primer degeneracy is not the main issue affecting library diversity in PCR-based saturation mutagenesis protocols (Supplementary Figure S3).

Influence of Other Parameters on the Quality of 22c-trick Libraries. The traditional process of library creation *via* saturation mutagenesis has proven to be successful, but it is far from being perfect. Any improvement in library quality adds to the importance of this gene mutagenesis method. A critical factor for achieving a good quality library is the method of randomization. Most of the tools to generate saturation mutagenesis libraries rely on PCR. The most common one is QC due to its simplicity as one needs only a sense and

dx.doi.org/10.1021/sb300037w | *ACS Synth. Biol.* 2013, 2, 83−92

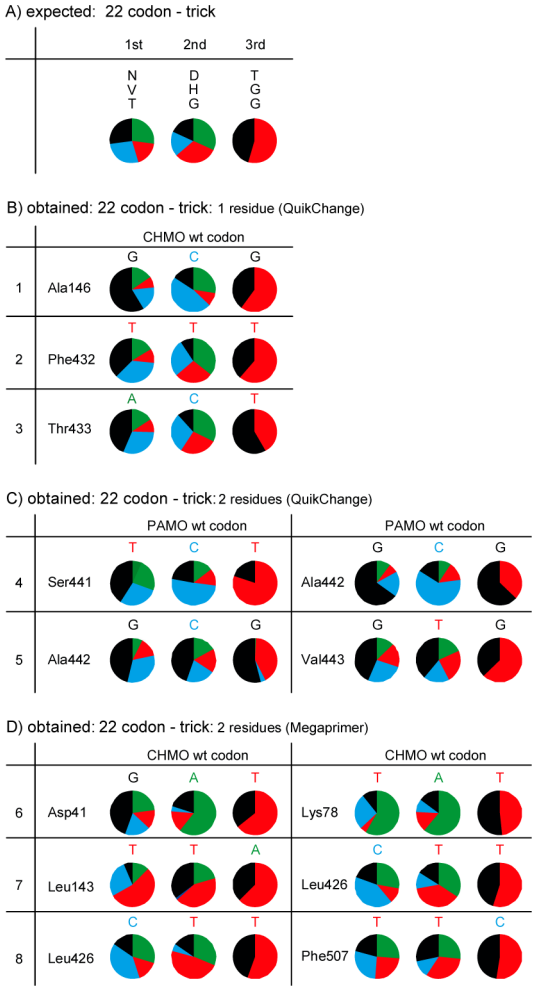


Figure 5. Quick quality controls of 8 libraries generated with the 22c-trick. (A) Expected distribution of nucleotides. (B) Obtained distribution for single residues randomization on CHMO with QuikChange (QC). (C) Two-site PAMO libraries created with QC. (D) Two-site CHMO libraries created with MegaPrimer PCR. The WT codon is presented above the pie diagrams. The nucleotide base guanidine (G) is depicted in black, adenosine (A) in green, threonine (T) in red, and cytosine (C) in blue.

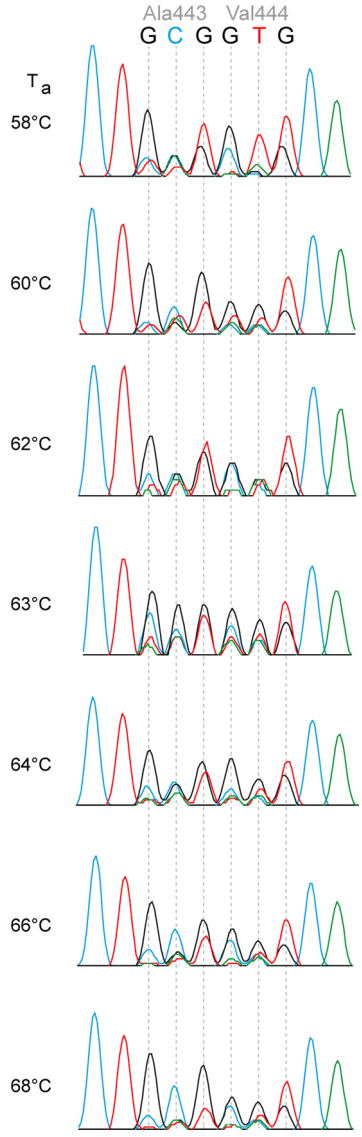


Figure 6. Influence of annealing temperature on target codon randomized with the 22c-trick. Residues Ala443(GCG) and Val444-(GTG) of PAMO were randomized simultaneously via QuikChange. The DNA electropherograms are the result of a QQC upon pooling more than 1000 colonies. The nucleotide base guanidine (G) is depicted in black, adenosine (A) in green, threonine (T) in red, and cytosine (C) in blue.

antisense primer. Other methods include MP or OE-PCR, where one, two, or more primers can be degenerate.

Our experience has taught us that it is not possible to obtain a perfect library, irrespective of the method used. Depending on how well the overall PCR process is optimized, we have observed libraries with very good quality (Figure 5) but also very low qualities (Supplementary Figure S3) as judged by our QCC. Several factors affect library quality, for example, gene size and GC content,[57] the position and sequence of the target codon in the gene,[60,62] melting and consequently annealing temperature,[60,61] as well as primer length and synthesis.[59,63]The adjustment of the annealing temperature (T_a) is a key factor for optimizing library quality, as we have observed when creating a two-residue PAMO library at Ala443(GCG) and Val444(GTG) using different T_a's (Figure 6). The T_a depends on the primer melting temperature (T_m). Since degenerate bases at specific locations are mixtures, the T_m is in reality a range of T_m's. In this example, the difference between the primer with the lowest and that with the highest T_m was calculated to be 7 °C. We observed that changing the T_a of the PCR has important consequences in terms of quality: The codons with GXG and GXT anneal predominantly at both positions but more strongly at residue 443. The first position of both residues, in particular, is highly dominated by G. This kind of WT-codon controlled annealing bias has been reported by Airaksinen et al.[60] and has a pronounced influence on the outcome of the saturation mutagenesis library. Bias toward the WT and WT-related codons can be overcome by modifying the ratio of phosphoramidite mixtures to deplete the WT codon from the resulting degeneracy.[60] Also apparent is the rare appearance of the adenosine bases in the first position of codon 443, resulting in rare sequences for all AXX codons (here Ile, Asn, Ser, Met, Thr, and Lys; Figure 6). Therefore, during the optimization of methodological conditions for library creation and indeed whenever applying any kind of saturation mutagenesis, it is necessary to check the quality of the obtained mutant libraries.

dx.doi.org/10.1021/sb300037w | ACS Synth. Biol. 2013, 2, 83−92

CONCLUSION

When creating a molecular diverse combinatorial library for any kind of study involving proteins, parameters such as codon degeneracy, library completeness, and oversampling, all essential for correct data interpretation, must be seriously considered for ensuring maximum efficiency.

In the present study we introduce the 22c-trick, which consists of a 22:20 codons to amino acid mixture of two primers bearing degenerate bases and one containing the TGG codon, as an efficient strategy to reduce codon redundancy in saturation mutagenesis. We have also compared our approach to Tang's "small intelligent" strategy, both very closely related. In the latter method, the elimination of amino acid bias goes further than our strategy, but it requires a higher number of primers, especially when randomizing sites composed of two or more amino acid positions. It is in the hands and budget of the experimenter to decide which strategy to choose since both have similar advantages and disadvantages. Using our trick, nevertheless, we demonstrated the significant reduction of screening effort by removing *almost* all codon redundancy. Our 22c-trick has a significant advantage compared to classical randomization with NNK/S. Unrestricted sequence space exploration with all 20 canonical amino acids is possible with a >50% reduced screening effort for two or three residues. This enables researchers, especially in the fields of protein engineering and specifically directed evolution, to screen faster and explore more efficiently the sequence space of important proteins. The 22c-trick provides an alternative to the currently standard approach for reducing screening effort, which involves a limitation of sequence space by using a smaller set of amino acids as defined by the respective codon degeneracy (e.g., 12 amino acids as given by NDT).[18,22,42,43] In addition, the balanced set of codons in the 22c-trick with its lower redundancy will always create libraries with a higher diversity, because more unique variants are present in a fixed number of probed samples. Therefore applied with FACS (1×10^7) or growth-based (1×10^{12}) selection systems,[5] it will allow the full randomization of one additional residue compared to NNK/S, increasing the total number of residues manageable for saturation to five and eight, respectively.

Library design is a balanced act between library quality, library diversity, and library completeness. With this in mind, we additionally constructed several single and double saturation mutagenesis libraries with our 22c-trick degeneracy and investigated their quality with a reliable and cost-efficient (only one sequencing run) Quick Quality Control. The utility of this test shows that saturation mutagenesis seldom creates the complete desired diversity in a perfect manner. Of course, this applies to any form of saturation mutagenesis, some more, some less. The important point for those engaging in applications is simple: Invest your time and efforts optimally by choosing the best strategy. Therefore, we emphasize again how essential the quality control of mutant libraries is. Successful diversity creation by saturation mutagenesis should not be taken blindly.

METHODS

KOD Hot Start DNA polymerase was purchased from Novagen (Merck KGaA, Darmstadt, Germany), *Dpn*I was from New England Biolabs GmbH (Frankfurt am Main, Germany), and desalted primers were obtained from Life Technologies GmbH (Darmstadt, Germany).

Mutant Libraries of the CHMO (Cyclohexanone Monooxygenase Gene from *Acinetobacter* sp. strain NCIMB 9871).[64] Library creation Leu426-NNK was performed using the QuikChange PCR method with pET22b(+)-CHMO-V40 template and mutagenic primers (desalted, Life Technologies) as described in Supplementary Table S4. The total reaction volume of the PCR reaction was 20 μL. To a volume of 11.3 μL of Millipore-Q water were added in this order 2 μL of 10x-KOD Hot Start DNA polymerase buffer, 0.8 μL of $MgSO_4$ (25 mM), 2 μL of dNTP mix (2 mM each), 0.7 μL of forward primer (10 μM), 0.7 μL of reverse primer (10 μM), 2 μL of template (25 ng/μL), and 0.5 μL of KOD Hot Start DNA Polymerase (1.0 U/μL). The PCR temperature program was 3 min at 95 °C, followed by 27 cycles of 1 min, 95 °C denaturing; 1 min, 55 °C annealing and 8 min, 68 °C extension. Final extension was carried out for 10 min at 72 °C. Methylated template was removed by *Dpn*I digestion (2.5 h, 37 °C, 16 μL of PCR sample, 1 μL of *Dpn*I (20 kU/μL), 1 μL of NEB4, 3 μL of water). The sample was dialyzed against Millipore-Q water for 30 min on Millipore MF-membrane filters (0.05 μm).

22c-trick libraries were created for single-residue saturation by QuikChange and for double-residue sites by the MegaPrimer method using recombinant plasmid pET22b(+)-CHMO or pET22b(+)-CHMO-V40 for Leu426 saturation as the template and mutagenic primers as described in Supplementary Table S4. The PCR reaction mixture (25 μL final volume) contained 2.5 μL of 10x-KOD Hot Start DNA polymerase buffer, 2.5 μL of dNTP mix (2 mM each), 0.8 μL of $MgSO_4$ (25 mM), 0.5 μL of template (30 ng/μL), 0.7 μL of primers (10 μM each mix), and 1 μL of KOD Hot Start DNA polymerase (1.0 U/μL). The PCR temperature program consisted of an initial cycle at 95 °C for 3 min, followed by 20 cycles of denaturing at 95 °C for 1 min, annealing at temperatures given below for 1 min, and elongation at 72 °C for 8 min with a final extension at 72 °C for 12 min. The following annealing temperatures were used for the creation of libraries: Ala146: 52, 53, and 55 °C; Leu426: 52 and 54 °C; Phe432: 52, 54, and 56 °C; Thr433: 52, 54, and 56 °C; Asp41/Lys78: 52 and 54 °C; Leu143/Leu426: 52, 54, and 57 °C and library Leu426/Leu505: 52 and 54 °C. The PCR samples were mixed, and the template was removed by adding 1 μL of *Dpn*I (20 U/μL) to the sample followed by incubation at 37 °C for 4 h. After transformation of *E. coli* BL21-Gold(DE3) cells (25 μL) by electroporation with 1 μL of sample, cells were suspended in 1 mL of SOC medium, incubated for 1 h at 37 °C, and plated on LB-agar containing 100 μg/mL carbenicillin. For CHMO library Leu426-22c-trick and Leu426-NNK, 144 and 92 colonies were picked, and plasmid DNA preparation and sequencing analysis was performed in 96-well plate format by GATC biotech.

Mutant Libraries of PAMO (Phenylacetone Monooxygenase Gene from *Thermobifida fusca*).[65] Library creation was performed using the QuikChange PCR method, saturating 2 residues simultaneously with the mutagenic primers described in Supplementary Table S4. The recombinant plasmid pBAD-PAMO-VQ was used as template. The amplification reaction contained in a total volume of 20 μL was 2 μL of 10x-KOD Hot Start DNA polymerase buffer, 2 μL of dNTP mix (2 mM each), 0.8 μL of $MgSO_4$ (25 mM), 0.7 μL of primers (10 μM each mix), template plasmid (50 ng), and 0.5 μL of KOD Hot Start polymerase (1.0 U/μL). The PCR conditions were 1 cycle at 95 °C for 3 min; 27 cycles of

dx.doi.org/10.1021/sb300037w | *ACS Synth. Biol.* 2013, 2, 83−92

denaturing at 95 °C for 1 min, annealing for 1 min with temperatures stated below, and extension at 68 °C for 8 min. Final extension step was carried out at 68 °C for 16 min. Eight different temperatures were assayed for the annealing step for library PAMO-Ala443/Val444: 58, 60, 62, 63, 64, 66, and 68 °C. PAMO-Ser442/Ala443 library was generated with an annealing temperature of 65 °C. The PCR products were digested with *Dpn*I by adding 1 μL of *Dpn*I (20 U/μL) to the PCR sample and incubating the reaction at 37 °C for 1.5 h, followed by another 1 μL of *Dpn*I addition to the reaction, and the incubation was continued for 1.5 h. After transformation of *E. coli* TOP10 cells (25 μL) by electroporation with 2 μL of sample, cells were suspended in 1 mL of SOC medium, incubated for 1 h at 37 °C, and plated on LB-agar containing 100 μg/mL carbenicillin.

Generation of P450-BM3 Libraries. The P450-BM3 from *Bacillus megaterium* was randomized via a one-step MegaPrimer protocol as reported elsewhere.[43] Primers were ordered cartridge-purified from Metabion (Martinsried, Germany). Briefly, in each of the four 50 μL PCR reactions, the following components were added: 32.5 μL of Millipore-Q water, 5 μL of 10x-KOD Hot Start DNA polymerase buffer, 3 μL of $MgSO_4$ (25 mM), 5 μL of dNTP mix (2 mM each), 2 μL of forward (silent) primer (20 μM), 4 μL of reverse (mutagenic) primer (20 μM), 1 μL of template (25 ng/μL), and 0.5 μL of KOD Hot Start DNA Polymerase (1.0 U/μL). The program started with 3 min at 95 °C, followed first by 5 cycles of 95 °C (30 s), 50 or 60 °C (1 min), and 72 °C (5 min) and then 20 cycles of 95 °C (1 min) and 72 °C (12 min), ending up with 72 °C (10 min) and subsequent cooling. The methylated template was digested with 1 μL of DpnI (20 kU/μL) for 2 h at 37 °C in suitable buffer, followed by dialysis against Millipore-Q water using Millipore MF membrane filters (0.05 μm) for 30 min. Finally, 25 μL of *E. coli* BL21-Gold(DE3) cells were electroporated with 2 μL of dialyzed DNA sample, suspended in 1 mL of SOC medium, incubated for 1 h at 37 °C, plated on LB-kanamycin (30 μg/mL) agar plates, and incubated overnight at 37 °C.

Obtaining the 22c-trick Mixture of Primers. The single primers were mixed according to the ratios described in Supplementary Table S1.

Determination of Annealing Temperature for Degenerated Primers. Adjusting the annealing temperature (T_a) is a key factor for achieving the desired degeneracy. The T_a depends on the primer melting temperature (T_m) and the overall salt concentration in a PCR. Since degenerated primers are mixtures of sequences, their melting temperatures cover a range of temperatures. The minimal and maximal T_m of the range can be calculated with the codons representing the lowest and highest GC content existing in the particular degeneracy, e.g., TTT (low T_m) and GCG (high T_m) for the 22c-trick. Promegas T_m Calculator for Oligos with standard program settings was used, and the salt-adjusted T_m was utilized for determination of T_a, e.g., the primer sets PAMO_443444_rv and PAMO_443444_fw had T_m values ranging from 58 to 65 °C. The annealing temperatures were probed at 58, 60, 62, 63, 64, and 68 °C. The library at 63 °C was judged best based on the QQC.

Quick Quality Control. QQC was performed as reported elsewhere.[57] Briefly, obtained colonies after transformation were scratched with a Drigalsky spatula from plate after adding 1 mL of water to the plate. The pool of plasmid DNA was extracted from the collected cells using the QIAprep Miniprep Kit and analyzed by sequencing.

■ ASSOCIATED CONTENT

Supporting Information

This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: jpacevedo@uandes.cl.

Author Contributions

J.P.A. and D.J.O. developed the concept. S.K. and C.G.A.-R. wrote the manuscript with help and edits from J.P.A. and M.T.R.. S.K. generated the model libraries and analyzed and interpreted the data, L.P.P. generated PAMO libraries, Z.-G.Z. generated CHMO libraries, and C.G.A.-R. generated BM3 libraries. All authors revised the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Yosephine Gumulya for helpful discussions as well as Rubén Agudo and Hajo Holzmann for statistical advice. Support by the Max-Planck-Society and the Arthur C. Cope Foundation is gratefully acknowledged.

■ ABBREVIATIONS

MP: MegaPrimer; OE-PCR: Overlap-Extension PCR; QC: QuikChange; QQC: Quick Quality Control

■ REFERENCES

(1) Wells, J. A., Vasser, M., and Powers, D. B. (1985) Cassette mutagenesis - an efficient method for generation of multiple mutations at defined sites. *Gene 34*, 315−323.

(2) Derbyshire, K. M., Salvo, J. J., and Grindley, N. D. (1986) A simple and efficient procedure for saturation mutagenesis using mixed oligodeoxynucleotides. *Gene 46*, 145−152.

(3) Oliphant, A. R., Nussbaum, A. L., and Struhl, K. (1986) Cloning of random-sequence oligodeoxynucleotides. *Gene 44*, 177−183.

(4) Sidhu, S. S., and Kossiakoff, A. A. (2007) Exploring and designing protein function with restricted diversity. *Curr. Opin. Chem. Biol. 11*, 347−354.

(5) Bommarius, A. S., Blum, J. K., and Abrahamson, M. J. (2011) Status of protein engineering for biocatalysts: how to design an industrially useful biocatalyst. *Curr. Opin. Chem. Biol. 15*, 194−200.

(6) Dalby, P. A. (2011) Strategy and success for the directed evolution of enzymes. *Curr. Opin. Struct. Biol. 21*, 473−480.

(7) Reetz, M. T. (2011) Laboratory evolution of stereoselective enzymes: A prolific source of catalysts for asymmetric reactions. *Angew. Chem., Int. Ed. 50*, 138−174.

(8) Kazlauskas, R. J., and Bornscheuer, U. T. (2009) Finding better protein engineering strategies. *Nat. Chem. Biol. 5*, 526−529.

(9) Turner, N. J. (2009) Directed evolution drives the next generation of biocatalysts. *Nat. Chem. Biol. 5*, 567−573.

(10) Shivange, A. V., Marienhagen, J., Mundhada, H., Schenk, A., and Schwaneberg, U. (2009) Advances in generating functional diversity for directed protein evolution. *Curr. Opin. Chem. Biol. 13*, 19−25.

(11) Brustad, E. M., and Arnold, F. H. (2011) Optimizing non-natural protein function with directed evolution. *Curr. Opin. Chem. Biol. 15*, 201−210.

(12) Lutz, S., and Patrick, W. M. (2004) Novel methods for directed evolution of enzymes: quality, not quantity. *Curr. Opin. Biotechnol. 15*, 291−297.

(13) Neylon, C. (2004) Chemical and biochemical strategies for the randomization of protein encoding DNA sequences: library construction methods for directed evolution. *Nucleic Acids Res. 32*, 1448−1459.
(14) Siloto, R. M. P., and Weselake, R. J. (2012) Site saturation mutagenesis: methods and applications in protein engineering. *Biocatal. Agric. Biotechnol.*, 181−189.
(15) Mildvan, A. S. (2004) Inverse thinking about double mutants of enzymes. *Biochemistry 43*, 14517−14520.
(16) Reetz, M. T., and Sanchis, J. (2008) Constructing and analyzing the fitness landscape of an experimental evolutionary process. *ChemBioChem 9*, 2260−2267.
(17) Smith, J. M. (1970) Natural selection and concept of a protein space. *Nature 225*, 563−564.
(18) Reetz, M. T., Kahakeaw, D., and Lohmer, R. (2008) Addressing the numbers problem in directed evolution. *ChemBioChem 9*, 1797−1804.
(19) Aharoni, A., Griffiths, A. D., and Tawfik, D. S. (2005) High-throughput screens and selections of enzyme-encoding genes. *Curr. Opin. Chem. Biol. 9*, 210−216.
(20) Baker, M. (2011) Protein engineering: navigating between chance and reason. *Nat. Methods 8*, 623−626.
(21) Reetz, M. T., Kahakeaw, D., and Sanchis, J. (2009) Shedding light on the efficacy of laboratory evolution based on iterative saturation mutagenesis. *Mol. BioSyst. 5*, 115−122.
(22) Balint, R. F., and Larrick, J. W. (1993) Antibody Engineering by Parsimonious Mutagenesis. *Gene 137*, 109−118.
(23) Reetz, M. T., and Wu, S. (2008) Greatly reduced amino acid alphabets in directed evolution: making the right choice for saturation mutagenesis at homologous enzyme positions. *Chem. Commun. (Cambridge)*, 5499−5501.
(24) Chica, R. A., Doucet, N., and Pelletier, J. N. (2005) Semi-rational approaches to engineering enzyme activity: combining the benefits of directed evolution and rational design. *Curr. Opin. Biotechnol. 16*, 378−384.
(25) Morley, K. L., and Kazlauskas, R. J. (2005) Improving enzyme properties: when are closer mutations better? *Trends Biotechnol. 23*, 231−237.
(26) Reetz, M. T., Prasad, S., Carballeira, J. D., Gumulya, Y., and Bocola, M. (2010) Iterative saturation mutagenesis accelerates laboratory evolution of enzyme stereoselectivity: rigorous comparison with traditional methods. *J. Am. Chem. Soc. 132*, 9144−9152.
(27) Chen, M. M., Snow, C. D., Vizcarra, C. L., Mayo, S. L., and Arnold, F. H. (2012) Comparison of random mutagenesis and semi-rational designed libraries for improved cytochrome P450 BM3-catalyzed hydroxylation of small alkanes. *Protein Eng. Des. Sel. 25*, 171−178.
(28) Parikh, M. R., and Matsumura, I. (2005) Site-saturation mutagenesis is more efficient than DNA shuffling for the directed evolution of beta-fucosidase from beta-galactosidase. *J. Mol. Biol. 352*, 621−628.
(29) Gray, Kevin A., Richardson, Toby, H., Kretz, K., Short, Jay M., Bartnek, F., Knowles, R., Kan, L., Swanson, Paul E., and Robertson, Dan E. (2001) Rapid evolution of reversible denaturation and elevated melting temperature in a microbial haloalkane dehalogenase. *Adv. Synth. Catal. 343*, 607−617.
(30) O'Maille, P. E., Bakhtina, M., and Tsai, M. D. (2002) Structure-based combinatorial protein engineering (SCOPE). *J. Mol. Biol. 321*, 677−691.
(31) Reetz, M. T., Bocola, M., Carballeira, J. D., Zha, D. X., and Vogel, A. (2005) Expanding the range of substrate acceptance of enzymes: Combinatorial active-site saturation test. *Angew. Chem., Int. Ed. 44*, 4192−4196.
(32) Reetz, M. T., and Carballeira, J. D. (2007) Iterative saturation mutagenesis (ISM) for rapid directed evolution of functional enzymes. *Nat. Protoc. 2*, 891−903.
(33) Hayes, R. J., Bentzien, J., Ary, M. L., Hwang, M. Y., Jacinto, J. M., Vielmetter, J., Kundu, A., and Dahiyat, B. I. (2002) Combining computational and experimental screening for rapid optimization of protein properties. *Proc. Natl. Acad. Sci. U.S.A. 99*, 15926−15931.
(34) Treynor, T. P., Vizcarra, C. L., Nedelcu, D., and Mayo, S. L. (2007) Computationally designed libraries of fluorescent proteins evaluated by preservation and diversity of function. *Proc. Natl. Acad. Sci. U.S.A. 104*, 48−53.
(35) Damborsky, J., and Brezovsky, J. (2009) Computational tools for designing and engineering biocatalysts. *Curr. Opin. Chem. Biol. 13*, 26−34.
(36) Privett, H. K., Kiss, G., Lee, T. M., Blomberg, R., Chica, R. A., Thomas, L. M., Hilvert, D., Houk, K. N., and Mayo, S. L. (2012) Iterative approach to computational enzyme design. *Proc. Natl. Acad. Sci. U.S.A. 109*, 3790−3795.
(37) Fox, R. J., Davis, S. C., Mundorff, E. C., Newman, L. M., Gavrilovic, V., Ma, S. K., Chung, L. M., Ching, C., Tam, S., Muley, S., Grate, J., Gruber, J., Whitman, J. C., Sheldon, R. A., and Huisman, G. W. (2007) Improving catalytic function by ProSAR-driven enzyme evolution. *Nat. Biotechnol. 25*, 338−344.
(38) Denault, M., and Pelletier, J. N. (2007) Protein library design and screening: working out the probabilities. *Methods Mol. Biol. 352*, 127−154.
(39) Hughes, M. D., Nagel, D. A., Santos, A. F., Sutherland, A. J., and Hine, A. V. (2003) Removing the redundancy from randomised gene libraries. *J. Mol. Biol. 331*, 973−979.
(40) Neuner, P., Cortese, R., and Monaci, P. (1998) Codon-based mutagenesis using dimer-phosphoramidites. *Nucleic Acids Res. 26*, 1223−1227.
(41) Ono, A., Matsuda, A., Zhao, J., and Santi, D. V. (1995) The synthesis of blocked triplet-phosphoramidites and their use in mutagenesis. *Nucleic Acids Res. 23*, 4677−4682.
(42) Mena, M. A., and Daugherty, P. S. (2005) Automated design of degenerate codon libraries. *Protein Eng., Des. Sel. 18*, 559−561.
(43) Patrick, W. M., and Firth, A. E. (2005) Strategies and computational tools for improving randomized protein libraries. *Biomol. Eng. 22*, 105−112.
(44) Jochens, H., and Bornscheuer, U. T. (2010) Natural diversity to guide focused directed evolution. *ChemBioChem 11*, 1861−1866.
(45) Kille, S., Zilly, F. E., Acevedo, J. P., and Reetz, M. T. (2011) Regio- and stereoselectivity of P450-catalysed hydroxylation of steroids controlled by laboratory evolution. *Nat. Chem 3*, 738−743.
(46) Nov, Y. (2012) When second best is good enough: another probabilistic look at saturation mutagenesis. *Appl. Environ. Microbiol. 78*, 258−262.
(47) Cornishbowden, A. (1985) Nomenclature for incompletely specified bases in nucleic-acid sequences - Recommendations 1984. *Nucleic Acids Res. 13*, 3021−3030.
(48) Bougioukou, D. J., Kille, S., Taglieber, A., and Reetz, M. T. (2009) Directed evolution of an enantioselective enoate-reductase: Testing the utility of iterative saturation mutagenesis. *Adv. Synth. Catal. 351*, 3287−3305.
(49) Hogrefe, H. H., Cline, J., Youngblood, G. L., and Allen, R. M. (2002) Creating randomized amino acid libraries with the QuikChange Multi Site-Directed Mutagenesis Kit. *BioTechniques 33*, 1158−1160 , 1162, 1164−1165.
(50) Sarkar, G., and Sommer, S. S. (1990) The "megaprimer" method of site-directed mutagenesis. *BioTechniques 8*, 404−407.
(51) Ho, S. N., Hunt, H. D., Horton, R. M., Pullen, J. K., and Pease, L. R. (1989) Site-directed mutagenesis by overlap extension using the polymerase chain-reaction. *Gene 77*, 51−59.
(52) Tang, L., Gao, H., Zhu, X., Wang, X., Zhou, M., and Jiang, R. (2012) Construction of "small-intelligent" focused mutagenesis libraries using well-designed combinatorial degenerate primers. *BioTechniques 52*, 149−158.
(53) Bosley, A. D., and Ostermeier, M. (2005) Mathematical expressions useful in the construction, description and evaluation of protein libraries. *Biomol. Eng. 22*, 57−61.
(54) Patrick, W. M., Firth, A. E., and Blackburn, J. M. (2003) User-friendly algorithms for estimating completeness and diversity in randomized protein-encoding libraries. *Protein Eng. 16*, 451−457.

(55) Reymond, J. L. (2006) *Enzyme Assays: High-throughput Screening, Genetic Selection and Fingerprinting*, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim.
(56) Polizzi, K. M., Parikh, M., Spencer, C. U., Matsumura, I., Lee, J. H., Realff, M. J., and Bommarius, A. S. (2006) Pooling for improved screening of combinatorial libraries for directed evolution. *Biotechnol. Prog. 22*, 961−967.
(57) Sanchis, J., Fernandez, L., Carballeira, J. D., Drone, J., Gumulya, Y., Hobenreich, H., Kahakeaw, D., Kille, S., Lohmer, R., Peyralans, J. J., Podtetenieff, J., Prasad, S., Soni, P., Taglieber, A., Wu, S., Zilly, F. E., and Reetz, M. T. (2008) Improved PCR method for the creation of saturation mutagenesis libraries in directed evolution: application to difficult-to-amplify templates. *Appl. Microbiol. Biotechnol. 81*, 387−397.
(58) Iyidogan, P., and Lutz, S. (2008) Systematic exploration of active site mutations on human deoxycytidine kinase substrate specificity. *Biochemistry 47*, 4711−4720.
(59) Breslauer, K. J., Frank, R., Blocker, H., and Marky, L. A. (1986) Predicting DNA duplex stability from the base sequence. *Proc. Natl. Acad. Sci. U.S.A. 83*, 3746−3750.
(60) Airaksinen, A., and Hovi, T. (1998) Modified base compositions at degenerate positions of a mutagenic oligonucleotide enhance randomness in site-saturation mutagenesis. *Nucleic Acids Res. 26*, 576−581.
(61) Lueders, T., and Friedrich, M. W. (2003) Evaluation of PCR amplification bias by terminal restriction fragment length polymorphism analysis of small-subunit rRNA and mcrA genes by using defined template mixtures of methanogenic pure cultures and soil DNA extracts. *Appl. Environ. Microbiol. 69*, 320−326.
(62) Reidhaar-Olson, J. F., Bowie, J. U., Breyer, R. M., Hu, J. C., Knight, K. L., Lim, W. A., Mossing, M. C., Parsell, D. A., Shoemaker, K. R., and Sauer, R. T. (1991) Random mutagenesis of protein sequences using oligonucleotide cassettes. *Methods Enzymol. 208*, 564−586.
(63) Palfrey, D., Picardo, M., and Hine, A. V. (2000) A new randomization assay reveals unexpected elements of sequence bias in model 'randomized' gene libraries: implications for biopanning. *Gene 251*, 91−99.
(64) Chen, Y. C. J., Peoples, O. P., and Walsh, C. T. (1988) Acinetobacter cyclohexanone monooxygenase - gene cloning and sequence determination. *J. Bacteriol. 170*, 781−789.
(65) Fraaije, M. W., Wu, J., Heuts, D. P. H. M., van Hellemond, E. W., Spelberg, J. H. L., and Janssen, D. B. (2005) Discovery of a thermostable Baeyer-Villiger monooxygenase by genome mining. *Appl. Microbiol. Biotechnol. 66*, 393−400.