Cite this: *Chem. Commun.*, 2015, 51, 9760

Directed evolution 2.0: improving and deciphering enzyme properties

Feng Cheng,^{†a} Leilei Zhu^{†a} and Ulrich Schwaneberg^{*ab}

Directed evolution has matured to a routinely applied algorithm to tailor enzyme properties to meet the demands in various applications. In order to free directed enzyme evolution from methodological restraints and to efficiently explore its potential, many different strategies have been used in directed evolution campaigns. Analysis of directed evolution campaigns reveals that traditional approaches, in which several iterative rounds of diversity generation and screening are performed, are gradually replaced by strategies which require less time, less screening efforts, and generate a molecular understanding of the targeted properties. In this review, conceptual advances in knowledge generating directed evolution strategies are summarized, compared to each other and to traditional directed evolution strategies. Finally, a 'KnowVolution' (knowledge gaining directed evolution) termed strategy is proposed.

Received 22nd February 2015,
Accepted 2nd April 2015

DOI: 10.1039/c5cc01594d

www.rsc.org/chemcomm

1. Introduction

Several excellent reviews^{1–6} have been published on directed protein evolution covering advances and remaining challenges in enabling technologies (diversity generation, high-throughput screening, and modelling) and lessons learned from many successful directed evolution campaigns. In the first slow pick up phase, the number of directed evolution reports from the late

1960s (extracellular evolution of a self-replicating nucleic acid molecule⁷) to approximately 1996 was less than 40 according to ISI Web of Science (Core collection; search term "Directed Evolution"). Until 2006, a rapid boom in the field was observed (~320 reports in 2006), followed by slowdown in growth to around 450 reports per year since 2012 (2012: ~450; 2013: ~480; 2014: ~450). Nowadays, protein engineering by directed evolution has become a standard method to tailor enzyme properties in academia and industry (mainly for biocatalysis⁸ purposes). Novel application fields in material science⁹ and medical science^{10–12} will likely lead to a further growth in directed evolution reports despite that the development time for improving enzymes by directed evolution is often still too long to be implemented as a standard operation in process development.

^a Lehrstuhl für Biotechnologie, RWTH Aachen University, Worringerweg 3, 52074 Aachen, Germany. E-mail: u.schwaneberg@biotec.rwth-aachen.de

^b DWI-Leibniz Institute for Interactive Materials, Forckenbeckstraße 50, 52056, Aachen, Germany

[†] These authors contributed equally.

Feng Cheng

Feng Cheng was born in Zhejiang Province, P. R. China, in 1986. He received his bachelor's degree in 2008 and master's degree in 2011 from the Zhejiang University of Technology, China. Currently, he is pursuing his doctoral degree under the supervision of Prof. Ulrich Schwaneberg at RWTH Aachen University, Germany. His current research interests focus on tailoring biocatalysts for industrial and medical applications by laboratory evolution and rational design.

Leilei Zhu

Leilei Zhu received her bachelor's degree (biotechnology) in 2005 and master's degree (biochemical engineering) in 2007 from Jiangnan University in Wuxi, China. Then she joined Prof. Schwaneberg's group and obtained her doctoral degree in biotechnology in 2010 at RWTH Aachen University, Germany. Afterwards, she stayed as a subgroup leader in the Schwaneberg group where she continued to pursue her interest in protein engineering. Her current research interests lie in protein engineering for therapeutic proteins and biohybrid systems.

Till today most directed evolution campaigns are performed in a traditional manner, in which usually random mutagenesis libraries are generated through epPCR with low mutation frequencies (1 to 3 mutations per 1000 bp) and screened (MTP format; often 1000–2000 variants) in order to obtain improved variants. Gene(s) encoding the best or the best few variants are then isolated and used in an iterative process for the next round of directed enzyme evolution; often ≥ 4 rounds are required to achieve significant improvements. Improved variants can usually be found after screening of a few thousand clones, which represent only a negligible fraction of theoretical protein sequence space (e.g. a peptide with 6 amino acids can be composed of 64 million different variants). Being able to find improved variants in a few thousand variants was a great success, which has resulted in a lot of excitement and expectations. Meanwhile it was overlooked that directed evolution campaigns got hit by the experimental setup (bias of polymerase in diversity generation; the number of screened variants) which determines, in a similar manner to Heisenberg's observer effect, the outcome of directed evolution experiments. The excitement and result-oriented view is exemplarily summarized in a research news article in 1998 entitled "When blind is better: Protein design by evolution".¹³ Retrospectively, we have to ask what have we learned from thousands of success stories in directed evolution? If we are honest, only a few lessons and experimentally derived rules/trends were reported such as "you only get what you screen for".¹⁴ A molecular understanding of improved enzyme properties and underlying principles to reengineer enzymes was usually not pursued. In the last ten years directed evolution matured, since there are more and more researchers excited to understand improvements of enzyme properties on the molecular level, and to discover the underlying interaction principles. In this review, we summarized advances in knowledge-generating directed evolution strategies which aim to generate a deeper molecular understanding of enzyme properties and

compared them to each other as well as to traditional directed evolution. We thereby aim to summarize the state of the art for experimentalists and to equip them with information to design efficient directed evolution strategies for their specific directed evolution case. This review does not summarize the advances in the development of new screening/selection systems and the *de novo* synthesis of enzymes combined with directed evolution, which have been excellently summarized in recent reviews^{15–17} and articles.^{18,19}

2. Protein evolution strategies

This review is divided into three parts: traditional directed evolution, sequence and/or structure based enzyme evolution strategies as well as emerging combined strategies. Traditional directed evolution has been successfully used thousands of times to improve enzyme properties. The traditional directed evolution campaign is goal-oriented and does not usually result in a molecular understanding of the obtained improvement. Sequence and/or structure based strategies have been developed in the past decade to make directed protein evolution faster and more efficient by taking better advantage of available structural and sequence information. Combined strategies, such as CASTing, FRESCO, ProSAR, MORPHING, and KnowVolution, aim to improve enzyme properties with minimized screening efforts and gain a molecular understanding of improved properties. Deciphering molecular principles that determine enzyme properties delivers on the long run valuable knowledge to guide protein engineering campaigns.

2.1. Traditional directed evolution

Traditionally, directed evolution includes iterative rounds (often ≥ 4) of diversity generation (usually epPCR) with a mutation frequency of 1 to 3 mutations per 1000 bp and screening.²⁰ Microtiter plates are classically used as screening formats and often 1000–2000 clones are screened per one round of directed evolution. Gene(s) encoding either the best or the best few variants are used for the subsequent round of directed evolution and significantly improved enzymes often containing 6 to 12 substitutions.²¹ The latter makes it in general difficult to elucidate the role of each amino acid substitution and to gain a molecular understanding of the improved properties. Furthermore, many amino acid exchanges are, due to statistical reasons and as a consequence of polymerase bias as well as the number of screened clones, located on the protein surface, and cannot be analyzed computationally (require ms of MD-simulation time). It has therefore been a puzzling question why beneficial variants can be found after screening of only a few thousand clones which represent a negligible amount of the theoretically possible diversity. A detailed analysis of 3000 mutations in the lipase gene *bsla*, which were generated using three random mutagenesis methods (1000 mutations each: epPCR-low, 1000 PCR-high, and SeSaM-Tv P/P) and their comparison to a site saturation mutagenesis (SSM) library of *bsla* which covers the natural diversity at each amino acid position of BSLA (in total 181 positions), yielded first

Ulrich Schwaneberg

Ulrich Schwaneberg received his doctoral degree in chemistry in 1999 at the Institute of Technical Biochemistry, University of Stuttgart, Germany. Afterwards, he joined CalTech (California Institute of Technology) as a postdoctoral researcher for two years. In 2002, he was appointed professor of biochemical engineering at International University Bremen. In 2009, he moved to RWTH Aachen University as chair of Institute of Biotechnology. In 2010, he was co-appointed at the DWI (Leibniz Institute for Interactive Materials). He represents the RWTH Aachen in the Bioeconomy Science Centre as a director and is a speaker of HICAST (Henkel Innovation Campus for Advanced and Sustainable Technologies).

experimental and quantitative data on the obtainable diversity.^{22,23} The SSM-BSLA library was generated in 181 site saturation mutagenesis experiments, subsequent sequencing of 18 547 clones and subsequent site-directed mutagenesis experiments to obtain missing substitutions and to ensure that the full natural diversity is available.²³ BSLA resistances towards an ionic liquid [C₄mim][TfO] were studied by screening the SSM-BSLA library and in total 292 beneficial substitutions at 104 different amino acid positions were identified. Among the beneficial positions, 15 positions were obtained by epPCR-low and 18 by epPCR-high, which proved that ~80% beneficial amino acid positions were missed in a traditional directed evolution experiment with BSLA (partially unpublished data).²² To increase the quality of mutant libraries (e.g. more diversity at the protein level, control over biases) and to simplify procedures, new diversity generating methods² were developed in the last 5 years, such as TRINS²⁴ (tandem repeat insertion) and SeSaM-Tv-II²⁵ (sequence saturation mutagenesis). Furthermore, computational tools which assist research in their directed evolution campaigns by *in silico* analysis of the generated diversity have been reported (e.g. MAP²⁶ and PEDAL-AA²⁷). Thereby, mutational biases can for instance be varied in the iterative round of evolution or randomization strategies for library constructions (Codon Calculator²⁷ and AA-Calculator²⁷) can be used, and the saturation mutagenesis experiment (TopLib²⁸) can be designed. Such computer-aided tools for directed protein evolution have well been reviewed by Roccatano *et al.*²⁹

2.2. Sequence- and/or structure-based protein evolution strategies

In the last few years, structure-based and sequence-based semi-rational strategies have been developed and successfully applied in improving enzyme properties. These strategies utilize information on protein sequences, structure–function relationships, as well as computational predictive algorithms to preselect promising amino acid positions for property improvements. In most cases, the focus is on identifying a small number of amino acid positions (4–50 positions) to reduce library sizes and time requirements for protein reengineering.

2.2.1. Structural analysis based protein evolution strategy. Highlights of structural analysis-guided directed protein evolution strategies comprise SCOPE, CASTing, PTRec, and FRESCO.

SCOPE (structure-based combinatorial protein engineering)³⁰ is based on the concept of exon shuffling and generates multiple crossover libraries from distantly related genes. The strategy SCOPE was successfully used to evolve two DNA polymerases (rat DNA polymerase beta (Pol β) and African swine fever virus DNA polymerase X (Pol X)). Both polymerases share similar folds but differ in size (Pol β- 335 vs. Pol X- 174 amino acid residues) and have a low sequence identity (18%).³⁰ Corresponding ‘structural elements’ in Pol X were identified based on an analysis of a homology model of Pol X and a sequence alignment between Pol β and Pol X. Libraries of chimeric genes were synthesized in a series of PCRs employing hybrid oligonucleotides that code for variable connections between structural elements. As a result, several novel DNA polymerases with an enhanced complementation phenotype were finally identified.³⁰

CASTing (combinatorial active-site saturation test) uses the information derived from structural/functional data to identify amino acids in the substrate binding pocket, which are subsequently and in a sequential manner subjected to mutagenesis (usually site-saturation and multi-site saturation with one primer based on the QuikChange mutagenesis method). The selected positions are located in the enzyme’s binding pocket in order to usually improve the activity and selectivity of targeted enzymes.³¹ One prominent example is the lipase from *Pseudomonas aeruginosa*, in which five sets of amino acid pairs are located in the binding pocket and in close proximity to each other. In each of the five libraries, two amino acid positions were saturated simultaneously and the hydrolytic activity of the lipase was significantly enhanced for 11 esters.³¹ A further highlight is the epoxide hydrolase from *Aspergillus niger* (ANEH) with an increase in the *E* value (10.8 to 115 for hydrolysis of glycidyl phenyl ether) after five rounds of site saturation mutagenesis.³²

The PTRec (phosphorothioate-based DNA recombination) approach was developed for the combinatorial assembly of multiple DNA fragments with rationally preselected cross-over points and can be used for recombining genes with less than 50% sequence identity on the protein level.³³ PTRec includes four experimental steps: (1) amplification of individual domains with primers harboring 12 subsequent phosphorothioated nucleotides, (2) multiple chemical cleavage of phosphorothioated nucleotides to generate single-stranded overhangs (12 nts) at crossover points, (3) recombination and hybridization of all DNA fragments in a single tube, and (4) transformation into bacterial cells without further purification. As proof of concept, a recombination library with three phytase genes (with less than 50% sequence identity) and four crossover points was generated. Sequencing of 42 phytase genes confirmed a nearly equal distribution of recombined domains without frameshifts. In all 42 cases, the recombined domains were in the expected order. PTRec is a ligase-free and restriction site-independent method which enables recombination of more than two genes and requires only 4 identical amino acids to define a crossover point. The latter enables to efficiently recombine genes at much lower sequence identity than random recombination methods such as gene-shuffling³⁴ or the STEP (staggered extension process).³⁵ Gene shuffling requires at least 80% sequence identity³⁶ to generate ~0–3 crossover points, and even the most diverse random recombination methods such as ITCHY,³⁷ and SCRATCHY³⁸ generate at 50% sequence identity ~0–3 crossover points and are limited to two genes.³⁶

FRESCO (framework for rapid enzyme stabilization by computational library strategy)³⁹ was developed for targeting enzyme stability. FRESCO identifies potential amino acid positions and individual amino acid substitutions using the programs Rosetta and FoldX. Substitutions will be analyzed in terms of free energy changes ($\Delta\Delta G^{\text{Fold}} < -5 \text{ kJ mol}^{-1}$ or $\Delta\Delta G^{\text{Fold}}$) in the range of -5 to $+5 \text{ kJ mol}^{-1}$. As a general trend, one could observe that amino acid substitutions that belong to certain types (*i.e.* XXX to Arg, XXX to Pro, and Gly to XXX) are often observed to improve thermal resistance of enzymes. Disulfide bonds will also be considered and conformational space of the enzyme backbone will be analyzed to ensure sufficient flexibility, thereby maintaining a high activity.

Orthogonal *in silico* screening steps will be performed next, to exclude substitutions that are chemically unreasonable (e.g. a surface exposed hydrophobic side chain or an unsatisfied H-bond donor or acceptor) or are predicted to increase enzyme flexibility. After the selection, the predicted potentially beneficial substitutions will be subjected to MD simulations to exclude mutations that increase the local flexibility of the targeted enzyme. At last, the selected amino acid positions will be subjected to site directed mutagenesis (SDM) and subsequent screening. Obtained beneficial substitutions will be recombined subsequently using SDM. Using FRESCO, the thermostability of a limonene epoxide hydrolase (LEH)³⁹ was significantly increased (T_m increase by 25 °C) with minimal screening efforts (only 64 variants were experimentally screened). In the case of the haloalkane dehalogenase LinB,⁴⁰ the T_m was increased by 23 °C with an over 200-fold prolonged half-life at 60 °C after screening of <80 clones.

Furthermore, a number of useful computational programs (e.g. SiteComp,⁴¹ 3DLigandSite,⁴² and ProBiS-ligands⁴³) were developed for the analysis of ligand binding sites in proteins and have been reviewed.⁴⁴

2.2.2. Sequence based protein evolution strategy. Multiple sequence alignments (MSA) represent a common approach to identify functionally significant or evolutionarily variable amino acids or regions in a protein. MSA-based methods like ConSurf 2010⁴⁵ have been intensively summarized in other reviews.^{29,46}

The REAP (reconstructed evolutionary adaptive path) is a notable highlight, and exploits sequence data from ancestral proteins for the generation of focused and functionally enriched libraries.⁴⁷ In an example of evolving DNA polymerase, Chen *et al.* explored the split in the phylogenetic tree of the DNA polymerase family A into viral and non-viral subgroups to identify mutations in gene sequences that emerged during functional divergence from a common universal ancestor. Viral polymerases are known for their broad substrate promiscuity compared to non-viral enzymes; hence differences in the sequences of the predicted ancestral polymerase at an evolutionary branching point were postulated to account for substrate specificity changes without reducing catalytic performance. By comparing the sequence information of viral polymerases with non-viral polymerases, a library of only 93 Taq variants targeting 35 amino acid positions close to the active site were screened and yielded eight variants with improved ability to accept dNTP-ONH₂ as substrates.

2.2.3. Structure and sequence based protein evolution strategy. Combining structural analysis and sequence information has further been used to identify functionally important residues. Broadly and successfully applied software packages are among others, 3DM⁴⁸ and HotSpot Wizard.⁴⁹ A further interesting structure-sequenced based approach was used to alter the pH profile of proteases through surface charge engineering.⁵⁰

The 3DM⁴⁸ information system performs structure based multiple sequence alignments (MSA) of the members from a protein superfamily and provides information of structure-function relationships⁵¹ which is collected from PDB, GenBank, PubMed and Swiss-Prot. Notably, the enantioselectivity of a *Pseudomonas fluorescens* esterase was improved by simultaneous multi-site saturation at four amino acid positions, resulting in

the identification of variants with improved activity (up to 240-fold) and increased enantioselectivity (up to $E_{true} = 80$) towards 3-PBA-ethyl ester.⁵²

HotSpot Wizard⁴⁹ integrates several bioinformatics databases for evolutionary analysis and computational tools for structural analysis, representing an efficient approach to identify highly conserved sites and potential positions/substitutions which can be altered to improve substrate specificity, activity or enantioselectivity. For instance, six potential positions that influence the activity of NfsB (bacterial type I nitroreductase) towards 7-amino-benzodiazepine production (clinical sedative hypnotic drugs) were identified using the HotSpot Wizard.⁵³ SDM experiments showed that especially the positions N71 and F124 were beneficial; the combined substitutions N71S/F124W yielded a variant with increased production under anaerobic (290 ng vs. 51 ng(WT)) and aerobic (348 ng vs. 30 ng(WT)) conditions.⁵³ More examples can be found in the review.⁵⁴

A combined (sequence-structure) based strategy for shifting pH-activity profiles and pH optima through surface charge engineering was reported for a high alkaline protease (*Bacillus gibsonii* alkaline protease (BgAP)) by Jakob *et al.*⁵⁰ Nine residues (Asn and Gln) were identified on the BgAP surface, which fulfill all three defined selection criteria (i. the Asn or Gln residues are not conserved within the enzyme family, ii. are surface exposed, and iii. neighbored by glycine which promotes spontaneous deamidation⁵⁵). Eight positions were finally selected to substitute Asn and/or Gln by the corresponding Asp and Glu since one position was closed to N-terminus. Seven out of 8 variants showed a higher activity at pH 8.6 and a final variant (N253D/Q256E) had a shift in pH optimum by 1.5 units and a doubled activity at pH 8.5 when compared to BgAP WT.

2.3. Conceptual comprehensive directed evolution strategies with combined methods

Comprehensive protein engineering strategies for directed evolution with combined methods are currently emerging. All have in common to reduce experimental efforts in high-throughput screening and thereby gain a molecular understanding of the improved properties. Highlights comprise the ProSAR strategy, the MORPHING strategy, and the KnowVolution strategy, which we propose based on several successful directed evolution campaigns.

The ProSAR⁵⁶ (protein sequence activity relationship analysis based strategy) is an iterative process of diversity generation (random mutagenesis, site saturation mutagenesis, and gene shuffling), screening followed by statistical analysis on sequence-activity datasets derived from one or more combinatorial libraries per round. The sequence-activity data, obtained from the previous library screening and sequencing, are used to build a statistical model that assigns a regression coefficient to each substitution corresponding to the impact of substitutions on the activity. The ProSAR approach enables improvement by the combination of substitutions that do not substantially improve an enzyme property individually. Using the outlined strategy, a halohydrin dehalogenase was successfully improved (4000-fold augmented volumetric productivity in the synthesis of the side chain of atorvastatin (Lipitor)).⁵⁶ It is difficult without detailed

insights to evaluate the amount of data and experimental effort required to obtain statistical-relevant datasets for a successful ProSAR analysis.

MORPHING⁵⁷ (mutagenic organized recombination process by homologous *in vivo* grouping) combines structural analysis with random mutagenesis and enables the random introduction of mutations in specific protein segments in one-pot. Firstly, the gene encoding the targeted enzyme is analyzed and divided into two categories of fragments: (1) fragments that are subjected to random mutagenesis and (2) fragments that are not mutagenized. At the 5' and 3'-end of the fragments, ~50 bp long sequences are introduced through PCRs to ensure an efficient homology recombination of the fragments. MORPHING was validated by Alcalde and co-workers in two case studies employing a versatile peroxidase (VP) and an unspecific peroxygenase (UPO). For example, in the case of VP, three fragments of the target areas were mutated by random mutagenesis (mutation loads: 1 to 5 mutations per segment) and three fragments were amplified by a high-fidelity polymerase. The assembly of these segments in three consecutive rounds of homologous recombination and screening yielded the 3G10 variant which showed a 2.1-fold improvement in stability in the presence of H₂O₂ and a 5.5 °C increase in the *T*₅₀.⁵⁷

KnowVolution, knowledge gaining directed evolution, comprises four phases (see Fig. 1). In phase I, a standard directed

evolution experiment is performed followed by sequencing of at least 20 of the most beneficial variants (depending on the gene length and the mutation frequency). As an outcome, around 12 positions should be identified that potentially contribute to the improved properties. Beneficial amino acid exchanges are thereby often clustered in close proximity within two to three regions of the targeted protein. In phase II, ~12 selected positions of phase I are subjected to saturation mutagenesis. Often less than 50% of the saturated positions contribute to the improvement of the targeted properties. Since the random mutagenesis methods usually generate 0 to 4 amino acid substitutions per amino acid position,⁵⁸ one would expect that saturation mutagenesis should nearly always yield more improved variants. The latter is however not the case since usually 2 to 3 of five saturated positions yield further improvements. Sequencing is again performed at each saturated position (~16 clones sequenced if beneficial variants are found; fewer if no beneficial variant is found in order to ensure that the site-saturation mutagenesis experiment worked well (no wild-type background or faulty primer)). Subsequent analysis of the obtained amino acid substitutions provides insights into the type of chemical interaction (charge, size, H-bond, and hydrophobicity) which contributes to the improvement of the targeted properties. The site saturation experiments should yield 4 to 6 amino acid positions with known substitution that

*Screening system is established

Fig. 1 Overview of the KnowVolution strategy which comprises four phases: (I) Identification of potentially beneficial amino acid positions, (II) determination of beneficial amino acid positions and substitutions, (III) computational analysis and a group of amino acid substitution which might interact with each other, and (IV) recombination of beneficial substitutions in a simultaneous or iterative manner. The KnowVolution strategy can also be performed in an iterative manner to further improve targeted enzyme properties.

contribute to an improvement of the targeted properties. In phase III, a computational analysis by structural inspection is performed to see in a 3D model whether amino acid substitutions are in close proximity or might even interact with each other. Based on the computer-assisted structural analysis, a decision is made how the beneficial substitutions should be combined; how these positions should be grouped and which amino acid substitutions should be generated at the selected positions. In phase IV, the recombination is performed depending on the results of phase III. Multi-site saturation mutagenesis experiments should preferentially be performed if several positions are in close proximity to each other. The number of targeted positions and the diversity of generated amino acid substitutions should be limited in order to ensure an efficient screening. For instance, “only” beneficial substitutions from the site-saturation mutagenesis libraries are included in the multi-site saturation experiment (see case I-glucose oxidase). In case that there are only four positions with one amino acid exchange, a one by one recombination or order of a synthetic gene with all four substitutions are time-efficient alternatives. The saturation mutagenesis experiments in phase II and recombination experiments in phase IV yield often improvements that are comparable or better than those obtained through a whole round of directed evolution despite significantly reduced screening efforts. Highlights of the gained knowledge are summarized in six directed evolution examples (glucose oxidase (GOx), phytase (Ymphytase), alkaline protease (BgAP), arginine deiminase (PpADI), serine alkaline protease (subtilisin E), and cellulase (CelA)), which followed the outlined KnowVolution strategy. The key in the KnowVolution process is to identify a sufficient amount of beneficial positions in phases I–III for efficient recombination in phase IV.

Further iterative rounds of the KnowVolution strategy (see Fig. 1; identification (phase I), determination (phase II), selection (phase III) and recombination (phase IV)) can be performed to further improve targeted enzyme properties.

Table 1 summarizes 6 case studies in which KnowVolution and related strategies were used in protein engineering. In all cases, potentially beneficial positions (phase I) were identified in one to three random mutagenesis experiments employing diversity generation epPCR and/or SeSaM and subsequent screening of 1056 to 6800 clones per round of directed evolution. In total 2000 to 12 700 clones were screened. In all cases, 11 to 19 potentially beneficial positions were identified and subsequent site saturation mutagenesis (phase II) yielded 4 to 9 amino acid positions which improved a targeted property. As a general trend, one could observe that the number of potentially beneficial positions is usually ≥ 2 times higher than the number of positions which finally contributed to a significant property improvement in phase II. The recombination of beneficial substitutions (phase IV) was performed after structural inspection *via* OmniChange⁵⁹ (simultaneous site-saturation mutagenesis; 2 examples: case 1-GOx and case 2-Ymphytase) and site directed mutagenesis (4 examples: case 3-BgAP, case 4-PpADI, case 5-subtilisin E and case 6-CelA). In most cases, four to six amino acid exchanges were sufficient to obtain dramatic improvements in targeted properties (case 1: GOx V7 with four substitutions displayed

37-fold decreased oxygen dependency; case 3: BgAP MF1 with six substitutions showed over 100 times prolonged half-life at 60 °C, and case 5: subtilisin E M6 harboring six substitutions exhibited a 193-fold increased half-life in 1 M GdmCl).

Case 1. Reengineering of glucose oxidase (GOx) for amperometric glucose determination by reducing oxygen dependency and increasing specific activity. GOx from *Apergillus niger* has been used in amperometric systems for diabetes care to determine the β -D-glucose concentration in blood.⁶⁰ The activity of GOx is highly dependent on oxygen which makes it difficult to determine with one enzyme β -D-glucose levels in arterial and venous blood. By employing the KnowVolution strategy, Gutierrez *et al.*⁶¹ gained a molecular understanding of (a) increased specific activity for the mediator (quinone diimine) and (b) the oxygen binding site. In two random mutagenesis experiments (1 $\times$ epPCR followed by 1 $\times$ SeSaM), 13 potentially beneficial sites were identified to increase specific activity or reduce oxygen dependency (phase I). Individual site saturation mutagenesis yielded four beneficial positions (A173, A332, F414, and V560), in which some substitutions (*e.g.* A332S, V560A) contributed to decreased oxygen activity and slightly increased mediator activity (phase II). Structural analysis showed that these four positions clustered around the FAD domain and at the entrance of a putative oxygen channel (phase III). The final variant GOx V7 (A173V/A332S/F414I/V560T) with a 37-fold decreased oxygen dependency was obtained from an OmniChange library by screening only 2112 variants (phase IV). Computational analysis and individual SSM experiments revealed on the molecular level several structural determinants. The position V560, next to FAD in the substrate binding pocket, is important for ‘oxygen activity’ and substitutions to L, P or T drastically reduced oxygen activity. Position F414 is responsible for increased mediator activity by substitutions from F to M, V, L or I. F414 is located at 8.2 Å above the FAD in the substrate binding pocket and involved in β -D-glucose binding, which indicated that the mediator can directly bind in the binding pocket to reoxidize FADH₂ efficiently. The substitutions on A173 (located on the surface of GOx) and A332 (located at the substrate entrance channel) lead to an increased activity (up to 3-fold).

Case 2. Reengineering of the phytase (Ymphytase) for improved thermal resistance, pH stability, and increased specific activity. Ymphytase (47 kDa) from *Yersinia mollaretii* is a feed supplement in animal (*e.g.* pigs) feeding for improving phytate-rich diets by catalyzing the hydrolysis of phosphate from phytic acid.⁷⁰ Phytases have to withstand a heat incubation step during the pelleting process. Shivange *et al.*^{62,63} used the ‘KnowVolution’ strategy to improve and to gain a molecular understanding of (a) thermal resistance and (b) specific activity of Ymphytase. In phase I, two random mutagenesis libraries were generated (epPCR and SeSaM), in total 9150 clones were screened and 9 potentially beneficial positions were identified. Subsequently, in phase II, individual site saturation mutagenesis to identify the beneficial positions was performed and five positions (D52, T77, K139, G187, and V298) were discovered that were either beneficial for improving the thermal resistance or enhancing the specific activity. Computational analysis showed that D52 is located next to the active site loop (S42-T47), K139 is located on the

Table 1 Six case studies in which KnowVolution and related strategies were performed in order to gain a molecular understanding of targeted properties. The six case studies comprise a glucose oxidase from *A. niger*, a phytase from *Y. mollaretii*, an alkaline protease from *B. gibsonii*, an arginine deiminase from *P. plecoglossicida*, a serine alkaline protease from *B. subtilis* and a cellulase CelA2 isolated from a metagenomic library. Table 1 is organized according to the four phases (P.I identification; P.II selection and P.IV recombination). Computational analysis by structure guided selection of recombined amino acids (phase III) was performed in all six cases

Case no./targeted enzyme	P.I directed evolution (number of screened clones)	P.I number of potentially beneficial amino acid positions (number of sequenced clones)		P.II by SSM identified beneficial amino acid positions	P.IV recombination of identified beneficial substitutions (codon subsets; screened and sequenced clones)	The final variant/obtained improvement	Ref.
		1 × SeSaM (1056)	9 (31)				
1/glucose oxidase (GOx)	1 × epPCR (1232)	4 (22)		4 positions (A173, A332, F414, and V560)	Multi-site saturation with a reduced subset of amino acids using OmniChange (RVT codon at 173, ART codon at 332, NDT codon at 414, and VYK codon at 560; 2112 clones were screened and 5 were sequenced)	GOx V7 (A173V/A332S/F414I/V560T)/37-fold decreased oxygen dependency and 5.7-fold improved activity	61
2/phytase (Ymphytase)	1 × epPCR (6800) 1 × SeSaM (2350)	3 (30) 8 (20)		5 positions (D52, T77, K139, G187, and V298)	Multi-site saturation mutagenesis using OmniChange (NNK codon, 1110 clones were screened and 30 were sequenced)	Ymphytase Omni1 (D52E/K139T/G187S/V298F) 32% improved residual activity, 2 °C increased apparent T_m , and 2-fold higher pH stability	62 63
3/alkaline protease (BgAP)	3 × SeSaM (12 700)	11 (39)		6 positions with most beneficial substitution (I21V, S39E, N74D, D87E, M122L, and N253D)	Site-directed mutagenesis	BgAP MF1 (I21V/S39E/N74D/D87E/M122L/N253D)/1.5-fold increased specific activity at 15 °C and over 100 times prolonged half-life at 60 °C	64
4/arginine deiminase (PpADI)	3 × epPCR (6300) 1 × SeSaM (3000)	12 (31) 7 (23)		9 positions with most beneficial substitution (K30R, C37R, D38H, D44E, A128T, L148P, V291L, E296K, and H404R)	Site-directed mutagenesis	PpADI M19 (K5T/K30R/C37R/D38H/D44E/A128T/L148P/V291L/E296K/H404R)/105.5-fold increased k_{cat} value and 97-fold reduced IC_{50} value for a melanoma cell line (G361)	11 65 66
5/serine alkaline protease (subtilisin E)	2 × epPCR (2600) 1 × SeSaM (1300)	7 (20) 5 (10)		6 positions with most beneficial substitution (S62I, A153V, G166S, I205V, N218S, and T224A)	Site-directed mutagenesis	Subtilisin E M6 (S62I/A153V/G166S/I205V/N218S/T224A)/193-fold increased half-life from <2 to 385 min, 1 M GdmCl, and 73-fold increased from <2 to 146 min in 0.5% SDS	67 68
6/cellulase (CelA2)	1 × epPCR (2000)	12 (18)		6 positions with most beneficial substitution (L21P, L184Q, H288R, K299I, D330G, and N442D)	Site-directed mutagenesis	CelA2 4D1 (L21P/L184Q/H288R/K299I/D330G/N442D)/7.5-fold increase in specific activity in 30% (v/v) ChCl: Gly and 1.6-fold in concentrated seawater	69

Ymphytase surface, G187 is in the α -domain and V298 is located on a helix in the $\alpha\beta$ -domain and surrounded by hydrophobic amino acids (phase III). In phase IV, an OmniChange library of these five positions was generated and screening of 1100 clones yielded a variant Omni1 that showed a 41 U mg⁻¹ improved activity, a 2 °C increased T_m and a 2-fold higher pH stability (at pH 2.8).⁶³ Furthermore, MD simulation analysis of Ymphytase WT and Omni1 revealed that a strengthened the hydrogen bonding network (e.g. G187S, K289E/K289Q) and a salt-bridge (T77K) reduced the overall dynamics of Ymphytase, especially the flexibility in the loops (loop-B4: L92-T102, loop-K10: Q304-Q317, and loop-J1: N247-P252) located next to helices (B, F, and K). Interestingly, the local flexibility in the active site loop (S42-T47) was increased by the substitution D52N, which likely contributed to the higher specific activity.

Case 3. Reengineering of the alkaline protease (BgAP) for increased activity at low temperatures (15 °C). BgAP (27 kDa) from *Bacillus gibsonii* is a high-alkaline protease and attractive for laundry applications. Proteases with improved specific activities at 20 °C and 40 °C would enable consumers to reduce washing temperatures, energy consumption, and CO₂ emissions. Improving the specific activity of BgAP at low temperatures has to be performed without losing BgAPs' storage stability which correlates well to its thermal resistance. Martinez *et al.*⁶⁴ used the KnowVolution strategy to improve (a) specific activity of BgAP at low temperatures and (b) to increase its thermal resistance. Three iterative rounds of random mutagenesis with SeSaM yielded 11 potential substitutions after screening of 12 700 clones (phase I). In phase II, 6 out of 11 beneficial amino acid positions were identified by performing saturation mutagenesis. Structural inspection (phase III) resulted in grouping the positions into two subsets. One subset contained the two positions (I21 and M122) which were close to the active site of BgAP, whereas S39, N74, D87, and N253 were placed in the other subset (located at the surface of BgAP). Recombination (phase IV) of two subsets of these substitutions using SDM resulted in the final variant MF1 with a 1.5-fold increased specific activity at 15 °C and >100-times prolonged half-life at 60 °C (224 min compared to 2 min of the WT BgAP). The four surface amino acid positions were mutated to negatively charged amino acids (Glu and Asp). In the final variant MF1 (S39E, N74D, D87E, and N253D), a high activity and high thermal resistance of BgAP could be obtained.

Case 4. Reengineering of the arginine deiminase (PpADI) for increased activity under physiological conditions. PpADI from *Pseudomonas plecoglossicida* is an arginine-metabolizing enzyme and a potential anti-tumor enzyme for treating arginine-auxotrophic tumors such as hepatocellular carcinomas (HCCs) and melanomas.⁷¹ Zhu and Cheng *et al.*^{11,65,66} used the 'KnowVolution' strategy to improve PpADI's specific activity at low arginine concentrations (0.1 mM under physiological conditions in blood) and to gain molecular understandings to shift the pH-optimum and to decrease the $S_{0.5}$ ($'K_m'$) value of PpADI. In phase I, three iterative rounds of ePCR and one round of SeSaM were performed and in total 9300 clones were screened. Screening yielded 13 potentially beneficial substitutions for decreased $S_{0.5}$ ($'K_m'$) values or increased

k_{cat} values. Subsequent SSM (phase II) results showed that 6 positions were responsible for increased activity of PpADI under physiological conditions. Structural inspection (phase III) showed that the positions of K30, C37, D44E, and H404R are located on two loops (loop 1 (from R30 to I46) and loop 4 (from 393 to 404)) which are next to the active site and undergo a conformational transition upon arginine binding;⁷² positions L148 and E296 were located at the interface between two monomers of PpADI. The identified beneficial substitutions from SSM libraries were recombined one by one using SDM (phase IV), and resulted in final variant PpADI M19 with a 105.5-fold higher k_{cat} , reduced $S_{0.5}$ (1.3 mM (WT) to 0.35 mM), and a pH optimum shift from 6.5 to 7.5. Computational analysis and individual substitution experiments identified on the molecular level several structural determinants. The substitutions on the loops 1 and 4, for example C37R, lead to an increase in flexibility of the side chain of a gating residue R400 (break of a hydrogen bond between H38 and R400) and are important for the specific deiminase activity. The pH-optimum of PpADI could be shifted by a histidine to arginine substitution (H404R). Interestingly, two substitutions (L148P and E296K) located at the interface of the dimeric PpADI can yield nearly exclusive PpADI tetramers with a lowered $S_{0.5}$ value. Recently, a 'KnowVolution' derived strategy was employed to improve the thermal resistance⁷³ of PpADI and proved that several properties of PpADI can be improved.

Case 5. Reengineering of the protease subtilisin E for improved tolerance towards chaotropic salts. Subtilisin E (36 kDa) from *Bacillus subtilis* was identified to possess a high resistance toward chaotropic salts. In order to efficiently extract DNA from blood samples, proteins have to be efficiently degraded in the presence of high guanidinium chloride concentrations (GdmCl; 1 M).⁷⁴ Li *et al.* used the KnowVolution strategy to improve subtilisin E's resistance against GdmCl.^{67,68} In phase I, three random mutagenesis libraries were generated (2 × ePCR and 1 × SeSaM) and after screening of 3900 clones, 10 potentially beneficial positions were identified. Saturation mutagenesis at each position of the wild-type subtilisin E (phase II) yielded four positions which contributed to improved chaotolerance. All the four beneficial positions were either located in close proximity to the substrate binding pocket or on the surface. Finally, the variant M6 generated by combination of the most beneficial single substitutions showed a significantly improved activity and resistance in the presence of 1 M GdmCl (half-life increased 193-fold from <2 min to 385 min (phase IV)). Finally, MD simulations suggested that Gdm⁺ binding happens in two stages: firstly an interaction with His64 which results in the displacement of His64 and secondly a tighter binding of Gdm⁺ close to the active site. The obtained substitution S62I prevents His64 displacement and Gdm⁺ disturbance within the active site.⁶⁸

Case 6. Reengineering of the cellulase (CelA2) for improved resistance towards ionic liquids (ILs), deep eutectic solvents (DES) and concentrated seawater. Cellulase CelA2 (69 kDa; GenBank: JF826524.1) was isolated by Prof. Streit's group from a metagenomic library for the degradation of cellulose and hemi-cellulose in the presence of ionic liquids (ILs), deep eutectic solvents (DESS) or concentrated seawater as solvents.

A KnowVolution strategy was pursued in order to further improve CelA2 resistance in the presence of ILs (e.g. [BMIM]Cl), DES (e.g. ChCl: Gly) and concentrated seawater. In phase I, a random mutagenesis library with high mutation frequency was generated, screened (2000 clones), and 6 potentially beneficial positions were identified in the variant 4D1. Only one position, position 288, significantly contributed to the improved IL, DES, and seawater resistance (phase II). Structural analysis (phase III) and further site saturation experiments yielded three positions (A285, H288, and S300).⁷⁵ The beneficial positions are mainly located at the entrance of the active site and in close proximity to the modelled substrate. Specific activities were 7.5-fold increased in 30% (v/v) ChCl: Gly (0.4 to 3.0 U mg⁻¹) and 1.6-fold in concentrated seawater (5.5 to 9.3 U mg⁻¹).

As a highlight a CelA2 variant which is nearly inactive in buffer and activated upon supplementing salt was discovered. The substitution to Arg300 forms with Asp287 a salt bridge next to the active site. When the salt bridge is formed, CelA2 becomes nearly inactive. Supplement of IL, DES, and saltwater neutralizes the salt bridge and activates CelA2. MD simulations supported that Arg300 acts as a key residue for the ionic strength activation.⁷⁵

Main lessons from KnowVolution cases are that in phase II (determination) less than half of the positions from phase I contribute to property improvement and that significant improvements can be achieved by site saturation mutagenesis (phase II) in all six KnowVolution campaigns which are comparable to improvements from a whole round of directed evolution. In phase IV (recombination), further significant improvements can be achieved in five of the six examples so that the project objectives were already achieved in all examples after a single round of KnowVolution. One further important lesson is that limiting the number of amino acid substitution through phase II is crucial to not reduce the 'application' fitness of the evolved enzyme (e.g. storage stability, thermal resistance, process performance).

3. Comparison of the methods for recombining amino acid substitutions

The above mentioned examples demonstrated the success of the KnowVolution strategy for improving enzyme properties of five hydrolases (specific activity, thermal resistance, IL/DES resistance, and substrate affinity) and a glucose oxidase (specific activity; oxygen dependency). A key issue in the KnowVolution approach is the recombination strategy of beneficial substitutions in phase IV. In cases 1 (GOx) and 2 (Ymphytase), multi-site saturation mutagenesis using OmniChange⁵⁹ was performed at 4 or 5 positions. After screening of only ≤ 3000 clones, significantly improved variants were obtained, e.g. GOx variant V7 (37-fold decreased oxygen dependency; 5.7-fold improved activity). In these two cases (1 and 2), the most beneficial variant carried different substitutions compared to the ones identified in the individual site saturation mutagenesis libraries. For instance, GOX V7 (A173V/A332S/F414I/V560T) showed a ~ 3.5 -fold improved oxygen and

mediator activity compared to the individual best variant V01 (A173T). This indicates that simultaneous multi-site saturation mutagenesis yields novel substitutions to exploit potential cooperative effects. SDM was performed in cases 3 to 6 to recombine beneficial substitutions. In cases 3, 4, 5, and 6, substitutions were introduced one by one after all the individual SSM libraries and their subsequent screening. In case 5, the SSM libraries and subsequent SDM recombinations were performed in between rounds of random mutagenesis. Selection of recombined substitutions was mainly based on obtained property improvements (e.g. >1.5 fold BgAP's activity at low temperature than WT, and >2 times PpADI's activity under physiological conditions compared to WT). In general, it was proved that four to six amino acid substitutions were sufficient to generate significantly improved hydrolase variants (case 3: BgAP MF1 with six substitutions showed over 100 times prolonged half-life at 60 °C; case 5: subtilisin E M6 with six substitutions displayed 193-fold increased half-life in 1 M GdmCl).

Table 2 compares common mutagenesis methods in terms of diversity, and ability to generate amino acid substitutions in an iterative manner or simultaneously. All mutagenesis methods in Table 2 can be performed in one day with comparable time requirements. Interestingly, there is a huge difference in the number of different protein variants that can be generated *via* iterative or simultaneous site saturation mutagenesis. In five subsequent rounds of SSM, one can generate $20 + 20 + 20 + 20 + 20 = 100$ different protein variants. CASTing which uses QuikChange based mutagenesis protocols can harbor two to three simultaneously mutated positions in one primer if the targeted positions are in close proximity to each other (≤ 20 amino acids). A QuikChange experiment would be sufficient to generate 400 (20×20 ; two positions) or 8000 (three positions) different protein variants. Due to primer length limitations, saturation of more than three positions has to our best knowledge not been reported for the CASTing strategy.^{31,32,76–79} QuikChange based mutagenesis methods were further advanced to saturate 4 (PFunkel⁸⁰) to 6 (PFLF-MSDM⁸¹) positions simultaneously. In PFLF-MSDM,⁸¹ the multiple site saturation libraries are obtained in a four step procedure in which the first step of QuikChange is replaced by a two-step PCR procedure in which mutations are introduced first through megaprimer formation (step 1) and subsequent whole plasmid amplification (step 2). In total, 40% of the picked clones (30 clones) had all 6 positions mutated.⁸¹ In the PFunkel⁸⁰ method, multiple site saturation libraries are obtained by using 5'-phosphorylated mutagenic primers and an uracil-containing ssDNA template (similar to Kunkel mutagenesis⁸²) and a ligase is employed in the multi-step procedure. In the case of the PFunkel focused mutagenesis method, $\sim 70\%$ of the picked clones (279 clones) were reported to have 4 distant amino acid positions (A42, E104, M182, and G238) mutated.⁸⁰ In the case of the QuikChange Multi Site-Directed Mutagenesis Kit, $\sim 55\%$ of 40 randomly picked clones had 4 mutated positions.⁸³ A conceptually different method is OmniChange⁵⁹ which employs chemically cleavable phosphorothioate primers and saturates up to five positions simultaneously independent of their distance to each other. OmniChange is a simple four step

Table 2 Comparison and performance of mainly used focused mutagenesis methods and their application in recombining amino acid beneficial substitutions

Method	Number of different variants in one round	Number of variants in three iterative rounds	Amino acid substitutions or targeted positions (% of targeted positions)	Ref.
QuikChange based site-directed mutagenesis	1	3	One specific amino acid exchange	84
QuikChange based site saturation mutagenesis	20 (NNK codon) ^a	60	One position per experiment	84
QuikChange multi site-directed mutagenesis kit	16 000 (4 amino acid positions; NNK codon)	48 000	Up to three positions per experiment (55%; 3 positions)	83
QuikChange based: PFLF-MSDM	64 million (6 amino acid positions; NNK codon)	192 million	Up to six positions per experiment (40%; 6 positions)	81
QuikChange based: PFunkel	160 000 (4 amino acid positions; NNK codon)	480 000	Efficient up to four positions per experiment (70%; 4 positions)	80
OmniChange multi-site saturation method	3.2 million (5 amino acid positions; NNK codon)	9.6 million ^b	Up to five positions simultaneously in one experiment (92%; 5 positions)	59

^a NNK, NNS, and NDT codons are commonly used for saturation mutagenesis. ^b OmniChange enables simultaneous site-saturation of five codons. Usually multiple rounds of OmniChange are not necessary. More detailed comparison can be found in Table 5 of ref. 85

method which does not require any gel extraction or final PCR amplification steps. Analysis of 48 sequenced clones showed that 92% had all 5 positions mutated.

In general, multi-site saturation mutagenesis methods offer the possibility to dramatically change binding pockets by generating millions of variants to alter especially activity and selectivity of enzymes.

4. Comparison of the protein evolution strategies

Traditional directed evolution campaigns in which enzymes are improved in iterative rounds of diversity generation and screening were very successful and yielded thousands of success stories. Nevertheless, protein engineering by directed evolution has not

become a standard tool in process design. This can be attributed to time requirements of directed evolution campaigns, which often require more than one year of development time and improved enzymes often failed under application conditions. The failure of the improved enzymes in application often occurs when the selected conditions were not sufficiently close to application conditions or a loss of thermal resistance resulted in reduced process or storage stability. Based on the fact that the theoretical protein sequence space cannot be sampled, several knowledge gaining or driven strategies have been developed to reengineer enzyme more efficiently. Table 3 summarizes and benchmarks prominent or recently reported protein engineering strategies (CASTing, FRESO, ProSAR, MORPHING, and KnowVolution) to traditional directed enzyme evolution.

Traditional directed enzyme evolution experiments employ usually epPCR mutant libraries with a low mutation frequency

Table 3 Knowledge gaining protein engineering strategies

	Requirements ^a	Applicability/ number of reports ^b	The method for identifying potentially beneficial positions	The method for identifying substitutions	The method for combining of beneficial substitutions	Number of screened clones
Traditional directed evolution	No	Broad/>1000	Random mutagenesis	—	Identified positions as a parent for the next round	1000–15 000
CASTing ³¹	Structural model	'Localized' properties in the binding pocket/ ~30 ^c	Focused mutagenesis	SSM	SDM	500–5000
FRESO ³⁹	Structural model and MD simulations	Thermostability/2	<i>In silico</i> screening by $\Delta\Delta G$ calculation and MD simulation	SDM	SDM	109/1634
ProSAR ⁵⁶	Sequence-activity dataset	Broad/6	Random and focused mutagenesis	ProSAR analysis	Semisynthetic shuffling ⁸⁶	60 000
KnowVolution	Structural model	Broad/new	Random mutagenesis	SSM	SDM or OmniChange	2000–12 700
MORPHING ⁵⁷	Structural model	Broad/new	Random mutagenesis	SSM	SDM	500

^a Besides functional expression and screening. ^b The number of reports in ISI Web of Science core collection with the search terms "CASTing & Directed evolution", "FRESO", and "ProSAR" in February 2015. ^c Review papers are not included.

and usually <15 000 clones (often 1000–2000 clones) are screened to find beneficial variants. It is characterized to improve enzyme properties in iterative rounds of diversity generation and screening. Thereby often ~50% of the obtained amino acid substitutions are not beneficial for the targeted properties and can negatively affect properties such as thermal resistance. The accumulation of beneficial and non-beneficial amino acid substitutions in iterative rounds (often more than 10 amino acid substitution) makes it challenging to pinpoint improved properties to individual amino acids. Additionally, there is often no systematic recombination step for the identified substitutions and cooperative effects will likely be missed due to the low mutation frequency. In essence, traditional directed evolution does often not enable a molecular understanding of the improved properties through beneficial amino acid substitutions. Furthermore, the most beneficial substitutions at beneficial positions are often not identified and the accumulation of beneficial and non-beneficial substitutions often affects properties which are not under selective pressure (e.g. thermal resistance).

In the last 15 years, researchers have learned a key lesson that recombination of beneficial positions before the next round of directed evolution yields significantly improved variants with minimized screening efforts.^{17,66,87} The obtained improvements from recombination were often higher than those obtained in a whole round of directed evolution. Strategies to recombine or to generate more beneficial variants (see Table 2), which allow to simultaneously mutagenize up to six amino acid positions, offer nowadays unique possibilities to reshape for example whole substrate binding sites and thereby to go far beyond the diversity which can be generated by random mutagenesis libraries at these positions. Directed evolution 2.0 strategies enable a more rapid enzyme evolution through key beneficial positions which assist researchers to generate a deep molecular understanding of enzyme properties (see prominent examples in Table 3). A key question in directed evolution is which clone(s) has/have to be selected for the next round of directed evolution. In traditional directed evolution campaigns, the most beneficial clone(s) was/were used for the next round of evolution and many iterative rounds (often more than five) were performed to generate significantly improved enzyme variants. Often, amino acid substitutions were located on surfaces of enzymes which prevented a molecular understanding by computational analysis. The proposed KnowVolution strategy overcomes the above mentioned drawbacks and generates improved enzyme variants and a molecular understanding in shorter time. The site saturation mutagenesis step (phase II; Fig. 1) at the beneficial positions identifies the most beneficial substitutions and sequencing yields the type(s) and chemical 'characteristics' of amino acid substitutions that can contribute to the improved properties. With this information, computational analysis by structural inspection of the location/molecular interaction of the identified positions will guide the selection and diversity generation of targeted positions (phase III) for subsequent recombination. Recombination (phase IV) through multi-site saturation mutagenesis or site directed mutagenesis allows to explore efficiently the beneficial sequence space and to yield further improved variants (see Table 1) with minimized screening efforts. Structurally guided diversity generation has the

general advantages of minimizing screening efforts, e.g. <2000 clones were screened in a FRESCO example to yield a dramatic improvement (α >250-fold longer half-life of LEH). A key challenge in computationally assisted protein engineering campaigns is the knowledge and/or experimental data required for efficient analysis which hampers their general applicability. For example, FRESCO was designed for improving enzyme thermal stability and computational analysis including calculations of free energies and MD simulations are required (Table 3). ProSAR (as a combined strategy) however requires a high screening capacity for generating a sequence-activity dataset, for example, 60 000 clones were screened in the reported case (see ref. 88 for further details).

5. Conclusions

Currently we are witnessing in protein engineering a paradigm shift from an engineer's approach to improve enzymes in iterative rounds of diversity generation and screening a knowledge-based approach in order to gain a molecular understanding of enzyme properties. The latter is driven by time requirements of directed evolution campaigns, which are today often too long to implement directed evolution as a standard operation in process development. On the long term, the knowledge-based approach will offer to significantly shorten enzyme development time and it is likely that technological advances such as cell-free expression systems in combination with compartmentalized HTS systems (e.g. microfluidic or flow cytometry) will make it possible to achieve a whole round of evolution in one to two days. Combination of knowledge-based developments and advances in directed evolution technologies will likely enable to reduce the development time of enzyme reengineering to one to two months in the upcoming decade so that enzyme engineering can become a central and routine part of process developments.

Consequently, protein engineers have accepted that it is technically impossible to explore the theoretical protein sequence space and several protein engineering strategies emerged in which beneficial positions are identified through computer-assisted methods and/or through random mutagenesis experiments with subsequent analysis of regions in which beneficial substitutions cluster. In order to efficiently explore the selected or identified positions, focused mutagenesis methods advanced so that five to six amino acid positions can simultaneously be saturated and thereby up to 3.2 million (five amino acid positions) and 64 million (six amino acid positions) enzyme variants are generated in half a day. These focused mutagenesis libraries will likely have a significant impact on generating superior enzymes in short time frames.

Acknowledgements

This work was supported through funds from RWTH Aachen University, DWI-Leibniz Institute for Interactive Materials, the German Federal Ministry of Education and Research (BMBF) (FKZ 031A227F), the Deutsche Forschungsgemeinschaft through

the Cluster of Excellence “Tailor-Made Fuels from Biomass” (TMFB), and Chinese Academy of Sciences Visiting Professorships for Senior International Scientists (Y3J8041101).

Notes and references

- 1 T. S. Wong, D. Roccatano and U. Schwaneberg, *Environ. Microbiol.*, 2007, **9**, 2645–2659.
- 2 A. J. Ruff, A. Dennig and U. Schwaneberg, *FEBS J.*, 2013, **280**, 2961–2978.
- 3 M. Goldsmith and D. S. Tawfik, *Methods Enzymol.*, 2013, **523**, 257–283.
- 4 P. A. Romero and F. H. Arnold, *Nat. Rev. Mol. Cell Biol.*, 2009, **10**, 866–876.
- 5 M. T. Reetz, D. Kahakeaw and R. Lohmer, *ChemBioChem*, 2008, **9**, 1797–1804.
- 6 S. Lutz, *Curr. Opin. Biotechnol.*, 2010, **21**, 734–743.
- 7 D. R. Mills, R. L. Peterson and S. Spiegelman, *Proc. Natl. Acad. Sci. U. S. A.*, 1967, **58**, 217–224.
- 8 G. Guven, R. Prodanovic and U. Schwaneberg, *Electroanalysis*, 2010, **22**, 765–775.
- 9 N. L. Wagner, J. A. Greco, M. J. Ranaghan and R. R. Birge, *J. R. Soc., Interface*, 2013, **10**, 20130197.
- 10 F. Cheng, T. Kardashliev, C. Pitzler, A. Shehzad, H. Lue, J. Bernhagen, L. Zhu and U. Schwaneberg, *ACS Synth. Biol.*, 2015, DOI: 10.1021/sb500343g.
- 11 F. Cheng, L. Zhu, H. Lue, J. Bernhagen and U. Schwaneberg, *Appl. Microbiol. Biotechnol.*, 2015, **99**, 1237–1247.
- 12 P. A. Romero, E. Stone, C. Lamb, L. Chantranupong, A. Krause, A. E. Miklos, R. A. Hughes, B. Fechtel, A. D. Ellington, F. H. Arnold and G. Georgiou, *ACS Synth. Biol.*, 2012, **1**, 221–228.
- 13 F. H. Arnold, *Nat. Biotechnol.*, 1998, **16**, 617–618.
- 14 K. Q. Chen and F. H. Arnold, *Proc. Natl. Acad. Sci. U. S. A.*, 1993, **90**, 5618–5622.
- 15 M. D. Lane and B. Seelig, *Curr. Opin. Chem. Biol.*, 2014, **22**, 129–136.
- 16 M. Schallmeyer, J. Frunzke, L. Eggeling and J. Marienhagen, *Curr. Opin. Biotechnol.*, 2014, **26**, 148–154.
- 17 R. Martinez and U. Schwaneberg, *Biol. Res.*, 2013, **46**, 395–405.
- 18 J. Karanicolas, J. E. Com, I. Chen, L. A. Joachimiak, O. Dym, S. H. Peck, S. Albeck, T. Unger, W. X. Hu, G. H. Liu, S. Delbecq, G. T. Montelione, C. P. Spiegel, D. R. Liu and D. Baker, *Mol. Cell*, 2011, **42**, 250–260.
- 19 R. Blomberg, H. Kries, D. M. Pinkas, P. R. E. Mittl, M. G. Grutter, H. K. Privett, S. L. Mayo and D. Hilvert, *Nature*, 2013, **503**, 418–421.
- 20 K. L. Tee and U. Schwaneberg, *Comb. Chem. High Throughput Screening*, 2007, **10**, 197–217.
- 21 R. Fasan, Y. T. Meharena, C. D. Snow, T. L. Poulos and F. H. Arnold, *J. Mol. Biol.*, 2008, **383**, 1069–1080.
- 22 J. Zhao, T. Kardashliev, A. J. Ruff, M. Bocola and U. Schwaneberg, *Biotechnol. Bioeng.*, 2014, **111**, 2380–2389.
- 23 V. J. Frauenkron-Machedjou, A. Fulton, L. Zhu, C. Anker, M. Bocola, K.-E. Jaeger and U. Schwaneberg, *ChemBioChem*, 2015, **16**, 937–945.
- 24 Y. Kipnis, E. Dellus-Gur and D. S. Tawfik, *Protein Eng., Des. Sel.*, 2012, **25**, 437–444.
- 25 H. Mundhada, J. Marienhagen, A. Scacioc, A. Schenk, D. Roccatano and U. Schwaneberg, *ChemBioChem*, 2011, **12**, 1595–1601.
- 26 R. Verma, U. Schwaneberg and D. Roccatano, *ACS Synth. Biol.*, 2012, **1**, 139–150.
- 27 A. E. Firth and W. M. Patrick, *Nucleic Acids Res.*, 2008, **36**, W281–W285.
- 28 Y. Nov, *Appl. Environ. Microbiol.*, 2012, **78**, 258–262.
- 29 R. Verma, U. Schwaneberg and D. Roccatano, *Comput. Struct. Biotechnol. J.*, 2012, **2**, e201209008.
- 30 P. E. O'Maille, M. Bakhtina and M. D. Tsai, *J. Mol. Biol.*, 2002, **321**, 677–691.
- 31 M. T. Reetz, M. Bocola, J. D. Carballeira, D. Zha and A. Vogel, *Angew. Chem., Int. Ed.*, 2005, **44**, 4192–4196.
- 32 M. T. Reetz, L. W. Wang and M. Bocola, *Angew. Chem., Int. Ed.*, 2006, **45**, 1236–1241.
- 33 J. Marienhagen, A. Dennig and U. Schwaneberg, *BioTechniques*, 2012, **1**–6.
- 34 W. P. C. Stemmer, *Proc. Natl. Acad. Sci. U. S. A.*, 1994, **91**, 10747–10751.
- 35 H. M. Zhao, L. Giver, Z. X. Shao, J. A. Affholter and F. H. Arnold, *Nat. Biotechnol.*, 1998, **16**, 258–261.
- 36 C. Neylon, *Nucleic Acids Res.*, 2004, **32**, 1448–1459.
- 37 M. Ostermeier, A. E. Nixon, J. H. Shim and S. J. Benkovic, *Proc. Natl. Acad. Sci. U. S. A.*, 1999, **96**, 3562–3567.
- 38 S. Lutz, M. Ostermeier, G. L. Moore, C. D. Maranas and S. J. Benkovic, *Proc. Natl. Acad. Sci. U. S. A.*, 2001, **98**, 11248–11253.
- 39 H. J. Wijma, R. J. Floor, P. A. Jekel, D. Baker, S. J. Marrink and D. B. Janssen, *Protein Eng., Des. Sel.*, 2014, **27**, 49–58.
- 40 R. J. Floor, H. J. Wijma, D. I. Colpa, A. Ramos-Silva, P. A. Jekel, W. Szymanski, B. L. Feringa, S. J. Marrink and D. B. Janssen, *ChemBioChem*, 2014, **15**, 1659–1671.
- 41 Y. J. Lin, S. Yoo and R. Sanchez, *Bioinformatics*, 2012, **28**, 1172–1173.
- 42 M. N. Wass, L. A. Kelley and M. J. E. Sternberg, *Nucleic Acids Res.*, 2010, **38**, W469–W473.
- 43 J. Konc and D. Janezic, *Nucleic Acids Res.*, 2014, **42**, W215–W220.
- 44 P. A. Dalby, *Curr. Opin. Struct. Biol.*, 2011, **21**, 473–480.
- 45 H. Ashkenazy, E. Erez, E. Martz, T. Pupko and N. Ben-Tal, *Nucleic Acids Res.*, 2010, **38**, W529–W533.
- 46 U. T. Bornscheuer, *Adv. Biochem. Eng./Biotechnol.*, 2013, **137**, 25–40.
- 47 F. Chen, E. A. Gaucher, N. A. Leal, D. Hutter, S. A. Havemann, S. Govindarajan, E. A. Ortlund and S. A. Benner, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**, 1948–1953.
- 48 R. S. Kuipers, H. J. Joosten, W. J. H. van Berkel, N. G. H. Leferink, E. Rooijen, E. Ittmann, F. van Zimmeren, H. Jochens, U. Bornscheuer, G. Vriend, V. A. P. M. dos Santos and P. J. Schaap, *Proteins*, 2010, **78**, 2101–2113.
- 49 A. Pavelka, E. Chovancova and J. Damborsky, *Nucleic Acids Res.*, 2009, **37**, W376–W383.
- 50 F. Jakob, R. Martinez, J. Mandawe, H. Hellmuth, P. Siegert, K. H. Maurer and U. Schwaneberg, *Appl. Microbiol. Biotechnol.*, 2013, **97**, 6793–6802.
- 51 R. K. P. Kuipers, H. J. Joosten, E. Verwiel, S. Paans, J. Akerboom, J. van der Oost, N. G. H. Leferink, W. J. H. van Berkel, G. Vriend and P. J. Schaap, *Proteins*, 2009, **76**, 608–616.
- 52 H. Jochens and U. T. Bornscheuer, *ChemBioChem*, 2010, **11**, 1861–1866.
- 53 S. W. Linwu, C. A. Wu, F. C. Peng and A. H. J. Wang, *Biochem. Pharmacol.*, 2012, **83**, 1690–1699.
- 54 E. Cukuroglu, H. B. Engin, A. Gursoy and O. Keskin, *Prog. Biophys. Mol. Biol.*, 2014, **116**, 165–173.
- 55 R. Bischoff and H. Schluter, *J. Proteomics*, 2012, **75**, 2275–2296.
- 56 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, J. Grate, J. Gruber, J. C. Whitman, R. A. Sheldon and G. W. Huisman, *Nat. Biotechnol.*, 2007, **25**, 338–344.
- 57 D. Gonzalez-Perez, P. Molina-Espeja, E. Garcia-Ruiz and M. Alcalde, *PLoS One*, 2014, **9**, e90919.
- 58 J. Zhao, T. Kardashliev, A. J. Ruff, M. Bocola and U. Schwaneberg, *Biotechnol. Bioeng.*, 2014, **111**, 2380–2389.
- 59 A. Dennig, A. V. Shivange, J. Marienhagen and U. Schwaneberg, *PLoS One*, 2011, **6**, e26222.
- 60 G. Freckmann, C. Schmid, A. Baumstark, S. Pleus, M. Link and C. Haug, *J. Diabetes Sci. Technol.*, 2012, **6**, 1060–1075.
- 61 E. A. Gutierrez, H. Mundhada, T. Meier, H. Duefel, M. Bocola and U. Schwaneberg, *Biosens. Bioelectron.*, 2013, **50**, 84–90.
- 62 A. V. Shivange, A. Serwe, A. Dennig, D. Roccatano, S. Haefner and U. Schwaneberg, *Appl. Microbiol. Biotechnol.*, 2012, **95**, 405–418.
- 63 A. V. Shivange, A. Dennig and U. Schwaneberg, *J. Biotechnol.*, 2014, **170**, 68–72.
- 64 R. Martinez, F. Jakob, R. Tu, P. Siegert, K. H. Maurer and U. Schwaneberg, *Biotechnol. Bioeng.*, 2013, **110**, 711–720.
- 65 L. Zhu, K. L. Tee, D. Roccatano, B. Sonmez, Y. Ni, Z. H. Sun and U. Schwaneberg, *ChemBioChem*, 2010, **11**, 691–697.
- 66 L. Zhu, R. Verma, D. Roccatano, Y. Ni, Z. H. Sun and U. Schwaneberg, *ChemBioChem*, 2010, **11**, 2294–2301.
- 67 Z. Li, D. Roccatano, M. Lorenz and U. Schwaneberg, *ChemBioChem*, 2012, **13**, 691–699.
- 68 Z. Li, D. Roccatano, M. Lorenz, R. Martinez and U. Schwaneberg, *J. Biotechnol.*, 2014, **169**, 87–94.
- 69 C. Lehmann, F. Sibilla, Z. Maugeri, W. R. Streit, P. D. de Maria, R. Martinez and U. Schwaneberg, *Green Chem.*, 2012, **14**, 2719–2726.
- 70 S. Sebastian, S. P. Touchburn and E. R. Chavez, *World's Poult. Sci. J.*, 1998, **54**, 27–47.
- 71 Y. Ni, U. Schwaneberg and Z. H. Sun, *Cancer Lett.*, 2008, **261**, 1–11.

- 72 A. Galkin, X. F. Lu, D. Dunaway-Mariano and O. Herzberg, *J. Biol. Chem.*, 2005, **280**, 34080–34087.
- 73 L. Zhu, F. Cheng, V. Piatkowski and U. Schwaneberg, *ChemBioChem*, 2014, **15**, 276–283.
- 74 R. Gupta, Q. K. Beg and P. Lorenz, *Appl. Microbiol. Biotechnol.*, 2002, **59**, 15–32.
- 75 C. Lehmann, M. Bocola, W. R. Streit, R. Martinez and U. Schwaneberg, *Appl. Microbiol. Biotechnol.*, 2014, **98**, 5775–5785.
- 76 J. D. Carballeira, P. Krumlinde, M. Bocola, A. Vogel, M. T. Reetz and J. E. Backvall, *Chem. Commun.*, 2007, 1913–1915.
- 77 M. T. Reetz, J. D. Carballeira, J. Peyralans, H. Hobenreich, A. Maichele and A. Vogel, *Chemistry*, 2006, **12**, 6031–6038.
- 78 S. Kille, F. E. Zilly, J. P. Acevedo and M. T. Reetz, *Nat. Chem.*, 2011, **3**, 738–743.
- 79 S. Hoebenreich, F. E. Zilly, C. G. Acevedo-Rocha, M. Zilly and M. T. Reetz, *ACS Synth. Biol.*, 2015, **4**, 317–331.
- 80 E. Firnberg and M. Ostermeier, *PLoS One*, 2012, **7**, e52031.
- 81 W. C. Tseng, J. W. Lin, X. G. Hung and T. Y. Fang, *Anal. Biochem.*, 2010, **401**, 315–317.
- 82 T. A. Kunkel, *Proc. Natl. Acad. Sci. U. S. A.*, 1985, **82**, 488–492.
- 83 H. H. Hogrefe, J. Cline, G. L. Youngblood and R. M. Allen, *Biotechniques*, 2002, **33**, 1158–1165.
- 84 C. Papworth, J. C. Bauer, J. Braman and D. A. Wright, *Strategies*, 1996, **9**, 3–4.
- 85 K. L. Tee and T. S. Wong, *Biotechnol. Adv.*, 2013, **31**, 1707–1721.
- 86 K. Stutzman-Engwall, S. Conlon, R. Fedechko, H. McArthur, K. Pekrun, Y. Chen, S. Jenne, C. La, N. Trinh, S. Kim, Y. X. Zhang, R. Fox, C. Gustafsson and A. Krebber, *Metab. Eng.*, 2005, **7**, 27–37.
- 87 R. A. Chica, N. Doucet and J. N. Pelletier, *Curr. Opin. Biotechnol.*, 2005, **16**, 378–384.
- 88 N. J. Turner, *Nat. Chem. Biol.*, 2009, **5**, 568–574.