Review

Pharmaceutical BIOPROCESSING

Pharm. Bioprocess.
(2013) 1(2), 179–196

Recent advances in biocatalyst development in the pharmaceutical industry

Biocatalysts are increasingly employed as more efficient and environmentally safer alternatives to traditional chemical catalysts in the manufacturing of fine chemicals. This is driven by advances in recombinant DNA technology, protein engineering and bioinformatics, all of which are critical in the discovery, tailoring and optimization of enzymes for industrial processes. In this article we review these key technological innovations, as well as highlight the strategic application of these tools in the development of biocatalysts in the production of advanced pharmaceutical intermediates.

Mingzi M Zhang[†1], Xiaoyun Su[†1], Ee Lui Ang[1] & Huimin Zhao[*1,2]

[1]Metabolic Engineering Research Laboratory, Institute of Chemical & Engineering Sciences, Agency for Science, Technology & Research, Singapore

[2]Departments of Chemical & Biomolecular Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

*Author for correspondence:
Tel. : +1 217 333 2631
E-mail: zhao5@illinois. edu

[†]Authors contributed equally

The synthesis of pharmaceutical products and therapeutic agents using biological systems is becoming increasingly important for the healthcare industry. Integral to pharmaceutical bioprocessing is the development of biocatalysts for synthesis of desired compounds. While biocatalysts are capable of accepting a wide range of substrates with high enantioselectivity and regioselectivity [1], they are often not very active or stable under the process conditions frequently required for commercial-scale production, making many biocatalytic processes unscalable. However, technical advances have enabled biocatalysis to become economically feasible and a greener alternative to traditional chemical synthesis routes, especially in the pharmaceutical sector [2].

One of the key technological breakthroughs that aided the development of many industrial biocatalytic processes is the introduction of directed evolution strategies in the mid to late 1990s [3–5]. Since then, directed evolution has found many applications in the healthcare industry including areas such as therapeutic protein development [6,7], and active pharmaceutical ingredient synthesis [2]. In recent years, state-of-the-art gene discovery tools and DNA synthesis technologies have further accelerated biocatalyst discovery and development for pharmaceutical processes. For a comprehensive summary of recent examples of biocatalyst application in the pharmaceutical industry, readers can refer to the article by Bornscheuer *et al.* [2].

This review focuses on the recent technological innovations for biocatalyst discovery and engineering **(Figure 1)**. Selected case studies highlight how these methods were integrated and applied to develop enzyme-based and whole-cell biocatalysts for the production of active pharmaceutical ingredients. Emerging trends that promise to shape the future of biocatalyst development are also presented in this review.

Tools employed in biocatalyst discovery & engineering

» Biocatalyst discovery tools

DNA sequencing & genomics/ metagenomics

The biocatalysts used for pharmaceutical purposes can be individual or a combination of enzymes. Alternatively, whole cells can also be used as a biocatalyst, in which several enzymes coordinate to bring about chemical transformations. In any case, enzymes are the core of biocatalysis and there is a huge demand in the pharmaceutical industry for novel enzymes with more desirable

10.4155/PBP.13.14 ISSN 2048-9145

Key Terms

Enantioselectivity: Preference of a chemical reaction for one enantiomer over the other. Enantiomers are two molecules with non-superimposable chemical structures that are mirror images of each other.

Regioselectivity: Preference of a chemical reaction (bond breaking or formation) for a single isomer over the other products in the reaction.

Metagenome: Genetic material directly extracted from environmental samples. Metagenomics is the study of metagenomes. This approach is useful for identifying genes from unculturable microorganisms, which represent the vast majority (>99%) of microbial diversity.

characteristics in terms of substrate specificity, and regio- and enantioselectivities. The exponentially increasing genome and metagenome sequencing datasets allow access to vast biodiversity, providing an invaluable resource for such novel biocatalysts.

For two decades, the Sanger method has been a workhorse technology for DNA sequencing before the appearance of next-generation sequencing (NGS), which is widely employed in recent genomic, metagenomic and transcriptomics endeavors [8]. Three of the most popular methods include the 454™ Genome Sequencer FLX instrument from Roche Applied Science (Germany), the Solexa™ from Illumina (CA, USA), and SOLiD™ from Applied Biosystems (CA, USA). The 454 DNA sequencing method and the Illumina Solexa method are sequencing-by-synthesis methods. The 454 method is based on emulsion polymerase chain reaction (PCR) [9] to amplify the DNA followed by pyrosequencing technique [10] and can acquire over 1 million reads with read lengths of up to 1000 base pairs (bp) in 23 h [201]. For the Illumina Solexa method, the HiSeq2500/1500 system can acquire a maximum of 600 gigabase (Gb) of sequences with 6 billion reads with a read length of 2 × 100 bp in paired-end reads [202]. The SOLiD system involves sequencing by ligation instead of synthesis and can obtain sequence data above 20 Gb per day with a 50 bp average read length [203]. While these methods use clonally amplified templates, other methods, such as those developed by Helicos BioSciences (MA, USA) [11] and Pacific Biosciences (CA, USA) [12], utilize single-molecule templates, which are more demanding in image sensing. These technologies are called third-generation DNA sequencing. The Nanopore DNA sequencing technology also belongs to the third-generation DNA sequencing, but it does not require fluorescent labeling and the expensive charge-coupled device cameras [13]. The Helicos BioSciences HeliScope Single Molecule Sequencer can obtain over 1 Gb of usable nucleotide sequence per day [204].

Genome mining based on the rapidly evolving DNA sequencing technologies mentioned above uncovers novel enzymes with alternative characteristics, such as different substrate specificity, regio- and enantioselectivities. For example, cytochrome P450 enzymes (P450s) are a large and highly diverse superfamily of heme-containing enzymes that can transfer an oxygen atom from molecular oxygen to an organic molecule. P450s have two prominent characteristics that make them attractive candidates for pharmaceutical biocatalysts. First, they catalyze insertion of oxygen in less reactive carbon–hydrogen bonds under mild conditions. Second, they act on diverse substrates, including fatty acids, terpenes, steroids, prostaglandins, polyaromatic and heteroaromatic compounds [14]. Since many drugs or their intermediates are also P450 substrates, there have been efforts in mining sequenced genomes for novel P450s. For example, 18 putative P450 genes were identified from the genome of *Streptomyces coelicolor* A3 [15], 21 from *Sorangium cellulosum*, 18 from *Stigmatella aurantiaca*, 17 from *Haliangium ochraceum*, 7 from *Myxococcus xanthus* [16] and a total of 12,456 putative P450 genes were predicted from all kingdoms of life [17]. Prediction of the functions of P450s can be difficult because of their diversity. The genome mining method also presents a way of prediction by researching into the genome context of the P450s [16].

The plant and microbial genomes also encode a vast number of non-ribosomal peptide synthetases (NRPSs) and polyketide synthases (PKSs), two important families of enzymes involved in producing pharmaceutically important non-ribosomal peptide natural products and polyketides. Many of these natural products, such as erythromycin and vancomycin, are useful in antibiotic, anticancer and immunosuppressant treatments. Similar to P450s, elucidation of the exact roles of putative NRPSs or PKSs is not easy and is often time-consuming. Lautru *et al.* described the prediction of substrates recognized by the adenylation domains that are commonly present in NRPS. The prediction guided the identification of a *tris*-hydroxamatetetrapeptide iron chelatorcoelichelin produced

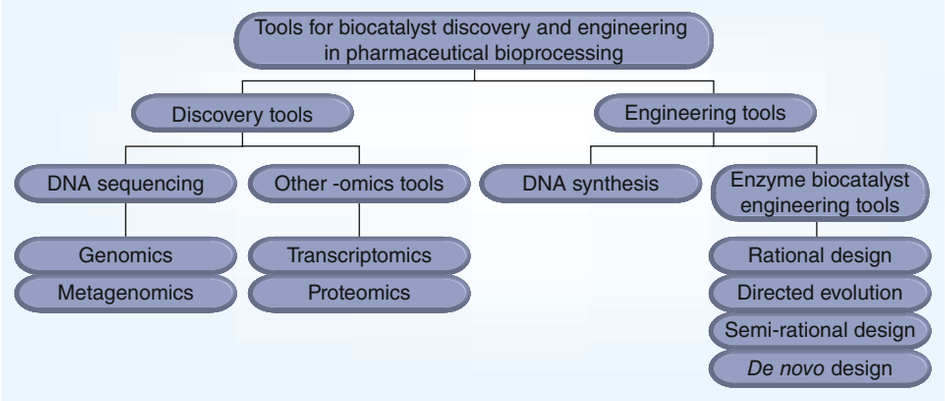

Figure 1. Tools for biocatalyst discovery and engineering in pharmaceutical bioprocessing.

by the *Streptomyces coelicolor* CchH [18]. Functional assignment to a putative NRPS or PKS is also hindered by silencing of the corresponding genes under standard fermentation conditions. In *Aspergillus nidulans*, no PKS–NPRS metabolites could be detected in the extracts from 40 different culture conditions. Bergmann *et al.* noted that, in this microorganism, one PKS–NPRS hybrid gene *apdA* is clustered with a putative transcription activator gene *apdR* [19]. Induced overexpression of ApdR drives expression of the silent ApdA metabolic pathway components, which allows them to identify new metabolites – aspyridones A and B. High-throughput selection, such as a phage-display selection method [20], takes advantage of post-translational modification of peptidyl carrier protein or acyl carrier proteins within NRPS or PKS, or screening such as a mass spectrometry-guided method, facilitated genome mining of NRPS and PKS enzymes with unprecedented high efficiency [21].

Transaminases are a family of enzymes that reversibly transfer amino and keto groups between two substrates. Transaminases are of interest to the pharmaceutical industry because they can be used to generate optically pure amines by either asymmetric synthesis [22] or kinetic resolution of racemic amines [22], and to synthesize non-proteinogenic amino acids [23], which can be used for synthesis of peptidomimetic pharmaceuticals. For example, transaminases have been coupled with pyruvate decarboxylase for asymmetric synthesis of chiral amines [24] or with amino acid oxidase for resolution of racemic amines [22]. Hohne *et al.* developed an algorithm that recognizes key amino acid residues indicating enantiopreference. An *in silico* screen employing this algorithm of a protein library composed of 5700 L-branched chain amino acid aminotransferases and 280 pyridoxal-5′-phosphate-dependent fold class IV proteins from NCBI protein database identified 21 proteins with putative (*R*)-selective amine transaminase activity [25]. The prokaryotic expression, purification and characterization of these enzymes indicated that 17 of them are true (*R*)-selective amine transaminases. Seven of these enzymes show potential in asymmetric synthesis of aliphatic, aromatic and arylaliphatic (*R*)-amines starting from the corresponding ketones [26].

Nitrilases catalyze conversion of organonitriles directly to the corresponding carboxylic acids, which are important intermediates in the production of pharmaceuticals and fine chemicals. Screening of metagenomic libraries from varying global environmental habitats yielded over 200 nitrilases that selectively hydrolyze organonitriles to carboxylic acid derivatives, including four enzymes with the ability to catalyze the desymmetrization of prochiral 3-hydroxyglutaryl nitrile to form (*R*)-4-cyano-3-hydroxybutyric acid, a key intermediate of atorvastatin [27]. Similarly, 137 nitrilases were identified from screening over 600 biotope-specific environmental DNA libraries [28]. By mining of published genomes, various nitrilases targeting different substrates have been identified from bacteria [29–31] and filamentous fungi [32,33].

Other -omics tools

Transcriptomics

The transcriptome of an organism is more dynamic than its genome and differential expression patterns of genes often provide insight into their function and associated regulatory networks. Analysis of the transcriptome of an organism can help identify putative biocatalysts that are components of a certain metabolic pathway and guide the metabolic engineering of a whole-cell biocatalyst. Traditionally, global analysis of a transcriptome employed the DNA microarrays [34], however, the emergence of NGS-based RNA-seq revolutionized transcriptomics, providing deep coverage and base-level resolution. In addition, RNA-seq does not need the genomic information and can be used to quantitatively analyze the transcript levels. In theory, any high-throughput DNA sequencing method can be used in RNA-seq. The Illumina system has been used to sequence the transcriptomes of *Saccharomyces cerevisiae* [35], *Schizosaccharomyces pombe* [36], *Arabidopsis thaliana* [37] and mouse [38], while the 454 system has been applied in maize [39], butterfly [40] and coral larval [41], and the SOLiD system has been applied to mouse [42].

Transcriptomics can be a powerful tool in identifying novel enzymes with potential as industrial biocatalysts. Geu-Flores *et al.* applied a co-regulation criterion to search for an iridoid synthase from the transcriptomic data of *Catharanthus roseus* (Apocynaceae), based on the assumption that expression profile of the unidentified enzyme would be similar to geraniol 10-hydroxylase (G10H), a known enzyme upstream in the iridiod biosynthesis pathway. This allowed them to narrow down candidate genes to 20, among which they found two NADPH-using enzymes. One of these two enzymes shares high similarity to progesterone-5β-reductase (P5βR) and was proven to be a new iridoid synthase [43].

Proteomics

Currently, both gel-based techniques, 2D sodium-dodecyl-sulfate polyacrylamide gel electrophoresis and high-throughput shotgun proteomics, by combining high-resolution liquid chromatography and mass spectrometry, are widely used [44]. Proteomics can be used in the identification of the amino acid sequences and even the *de novo* sequencing of target pharmaceutical biocatalysts, such as polyhydroxyalkanoate depolymer-

Key Term

Screen and selection: Compared with selections that can be easily used to assay libraries of approximately 10^8–10^{10} variants, screens tend to be limited to library sizes of approximately 10^4–10^6 since all variants have to be characterized. In both cases, assay design is the key and usually determines the choice between the strategies.

ases [45]. Proteomic analysis is also useful in engineering the whole-cell biocatalysts because the protein levels are also dynamic in addition to transcript levels within a cell.

Furthermore, functional proteomics is used for the discovery of carbohydrate-related enzymatic activities, which has been extensively reviewed [46]. Activity-based proteomics is another specialized method developed for global analysis of protein function in native biological systems [47]. Although it is widely used for comparative enzyme profiling, it can also be used for novel biocatalyst discovery. For example, 9-*O*-acetyl-esterase was identified as a candidate serine hydrolase by using an activity-based fluorophopsphonate probe [48]. This method relies on a specially designed probe that recognizes and covalently binds to the active site of specific enzymes with its reactive group and can be detected by radioactivity, fluorescence or color development with the reporter tag. Activity-based proteomics has been applied to study a variety of enzymes such as serine proteases, cysteine proteases, threonine proteases, lipases, glycosidases, tyrosine phosphatases and kinases [47,49].

The advent of gene discovery tools from the genomic to the proteomic level is critical for generating starting points for biocatalysts development. Novel enzymes identified through these methods may have new sequences not claimed in patents, which is useful when developing a bioprocess in an area with an established intellectual property landscape [50]. The diversity can also be combined into a screening kit for rapid identification of starting enzymes [51]. To effectively tap into this diversity, the biggest challenge is to synthesize and express the genes in heterologous systems. Following functional expression, promising enzymes will most likely need to be engineered to fit process requirements. The following section will focus on recent developments in DNA synthesis and protein engineering, which has provided much promise in bringing these enzymes into industrial applications.

» Biocatalyst engineering tools

DNA synthesis

Engineering of individual enzymes, metabolic pathways and whole-cell biocatalysts require manipulation of genes. Although PCR is routinely used to clone or evolve gene(s) of interest, the cost–competitiveness and productivity of gene synthesis are continually improving [52], impacting our way of working with genes, genetic pathways and networks. *De novo* synthesis of individual genes has been implemented for the *Bacillus thuringiensis vip3A* gene [53], human protein kinase genes *PKB2*, *S6K1* and *PDK1* [54]. At the same time, chemical synthesis has also been applied to synthetic pathways such as a polyketide synthase gene cluster [55] and complex whole genomes including a poliovirus cDNA [56] and a ØX174 bacteriophage [57]. Groundbreaking progress took place when J Craig Venter's group synthesized the entire 1.08 Mbp *Mycoplasma mycoides* genome and showed it governed the regenerated cell as an 'operating system' [58]. While these syntheses utilize different strategies to assemble intermediate fragments into larger and the ultimate DNA constructs, all of them involve assembly of short oligonucleotides with overlapping termini sequences into intermediate DNA fragments. In these endeavors, the high-throughput synthesis of DNA on a microarray chip, which can be controlled by methods such as inkjet printing and photolithography [59–61], proved to be a promising gene synthesis method. This method can be used in massive parallel synthesis of high-density oligonucleotide arrays. The unpurified pooled mixture of oligonucleotides can be used to synthesize genes of interest by the same assembly methods in other traditional gene synthesis methods. Softwares such as GeMS [62] have been developed to guide and facilitate experimental design. Although traditional chemical synthesis of oligonucleotides does not exceed 100 bp, by optimization of reagent flows, the commonly observed depurination in DNA synthesis is minimized, thereby allowing up to 150 bp oligonucleotides to be synthesized on microchips [63]. Compared with the oligonucleotides synthesized on columns, DNA synthesis on microchips is more prone to errors. By application of the state-of-the-art NGS to control the integrity of DNA synthesized on microchips, Matzas *et al.* reduced the error rate by a factor of 500 compared with one error in 40 bp in the starting oligonucleotides [64]. More recently, direct gene synthesis chips have been established with an error rate of 0.19 error per kb [65]. Since most, if not all, of these methods are automated, engineering the genes encoding the biocatalysts can be easily integrated with DNA synthesis. The pharmaceutical industry can benefit from cost-effective, accurate and high-throughput gene synthesis.

Protein engineering

Most naturally existing enzymes are unsuitable for practical applications in various aspects, such as substrate specificity, stability, activity and enantioselectivity. Protein engineering, involving rational design, directed evolution and semi-rational design tools, is critical for the development and optimization of biocatalysts for industrial processes.

Rational design

Rational design is based on our knowledge of the enzymes, particularly of their 3D structures. Nature has evolved diverse enzymes with different biochemical characteristics but similar 3D structures, which can be learned for rational design of biocatalysts. For example, gain-of-function mutants of isocitrate dehydrogenase (IDH) in cancer cell are able to convert 2-oxoglutarate to (*R*)-2-hydroxyglutarate. Structural superposition and amino acid sequence alignment to *S. pombe* homo IDH (HIDH) analogues of human cytoplasmic IDHs revealed residues homologous to the human mutational hotspots. Based on this, four mutants, R114Q, R143C, R143H and R143K of the *S. cerevisiae* HIDH (ScHIDH) were produced and biochemical characterization indicated that ScHIDH^{R143H}, ScHIDH^{R143C} and ScHIDH^{R143K} are (*R*)-2-hydroxyadipate dehydrogenases with excellent enantioselectivity [66].

Amino acid residues in the substrate binding pockets of enzymes are well-known for their importance in substrate selectivity and are good engineering targets. Transketolases are a family of enzymes that catalyze the transfer of a two-carbon keto group from a ketose sugar to aldose. To improve the acceptance of aromatic aldehydes by transketolases for biosynthesis of novel antibiotics, a series of mutants focusing in the phosphate binding pocket of *Escherichia coli* transketolase were designed, characterized and computationally docked with 3-formylbenzoic acid and 4-formylbenzoic acid. One of the mutants exhibited 250-fold higher specific activity in producing α,α-dihydroxyketone, showcasing how an understanding of a substrate binding pocket facilitates redesigning of enzyme substrate specificity [67].

Enzymes with high amino acid sequence similarity but distinct enzymatic activities are good candidates for rational design. Human aldo-keto 1D1 (AKR1D1) and AKR1C are such highly similar enzymes: AKR1D1 catalyzes the 5β-reduction of Δ^4-3-ketosteroids, AKR1C is a hydroxysteroid dehydrogenase. Changing a single amino acid residue His^{120} in AKR1D1 with glutamate converted this enzyme into a 3β-hydroxysteroid dehydrogenase [68].

Directed evolution

Directed evolution has been developed as a strategy to improve biocatalysts by combining the generation of a mutant protein library with identification of a mutated protein with desirable function [69]. Although initially designed for individual enzymes, the same strategy has been adapted for metabolic pathways [70], which makes engineering whole-cell biocatalysts possible. In principle, the directed evolution strategy involves four steps: choosing a parent protein, creating a mutant library from the parent protein, screening or selecting for proteins with desired properties and repeating steps 1–3 [71].

To generate diversity, either random mutagenesis or DNA recombination can be utilized. Currently, error-prone PCR (epPCR) employing low-fidelity DNA polymerases is still widely used for random mutagenesis of biocatalysts used in pharmaceutical processes. In addition, other modifications of routine PCR, such as replacing Mg^{2+} with varying concentrations of Mn^{2+} and using an uneven mixture of the four deoxynucleotide triphosphates, are adopted to fine-tune the mutation rate of the target gene. epPCR is limited to creating small numbers of mutations through the genes. DNA shuffling, a method developed for *in vitro* DNA recombination, is able to generate multiple mutations within a gene and is, therefore, also employed for diversity generation [72]. The DNA from the gene of interest and its random mutants or from homologous genes are fragmented by DNaseI. The DNA fragments are combined and homologous regions allow annealing to take place between these fragments, which will be assembled into full-length genes. Variations are reported for both random mutagenesis and DNA shuffling, such as sequence saturation mutagenesis [73] and staggered extension process [74], respectively. DNA shuffling methodologies that are able to assemble DNA fragments independent of homology are also available, such as incremental truncation for the creation of hybrid enzymes [75], random multi-recombinant PCR [76] and non-homologous random recombination [77]. Complementary to these *in vitro* evolution techniques, *in vivo* evolution methods, such as Heritable Recombination in yeast [78] and phage-assisted continuous evolution in *E. coli* [79], may also be adapted for engineering of single enzymes, or even whole-cell biocatalysts, for pharmaceutical processes.

The key to successful identification of improved mutants by directed evolution includes the generation of a library of high-quality, as well as efficient screen and selection. A traditional screening method employing 96-well microplates is still used currently, but is limited by its capacity to analyze the mutants (less than a few thousand mutants per round [80]). However, there is already tremendous progress in high-throughput or even ultra high-throughput screening methods, which greatly augment directed evolution approaches. By labeling each of the two enantiomers with a different fluorescent dye, Becker *et al.* established a single-cell high-throughput screening method to identify enantioselective hydrolytic enzymes [81]. A drop-based microfluidics screening method in combination with fluorescence-activated cell sorting gains ultra high-throughput efficiency by dispersing aqueous drops in oil at picoliter-volume [82]. This versatile method has been applied to screening of yeast cells displaying a horseradish peroxidase [82] or single genes of 30,000

Key Term

Saturation mutagenesis: A strategy involving the generation of all (or most) possible mutations at a specific site or narrow region of a gene.

copies expressing a β-galactosidase [83]. Compared with traditional microplate-based screening, selection appears to be more favorable, featuring a higher throughput of 10^5–10^8 clones assessed each round [80]. The selection strategy, however, generally requires coupling the enzymatic reaction of interest with cell viability. Fernandez-Alvaro linked one enantiomer of 3-phenyl butyric acid to glycerol and the other to 2,3-dibromopropanol [80]. Esterase mutants that release glycerol will support the growth of the *E. coli* host, and those that release the toxic 2,3-dibromopropanol will kill the cells. By coupling this *in vivo* selection method with flow cytometry, they obtained ultra high-throughput identification of esterase mutants with altered enantioselectivity.

Semi-rational design

Rational design saves benchwork time because it only requires dealing with small numbers of mutants. However, the mutants may not be completely desirable. On the other hand, directed evolution has the power to create desirable mutants but suffers from time-consuming library screening. Semi-rational design, also known as combinational rational design, integrates rational design with directed evolution to reduce library size while maintaining diversity of functional mutants. Several semi-rational design methods have been successfully employed for engineering enzyme biocatalysts. The Arnold group developed a SCHEMA algorithm that can be used to guide the construction of mutant library with reduced size [84]. Traditionally, the DNA shuffling creates libraries containing a large number of incorrectly folded and non-functional protein mutants. SCHEMA uses structural information to identify interacting amino acid residue pairs and interactions that are broken during recombination [85], thereby increasing the proportion of properly folded proteins in the library. It is successfully applied to generate libraries from distantly [85] to moderately related proteins and recover enzyme mutants with improved biochemical properties such as thermostability [86,87]. Library sizes are reduced due to increased efficiency in obtaining functional mutants.

The PROtein Sequence Activity Relationships (PROSAR) method analyzes mutations from different sources including homology guided mutagenesis, random mutagenesis, saturation mutagenesis and rational design [88]. The method defines mutations as 'beneficial', 'potentially beneficial', 'neutral' and 'deleterious' by computer-aided statistics. Beneficial mutations are combined, potentially beneficial mutants are re-tested, while neutral and deleterious mutations are discarded. The PROSAR method stresses the accumulation of multiple mutations in improving an enzyme. Notably, mutations that appear to harm enzyme activity but contribute to the overall improved enzyme activity can be selected using this method [88].

The third widely used strategy of semi-rational design is iterative saturation mutagenesis. In this method, only amino acid residues that are critical for the biological functions of an enzyme, such as ligand-binding [89] or thermostability [90], are selected and saturation mutagenesis is carried out for each of these residues. For example, Hoffmann *et al.* focused on the active site of a P450cam monooxygenase when generating mutant libraries. Of the 13 key amino acid residues in the active site, nine were manually selected for mutant library construction. For each residue, three to six amino acid changes were made, yielding a focused library of 291,600 variants. Monitoring NADH depletion, they found five mutants that were able to transform diphenylmethane and one mutant that could convert diphenylmethane into 4-hydroxydiphenylmethane [91].

Deep sequencing can be combined with computer-aided protein design [92]. Iterative directed evolution of enzymes may not always improve enzymes because the method can reach a plateau after certain rounds. For example, a hemaglutinin influenza virus binding inhibitor can be artificially evolved by epPCR to have an affinity at the nanomolar level. However, further directed evolution failed to discover mutants of higher affinity. Whitehead *et al.* hypothesized that small but combined contributions from mutations should allow further improvement [92]. For this purpose, they generated libraries containing approximately 1000 unique single point mutations, displayed the mutants on yeast and collected the mutants that were able to bind to hemagglutinin epitopes with different stringencies by cell sorting. The plasmids in collected yeast cells were extracted and subjected to deep-sequencing. Sequence–function landscapes were drawn and used to guide the design of mutants with combined enriched substitutions. The resulting mutants have higher subnanomolar binding affinities.

De novo design

In addition to the engineering methods mentioned above, there are also endeavors to design enzymes *de novo* [93]. A Rosetta method has been successfully and widely used to construct novel retro-aldol enzymes [94], Kemp elimination catalysts [94], an enzyme catalyzing a stereoselective Diels-Alder reaction [95] and a triosephosphate isomerase [96]. As described by Richter *et al.*, enzyme design using the Rosetta method essentially includes four stages: choice of a catalytic mechanism and corresponding minimal model active site;

identification of sites in a set of scaffold proteins where this minimal active site can be realized; optimization of the identities of the surrounding residues for stabilizing interactions with the transition state and primary catalytic residues; and evaluation and ranking the resulting designed sequences [96]. The model with the best overall score, the best ligand score, or the best constraint score is be chosen for experimental validation. The Rosetta method has been used for the design of a homodimer starting from a 5 kDa helical hairpin [97]. In this *de novo* design of zinc-mediated protein–protein interaction, Der *et al.* found that the protein–protein interface may form clefts that can hydrolyze carboxyesters and phosphoesters and are, therefore, candidates for enzyme engineering [98]. Notably, although the newly generated enzymes may be able to catalyze the reactions as expected, the activities may not be high enough for industrial applications. Therefore, a directed evolution method is often coupled with such *de novo* enzyme design for improvement.

The advantages and disadvantages of each protein engineering strategy are summarized in Table 1. Although each strategy is suited for different circumstances, they are by no means exclusive in their application. Semi-rational design, which is thus far the most successful protein engineering approach for developing industrial biocatalysts, is a combination of rational design and directed evolution. The following section highlights how these methods were integrated to successfully develop biocatalysts for industrial applications.

Selected examples of biocatalysts in the pharmaceutical industry

Driven by the research and innovations described above, the last decade witnessed a surge in the number of biocatalysts developed for industrial-scale production of pharmaceutical intermediates. In the form of purified enzymes or engineered microorganisms, biocatalysts are shaping the future of the pharmaceutical industry by providing cost-effective and green alternatives to organic chemistry methodologies. The examples in this section are selected to highlight the combination of technologies used by academic and industrial laboratories for adapting enzymes to pharmaceutical manufacturing, as well as the potential of biocatalysts to transform the industry. A comprehensive list of recently developed biocatalysts in the pharmaceutical industry using similar technologies is available in a separate review [2].

» Enzyme-based biocatalysts

The manufacture of sitagliptin, the active compound of leading antidiabetic drug Januvia™, showcases the immense potential of engineered enzyme-based biocatalysts in the pharmaceutical industry in both cost–competitiveness and environmental sustainability. Conventional chemical synthesis of sitagliptin involves the asymmetric hydrogenation of prositagliptin at high-pressure with a rhodium-based catalyst, followed by carbon treatment for rhodium-removal and subsequent crystallization to isolate the desired product [99]. While enantioselective, this chemocatalytic route is circuitous and costly due to the need for specialized high-pressure vessels, as well as the use and removal of precious transition metal catalysts. These limitations are circumvented with the development of a transaminase biocatalyst, which enables efficient one-step conversion of prositagliptin to stereopure sitagliptin at substantially higher yields and volumetric productivities (Figure 2). Starting from a transaminase scaffold with no apparent activity towards prositagliptin ketone, Hughes and co-workers obtained an enzyme with marginal activity by combining rational *in silico* design with saturation mutagenesis of amino acid residues in the small and large catalytic pockets [100,101]. Subsequent rounds of directed evolution and screening improved enzyme activity and simultaneously optimized the enzyme for performance criteria that are important for large-scale industrial applications such as dimethyl sulfoxide tolerance, stability at high

Table 1. Comparison of the different protein engineering strategies.

Strategy	Advantages	Disadvantages
Rational design	Small set of mutants to test	Knowledge about the enzyme, particularly the structural information, is required. A limited sequence space is explored and beneficial mutations may be consequently missed
Directed evolution	Knowledge of the enzyme is not required	Can be labor-intensive. Throughput is highly dependent on the screening or selection method used
Semi-rational design	Less labor-intensive than directed evolution, more diversity than rational design	Structural information of enzyme or several rounds of screening is required
De novo design	Can be used to create novel biocatalysts not found in nature	The designed biocatalysts usually have relatively low activity and need further optimization/improvement

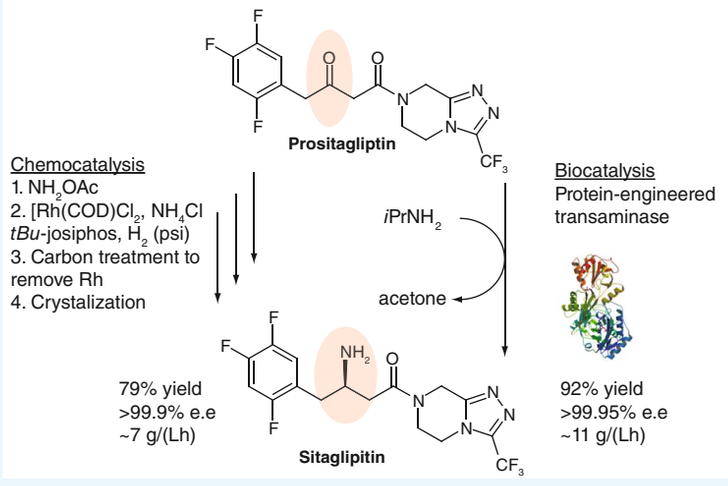

Figure 2. Improved biocatalytic asymmetric synthesis of sitagliptin. The one-step biocatalytic route involving an engineered transaminase results in a higher yield of optically pure sitagliptin compared with the chemocatalytic route involving chemical asymmetric hydrogenation at high pressure in the presence of a transitional metal catalyst.
e.e.: Enantiomeric excess.
Adapted with permission from [2] © Macmillan Publishers Ltd (2012).

temperatures and substrate concentrations, as well as expression in *E. coli* [100]. Beyond sitagliptin manufacture, the broad substrate range of these evolved transaminases also provides the framework for synthesis of chiral amines from prochiral ketones, complementing the more established approach of synthesizing chiral alcohols with biocatalysts such as ketoreductases (KREDs) in the manufacturing of key pharmaceutical intermediates. As scientific and technological advances in protein engineering and design drives the development and optimization of enzymes with broad utility, the expanding biocatalytic toolbox will be invaluable to the pharmaceutical industry [2].

When it comes to designing biocatalytic routes for pharmaceutical production, all roads lead to Rome, as epitomized by atorvastatin – the active ingredient of the top-selling cholesterol-lowering drug Lipitor®. The (3*R*,5*S*)-dihydroxyhexanoate side chain of atorvastatin, with its two chiral centers, is challenging to make by traditional chemical means. Driven by high market demand as well as the need for high chemical and stereochemical purity, a variety of chemozymatic routes involving different classes of enzymes and starting materials have been developed for enantioselective synthesis of the homochiral dihydrohexanoate pharmacophore common to many statins, including atorvastatin [102]. These chemoenzymatic approaches may be broadly grouped into two main strategies. The first strategy involves synthesis of the key chiral building block, ethyl (*R*)-4-cyano-3-hydroxybutyrate, on which the second stereocenter can be introduced at a later step. Most proposed biocatalytic routes, including the use of nitrilases to desymmetrize prochiral 3-hydroxyglutaryl nitrile (Figure 3) [27,103], fall under this category but are yet to be commercialized because of the high enzyme load needed, which complicates product recovery and increases production costs [102]. Leveraging directed evolution technologies, Codexis (CA, USA) scientists developed a two-step, three-enzyme process that is efficient, economical and environmentally friendly for industrial-scale production of ethyl (*R*)-4-cyano-3-hydroxybutyrate (Figure 3) [23,104]. Notably, this two-step process has been shown to proceed as a one-pot process on a laboratory scale [23]. In the first step, enantioselective reduction of ethyl 4-chloroacetate is performed using a KRED while an NADP-dependent glucose dehydrogenase regenerates the NADPH cofactor pool using glucose as a reductant. A halohydrin dehydrogenase (HHDH) then replaces the chloro group with a cyano substituent in the second step to form the desired product. For this biocatalytic process to be commercially relevant, gene shuffling was used to improve the stability and activity of KRED and glucose dehydrogenase under industrial conditions while maintaining their remarkable enantioselectivity (>99.5% enantiomeric excess) [104]. For HHDH, a semi-rational approach integrating multivariate optimization with semisynthetic gene shuffling yielded an enzyme with an impressive 4000-fold increase in volumetric productivity and near perfect enantioselectivity (>99.9% enantiomeric excess) [88]. Altogether, the combination of bioinformatics and directed evolution of individual biocatalysts in a multienzyme cascade enabled striking improvements in substrate and biocatalyst loading, volumetric productivity as well as overall yield, which in turn enabled industrial-scale enzymatic reactions that would otherwise be economically unfeasible.

The second strategy involves the generation of both chiral centers in a single step using enzyme-catalyzed aldol condensation of choloracetaldehyde and acetaldehyde to produce a more advanced six-carbon atorvastatin intermediate (Figure 3). The cross-aldolreaction was first described by Wong and co-workers, who showed that 2-deoxyribose-5-phosphate aldolase (DERA) has substrate specificity towards acetaldehyde as a donor aldehyde substrate in asymmetric aldol condensations [105–107]. Low production capacity, low resistance to aldehyde levels and poor acceptance of non-phosphorylated aldehyde substrates, however, limited the industrial practicality of the wild-type *E. coli* DERA enzyme [106]. As such, a wide range of strategies have been employed to develop and optimize DERA as an industrial biocatalyst in atorvastatin production. By screening environmental genomic libraries to isolate natural aldolase variants

with higher activities and overcoming substrate inhibition through an improved fed-batch reaction process, Burk and colleagues achieved a 400-fold increase in volumetric productivity and a tenfold improvement in catalyst load for the reaction between acetaldehyde and chloroacetaldehyde [108]. Rational design of the DERA yielded a S238D variant with expanded substrate selectivity over the parent enzyme [109]. The S238D mutant enzyme is capable of catalyzing aldol condensation between the non-physiological substrate 3-azidopropinaldehyde and two molecules of acetaldehyde to form azidoethyl pyranose, a strategic precursor in atorvastatin synthesis (Figure 3). Jennewein *et al.* employed a directed evolution strategy using epPCR and screening for DERA variants with desired properties under industrially relevant conditions, thereby identifying structural 'hotspots' affecting DERA resistance and catalytic activity towards the non-physiological acceptor substrate chloroacetaldehyde [110]. Recombining beneficial mutations yielded an enzyme with increased chloroacetaldehyde tolerance and productivity (tenfold) in the synthesis of (3*R*,5*S*)-6-chloro-2,4,6-trideoxyhexapyranoside. Overall, while there is certainly still room for improvement, engineered DERA variants enabled efficient and enantioselective formation of carbon–carbon bonds to give an advanced atorvastatin intermediate with both stereocenters in one step from cheap bulk chemicals, thereby dramatically enhancing the cost–competitiveness of this biocatalytic route.

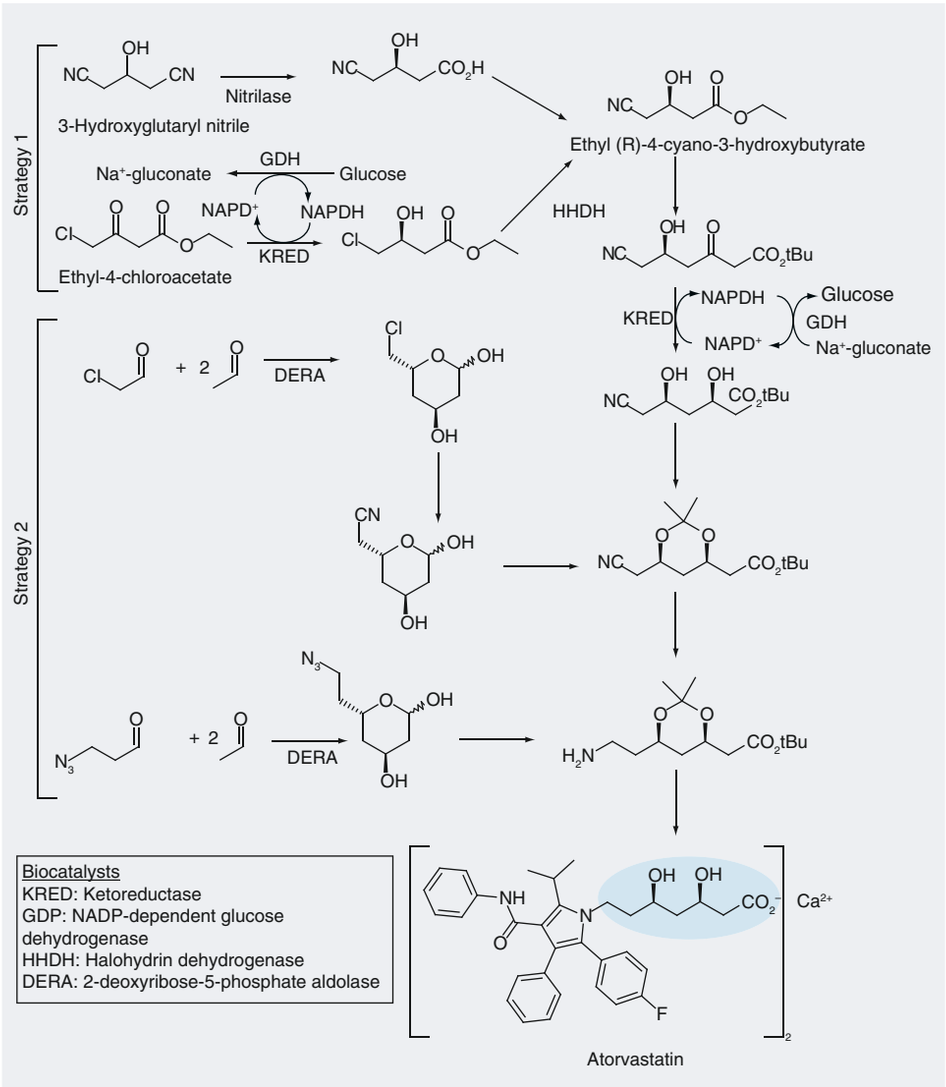

Figure 3. Multiple biocatalytic routes towards key intermediates of atorvastatin. The different processes can be classified into two main strategies. They differ not only in class of enzyme, but also in choice of (inexpensive) starting material, activity and selectivity of the biocatalyst, downstream processing, and yield and purity of final product. Strategy one creates one stereocenter of the homochiral atorvastatin side chain (shaded) and the second stereocenter in a subsequent step, while strategy two creates both stereocenters required for the advanced pharmaceutical intermediate. Identities of the biocatalysts are as indicated.
Adapted with permission [2] © Macmillan Publishers Ltd (2012).

Advances in protein engineering enable the conversion of naturally occurring enzymes into practical industrial biocatalysts with dramatically improved catalytic properties and tolerance to harsh processing conditions. Engineered biocatalysts with broad substrate ranges provide the framework for core enzymatic transformations and can be evolved to accept unnatural substrates, fit process specifications, or be integrated into multienzyme cascade processes, thereby accelerating biocatalyst and process development. Successful implementation of a biocatalytic manufacturing process will be determined by its overall economic viability, which takes into account factors such as scale, product yield and purity, starting material costs and follow-up steps.

» Whole-cell biocatalysts

Whole-cell biocatalysts represent the next logical step in biocatalysis: *in vivo* multienzyme cascade processes. Using a combination of metabolic and protein engineering tools to engineer and redirect biosynthetic pathways, complicated combinatorial reactions can

be carried out in a microbial host to overproduce a multitude of chemicals from cheap renewable raw materials [67,68]. This is particularly advantageous for the production of natural products, a major pharmaceutical source, whose structural complexity has hindered industrial production since chemical syntheses are challenging while extraction from natural sources is often plagued by meager yields. In particular, many plant terpenoids, such as artemisinin and paclitaxel have pharmaceutical properties, but chemical synthesis of these complex molecules is impractical and availability of the petroleum-derived starting materials is limited [111]. Biological production of natural product precursors or derivatives from renewable feedstocks using whole-cell biocatalysts, in these cases, may be more cost effective and sustainable.

Jay Keasling's semisynthetic anti-malarial artemisinin has been Exhibit A for the promise of whole-cell biocatalysts in the pharmaceutical industry. The current supply and affordability of artemisinin extracted from sweet wormwood *Artemisia annua* are inadequate to address the global burden of malaria, especially in the developing world. In comparison, the semi-synthesis of artemisinin from microbially produced precursors offers a more steady and cost–effective source. By transplanting biosynthetic genes from *A. annua* into yeast and re-directing of carbon flux to increase production, Keasling and co-workers showed that up to 1 g/l of artemisinic acid, an immediate precursor of artemisinin, may be produced from simple sugars **(Figure 4)** [112,113]. Another yeast strain was engineered to produce up to 40 g/l of amorphadiene, another artemisinin precursor [114]. Further bolstering this semi-synthetic approach is the development of the continuous-flow conversion of artemisinic acid to artemisinin, which has been a chemically challenging step. Since purification of intermediates is not needed, continuous-flow synthesis of artemisinin is efficient, scalable and inexpensive, ensuring a reliable industrial supply of the drug from yeast-derived artemisinic acid at lower costs [115].

In addition to rewiring metabolic pathways, protein engineering is also indispensable in the development and optimization of whole-cell biocatalysts. In the native semisynthetic route, the *A. annua* cytochrome P450 monooxygenase catalyzes the three-step oxidation of amorpha-4,11-diene to artemisinic acid [112]. A semi-rational approach integrating computer modeling and saturation mutagenesis to increase the substrate promiscuity of P450 from *Bacillus megaterium* yielded an enzyme that is capable of selective oxidation of amorpha-4,11-diene to artemisinic-11S,12-epoxide in cells, providing an alternative semisynthetic route for artemesinin production **(Figure 4)** [116]. Directed evolution involving active site mutagenesis, high-throughput P450 fingerprinting and multivariate analysis generated P450 variants with different regioselectivities, capable of hydroxylating artemisinin derivatives at inaccessible parts of the complex molecule at high yields (>90%) and preparative scales *in vitro* [117].

A similar approach was employed to overproduce the precursor of the widely used antineoplastic drug paclitaxel (Taxol®) in *E. coli* [118]. By introducing a heterologous terpenoid biosynthetic pathway coupled with multivariate-modular optimization of metabolic flux through the upstream isoprenoid pathway, Stephanopoulos' group obtained an engineered *E. coli* strain that produces taxadiene at approximately 1 g/l, an impressive 15,000-fold improvement over the parent strain before optimization **(Figure 4)**. Another key step in the development involves protein engineering of *Taxus* cytochrome P450 for regioselective hydroxylation of taxadiene to taxadien-5α-ol; transmembrane engineering and translational fusions with cognate redox partners yielded chimeric enzymes with improved solubilities and efficiencies. Nonetheless, before the identities of genes in the Taxol biosynthetic pathway are fully elucidated and engineered into industrial microorganisms with optimized metabolic flux, it will be difficult for biosynthetic or semisynthetic Taxol to compete with the current production method involving plant cell fermentation and direct extraction [119].

In addition to transplanting biosynthetic pathways and enzymes into heterologous industrial hosts, metabolic engineering also involves the redirection and balancing of metabolic flux to maximize yield and productivity [120,121]. Protein engineering of enzymes, such as P450s, remains relevant for generating new enzyme functions and expanding the scope and utility of these cellular factories in producing structurally diverse natural product precursors and derivatives [122]. Continued development of metabolic and protein engineering strategies, aided by bioinformatics, will be needed to fully realize the industrial potential of cell-based biocatalysts.

Moving forward: new enzymes, new chemistries

Directed evolution has been shown to dramatically improve desired enzyme properties such as stability and activity and in some cases, laboratory evolution has yielded biocatalysts from non-catalytic scaffolds [123,124]. Nonetheless, directed evolution generally requires, and is most effective, when starting with an enzyme with an existing, albeit marginal, activity of interest. While nature remains a rich and important source for new enzymes and chemistries [43], the development of bioinformatics tools backed by extensive empirical knowledge and/or technologies for strategic introduction of new

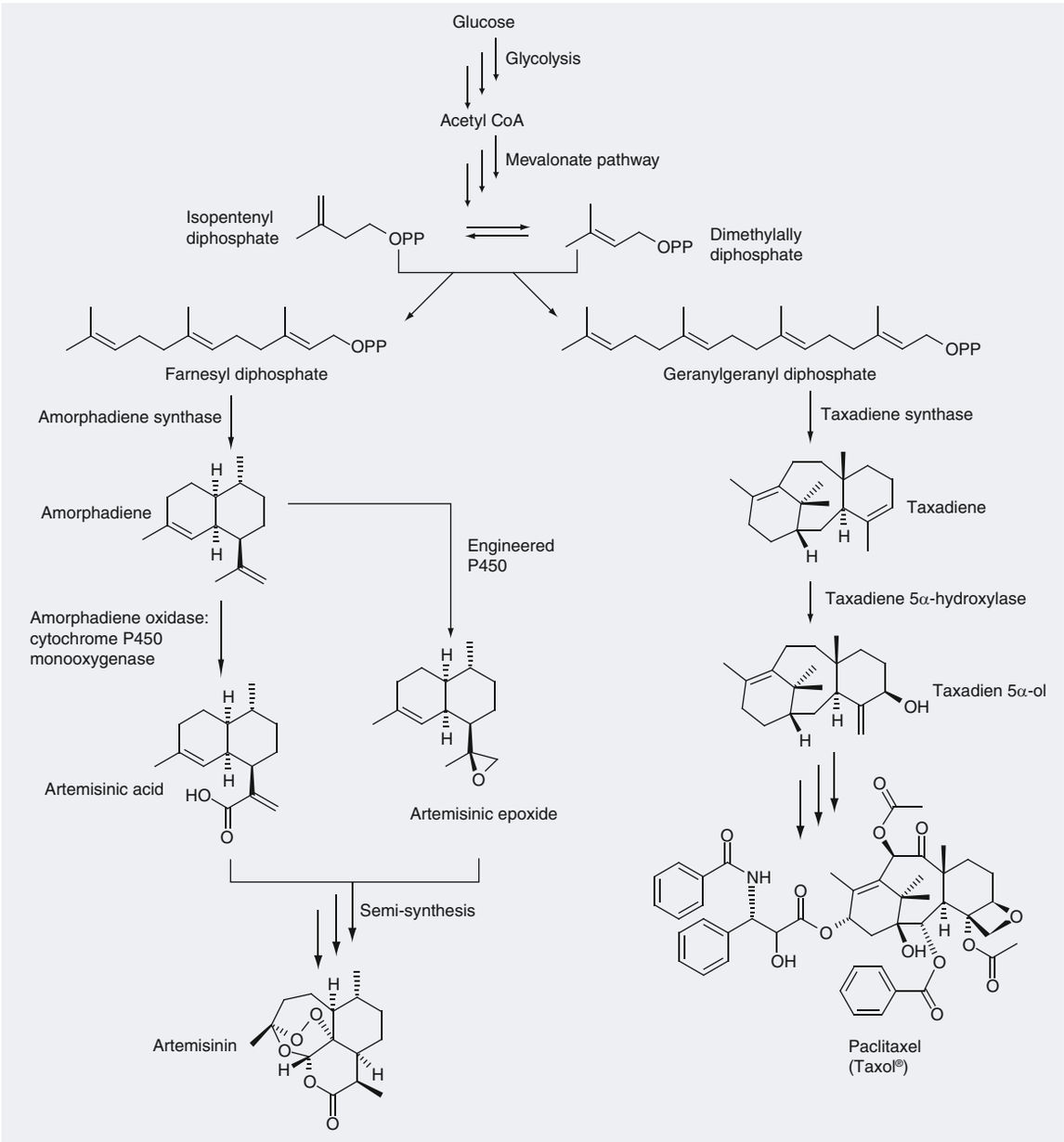

Figure 4. Biosynthetic strategies of artemisinin and paclitaxel using whole-cell biocatalysts. On the left are the semi-synthetic strategies for artemisinin. Introduction of *Artemisia annua* genes encoding the amorphadiene synthase, amorphodiene and cytochrome P450 monooxygenase resulted in microbial strains that produce artemisinic acid, which can be chemically converted to artemisinin [112]. In an alternative route, an engineered P450 produces artemisinic epoxide that can also be transformed into artemisinin by established chemical means [116]. On the right is the biosynthetic strategy for paclitaxel (Taxol®). Introduction of the taxadiene synthase and an engineered taxadiene 5α-hydroxylase from the *Taxus* host allowed the production of key Taxol intermediates, taxadiene and taxadien-5α-ol, in *Escherichia coli* [118].

chemical functionalities paves the way for the creation of artificial enzymes for novel functions.

With an improved structural and mechanistic understanding of how naturally occurring enzymes work, *de novo* design of enzymes is now possible. David Baker's laboratory has made notable progress in computational design of novel enzyme catalysts, generating enzymes from existing protein scaffolds whose backbone positions in the binding pockets align with the geometries of the ideal active site for a given reaction [125]. This approach has been successfully used to create enzymes for reactions for which no naturally occurring enzyme exists, including the Kemp elimination [94], retro-aldol [94] and Diels–Alder [95] reactions. For organometallic

catalysis, an Rh(III) metalloenzyme with high reactivity and selectivity was fashioned from a non-catalytic streptavidin scaffold by precise positioning of a carboxylate side chain close to the metallocenter in the active site [126]. In fact, the enantioselectivity of the artificial Rh(III) metalloenzyme in directed carbon–hydrogen bond functionalization is comparable to that of chiral cyclopentadienyl ligands [126,127], offering new opportunities for late transition metal asymmetric catalysis in industrial biotechnology, including compatibility with biocatalysts in concurrent cascade reactions [128]. For designed enzymes however, the resulting proteins are often less than ideal as strategic placement of catalytic residues is likely to compromise substrate and transition-state binding affinity, leading to low catalytic efficiency. These limitations can be overcome by directed evolution, which can be further guided by bioinformatics. For example, directed evolution of the rationally designed KE59 Kemp eliminase led to a >2000-fold increase in its catalytic efficiency, approaching that of naturally occurring enzymes [129]. Until the day we fully understand the principles of designing an ideal enzyme [130], the marriage of computational design and molecular evolution will be invaluable towards developing customized biocatalysts for a wide range of existing and novel chemical transformations.

Chemical modifications of specific amino acid residues can lead to enzymes with enhanced or novel physicochemical and biological properties. In fact, enzymes found in nature often require chemistries beyond those offered by the side chains of the twenty canonical proteinogenic amino acids for function in the form of cofactors or post-translational modifications [131–133]. Propelled by technological advances in protein ligation [134–136] and genetic code reprogramming [137,138], considerable progress has been made towards site-specific incorporation of unnatural building blocks into proteins. Pioneered by Tom Muir, expressed protein ligation and protein trans-splicing are intein-based methods that allow traceless ligation of recombinant and synthetic polypeptides, the latter fragment harboring almost any chemical modification(s) at the site of interest **(Figure 5)** [134,135]. This semisynthetic approach circumvents the size limitation associated with total chemical synthesis and allows chemical manipulation of the protein backbone, which is crucial for protein structure and function but inaccessible by standard mutagenesis [136]. In a complementary approach pioneered by Peter Schultz, a myriad of unnatural amino acids with distinct structures and chemistries are genetically encoded in response to 'blank' nonsense and frameshift quadruplet codons and introduced into proteins by engineered translational machinery in bacterial, yeast and mammalian systems **(Figure 6)** [137,138]. The use of genetically encoded unnatural amino acids expands the protein function space that can then be sampled and selected for by directed evolution [119,139,140]. Collectively, these emerging technologies advance the concept of protein engineering beyond nature's genetic code, expanding the catalytic capacities of enzymes or microorganisms and promising new biocatalytic solutions to the plethora of synthetic chemical transformations inaccessible to naturally occurring enzymes.

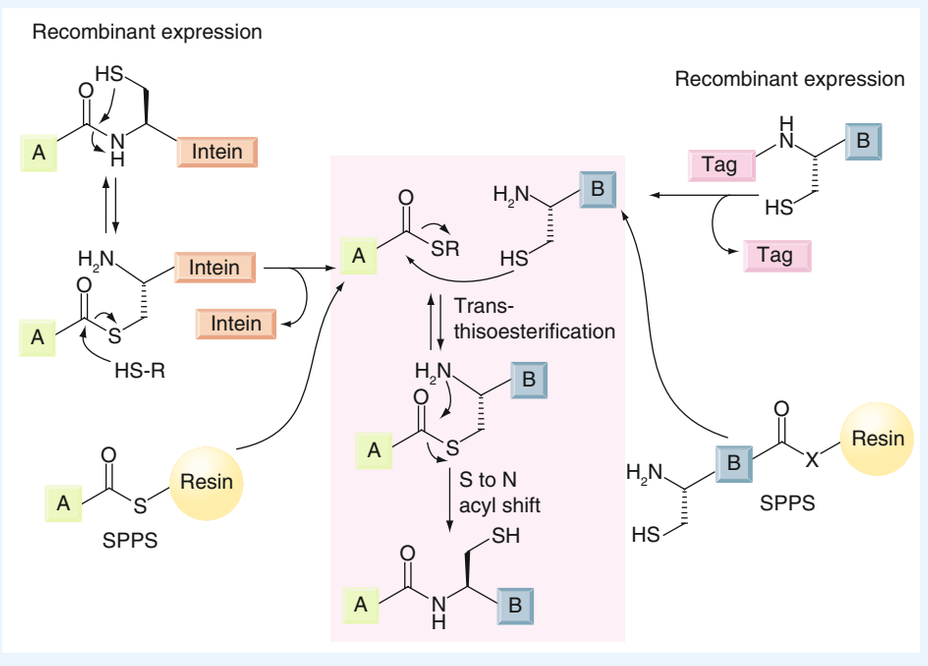

Figure 5. Expressed protein ligation. The boxed region designates the native chemical ligation reaction in which trans-thioesterification of the protein α-thioester by the N-terminal Cys polypeptide is followed by an S to N acyl shift to generate a new peptide bond linking the two polypeptides. α-thioesters can be obtained recombinantly, using engineered inteins, or by chemical synthesis. N-terminal Cys polypeptides can also be produced recombinantly or made using SPPS. SPPS: Solid-phase peptide synthesis methods.
Reproduced with permission from [136].

Future perspective

We envision a future where biocatalytic and chemocatalytic processes are seamlessly integrated in pharmaceutical manufacturing with biocatalysts being made-to-order: enzymes or microorganisms are designed and optimized for specified chemical conversions. Considerations to be made during the computer-aided design

process include choice of possible biocatalytic routes, compatibility of the biocatalysts with the manufacturing process as well as the costs of starting materials and downstream processing. In addition, the desired properties and target performance criteria of the biocatalyst will also be defined at the drawing board. Advanced low-cost high-throughput DNA synthesis will empower biocatalyst development by facilitating the creation of customized genes or gene libraries for directed evolution and screening. In the case of whole-cell biocatalysts, entire pathways or genomes will be synthesized and directly transplanted into industrial microorganisms, reprogramming them into microbial cell factories with desired metabolic characteristics to manufacture desired molecules from designated renewable feedstocks, aided by advanced genome and metabolic engineering tools. There may be a day when we can design and engineer industrial enzymes or microorganisms satisfying specified performance criteria without having to go through iterative cycles of optimization. Until then, a semi-rational approach involving directed evolution and multivariate optimization algorithms remains indispensable and key to biocatalyst development. Continued advances in computing as well as protein and metabolic engineering capabilities will be critical to harness nature's diverse catalytic toolkit and to realize the full potential of biocatalysts for chemical synthesis in the pharmaceutical industry.

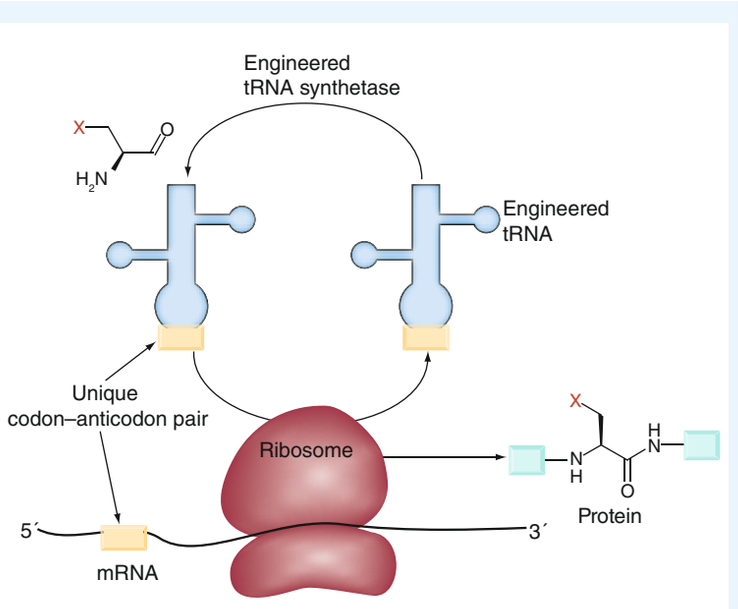

Figure 6. Genetically encoded unnatural amino acids. The incorporation position of unnatural amino acid is genetically encoded by 'blank' nonsense or frameshift quadruplet codons that are recognized during translation by the engineered tRNA with complementary anticodons. Charging of the tRNA with the unnatural amino acid is specifically carried out by an orthogonal engineered aminoacyl tRNA synthetase. 'X' represents unnatural amino acid side chains, that is, beyond the 20 canonical amino acids, including heavy atoms, photocrosslinkers, chemical handles such as keto, azide, thioester moieties and other novel chemical functionalities [137,138].

Executive summary

Biocatalyst discovery

» With unprecedented ability to sequence genomes, metagenomes and RNA libraries, the process of discovering novel biocatalysts has been greatly enhanced, offering researchers access to a vast number of potential starting enzymes for biocatalytic processes.

» The advent of powerful gene synthesis technologies has enabled researchers to tap into the diverse candidate genes generated from DNA sequencing and test for feasible starting enzymes.

Biocatalyst development by protein & pathway engineering

» Semi-rational design, which is an integration of rational protein design and directed evolution, is, to date, the most successful method in engineering biocatalysts for pharmaceutical processes.

» The biosynthesis of sitagliptin and key intermediates of atorvastatin are successful examples of the combination of bioinformatics and directed evolution to enable striking improvements in substrate and biocatalyst loading, volumetric productivity as weell as overall yield, which in turn enable industrial-scale enzymatic reactions.

» Whole-cell biocatalysis involving *in vivo* multienzyme catalytic processes is particularly advantageous for the production of natural products, which are often difficult to synthesize chemically or extract from natural sources. The biosynthetic routes to artemisinin and paclitaxel are promising examples of whole-cell biocatalytic processes developed through enzyme and pathway engineering.

Creating biocatalysts with novel functions

» *De novo* design of enzymes, which can create enzymes with novel functions, is now achievable with improved structural and mechanistic understanding of how natural enzymes work. However, as we currently do not fully understand the principles of designing an ideal enzyme, the coupling of computational design and molecular evolution will still be required for developing customized biocatalysts for a wide range of existing and novel chemical transformations.

» Continued advances in computing as well as protein and metabolic engineering capabilities will be critical to realize the full potential of biocatalysts for chemical synthesis in the pharmaceutical industry.

Financial & competing interests disclosure

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties. No writing assistance was utilized in the production of this manuscript.

References

Papers of special note have been highlighted as:
■ of interest
■■ of considerable interest

1 Schmid A, Dordick JS, Hauer B, Kiener A, Wubbolts M, Witholt B. Industrial biocatalysis today and tomorrow. *Nature* 409(6817), 258–268 (2001).

2 Bornscheuer UT, Huisman GW, Kazlauskas RJ, Lutz S, Moore JC, Robins K. Engineering the third wave of biocatalysis. *Nature* 485(7397), 185–194 (2012).

■■ **Technological and conceptual advances in protein engineering for biocatalyst design and development for industrial applications.**

3 Arnold FH. Combinatorial and computational challenges for biocatalyst design. *Nature* 409(6817), 253–257 (2001).

4 Turner NJ. Directed evolution drives the next generation of biocatalysts. *Nat. Chem. Biol.* 5(8), 567–573 (2009).

5 Cobb RE, Chao R, Zhao H. Directed evolution: past, present and future. *AIChE J.* 59(5), 1432–1440 (2013).

6 Buchanan A, Ferraro F, Rust S *et al.* Improved drug-like properties of therapeutic proteins by directed evolution. *Protein Eng. Des. Sel.* 25(10), 631–638 (2012).

7 Jones DS, Silverman AP, Cochran JR. Developing therapeutic proteins by engineering ligand–receptor interactions. *Trends Biotechnol.* 26(9), 498–505 (2008).

8 Metzker ML. Sequencing technologies – the next generation. *Nat. Rev. Genet.* 11(1), 31–46 (2010).

9 Dressman D, Yan H, Traverso G, Kinzler KW, Vogelstein B. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc. Natl Acad. Sci. USA* 100(15), 8817–8822 (2003).

10 Ronaghi M, Karamohamed S, Pettersson B, Uhlen M, Nyren P. Real-time DNA sequencing using detection of pyrophosphate release. *Anal. Biochem.* 242(1), 84–89 (1996).

11 Harris TD, Buzby PR, Babcock H *et al.* Single-molecule DNA sequencing of a viral genome. *Science* 320(5872), 106–109 (2008).

12 Eid J, Fehr A, Gray J *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* 323(5910), 133–138 (2009).

13 Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, Bayley H. Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.* 4(4), 265–270 (2009).

14 Urlacher VB, Girhard M. Cytochrome P450 monooxygenases: an update on perspectives for synthetic application. *Trends Biotechnol.* 30(1), 26–36 (2012).

15 Lamb DC, Skaug T, Song HL *et al.* The cytochrome P450 complement (CYPome) of *Streptomyces coelicolor* A3(2). *J. Biol. Chem.* 277(27), 24000–24005 (2002).

16 Khatri Y, Hannemann F, Perlova O, Muller R, Bernhardt R. Investigation of cytochromes P450 in myxobacteria: excavation of cytochromes P450 from the genome of *Sorangium cellulosum* So ce56. *FEBS Lett.* 585(11), 1506–1513 (2011).

17 Nelson DR. Progress in tracing the evolutionary paths of cytochrome P450. *Biochim. Biophys. Acta* 1814(1), 14–18 (2011).

18 Lautru S, Deeth RJ, Bailey LM, Challis GL. Discovery of a new peptide natural product by *Streptomyces coelicolor* genome mining. *Nat. Chem. Biol.* 1(5), 265–269 (2005).

19 Bergmann S, Schumann J, Scherlach K, Lange C, Brakhage AA, Hertweck C. Genomics-driven discovery of PKS-NRPS hybrid metabolites from *Aspergillus nidulans*. *Nat. Chem. Biol.* 3(4), 213–217 (2007).

20 Yin J, Straight PD, Hrvatin S *et al.* Genome-wide high-throughput mining of natural-product biosynthetic gene clusters by phage display. *Chem. Biol.* 14(3), 303–312 (2007).

21 Kersten RD, Yang YL, Xu Y *et al.* A mass spectrometry-guided genome mining approach for natural product peptidogenomics. *Nat. Chem. Biol.* 7(11), 794–802 (2011).

22 Truppo MD, Turner NJ, Rozzell JD. Efficient kinetic resolution of racemic amines using a transaminase in combination with an amino acid oxidase. *Chem. Commun. (Camb.)* (16), 2127–2129 (2009).

23 Ager DJ, Li T, Pantaleone DP, Senkpeil RF, Taylor PP, Fotheringham IG. Novel biosynthetic routes to non-proteinogenic amino acids as chiral pharmaceutical intermediates. *J. Mol. Catal. B Enzymatic* 11, 199–205 (2001).

24 Hohne M, Kuhl S, Robins K, Bornscheuer UT. Efficient asymmetric synthesis of chiral amines by combining transaminase and pyruvate decarboxylase. *ChemBioChem* 9(3), 363–365 (2008).

25 Hohne M, Schatzle S, Jochens H, Robins K, Bornscheuer UT. Rational assignment of key motifs for function guides *in silico* enzyme identification. *Nat. Chem. Biol.* 6(11), 807–813 (2010).

26 Schatzle S, Steffen-Munsberg F, Thontowi A, Hohne M, Robins K, Bornscheuer UT. Enzymatic asymmetric synthesis of enantiomerically pure aliphatic, aromatic and arylaliphatic amines with *(R)*-selective amine transaminases. *Adv. Synth. Catal.* 353, 2439–2445 (2011).

27 Desantis G, Zhu Z, Greenberg WA *et al.* An enzyme library approach to biocatalysis: development of nitrilases for enantioselective production of carboxylic acid derivatives. *J. Am. Chem. Soc.* 124(31), 9024–9025 (2002).

28 Robertson DE, Chaplin JA, Desantis G *et al.* Exploring nitrilase sequence space for enantioselective catalysis. *Appl. Environ. Microbiol.* 70(4), 2429–2436 (2004).

29 Zhu S, Gong C, Song D, Gao S, Zheng G. Discovery of a novel (+)-γ-lactamase from *Bradyrhizobium japonicum* USDA 6 by rational genome mining. *Appl. Environ. Microbiol.* 78(20), 7492–7495 (2012).

30 Zhu D, Mukherjee C, Biehl ER, Hua L. Discovery of a mandelonitrile hydrolase from *Bradyrhizobium japonicum* USDA110 by rational genome mining. *J. Biotechnol.* 129(4), 645–650 (2007).

31 Zhu D, Mukherjee C, Yang Y *et al.* A new nitrilase from *Bradyrhizobium japonicum* USDA 110. Gene cloning, biochemical characterization and substrate specificity. *J. Biotechnol.* 133(3), 327–333 (2008).

32 Kaplan O, Bezouska K, Malandra A *et al.* Genome mining for the discovery of new nitrilases in filamentous fungi. *Biotechnol. Lett.* 33(2), 309–312 (2011).

33 Basile LJ, Willson RC, Sewell BT, Benedik MJ. Genome mining of cyanide-degrading nitrilases from filamentous fungi. *Appl. Microbiol. Biotechnol.* 80(3), 427–435 (2008).

34 Guimil S, Chang HS, Zhu T *et al.* Comparative transcriptomics of rice reveals an ancient pattern of response to microbial colonization. *Proc. Natl Acad. Sci. USA* 102(22), 8066–8070 (2005).

35 Nagalakshmi U, Wang Z, Waern K *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320(5881), 1344–1349 (2008).

36 Wilhelm BT, Marguerat S, Watt S *et al.* Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* 453(7199), 1239–1243 (2008).

37 Lister R, O'Malley RC, Tonti-Filippini J *et al.* Highly integrated single-base resolution maps of the epigenome in *Arabidopsis. Cell* 133(3), 523–536 (2008).

38 Mortazavi A, Williams BA, Mccue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5(7), 621–628 (2008).

39 Barbazuk WB, Emrich SJ, Chen HD, Li L, Schnable PS. SNP discovery via 454 transcriptome sequencing. *Plant J.* 51(5), 910–918 (2007).

40 Vera JC, Wheat CW, Fescemyer HW *et al.* Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Mol. Ecol.* 17(7), 1636–1647 (2008).

41 Meyer E, Aglyamova GV, Wang S *et al.* Sequencing and *de novo* analysis of a coral larval transcriptome using 454 GSFlx. *BMC Genomics* 10, 219 (2009).

42 Cloonan N, Forrest AR, Kolle G *et al.* Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods* 5(7), 613–619 (2008).

43 Geu-Flores F, Sherden NH, Courdavault V *et al.* An alternative route to cyclic terpenes by reductive cyclization in iridoid biosynthesis. *Nature* 492, 138–142 (2012).

44 Wright PC, Noirel J, Ow SY, Fazeli A. A review of current proteomics technologies with a survey on their widespread use in reproductive biology investigations. *Theriogenology* 77(4), 738–765.e752 (2012).

45 Gangoiti J, Santos M, Prieto MA, De La Mata I, Serra JL, Llama MJ. Characterization of a novel subgroup of extracellular medium-chain-length polyhydroxyalkanoate depolymerases from actinobacteria. *Appl. Environ. Microbiol.* 78(20), 7229–7237 (2012).

46 Pohl NL. Functional proteomics for the discovery of carbohydrate-related enzyme activities. *Curr. Opin. Chem. Biol.* 9(1), 76–81 (2005).

47 Schmidinger H, Hermetter A, Birner-Gruenberger R. Activity-based proteomics: enzymatic activity profiling in complex proteomes. *Amino Acids* 30(4), 333–350 (2006).

48 Jessani N, Young JA, Diaz SL, Patricelli MP, Varki A, Cravatt BF. Class assignment of sequence-unrelated members of enzyme superfamilies by activity-based protein profiling. *Angew. Chem. Int. Ed. Engl.* 44(16), 2400–2403 (2005).

49 Uttamchandani M, Li J, Sun H, Yao SQ. Activity-based protein profiling: new developments and directions in functional proteomics. *ChemBioChem* 9(5), 667–675 (2008).

50 Lorenz P, Eck J. Metagenomics and industrial applications. *Nat. Rev. Microbiol.* 3(6), 510–516 (2005).

51 Homann MJ, Vail RB, Previte E *et al.* Rapid identification of enantioselective ketone reductions using targeted microbial libraries. *Tetrahedron* 60(3), 789–797 (2004).

52 Carlson R. The changing economics of DNA synthesis. *Nat. Biotechnol.* 27(12), 1091–1094 (2009).

53 Xiong AS, Yao QH, Peng RH *et al.* A simple, rapid, high-fidelity and cost-effective PCR-based two-step DNA synthesis method for long gene sequences. *Nucleic Acids Res.* 32(12), e98 (2004).

54 Gao X, Yo P, Keith A, Ragan TJ, Harris TK. Thermodynamically balanced inside-out (TBIO) PCR-based gene synthesis: a novel method of primer design for high-fidelity assembly of longer gene sequences. *Nucleic Acids Res.* 31(22), e143 (2003).

55 Kodumal SJ, Patel KG, Reid R, Menzella HG, Welch M, Santi DV. Total synthesis of long DNA sequences: synthesis of a contiguous 32-kb polyketide synthase gene cluster. *Proc. Natl Acad. Sci. USA* 101(44), 15573–15578 (2004).

56 Cello J, Paul AV, Wimmer E. Chemical synthesis of poliovirus cDNA: generation of infectious virus in the absence of natural template. *Science* 297(5583), 1016–1018 (2002).

57 Smith HO, Hutchison CA 3rd, Pfannkoch C, Venter JC. Generating a synthetic genome by whole genome assembly: phiX174 bacteriophage from synthetic oligonucleotides. *Proc. Natl Acad. Sci. USA* 100(26), 15440–15445 (2003).

58 Gibson DG, Glass JI, Lartigue C *et al.* Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* 329(5987), 52–56 (2010).

▪ **Demonstrates that a synthetic genome is sufficient to reprogram the recipient cells with appropriate phenotypes, pushing the boundaries and possibilities of genome engineering.**

59 Tian J, Gong H, Sheng N *et al.* Accurate multiplex gene synthesis from programmable DNA microchips. *Nature* 432(7020), 1050–1054 (2004).

60 Albert TJ, Norton J, Ott M *et al.* Light-directed 5′→3′ synthesis of complex oligonucleotide microarrays. *Nucleic Acids Res.* 31(7), e35 (2003).

61 Richmond KE, Li MH, Rodesch MJ *et al.* Amplification and assembly of chip-eluted DNA (AACED): a method for high-throughput gene synthesis. *Nucleic Acids Res.* 32(17), 5011–5018 (2004).

62 Jayaraj S, Reid R, Santi DV. GeMS: an advanced software package for designing synthetic genes. *Nucleic Acids Res.* 33(9), 3011–3016 (2005).

63 Leproust EM, Peck BJ, Spirin K *et al.* Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. *Nucleic Acids Res.* 38(8), 2522–2540 (2010).

64 Matzas M, Stahler PF, Kefer N *et al.* High-fidelity gene synthesis by retrieval of sequence-verified DNA identified using high-throughput pyrosequencing. *Nat. Biotechnol.* 28(12), 1291–1294 (2010).

65 Quan J, Saaem I, Tang N *et al.* Parallel on-chip gene synthesis and application to optimization of protein expression. *Nat. Biotechnol.* 29(5), 449–452 (2011).

66 Reitman ZJ, Choi BD, Spasojevic I, Bigner DD, Sampson JH, Yan H. Enzyme redesign guided by cancer-derived IDH1 mutations. *Nat. Chem. Biol.* 8(11), 887–889 (2012).

67 Payongsri P, Steadman D, Strafford J, Macmurray A, Hailes HC, Dalby PA. Rational substrate and enzyme engineering of transketolase for aromatics. *Org. Biomol. Chem.* 10(45), 9021–9029 (2012).

68 Chen M, Drury JE, Christianson DW, Penning TM. Conversion of human steroid 5 β -reductase (AKR1D1) into 5β-hydroxysteroid dehydrogenase by single point mutation E120H: example of perfect enzyme engineering. *J. Biol. Chem.* 287(20), 16609–16622 (2012).

69 Mclachlan MJ, Sullivan RP, Zhao H. Directed enzyme evolution and high-throughput screening. In: *Biocatalysis for the Pharmaceutical Industry: Discovery, Development, and Manufacturing.* Tao J, Lin GQ, Liese A (Eds). John Wiley and Sons Asia (Pte) Ltd, Singapore, Chapter 3, 45–64 (2009).

70 Du J, Yuan Y, Si T, Lian J, Zhao H. Customized optimization of metabolic pathways by combinatorial transcriptional engineering. *Nucleic Acids Res.* 40(18), e142 (2012).

71 Wang M, Si T, Zhao H. Biocatalyst development by directed evolution. *Bioresour. Technol.* 115, 117–125 (2012).

72 Zhang JH, Dawes G, Stemmer WP. Directed evolution of a fucosidase from a galactosidase by DNA shuffling and screening. *Proc. Natl Acad. Sci. USA* 94(9), 4504–4509 (1997).

73 Wong TS, Tee KL, Hauer B, Schwaneberg U. Sequence saturation mutagenesis (SeSaM): a novel method for directed evolution. *Nucleic Acids Res.* 32(3), e26 (2004).

74 Zhao H, Giver L, Shao Z, Affholter JA, Arnold FH. Molecular evolution by staggered extension process (StEP) *in vitro* recombination. *Nat. Biotechnol.* 16(3), 258–261 (1998).

75 Ostermeier M, Shim JH, Benkovic SJ. A combinatorial approach to hybrid enzymes independent of DNA homology. *Nat. Biotechnol.* 17(12), 1205–1209 (1999).

76 Tsuji T, Onimaru M, Yanagawa H. Random multi-recombinant PCR for the construction of combinatorial protein libraries. *Nucleic Acids Res.* 29(20), E97 (2001).

77 Bittker JA, Le BV, Liu JM, Liu DR. Directed evolution of protein enzymes using nonhomologous random recombination. *Proc. Natl Acad. Sci. USA* 101(18), 7011–7016 (2004).

78 Romanini DW, Peralta-Yahya P, Mondol V, Cornish VW. A heritable recombination system for synthetic Darwinian evolution in yeast. *ACS Synth. Biol.* 1, 602–609 (2012).

79 Esvelt KM, Carlson JC, Liu DR. A system for the continuous directed evolution of biomolecules. *Nature* 472(7344), 499–503 (2011).

80 Fernandez-Alvaro E, Snajdrova R, Jochens H, Davids T, Bottcher D, Bornscheuer UT. A combination of *in vivo* selection and cell sorting for the identification of enantioselective biocatalysts. *Angew. Chem. Int. Ed. Engl.* 50(37), 8584–8587 (2011).

81 Becker S, Hobenreich H, Vogel A *et al.* Single-cell high-throughput screening to identify enantioselective hydrolytic enzymes. *Angew. Chem. Int. Ed. Engl.* 47(27), 5085–5088 (2008).

82 Agresti JJ, Antipov E, Abate AR *et al.* Ultrahigh-throughput screening in drop-based microfluidics for directed evolution. *Proc. Natl Acad. Sci. USA* 107(9), 4004–4009 (2010).

83 Fallah-Araghi A, Baret JC, Ryckelynck M, Griffiths AD. A completely *in vitro* ultrahigh-throughput droplet-based microfluidic screening system for protein engineering and directed evolution. *Lab Chip* 12(5), 882–891 (2012).

84 Voigt CA, Martinez C, Wang ZG, Mayo SL, Arnold FH. Protein building blocks preserved by recombination. *Nat. Struct. Biol.* 9(7), 553–558 (2002).

85 Meyer MM, Hochrein L, Arnold FH. Structure-guided SCHEMA recombination of distantly related beta-lactamases. *Protein Eng. Des. Sel.* 19(12), 563–570 (2006).

86 Heinzelman P, Snow CD, Smith MA *et al.* SCHEMA recombination of a fungal cellulase uncovers a single mutation that contributes markedly to stability. *J. Biol. Chem.* 284(39), 26229–26233 (2009).

87 Heinzelman P, Komor R, Kanaan A *et al.* Efficient screening of fungal cellobiohydrolase class I enzymes for thermostabilizing sequence blocks by SCHEMA structure-guided recombination. *Protein Eng. Des. Sel.* 23(11), 871–880 (2010).

88 Fox RJ, Davis SC, Mundorff EC *et al.* Improving catalytic function by ProSAR-driven enzyme evolution. *Nat. Biotechnol.* 25(3), 338–344 (2007).

89 Chockalingam K, Chen Z, Katzenellenbogen JA, Zhao H. Directed evolution of specific receptor-ligand pairs for use in the creation of gene switches. *Proc. Natl Acad. Sci. USA* 102(16), 5691–5696 (2005).

90 Reetz MT, Carballeira JD, Vogel A. Iterative saturation mutagenesis on the basis of B factors as a strategy for increasing protein thermostability. *Angew. Chem. Int. Ed. Engl.* 45(46), 7745–7751 (2006).

91 Hoffmann G, Bonsch K, Greiner-Stoffele T, Ballschmiter M. Changing the substrate specificity of P450cam towards diphenylmethane by semi-rational enzyme engineering. *Protein Eng. Des. Sel.* 24(5), 439–446 (2011).

92 Whitehead TA, Chevalier A, Song Y *et al.* Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nat. Biotechnol.* 30(6), 543–548 (2012).

93 Degrado WF, Summa CM, Pavone V, Nastri F, Lombardi A. *De novo* design and structural characterization of proteins and metalloproteins. *Annu. Rev. Biochem.* 68, 779–819 (1999).

94 Jiang L, Althoff EA, Clemente FR *et al. De novo* computational design of retro-aldol enzymes. *Science* 319(5868), 1387–1391 (2008).

95 Siegel JB, Zanghellini A, Lovick HM *et al.* Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. *Science* 329(5989), 309–313 (2010).

96 Richter F, Leaver-Fay A, Khare SD, Bjelic S, Baker D. *De novo* enzyme design using Rosetta3. *PLoS ONE* 6(5), e19230 (2011).

97 Der BS, Machius M, Miley MJ, Mills JL, Szyperski T, Kuhlman B. Metal-mediated affinity and orientation specificity in a computationally designed protein homodimer. *J. Am. Chem. Soc.* 134(1), 375–385 (2012).

98 Der BS, Edwards DR, Kuhlman B. Catalysis by a *de novo* zinc-mediated protein interface: implications for natural enzyme evolution and rational enzyme engineering. *Biochemistry* 51(18), 3933–3940 (2012).

99 Hansen KB, Hsiao Y, Xu F *et al.* Highly efficient asymmetric synthesis of sitagliptin. *J. Am. Chem. Soc.* 131(25), 8798–8804 (2009).

100 Savile CK, Janey JM, Mundorff EC *et al.* Biocatalytic asymmetric synthesis of chiral amines from ketones applied to sitagliptin manufacture. *Science* 329(5989), 305–309 (2010).

101 Desai AA. Sitagliptin manufacture: a compelling tale of green chemistry, process intensification, and industrial asymmetric catalysis. *Angew. Chem. Int. Ed. Engl.* 50(9), 1974–1976 (2011).

▪ **Directly compares the biocatalytic and chemocatalytic routes in industrial sitagliptin synthesis.**

102 Muller M. Chemoenzymatic synthesis of building blocks for statin side chains. *Angew. Chem. Int. Ed. Engl.* 44(3), 362–365 (2005).

103 Desantis G, Wong K, Farwell B *et al.* Creation of a productive, highly enantioselective nitrilase through gene site saturation mutagenesis (GSSM). *J. Am. Chem. Soc.* 125(38), 11476–11477 (2003).

104 Ma SK, Gruber J, Davis SC *et al.* A green-by-design biocatalytic process for atorvastatin intermediate. *Green Chem.* 12, 81–86 (2010).

▪ **Describes the development of multiple biocatalysts for a significant overall improvement in the industrial production process of a key atorvastatin intermediate.**

105 Barbas CF 3rd, Wang YF, Wong CH. Deoxyribose-5-phosphate aldolase as a synthetic catalyst. *J. Am. Chem. Soc.* 112, 2013–2014 (1990).

106 Gijsen HJM, Wong CH. Unprecedented asymmetric aldol reactions with three aldehyde substrates catalyzed by 2-deoxyribose-5-phosphate aldolase. *J. Am. Chem. Soc.* 116, 8422–8423 (1994).

107 Chen L, Dumas PD, Wong CH. Deoxyribose-5-phosphate aldolase as a catalyst in asymmetric aldol condensation. *J. Am. Chem. Soc.* 114, 741–748 (1992).

108 Greenberg WA, Varvak A, Hanson SR *et al.* Development of an efficient, scalable, aldolase-catalyzed process for enantioselective synthesis of statin intermediates. *Proc. Natl Acad. Sci. USA* 101(16), 5788–5793 (2004).

109 Desantis G, Liu J, Clark DP, Heine A, Wilson IA, Wong CH. Structure-based mutagenesis approaches toward expanding the substrate specificity of D-2-deoxyribose-5-phosphate aldolase. *Bioorg. Med. Chem.* 11(1), 43–52 (2003).

110 Jennewein S, Schurmann M, Wolberg M *et al.* Directed evolution of an industrial biocatalyst: 2-deoxy-D-ribose 5-phosphate aldolase. *Biotechnol. J.* 1(5), 537–548 (2006).

111 Chang MC, Keasling JD. Production of isoprenoid pharmaceuticals by engineered microbes. *Nat. Chem. Biol.* 2(12), 674–681 (2006).

112 Ro DK, Paradise EM, Ouellet M *et al.* Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature* 440(7086), 940–943 (2006).

113 Ro DK, Ouellet M, Paradise EM *et al.* Induction of multiple pleiotropic drug resistance genes in yeast engineered to produce an increased level of anti-malarial drug precursor, artemisinic acid. *BMC Biotechnol.* 8, 83 (2008).

114 Westfall PJ, Pitera DJ, Lenihan JR *et al.* Production of amorphadiene in yeast, and its conversion to dihydroartemisinic acid, precursor to the antimalarial agent artemisinin. *Proc. Natl Acad. Sci. USA* 109(3), E111–E118 (2012).

115 Levesque F, Seeberger PH. Continuous-flow synthesis of the anti-malaria drug artemisinin. *Angew. Chem. Int. Ed. Engl.* 51(7), 1706–1709 (2012).

116 Dietrich JA, Yoshikuni Y, Fisher KJ *et al.* A novel semi-biosynthetic route for artemisinin production using engineered substrate-promiscuous P450(BM3). *ACS Chem. Biol.* 4(4), 261–267 (2009).

117 Zhang K, Shafer BM, Demars MD 2nd, Stern HA, Fasan R. Controlled oxidation of remote sp(3) C–H bonds in artemisinin via P450 catalysts with fine-tuned regio- and stereoselectivity. *J. Am. Chem. Soc.* 134(45), 18695–18704 (2012).

118 Ajikumar PK, Xiao WH, Tyo KE *et al.* Isoprenoid pathway optimization for Taxol precursor overproduction in *Escherichia coli*. *Science* 330(6000), 70–74 (2010).

▪ **Describes the use of a multivariate-modular approach to balance the metabolic flux and increase taxadiene production in engineered *Escherichia coli* cells.**

119 Liu CC, Mack AV, Tsao ML *et al.* Protein evolution with an expanded genetic code. *Proc. Natl Acad. Sci. USA* 105(46), 17688–17693 (2008).

120 Keasling JD. Manufacturing molecules through metabolic engineering. *Science* 330(6009), 1355–1358 (2010).

121 Khosla C, Keasling JD. Metabolic engineering for drug discovery and development. *Nat. Rev. Drug Discov.* 2(12), 1019–1025 (2003).

122 Jung ST, Lauchli R, Arnold FH. Cytochrome P450: taming a wild type enzyme. *Curr. Opin. Biotechnol.* 22(6), 809–817 (2011).

123 Cesaro-Tadic S, Lagos D, Honegger A *et al.* Turnover-based *in vitro* selection and evolution of biocatalysts from a fully synthetic antibody library. *Nat. Biotechnol.* 21(6), 679–685 (2003).

124 Seelig B, Szostak JW. Selection and evolution of enzymes from a partially randomized non-catalytic scaffold. *Nature* 448(7155), 828–831 (2007).

125 Zanghellini A, Jiang L, Wollacott AM *et al.* New algorithms and an *in silico* benchmark for computational enzyme design. *Protein Sci.* 15(12), 2785–2794 (2006).

126 Hyster TK, Knorr L, Ward TK, Rovis T. Biotinylated Rh(III) complexes in engineered streptavidin for accelerated asymmetric C–H activation. *Science* 338(6106), 500–503 (2012).

127 Ye B, Cramer N. Chiral cyclopentadienyl ligands as stereocontrolling element in asymmeteric C–H functionalization. *Science* 338(6106), 504–506 (2012).

128 Kohler V, Wilson M, Durrenberger M *et al.* Synthetic cascades are enabled by combining biocatalysts with artificial metalloenzymes. *Nat. Chem.* 5, 93–99 (2012).

129 Khersonsky O, Kiss G, Rothlisberger D *et al.* Bridging the gaps in design methodologies by evolutionary optimization of the stability and proficiency of designed Kemp eliminase KE59. *Proc. Natl Acad. Sci. USA* 109(26), 10358–10363 (2012).

130 Koga N, Tatsumi-Koga R, Liu G *et al.* Principles for designing ideal protein structures. *Nature* 491(7423), 222–227 (2012).

131 Walsh CT. *Posttranslational Modification of Proteins: Expanding Nature's Inventory.* Roberts, CO, USA, 576 (2005).

132 Yukl ET, Wilmot CM. Cofactor biosynthesis through protein post-translational modification. *Curr. Opin. Chem. Biol.* 16(1–2), 54–59 (2012).

133 Silverman RB. *Organic Chemistry of Enzyme-Catalyzed Reactions.* Academic Press, London, UK, 800 (2002).

134 Lockless SW, Muir TW. Traceless protein splicing utilizing evolved split inteins. *Proc. Natl Acad. Sci. USA* 106(27), 10999–11004 (2009).

135 Muralidharan V, Muir TW. Protein ligation: an enabling technology for the biophysical analysis of proteins. *Nat. Methods* 3(6), 429–438 (2006).

136 Vila-Perello M, Muir TW. Biological applications of protein splicing. *Cell* 143(2), 191–200 (2010).

137 Liu CC, Schultz PG. Adding new chemistries to the genetic code. *Annu. Rev. Biochem.* 79, 413–444 (2010).

138 Wang L, Xie J, Schultz PG. Expanding the genetic code. *Annu. Rev. Biophys. Biomol. Struct.* 35, 225–249 (2006).

139 Liu CC, Choe H, Farzan M, Smider VV, Schultz PG. Mutagenesis and evolution of sulfated antibodies using an expanded genetic code. *Biochemistry* 48(37), 8891–8898 (2009).

140 Brustad EM, Arnold FH. Optimizing non-natural protein function with directed evolution. *Curr. Opin. Chem. Biol.* 15(2), 201–210 (2011).

» Websites

201 Roche 454™ sequencing. www.454.com/index.asp

202 Illumina Solexa™ sequencing. www.illumina.com/systems/sequencing.ilmn

203 Applied Biosystems SOLiD™ ssequencing. www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing/next-generation-systems.html

204 Nanopore DNA sequencing. www.helicosbio.com/Products/HelicosregGeneticAnalysisSystem/HeliScopetradeSequencer/tabid/87/Default.aspx