CHEMISTRY
A European Journal

A Journal of
ChemPubSoc Europe

Accepted Article

Title: Computer-assisted Recombination (CompassR) teaches us how to recombine beneficial substitutions from directed evolution campaigns

Authors: Haiyang Cui, Hao Cao, Haiying Cai, Jaeger Karl-Erich, Mehdi Dolatabadi Davari, and Ulrich Schwaneberg

To be cited as: *Chem. Eur. J.* 10.1002/chem.201903994

Link to VoR: http://dx.doi.org/10.1002/chem.201903994

Supported by
ACES

WILEY-VCH

10.1002/chem.201903994

WILEY-VCH

FULL PAPER

Computer-assisted Recombination (CompassR) teaches us how to recombine beneficial substitutions from directed evolution campaigns

Haiyang Cui, Hao Cao, Haiying Cai, Karl Erich Jaeger, Mehdi D. Davari, Ulrich Schwaneberg*

Abstract: A main remaining challenge in protein engineering is how to recombine beneficial substitutions. Systematic recombination studies show that poorly performing variants are usually obtained after recombination of 3 to 4 beneficial substitutions. This limits researchers to exploit nature's potential in generating better enzymes. The Computer-assisted Recombination (CompassR) strategy provides a selection guide for beneficial substitutions that can be recombined to gradually improve enzyme performance by analysis of the relative free energy of folding ($\Delta\Delta G_{fold}$). The performance of CompassR was evaluated by analysis of 84 recombinants located on 13 positions of *Bacillus subtilis* lipase A. The finally obtained variant F17S/V54K/D64N/D91E had a 2.7-fold improved specific activity in 18.3 % (v/v) 1-Butyl-3-methylimidazolium Chloride ([BMIM][Cl]). In essence, the deducted CompassR rule allows to recombine beneficial substitutions in an iterative manner and empowers researchers to generate better enzymes in a time-efficient manner.

Introduction

Directed evolution of proteins has matured into a powerful methodology to improve enzyme properties, such as the stability, selectivity, and specific activity [1]. The Nobel Prize in chemistry in 2018 was awarded in recognition of the significant impact of directed evolution on both, gain of scientific knowledge and application in chemical industries and in medicine [2]. The beauty of directed protein evolution is that all kind of protein properties that can be reflected in the employed screening/selection system, can within physical boundaries be improved without any molecular understanding or hypothesis [3]. Subsequent analysis of the identified amino acid exchanges enables to discover new fundamental design principles of enzymes. The key technologies required for directed evolution are diversity generation and high-throughput screening. The diversity generation challenge is largely solved today with the development of random and multi-site saturation mutagenesis methods [2]. For example, already epPCR generates ~10^{12} variants under standard error-prone conditions within in two to three hours [4]. Remaining challenges are how to navigate through the huge protein sequence space (numbers problem in screening) and how to recombine beneficial substitutions. Beneficial substitution could be obtained from directed evolution experiments after screening a few thousand variants or by (semi-) rational design studies. Numerous reports point out that recombining more than two or three beneficial substitutions do not necessarily yield further improved enzymes variants [5]. Additionally, often the best performing variants are obtained in early stages of recombination, e.g. after one or two recombined substitutions [6,7]. Several studies also point out that beneficial substitutions drive each other to "extinction" [8]. Interestingly, the simultaneous site saturation of two sets of amino acids (each comprising five beneficial positions with four to five substitutions per recombined position) yielded after screening of 1500 variants a fraction of only 0.67% active clones for the phenylacetone monooxygenase (PAMO, 10 clones) [9]. Comparable results were reported for the alcohol dehydrogenase (cpADH5) [7d] with a fraction of 1.2% of active clones (4 simultaneously saturated positions, screening of 3500 variants). In another report, ten identified positions in limonene epoxide hydrolase (LEH) were simultaneously recombined using a multi-site directed mutagenesis method (one substitution per position) and after screening of 3320 clones, 533 active variants were obtained. The most beneficial variant with inverted enantioselectivity had only three substitutions [10]. All the latter reports confirm that rules and methods to guide recombination experiments are limited by a low fraction of active recombinants that are highly desirable in the field of protein engineering for generating better performing catalysts.

How can we ensure that enzymes are active after several iterative recombinations? Several factors affect the enzyme activity and function (e.g. substrate binding [11], product release [12], temperature [13], pH [14]), however it is generally accepted that enzymes must be able to fold stably in order to function properly [15]. The relationship between stability and function of an enzyme (referred to as stability-activity tradeoff) is well studied in respect to thermostability and catalytic activity [16]. The relative free energy of folding ($\Delta\Delta G_{fold}$) was employed as a measure of protein stability and to assess the relationship between stability and function in several enzymes (e.g. TEM-1 β-lactamase [17], cytochrome P450 BM3 [18], green fluorescent protein avGFP [19] and others [16c,20]). It is known that most proteins are marginally stable, and substitutions can be tolerated until the "robustness threshold" is reached [17b,d]. The variants that have higher stability tend to have higher protein fitness [17c] and extra stability could increase evolvability to accept a wider range of beneficial substitutions [18].

[*] H. Cui, Dr. H. Cao, Dr. H. Cai, Dr. M. D. Davari, Prof. Dr. U. Schwaneberg
Institute of Biotechnology RWTH Aachen University
Worringer Weg 3, 52074 Aachen (Germany)
E-mail: u.schwaneberg@biotec.rwth-aachen.de
Dr. H. Cao
Beijing Bioprocess Key Laboratory and College of Life Science and Technology
Beijing University of Chemical Technology
Beijing 100029 (China)
Prof. Dr. K. E. Jaeger
Institute of Molecular Enzyme Technology
Heinrich Heine University Düsseldorf and Research Center Jülich
Wilhelm Johnen Strasse, 52426, Jülich (Germany)
Prof. Dr. U. Schwaneberg
DWI Leibniz-Institute for Interactive Materials
Forckenbeckstrasse 50, 52074 Aachen (Germany)
Supporting information for this article is given via a link at the end of the document.

All these above studies indicate that the $\Delta\Delta G_{fold}$ is an important factor for predicting the evolvability and/or performance of proteins. In order to analyze the stability of all the single substitutions, researchers used the reported accuracy of $\Delta\Delta G_{fold}$ predictors to bin the $\Delta\Delta G_{fold}$ into several stabilizing/destabilizing categories [16c]. Computed $\Delta\Delta G_{fold}$ (in kcal/mol) of single substitutions are regarded as highly stabilizing (< −1.84), stabilizing (−1.84 to −0.92), slightly stabilizing (−0.92 to −0.46), neutral (−0.46 to +0.46), slightly destabilizing (+0.46 to +0.92), destabilizing (+0.92 to +1.84), and highly destabilizing (> +1.84) [12c]. Several computational protein stability predictors are available to calculate $\Delta\Delta G_{fold}$, e.g. FoldX [21], Rosetta [22], CUPSAT [23], PoPMuSiC [16b] and others [24]. Although stabilizing/destabilizing categories can be applied to classify single substitutions [16c], the thresholds of $\Delta\Delta G_{fold}$ values for recombination of single beneficial substitutions are still missing. We selected FoldX as it is a popular and reliable method for determining changes in the free energy of folding caused by substitutions. Compared with other predictors, FoldX achieved the highest correlation (r = 0.96) for binned data in a recent evaluation [25] and has often successfully been used for identifying beneficial positions [16c,24,26].

Bacillus subtilits Lipase A (BSLA) is a well-studied enzyme and was chosen to develop CompassR as a predictor for recombining substitutions. The "BSLA-SSM" library covers all the natural diversity with a single amino acid exchange at each position of BSLA (in total 181 positions; 3439 variants; "site-saturation mutagenesis" denoted as "SSM"). The "BSLA-SSM" library was constructed in our previous study [27] as well as screened towards improved 1-Butyl-3-methylimidazolium Chloride ([BMIM][Cl]) resistance. CompassR was developed by selecting 13 positions in three genes which encoded in total 39 substitutions that were recombined in a staggered extension process (StEP) library. Out of 39 substitutions, 13 beneficial substitutions (one substitution per position) were finally selected based on their $\Delta\Delta G_{fold}$ values for further recombination studies. The calculated $\Delta\Delta G_{fold}$ values of the 13 substitutions were used to place them into three categories. Three most stabilizing substitutions, F17S, V54K and G155P, were selected for two recombination campaigns ("intra-category" and "inter-category") with all other substitutions and up to four subsequent recombination experiments were performed generating in total 84 recombinants (see **Figure 2**). Analysis of activity of the BSLA recombinants and their corresponding $\Delta\Delta G_{fold}$ values was used to define the CompassR rule for recombination of beneficial substitutions.

Results

The results section is divided into four parts to illustrate how the CompassR rule was developed. The first part describes the results of standard recombination experiments in which 39 BLSA beneficial substitutions (39 = 3 substitutions × 13 positions) were distributed over three (synthetic) genes and recombined with the staggered extension process (StEP) method [28]. The analysis of the StEP recombination library demonstrated that the recombination challenge applies to BSLA in a similar manner than to reported enzymes (see introduction; e.g. *Pseudomonas aeruginosa* lipase [5], *β*-glucuronidase [8b], PAMO [9], cpADH5 [7d], LEH [10]). The second part describes the $\Delta\Delta G_{fold}$ calculation and recombination analysis. In detail, the 13 beneficial substitutions were placed in three categories based on their $\Delta\Delta G_{fold}$ and recombined in different modes ("intra-category" and "inter-category" recombination; in total 84 variants). In the third part the CompassR rule was postulated based on the obtained recombination results and in the concluding fourth part, 33 variants were analyzed in detail and a molecular understanding of BSLA's improved resistance towards the ionic liquid [BMIM][Cl] is provided.

BSLA recombination by the StEP method to obtain the fraction of active population after recombination. Thirty-nine beneficial substitutions at 13 positions were identified in the "BSLA-SSM" library [27]. The 13 mutated positions were selected that match the following criteria: i) the distance between each position was more than the minimum gap distance in the gene that can be resolved by the StEP method (~30 bp) [28], ii) the targeted positions were evenly distributed over on the whole *bsla* gene, and iii) at least 3 substitutions among 19 substitutions in each selected position were beneficial. In order to enable an efficient recombination by the StEP method [28], the 39 substitutions were distributed over three synthetic genes (three different substitutions per positions, **Figure S1** and **Table S1** in Supporting Information (SI)) and recombined with the BSLA wild type ("forth" substitution per position) employing the StEP method; the latter generates a theoretical diversity of 4^{13} ($\approx 10^8$) different variants. The recombination of the 13 selected positions at BSLA yielded mainly inactive variants (82 %) after screening of approximately 5000 clones. Sequencing of 30 randomly chosen variants showed that all eleven active variants harbored one to three substitution(s). In detail, 3 had one substitution, 6 had two, 2 had three (**Figure 1**, **Table S2** in SI). Inactive recombinants of BSLA harbored two to eleven substitutions. The high fraction of inactive recombinants of BSLA and the low number of substitution in active BSLA variants are well correlating with reports on the recombination challenge (see introduction; [8b,9-10]). The "best" variant obtained from the StEP-BSLA recombination experiment after screening of 5000 variants was the BSLA recombinant F17S/V54K/Y129M with a 1.7 times improved [BMIM][Cl] resistance when compared to BSLA wild type.

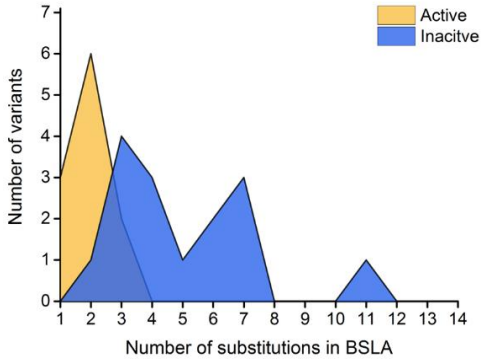


Figure 1. Overview of the diversity of the StEP recombination library in respect to the number of recombined substitutions determined by sequencing of 30 randomly picked variants. Yellow: active variants. Blue: inactive variants. Four picked variants were the BSLA wild type.

$\Delta\Delta G_{fold}$ calculations and analysis of intra-category and inter-category recombinations. $\Delta\Delta G_{fold}$ of the selected 39 beneficial substitutions were calculated using the FoldX method [21]. As shown in **Figure S2** in SI, substitutions were classified according to binned $\Delta\Delta G_{fold}$ values [16c] as follows: 20/39 substitutions (51.3 %) were highly destabilizing ($\Delta\Delta G_{fold}$ > +1.84 kcal/mol), 8/39 substitutions (20.5 %) were slightly destabilizing (+0.46 < $\Delta\Delta G_{fold}$ < +1.84 kcal/mol), 9/39 substitutions (23.1 %) showed the neural effect on the stability (-0.46 < $\Delta\Delta G_{fold}$ < +0.46 kcal/mol), only 2/39 substitutions (5.1 %) were stabilizing ($\Delta\Delta G_{fold}$ < -0.46 kcal/mol). As starting point for the CompassR rule, 13 substitutions (one substitution per position) with the lowest to highest $\Delta\Delta G_{fold}$ (-1.49 < $\Delta\Delta G_{fold}$ < +18.64 kcal/mol) were selected from the 39 beneficial substitutions and grouped in three categories (category A-five substitutions: $\Delta\Delta G_{fold}$ from -1.49 to +0.36 kcal/mol; category B-four substitutions: $\Delta\Delta G_{fold}$ from +1.83 to +4.89 kcal/mol; category C-four substitutions: $\Delta\Delta G_{fold}$ from +7.52 to +18.64 kcal/mol; see **Table 1**).

In order to identify the threshold values of $\Delta\Delta G_{fold}$ at which BSLA variants are active or inactive, two recombination campaigns ("**intra**-category" and "**inter**-category") were performed as follows:

In the first "**intra**-category" campaign substitutions among category A, category B, and category C were recombined. Main results in the Supporting Information show that all possible recombinants in category A yielded active variants until recombinants with five substitutions (F17S/V54K/D64N/D91E/G155P) were obtained in round IV (see **Figure S3** in SI; twelve variants had a reduced activity). In category B, except of one double substitution variant (A81E/L114E), all of the recombinants were inactive and in category C only inactive variants were obtained (already after recombining two beneficial substitutions). Overall, the fraction of active recombinants was 100 % (26/26) in category A, 13 % (1/8) in category B, and 0 % (0/6) in category C (**Table S3** in SI). All these results are in agreement with the common view (see Introduction; [17b-d,18,20a]) that protein stability and function often appear to trade off at the level of individual substituitons and prove that $\Delta\Delta G_{fold}$ is an excellent predictor for selecting beneficial substitutions that can be recombined.

In a second "**inter**-category" campaign three beneficial positions of category A (F17S, V54K, and G155P) were individually recombined in three sets of experiments with all substitutions of category A, B, and C until inactive BSLA variants were obtained (**Figure S4**, **S5** and **S6** in SI). **Figure 2** summarizes the results from the three "inter-category" recombination campaign with the beneficial substitutions F17S, V54K and G155P. The comparison of the three sets of experiments shows highly similar trends. Recombinants within category A yielded in all cases active variants; recombination within category B led to unpredictable recombination results (few active recombinants with two to three substitutions; none with four substitutions) and recombinants within category C were all inactive except one variant with a double substitution. Overall, the "inter-category" recombination campaign with F17S (**Figure S4** in SI) yielded in category A 100 % active recombinants (15/15), in category B 33 % (4/12), and in category C 14 % (1/7), respectively (**Table S3** in SI). The "inter-category" recombination campaign with V54K (**Figure S5** in SI) yielded in category A 100 % active recombinants (15/15), in category B 14 % (1/7), and in category C 0 % (0/7) (**Table S3** in SI). The "inter-category" recombination campaign with the most stabilized substitution G155P (**Figure S6** in SI) yielded in category A 100 % active recombinants (15/15), in category B 14 % (1/7), and in category C 0 % (0/4) (**Table S3** in SI).

Table 1. Thirteen selected substitutions at 13 positions of the BSLA grouped in three categories according to $\Delta\Delta G_{fold}$ values.

Category[a]	Substitution	$\Delta\Delta G_{fold}$ (kcal/mol)
A	G155P	-1.49
	F17S	-0.03
	D64N	+0.09
	V54K	+0.10
	D91E	+0.36
B	Y129N	+1.83
	L114E	+2.29
	A81E	+3.00
	V165E	+4.89
C	L36P	+7.52
	G104Q	+14.38
	P5W	+14.75
	G46H	+18.64

[a] Category A comprises five beneficial substitutions with the "lowest" $\Delta\Delta G_{fold}$ values, category B comprises four beneficial substitutions within the range of neutral $\Delta\Delta G_{fold}$ values and category C comprises four beneficial substitutions with the largest $\Delta\Delta G_{fold}$ values. The larger the $\Delta\Delta G_{fold}$ negative values, the higher the stability.

CompassR rule postulation. Based on the obtained results of in total 84 recombinants (**Figure 2** and **Table S3**) the following thresholds are postulated to place substitutions: in category A: "active recombinants" (substitutions with $\Delta\Delta G_{fold} \leq +0.36$ kcal/mol), in category B: "recombinants with unpredictable activity" (substitutions within +0.36 < $\Delta\Delta G_{fold}$ < +7.52 kcal/mol), in category C: "deactivating recombinants" ($\Delta\Delta G_{fold} \geq +7.52$ kcal/mol). In summary, the Computer-assisted Recombination (CompassR, **Figure 3**) rule guides experimentalists how to recombine beneficial substitutions based on $\Delta\Delta G_{fold}$ value of the beneficial substitutions. CompassR expects that active recombinants are generated by recombining amino acid substitutions that fall into category A ($\Delta\Delta G_{fold} \leq +0.36$ kcal/mol). Recombinations with beneficial substitutions in category C should be omitted and not used for recombinations. Recombination with beneficial positions in category B should be considered in case that only few beneficial substitutions are identified or used after recombining all beneficial substitutions from category A.

10.1002/chem.201903994

Ionic liquid resistance analysis of all active recombinants. The catalytic activity and ionic liquid ([BMIM][Cl]) resistance values of all active recombinants selected by CompassR are shown in **Figure S7** in SI. As a general trend, one can observe that for most BSLA recombinants in category A and B, the ionic liquid ([BMIM][Cl]) resistance increased with increasing number of substitutions (e.g. 1st round: F17S/D91E:1.3-fold, F17S/D64N: 1.5-fold / 2nd round: e.g. F17S/V54K/D64N: 2.4-fold, F17S/V54K/D91E: 2.2-fold / 3rd round: e.g. F17S/V54K/D64N/D91E: 2.7-fold, the best performing variant). The variant from the 4th round F17S/V54K/D64N/D91E/G155P had a 1.4-fold improved resistance against the ionic liquid [BMIM][Cl] and exhibited a high level of residual activity (approximately 96 % of the wild type activity).

Visualization of all substitutions of the best performing BLSA variant F17S/V54K/D64N/D91E from category A shows that they are located on the surface of BSLA (**Figure S8**). Among them, two substitutions pertain to charged amino acids (V54K, D91E) and two to polar ones (F17S, D64N). It is reported that the interaction of [BMIM][Cl] with the BSLA protein surface is the dominating factor that reduces BSLA activity [29]. The identified beneficial substitutions on the BSLA surface with changes to polar and charged residues are in accordance to these previous findings [29].

In directed evolution experiments more than ten beneficial positions are often identified in a single round of directed evolution after screening of only a few thousand variants [30]. As outlined in the introduction methodologies are missing that empower researches to recombine efficiently and quickly more than three amino acids and to capitalize on identified beneficial substitutions. The recombination challenge of beneficial variants from directed evolution experiments clearly represents a main challenge that hampers the design of efficient enzymes for biocatalysis. In the present work, the StEP recombination experiment (three *bsla* genes; each gene encoding 13 substitutions + wild type) confirmed that BSLA is not more tolerant recombinations than many other enzymes (18 % faction of active clones). Active BLSA variants carried three or less amino acid substitutions (see results section).

$\Delta\Delta G_{fold}$ analysis of substitutions in the StEP library indicated a clear trend that $\Delta\Delta G_{fold}$ is a predictor for the recombination experiments after analysis of active and inactive variants. “Intra-” and “inter-category” recombinations by sited-directed mutagenesis was performed in a stepwise manner as previous reports [31,32]. Based on the obtained results of in total 84 recombinants (**Figure 2** and **Table S3)** thresholds for recombining beneficial substitutions are postulated as CompassR rule in results section. CompassR expects that active recombinants are generated by recombining amino acid substitutions that fall into category A ($\Delta\Delta G_{fold} \leq +0.36$ kcal/mol)

Discussion

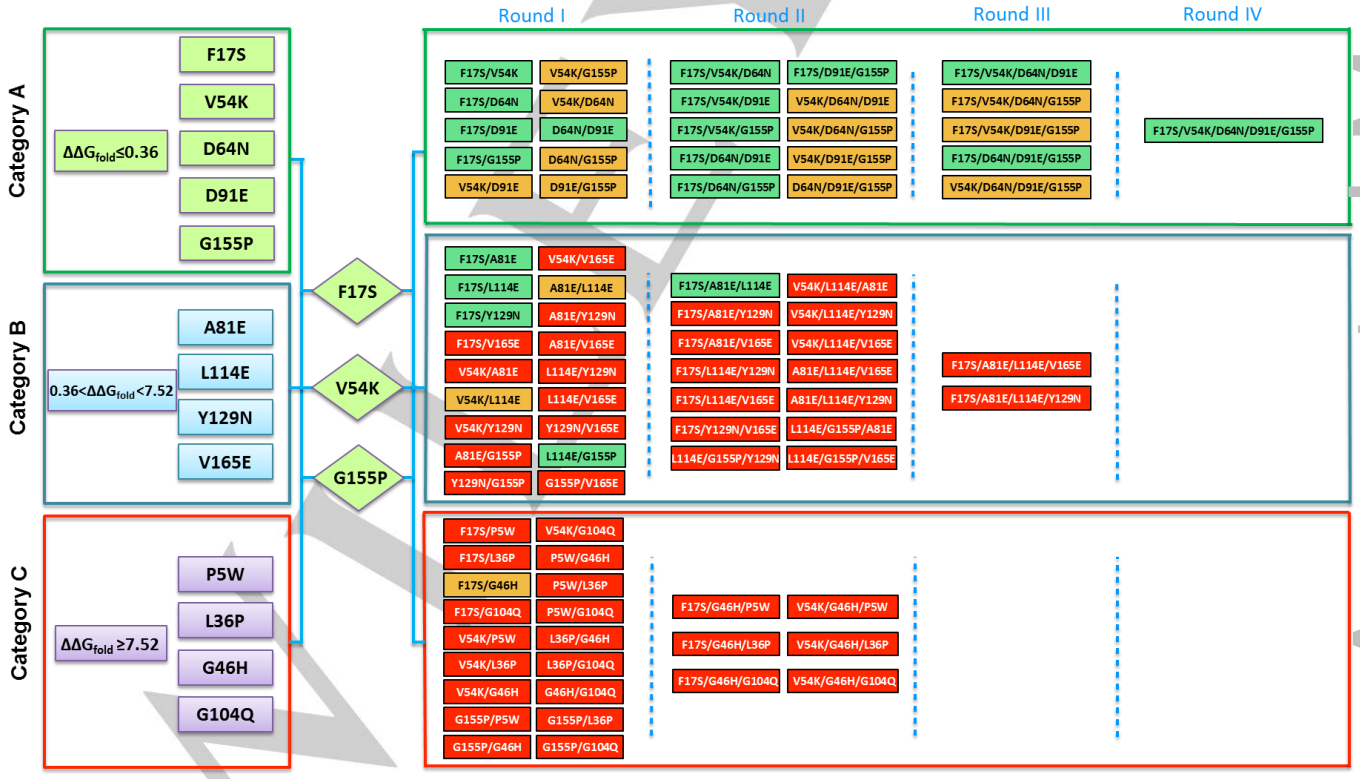


Figure 2. Overview of all BSLA recombinants generated in the recombination of each category (“intra-category”) and the beneficial substitutions F17S, V54K and G155P with beneficial substitutions from categories A (light green), B (light blue), and C (light purple) (“inter-category”). Categories (A, B, and C; on the left) are composed of 13 selected beneficial substitutions obtained from the BSLA-SSM library and grouped according to their $\Delta\Delta G_{fold}$ values. Notations of recombinants: dark green: residual activity (in buffer) ≥ 80 % of the BSLA wild type activity. Orange: residual activity (in buffer) between 10-80 % of the BSLA wild type activity. Red: residual activity (in buffer) is between 0-10 % of the BSLA wild type activity and referred to as “inactive” recombinant.

and recombinations with beneficial substitutions in category C should be omitted. Notably, only inactive variants were obtained for recombination experiments of all beneficial substitutions in category C ("intra-" and "inter-category") after recombination of three substitutions. Beneficial substitutions in category B yielded an unpredictable behavior (7 active variants: 27 inactive variants) and should, at least from our point of view be considered only in cases in which few beneficial positions are identified or after recombining all beneficial substitutions from category A.

The CompassR rule can be of high value for experimentalists enabling to generate small and highly active recombination libraries of substitutions that fall in the category A. The latter will significantly reduce experimental efforts; e.g. 5000 StEP variants of BSLA were screened in this study yielding the BSLA variant F17S/V54K/Y129M with a 1.7-fold ionic liquid resistance compared to four recombined BSLA variants in category A yielding the BSLA variant F17S/V54K/D64N/D91E with a 2.7-fold improved resistance. Interestingly, the improved variant F17S/V54K/Y129M found by StEP recombination experiment also comprised the stabilized single substitutions (F17S and V54K), indicating CompassR could find the substitutions obtained by StEP recombination experiment.

The CompassR rule enables to reduce screening efforts by recombining beneficial substitutions and generating highly functional variant libraries. CompassR is based on the relative free energy of folding calculations, but differs in comparison to sequence- and structure-based computational methods (e.g. FoldX [21], Rosetta [22], FireProt [24a], MuStab [33], I-Mutant2.0 [34], FuncLib [35], PoPMuSiC [16b], and others [32]) in its focus on beneficial recombinants. The mentioned methods concentrate mostly on the prediction of the effect of individual substitutions and their effect on protein stability; none of these methods has been used to categorize beneficial substitutions and to guide recombination of beneficial substitutions through experimentally determined beneficial positions. It is reported that inclusion of the most stable substitution (in our case G155P) is beneficial to compensate for destabilizing substitutions [18]. In order to see if the most stabilized substitution can compensate/perform in a better manner than two other substitutions (F17S and V54K) as parent, CompassR recombination experiment with 11 recombinants was performed (**Figure S6** and **Table S3**). Surprisingly, the "most" stabilized variant G155P did not increase the number of active clones after the first recombination in category B and C compared to F17S and V54K. In our BSLA experiments, the CompassR results from category A indicate that thermodynamic stability and enzymatic activity are not opposite sides of a coin and can by recombination jointly be improved as shown by the stepwise increased resistance against the ionic liquid ([BMIM]Cl). This finding agrees well with the general accepted concept that function of protein typically depends on its ability to fold to a sufficiently thermodynamically stable structure [18]. The thermodynamic stability varies from protein to protein according to the equilibrium stability, kinetic folding/unfolding processes and the temperature at which assays are conducted [36]. Thereby, the exact CompassR thresholds might slightly change depending on the type of protein and itself fold.

CompassR differs in respect to methods based on statistical analysis of protein sequence/activity relationships (e.g. ProSAR® [37] and MOSAIC® [38]) by establishing a correlation between $\Delta\Delta G_{fold}$ and catalytic activity. In addition, CompassR could be implemented in protein engineering strategies such as KnowVolution [8d] (4th recombination phase), CASTing [7b], or MORPHING [39] to guide recombination of beneficial substitutions and thereby speed up the design of significantly improved enzymes. CompassR could also be implemented as preselector in gene recombination experiments (e.g. gene shuffling [40] and StEP [28]) or in rational-guided methods like SCHEMA [41] (e.g. to select beneficial substitutions in parents and/or introduce substitutions which can rescue nonfunctional chimeric proteins) or PTRec [42] by limiting the recombination process (e.g. through synthetic genes) to encoded beneficial substitutions that fall into the category A.

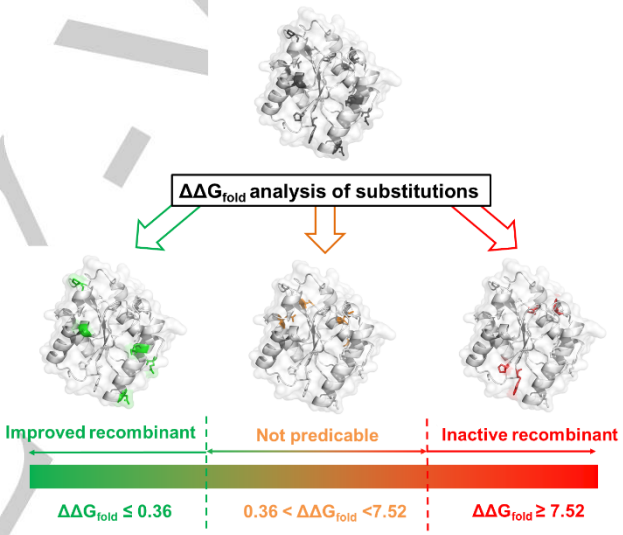


Figure 3. Computer-assisted Recombination (CompassR) rule for selecting beneficial substitutions in recombination experiments. When substitutions with $\Delta\Delta G_{fold}$ values ≤ +0.36 kcal/mol are recombined one can expect active and property improved recombinants (green). When beneficial substitutions are recombined with $\Delta\Delta G_{fold}$ values ranging from +0.36 to +7.52 kcal/mol one cannot predict whether the recombinants will be inactive or active (unpredictable behavior; orange). Recombination of beneficial substitutions with $\Delta\Delta G_{fold}$ ≥ +7.52 kcal/mol results in deactivated and in activity reduced recombinants (red). $\Delta\Delta G_{fold}$ is calculated by the FoldX method; surface representation of the BSLA (PDB ID: 1i6w, Chain A) is shown in grey. The highlighted substitutions in green, orange, and red are the selected 13 beneficial single substitutions that were obtained from the "BSLA-SSM" library.

Conclusion

CompassR enables the design of better enzymes with minimal experimental efforts through recombination of multiple beneficial substitutions that were previously identified by directed evolution and/or (semi-)rational design. The CompassR rule guides recombination of beneficial substitutions through analysis of the relative free energy of folding and an experimentally determined threshold; all BSLA recombinants in category A were active and improvements gradually increased with increasing the number of recombined beneficial substitutions. The latter is

contrast to standard recombination methods (e.g. StEP [28] or OmniChange [43]) which yield active populations ranging from 0.67 % to 16.55 %. CompassR is therefore of high value for experimentalists since highly active libraries are generated and screening efforts can be minimized to a few variants or even be omitted through gene synthesis by ordering genes that encode recombinants with multiple "category A" substitutions. Furthermore, the gradually increased ionic liquids resistance with increased number of substitutions (rounds of recombination) makes it likely that more than five beneficial substitutions can be recombined and much better performing enzymes can be designed in the future.

Experimental Section

Chemicals. All chemicals were of analytical grade or higher quality and purchased from Carl Roth (Karlsruhe, Germany), AppliChem (Darmstadt, Germany), and Sigma-Aldrich Chemie (Steinheim, Germany) unless specified. [BMIM][Cl] were synthesized by IoLiTec Ionic Liquids Technologies (Heilbronn, Germany) and were dissolved to 1.2 M by adding 18.3 % (v/v) Milli-Q water before use.

Strains and plasmids. The plasmid pET22b(+)-bsla WT was constructed in the previous work [27] and was used as the template for the polymerase chain reactions (PCRs) performed in the present work unless specified. Chemically competent *Escherichia coli* DH5a and *Escherichia coli* BL21-Gold (DE3) (Agilent Technologies; Santa Clara, USA) were used as hosts for plasmids amplification and protein expression, respectively.

StEP recombination library construction and expression. The StEP library of BSLA was generated using a modified StEP PCR protocol [28]. The two-step StEP PCR protocol is shown in **Tables S4-S5** in SI. Three *bsla* genes with 13 substitutions each were synthesized by Invitrogen (Germany). The BSLA StEP library was cloned into the pET22b(+) vector using the PLICing method [43], the specific primer are listed in **Table S6** in the SI. Then StEP recombination library was transformed and expressed in *Escherichia coli* BL21-Gold (DE3) using standard methods.

Site-directed mutagenesis. BSLA variants were stepwise constructed by PCR according to the QuikChange site directed mutagenesis method [31] using the primers listed in **Table S7** in SI.

Activity Assay in 96-well Microtiter Plate. The screening procedure and activity determinations with the *p*-nitrophenyl butyrate (*p*NPB) assay were performed as previously reported in 96-well MTPs [27,44]. BSLA resistance (wild type or variant) was evaluated as activity in the presence of ionic liquid divided by activity in absence of ionic liquid [27,44] (Infinite M200 PRO microtiter plate reader; Tecan, Maennedorf, Switzerland). Residual activity and background of an empty vector were determined and subtracted in all analysis. All data shown was at least measured in triplicates.

Computational procedures. The relative folding free energies ($\Delta\Delta G_{fold} = \Delta G_{fold,sub} - \Delta G_{fold,wt}$) were computed using FoldX version 3b5.1 [21] employing YASARA Plugin [45] in YASARA Structure version 13.9.8 [46]. The initial structure of the BSLA for analysis was taken from the BSLA crystal structure (PDB ID: 1i6w [47] Chain A, resolution 1.5 Å). Default FoldX parameters were used for temperature (298 K), ionic strength (0.05 M), and pH (7). The structure of the BSLA wild type was rotamerized and energy minimized using the "RepairObject" command to correct the residues that have non-standard torsion angles. Five FoldX runs were performed for each substitution to ensure that the minimum energy conformation of even large residues that possess many rotamers is identified. The accuracy of FoldX method in prediction of relative folding free energies is reported to be 0.46 kcal/mol (the standard deviation of the difference between $\Delta\Delta G_{fold}$ calculated by FoldX and the experimental values) [21]. Pymol [48] was used to visualize the BSLA structure.

Acknowledgements

Haiyang Cui was supported by a Ph.D. scholarship from China Scholarship Council (CSC No. 201604910840). We thank Mr. Subrata Pramanik, Dr. Gaurao V. Dhoke, Dr. Lingling Zhang, and Prof. Dr. Leilei Zhu for discussions. Calculations were performed with computing resources granted by JARA-HPC from RWTH Aachen University under projects JARA0169.

Keywords: Protein engineering • Directed evolution • Recombination • *Bacillus subtilis* lipase A • FoldX

Conflict of interest

The author declares no conflict of interest.

[1] a) F. H. Arnold, *Nat. Biotechnol.* **1998**, *16*, 617-618; b) F. H. Arnold, *Angew. Chem. Int. Ed.* **2018**, *57*, 4143-4148; *Angew. Chem.* **2018**, *130*, 4212-4218; c) F. Cheng, L. Zhu, U. Schwaneberg, *Chem. Commun.* **2015**, *51*, 9760-9772; d) K. L. Tee, D. Roccatano, S. Stolte, J. Arning, B. Jastorff, U. Schwaneberg, *Green Chem.* **2008**, *10*, 117-123.

[2] U. T. Bornscheuer, B. Hauer, K. E. Jaeger, U. Schwaneberg, *Angew. Chem. Int. Ed.* **2019**, *58*, 36-40; *Angew. Chem.* **2019**, *131*, 36-41.

[3] C. Jäckel, P. Kast, D. Hilvert, *Annu. Rev. Biophys.* **2008**, *37*, 153-173.

[4] a) K. Lan Tee, U. Schwaneberg, *Comb. Chem. High Throughput Screening* **2007**, *10*, 197-217; b) M. D. Lane, B. Seelig, *Curr. Opin. Chem. Biol.* **2014**, *22*, 129-136.

[5] K. Liebeton, A. Zonta, K. Schimossek, M. Nardini, D. Lang, B. W. Dijkstra, M. T. Reetz, K. E. Jaeger, *Chem. Biol.* **2000**, *7*, 709-718.

[6] C. G. Acevedo-Rocha, S. Hoebenreich, M. T. Reetz, in *Directed Evolution Library Creation*, Springer, **2014**, pp. 103-128.

[7] a) M. T. Reetz, P. Soni, L. Fernández, *Biotechnol. Bioeng.* **2009**, *102*, 1712-1717; b) M. T. Reetz, M. Bocola, J. D. Carballeira, D. Zha, A. Vogel, *Angew. Chem.* **2005**, *117*, 4264-4268; *Angew. Chem. Int. Ed.* **2005**, *44*, 4192-4196 c) S. Islam, D. Laaf, B. Infanzón, H. Pelantová, M. D. Davari, F. Jakob, V. Křen, L. Elling, U. Schwaneberg, *Chem. Eur. J.* **2018**, *24*, 17117-17124; d) Y. Ensari, G. V. Dhoke, M. D. Davari, A. J. Ruff, U. Schwaneberg, *ChemBioChem* **2018**, *19*, 1563-1569; eM. T. Reetz, J. D. Carballeira, J. Peyralans, H. Höbenreich, A. Maichele, A. Vogel, *Chem. Eur. J.* **2006**, *12*, 6031-6038.

[8] a) M.-W. Bhuiya, C.-J. Liu, *J. Biol. Chem.* **2010**, *285*, 277-285; b) L. A. Rowe, M. L. Geddie, O. B. Alexander, I. Matsumura, *J. Mol. Biol.* **2003**, *332*, 851-860; c) J. D. Bloom, M. M. Meyer, P. Meinhold, C. R. Otey, D. MacMillan, F. H. Arnold, *Curr. Opin. Chem. Biol.* **2005**, *15*, 447-452; d) K. Rübsam, M. D. Davari, F. Jakob, U. Schwaneberg, *Polymers* **2018**, *10*, 423.

[9] L. P. Parra, R. Agudo, M. T. Reetz, *ChemBioChem* **2013**, *14*, 2301-2309.

[10] Z. Sun, R. Lonsdale, X. D. Kong, J. H. Xu, J. Zhou, M. T. Reetz, *Angew. Chem. Int. Ed.* **2015**, *54*, 12410-12415; *Angew. Chem.* **2015**, *127*, 12587-12592.

[11] R. P.-A. Berntsson, S. H. Smits, L. Schmitt, D.-J. Slotboom, B. Poolman, *FEBS Lett.* **2010**, *584*, 2606-2617.

[12] J. A. Himmelberger, K. E. Cole, D. P. Dowling, in *Green Chem.*, Elsevier, **2018**, pp. 471-512.

[13] F. H. Arnold, A. A. Volkov, *Curr. Opin. Chem. Biol.***1999**, *3*, 54-59.

[14] K. M. Polizzi, A. S. Bommarius, J. M. Broering, J. F. Chaparro-Riggers, *Curr. Opin. Chem. Biol.* **2007**, *11*, 220-225.
[15] J. Echave, E. L. Jackson, C. O. Wilke, *Phys. Biol.* **2015**, *12*, 025002.
[16] a) B. K. Shoichet, W. A. Baase, R. Kuroki, B. W. Matthews, *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 452-456; b) Y. Dehouck, J. M. Kwasigroch, D. Gilis, M. Rooman, *BMC Bioinf.* **2011**, *12*, 151; c) R. A. Studer, P. A. Christin, M. A. Williams, C. A. Orengo, *Proc. Natl. Acad. Sci. U S A* **2014**, *111*, 2223-2228.
[17] a) N. Tokuriki, D. S. Tawfik, *Curr. Opin. Struct. Biol.* **2009**, *19*, 596-604; b) S. Bershtein, M. Segal, R. Bekerman, N. Tokuriki, D. S. Tawfik, *Nature* **2006**, *444*, 929-932; c) E. Firnberg, J. W. Labonte, J. J. Gray, M. Ostermeier, *Mol Biol Evol* **2014**, *31*, 1581-1592; d) M. Soskine, D. S. Tawfik, *Nat Rev Genet.* **2010**, *11*, 572.
[18] J. D. Bloom, S. T. Labthavikul, C. R. Otey, F. H. Arnold, *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 5869-5874.
[19] K. S. Sarkisyan, D. A. Bolotin, M. V. Meer, D. R. Usmanova, A. S. Mishin, G. V. Sharonov, D. N. Ivankov, N. G. Bozhanova, M. S. Baranov, O. Soylemez, *Nature* **2016**, *533*, 397.
[20] a) N. Tokuriki, F. Stricher, L. Serrano, D. S. Tawfik, *PLoS Comput. Biol.* **2008**, *4*, e1000002; b) H. Yu, Y. Yan, C. Zhang, P. A. Dalby, *Sci. Rep.* **2017**, *7*, 41212; c) C. Chen, J. Lin, Y. Chu, *BMC Bioinf.* **2013**, *14*, S5.
[21] R. Guerois, J. E. Nielsen, L. Serrano, *J. Mol. Biol.* **2002**, *320*, 369-387.
[22] E. H. Kellogg, A. Leaver-Fay, D. Baker, *Proteins: Struct., Funct., Bioinf.* **2011**, *79*, 830-838.
[23] V. Parthiban, M. M. Gromiha, D. Schomburg, *Nucleic Acids Res.* **2006**, *34*, W239-W242.
[24] a) M. Musil, J. Stourac, J. Bendl, J. Brezovsky, Z. Prokop, J. Zendulka, T. Martinek, D. Bednar, J. Damborsky, *Nucleic Acids Res.* **2017**, *45*, W393-W399; b) S. Khan, M. Vihinen, *Human mutation* **2010**, *31*, 675-684.
[25] V. Potapov, M. Cohen, G. Schreiber, *Protein Eng. Des. Sel.* **2009**, *22*, 553-560.
[26] a) O. Buß, J. Rudat, K. Ochsenreither, *Comput. Struct. Biotechnol. J.* **2018**, *16*, 25-33; b) N. J. Christensen, K. P. Kepp, *J. Chem. Theory Comput.* **2013**, *9*, 3210-3223.
[27] V. J. Frauenkron-Machedjou, A. Fulton, L. Zhu, C. Anker, M. Bocola, K. E. Jaeger, U. Schwaneberg, *ChemBioChem* **2015**, *16*, 937-945.
[28] H. Zhao, W. Zha, *Nat. Protoc.* **2006**, *1*, 1865-1871.
[29] E. M. Nordwald, G. S. Armstrong, J. L. Kaar, *ACS Catalysis* **2014**, *4*, 4057-4064.
[30] a) M. J. Thiele, M. D. Davari, M. König, I. Hofmann, N. O. Junker, T. Mirzaei Garakani, L. Vojcic, J. r. Fitter, U. Schwaneberg, *ACS Catalysis* **2018**, *8*, 10876-10887; b) Y. Ji, A. M. Mertens, C. Gertler, S. Fekiri, M. Keser, D. F. Sauer, K. E. Smith, U. Schwaneberg, *Chem. Eur. J.* **2018**, *24*, 16865-16872.
[31] L. J. Strategene, California, *Instruction Manual* **2003**.
[32] J. Viña-Gonzalez, D. Jimenez-Lalana, F. Sancho, A. Serrano, A. T. Martinez, V. Guallar, M. Alcalde, *Adv. Synth. Catal.* **2019**, *361*, 1-13.
[33] S. Teng, A. K. Srivastava, L. Wang, *BMC genomics* **2010**, *11*, 1.
[34] E. Capriotti, P. Fariselli, R. Casadio, *Nucleic Acids Res.* **2005**, *33*, W306-W310.
[35] O. Khersonsky, R. Lipsh, Z. Avizemer, Y. Ashani, M. Goldsmith, H. Leader, O. Dym, S. Rogotner, D. L. Trudeau, J. Prilusky, P. Amengual-Rigo, V. Guallar, D. S. Tawfik, S. J. Fleishman, *Mol Cell* **2018**, *72*, 178-186.e175.
[36] a) K. A. Luke, C. L. Higgins, P. Wittung-Stafshede, *FEBS J.* **2007**, *274*, 4023-4033; b) K. S. Siddiqui, *Crit. Rev. Biotechnol.* **2017**, *37*, 309-322.
[37] R. J. Fox, S. C. Davis, E. C. Mundorff, L. M. Newman, V. Gavrilovic, S. K. Ma, L. M. Chung, C. Ching, S. Tam, S. Muley, *Nat. Biotechnol.* **2007**, *25*, 338-344.
[38] I. Codexis, "MOSAIC" can be found under https://www.codexis.com/services/protein-engineering, **2002**.
[39] D. Gonzalez-Perez, P. Molina-Espeja, E. Garcia-Ruiz, M. Alcalde, *PLoS One* **2014**, *9*, e90919.
[40] W. P. Stemmer, *Nature* **1994**, *370*, 389.
[41] a) I. Mateljak, A. Rice, K. Yang, T. Tron, M. Alcalde, *ACS Synthetic Biology* **2019**, *8*, 833-843; b) C. R. Otey, M. Landwehr, J. B. Endelman, K. Hiraga, J. D. Bloom, F. H. Arnold, *PLoS biology* **2006**, *4*, e112; c) M. M. Meyer, L. Hochrein, F. H. Arnold, *Protein Eng. Des. Sel.* **2006**, *19*, 563-570; d) C. A. Voigt, C. Martinez, Z.-G. Wang, S. L. Mayo, F. H. Arnold, *Nat. Struct. Mol. Biol.* **2002**, *9*, 553-558.
[42] J. Marienhagen, A. Dennig, U. Schwaneberg, *BioTechniques* **2012**, *52*.
[43] A. Dennig, A. V. Shivange, J. Marienhagen, U. Schwaneberg, *PLoS One* **2011**, *6*, e26222.
[44] J. Zhao, N. Jia, K. E. Jaeger, M. Bocola, U. Schwaneberg, *Biotechnol. Bioeng.* **2015**, *112*, 1997-2004.
[45] J. Van Durme, J. Delgado, F. Stricher, L. Serrano, J. Schymkowitz, F. Rousseau, *Bioinformatics* **2011**, *27*, 1711-1712.
[46] E. Krieger, G. Koraimann, G. Vriend, *Proteins: Struct., Funct., Bioinf.* **2002**, *47*, 393-402.
[47] G. van Pouderoyen, T. Eggert, K.E. Jaeger, B. W. Dijkstra, *J. Mol. Biol.* **2001**, *309*, 215-226.
[48] W. L. DeLano, http://www.pymol.org **2002**.

10.1002/chem.201903994

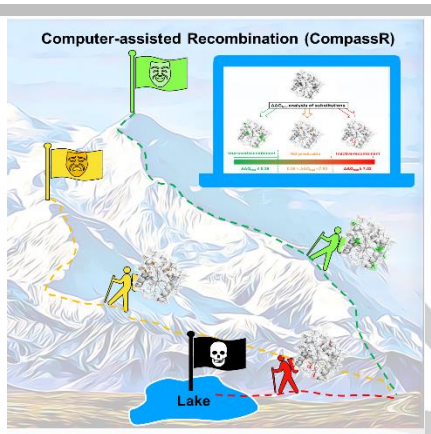
WILEY-VCH

FULL PAPER

Entry for the Table of Contents (Please choose one layout)

RESEARCH ARTICLE

The Computer-assisted Recombination (CompassR) strategy provides a selection guide for beneficial substitutions that can be recombined to gradually improve enzyme performance by analysis of the relative free energy of folding ($\Delta\Delta G_{fold}$)



*Haiyang Cui, Hao Cao, Haiying Cai, Karl Erich Jaeger, Mehdi D. Davari, Ulrich Schwaneberg**

Page No. – Page No.

Computer-assisted Recombination (CompassR) teaches us how to recombine beneficial substitutions from directed evolution campaigns