Published in final edited form as:
Biochemistry. 2011 June 21; 50(24): 5404–5418. doi:10.1021/bi200416g.

Constructing *de novo* biosynthetic pathways for chemical synthesis inside living cells[†]

Amy M. Weeks[||] and **Michelle C. Y. Chang**[*,||,§]

Michelle C. Y. Chang: mcchang@berkeley.edu

[||] Department of Chemistry, University of California, Berkeley, Berkeley California 94720-1460

[§] Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley California 94720-1460

Abstract

Living organisms have evolved a vast array of catalytic functions that make them ideally suited for the production of medicinally and industrially relevant small-molecule targets. Indeed, native metabolic pathways in microbial hosts have long been exploited and optimized for the scalable production of both fine and commodity chemicals. Our increasing capacity for DNA sequencing and synthesis has revealed the molecular basis for the biosynthesis of a variety of complex and useful metabolites and enables the *de novo* construction of novel metabolic pathways for the production of new and exotic molecular targets in genetically tractable microbes. However, the development of commercially viable processes for these engineered pathways is currently limited by our ability to quickly identify or engineer enzymes with the correct reaction and substrate selectivity as well as the speed by which metabolic bottlenecks can be determined and corrected. Efforts in understanding the relationship between sequence, structure, and function in the basic biochemical sciences can advance these goals for synthetic biology applications while also serving as an experimental platform to elucidate the *in vivo* specificity and function of enzymes and to reconstitute complex biochemical traits for study in a living model organism. Furthermore, the continuing discovery of natural mechanisms for the regulation of metabolic pathways has revealed new principles for the design of high-flux pathways with minimized metabolic burden and has inspired the development of new tools and approaches to engineer synthetic pathways in microbial hosts for chemical production.

Living systems have discovered diverse solutions to fundamental problems in chemical catalysis that have the potential to transform society if they could be tapped for synthetic chemistry. For example, the ability of autotrophs to fix and activate carbon dioxide from the atmosphere for use as a universal C_1 building block in biosynthesis has been a longstanding objective for human chemists and could find great utility in the industrial-scale production of commodity chemicals (1–3). With regard to production of complex bioactive compounds, the evolution of enzymes to regio- and stereoselectively utilize molecular oxygen to modify and functionalize complex hydrocarbon skeletons leads to extremely efficient and modular syntheses of entire families of drug-like structures with lower step counts (4–8). The synthetic capacity of organisms has long been adapted for the industrial production of commodity and fine chemicals that can be made in their native hosts at high yield and low

[†] Research in our laboratory is funded by UC Berkeley, the Chevron Career Development Chair, Camille and Henry Dreyfus Foundation, Arnold and Mabel Beckman Foundation, Agilent Foundation, Dow Sustainable Products and Solutions Program, Hellman Foundation, National Science Foundation, and Energy Biosciences Institute. A.M.W. also acknowledges the support of a National Institutes of Health NRSA Training Grant (1 T32 GMO66698) and National Science Foundation Graduate Research Fellowship.

[*] To whom correspondence should be addressed. mcchang@berkeley.edu. Telephone: (510) 642-8545. Fax: (510) 642-9863.

cost (9–12); however, the full combinatorial potential of cellular metabolism for designing new synthetic routes to novel targets has yet to be fully realized (13–16). With advances in DNA sequencing (17–19) and the resulting explosion in sequence information, we have collected a vast array of possible genetic components from which to assemble and construct pathways for *de novo* reaction sequences. In addition, our growing proficiency in large-scale DNA synthesis (20) and sequence manipulation (21) is beginning to provide the necessary tools to modify the chemical programming of cells at a genome level that would allow use of living cells for synthetic applications in medicine, alternative energy, and materials science.

Despite the enormous promise that synthetic biology offers for building new chemical function at an organism level, the development of technical tools to achieve these goals has outpaced our understanding of how chemistry works inside the cell and thus the fundamental design principles for the construction of new pathways. In contrast to traditional synthetic approaches, organismal chemistry must take place in the presence of the thousands of other chemical processes that occur simultaneously within the cell to maintain life. Naturally occurring pathways take advantage of the evolutionary optimization of connections between enzyme partners mediated by protein-protein interactions (22, 23), subcellular localization (24, 25), or complex homeostatic and regulatory networks to channel intermediates to product. In contrast, engineered pathways are built from individual components that have been extracted based on their native functions and reconstituted out of context within a new pathway or host and may produce metabolites and end products that are foreign to the cell. Despite these challenges, several chimeric pathways have been successfully constructed in tractable genetic hosts, such as *Escherichia coli* (26–29) and *Saccharomyces cerevisiae* (30), which are robust enough for the design of scalable industrial processes for commercial production.

These examples highlight the potential impact of synthetic pathway construction on the production of small molecule targets; however, each example has required significant resources to develop from the gene discovery stage and/or identification of key pathway bottlenecks to a genetically optimized strain that produces commercially viable product titers (Figure 1). One of the most significant roadblocks in this process is the initial pathway design and selection or engineering of enzymes capable of supporting sufficient flux through the desired synthetic route. The bottlenecks in a synthetic pathway can arise from many sources, including problems in protein folding, co-factor availability or assembly, the absence of protein partners and pathways that might be required for catalysis, the presence of unexpected promiscuous activities, and crosstalk of non-native metabolic intermediates with other pathways in the cell. A second obstacle is the elucidation of the factors that govern intrapathway flux and kinetic behavior that maintains carbon within the pathway rather than allowing it to be lost or dissipated to other pathways within the cell through metabolic crosstalk. This review will address biochemical developments related to these questions and goals and their application to synthetic pathway design.

IDENTIFICATION OF USEFUL CHEMICAL TRANSFORMATIONS

Many interesting and useful chemical phenotypes are apparent at the organism level; among the more industrially relevant transformations are multistep processes like the degradation of cellulose and lignin, fermentation of sugars to produce sustainable fuels, detoxification of environmental pollutants, and production of complex bioactive compounds (Figure 2). In addition, various specialized structural motifs can be found in natural products, such as strained ring structures (31–36) and unusual functional groups (37–39) or bond couplings (40, 41) that could be used as useful handles for chemical synthesis (Figure 3). However, these attributes need to be distilled into a minimal set of genes, including chaperones, activators, and upstream and downstream partners, required for robust enzymatic activity in a heterologous host in order to transplant the reaction of interest into a new host. Despite the wealth of genome information, it remains very difficult to pinpoint the genetic basis for these traits without biochemical insight to narrow the scope of possible candidates. Thus, the most successful gene identification approaches rely on work- or information-intensive routes such as protein purification and sequencing, comparative genomic studies, functional gene clustering, or analysis of gene expression patterns to successfully identify candidates of interest. Consequently, we need to further develop methods to exploit the wealth of sequence information to make predictions about the reactivity and specificity of new enzymes in order to expand our ability to more rapidly design new or enhanced pathways *in silico*. In this regard, recent advances in functional gene annotation have sought to improve our understanding of the evolutionary relationships between exotic enzymes and their well-studied counterparts and to leverage that understanding for assignment of activities to uncharacterized enzymes and for engineering new enzymatic activities. In addition to sequence-based analysis, physical and computational methods for substrate and/or transition state docking have aided in the identification of specific substrates by restricting the possibilities to a limited set of metabolites. These initial studies show promise for future applications if they can be generalized for the non-specialist user or used for high-throughput and accurate genome annotation so that the production of new small-molecule targets can be approached without requiring the commitment of expensive targeted gene discovery efforts.

The input and curation of reliable and specific gene annotations within the sequence database is essential to expanding the scope of synthetic pathway design, but inaccuracy and imprecision in functional assignments remains a major roadblock given the stringent requirements for enzyme behavior for synthetic biology purposes. For example, bottlenecks in the production of the target molecule (29, 42, 43) can result from the assignment of a minor catalytic activity to the enzyme of interest (44, 45) or non-specific and promiscuous activities that may cause carbon to exit the synthetic pathway. In addition to having reaction selectivity, synthetic pathway components also must be able to be expressed in a heterologous host with high productivity to prevent bottlenecks resulting from low functional expression. Beyond application to chemical industry, metabolic engineering studies also allow us to step beyond *in vitro* validation of gene annotations to begin testing enzyme function *in vivo* when placed within the context of a synthetic pathway.

Targeted gene identification

The discovery of new activities or genes responsible for the biosynthesis of a desired target has been most rapid in bacteria because their genomes are rich in operons and gene clusters that place functionally related genes into close physical proximity with one another (46). Indeed, the number of genes known to encode the production and tailoring of common microbial natural products such as polyketides, non-ribosomal peptides, and ribosomally synthesized peptides has dramatically increased since the development of high throughput sequencing techniques for genome and metagenome sequencing (47, 48). If the natural product of interest is structurally characterized and conditions are known for its production in the native host, the identification of any gene involved its biosynthesis allows the location of unknown classes of enzymes responsible for installation of the particular structural motif of interest by limiting the possibilities to those found within the gene cluster (49, 50). These approaches have identified ununusal and synthetically useful transformations such as aliphatic radical halogenation (49), cyclopropanation (51), enediyne formation (52), new sugar tailoring (53), nitro group installation (54), phosphonate biosynthesis (55), and others. Although these enzymes are typically highly specific for a particular biosynthetic intermediate, mechanistic studies aid in approaching the design of activities of interest from a related class of enzymes (56) or engineering the active site to accept new substrates (57–59). In some cases, mining of genome sequences can pinpoint new enzyme orthologs that naturally catalyze reactions with the desired specificity. For example, Höhne and coworkers predicted mutations that would reverse the enantioselectivity of the known (*S*)-specific transaminases and biochemically validated 17 previously undiscovered naturally occurring (*R*)-selective transaminases that were identified from protein databases by the presence of these mutations (60).

The proliferation of completed microbial genome sequences also allows the use of comparative genomics for gene discovery and has uncovered new patterns of analysis. The recent identification of a fatty aldehyde decarbonylase from the cyanobacteria, *Synechococcus* sp. PCC7002, relied on substractive genome analysis between alkane producing and non-producing cyanobacteria to limit the number of likely candidates (61). In fact, only a single hypothetical protein was left after analysis and its activity was validated *in vitro* and *in vivo* for the production of alkanes within the biodiesel range. Interestingly, this decarbonylase represents a different type of active site than those of the decarbonylases found in algae (62) or plants (63) and was clustered with a partner acyl-coenzyme A reductase. In addition to these unusual classes of transformations, ubiquitous enzymes, such as aldolases or dehydrogenases, can also be very useful for the construction of pathways for new targets and genomic context can help to identify their specific substrates. Even without a clear pathway defined for a particular gene of interest, functional correlation between genes that are consistently proximal to each other can be made if the pattern is found across many genomes (64–66). While this method has been used most widely in bacteria, recent evidence suggests that gene order conservation in eukaryotes also has implications for biological function (67–69). Genomic context mining has also been expanded to correlate evolutionary loss or gain of genes with their biological functions through the construction of phylogenetic profiles (70, 71), which correlate the presence or absence of a gene with the presence or absence of other genes across genomes, allowing a functional correlation to be drawn between genes that are lost and gained together.

Although microbial genomes have served an important function in the discovery of new biosynthetic genes, plants remain a large source for much of the diversity of bioactive small molecules at this time (72). For example, taxol, made by the Pacific Yew, is utilized as an important therapy for the treatment of human cancers; however, it is highly tailored and requires and estimated nineteen committed steps after formation of geranylgeranyl diphosphate (73). Despite its importance and the potential for lowering its cost through synthetic biology approaches utilized for other molecules of this family (74, 75), the identification of the necessary genes for engineering its production is a significant challenge and only nine genes have been found to date (74). Similar to taxol and other members of the isoprenoid class, entire families of medicinally relevant alkaloids remain cryptic with respect to their biosynthetic machinery (76, 77). Although the initiation of plant genome sequencing projects has accelerated, most studies rely on expressed sequence tag[1] (EST) libraries or RNA sequencing for biosynthetic gene identification (78–82). Part of the difficulty lies in the fact that tailoring enzymes are key for natural product biosynthesis but are widely represented across plant genomes. For instance, cytochrome P450s play an essential role in structural tailoring, but even a relatively biosynthetically silent organism, such as *Arabidopsis thaliana*, contains 272 of these enzymes within its genome. Functional genomic approaches such as phylogenetic and metabolite profiling and analysis of spatiotemporal patterns of gene expression and small molecule production hold great promise for addressing this problem (83–88) Thus, these methods are necessary for connecting known pathways of precursor production to important semisynthetic intermediates that may expand our synthetic capabilities such as in the production of artemisinin, an antimalarial isoprenoid (30).

Integrating sequence- and structure-based prediction of enzyme function

Given the almost overwhelming amount of sequence information now available, parallel work in functional gene annotation of biochemically uncharacterized proteins based on bioinformatics and computational approaches (Figure 4AB) can also contribute to providing new components for pathway design. The most common approaches operate at the sequence level and assume functional inheritance through sequence similarity (89) (Figure 4A). An obvious problem with this approach is that homologs arise through divergent evolution and may have developed different functions based on selective pressure. Because genes evolve at different rates depending on selection pressure, the species in which they are found (90), and the family to which they belong (91), it is difficult to establish a threshold for percent sequence identity above which functional annotation can be accurately transferred. Some studies have estimated that a minimum of 40% pairwise identity is required to transfer the first three digits of the Enzyme Commission (EC) number (representing the type of reaction, type of bond acted upon, and type of group acted upon) and that a minimum of 60% identity is required to transfer all four digits (which contain additional details about the substrate) (92, 93).

The relative inaccuracy of functional assignment based on pairwise identity alone has motivated the development of methods that incorporate position-specific information (94), manually curated family information (95), and structural data (96) to increase the precision of functional predictions. While the availability of structural information for a protein of unknown function increases the chances of successful function prediction, overall protein fold comparison is limited by the high structural similarity between divergent superfamily members and by the existence of ‘superfolds’ (97), which may span several superfamilies that are dissimilar on both the sequence and functional levels. However, if a range of activities are known within a superfamily, hidden Markov models constructed from structure-based alignments can help to pinpoint the precise reaction class and substrate class of an uncharacterized protein for mechanistically characterized superfamilies (98). To evaluate structures’ functional potential at a higher level of detail, methods have been developed to assess the catalytic potential of residues near the active site by scoring co-location of potentially catalytic groups (99), while others map sequences onto known 3D structures (100, 101) with manually annotated catalytic residues (102) for comparison or calculate protein electrostatic surfaces to predict interaction motifs for ligands or other proteins (103). Utilization of 3D structural models can also accelerate the directed evolution of new protein function by increasing the information content involved in library construction, thereby maximizing the number of positive hits from smaller library sizes (104–107).

Physical and computational methods can also be utilized to pinpoint the identity of a substrate for structurally characterized enzyme families by virtual screening of small molecule libraries, which can be computationally docked into the active site and scored based on the energetic favorability of their binding (Figure 4B). Adapted from a widely used method in medicinal chemistry for prediction of inhibitors for particular molecular targets (reviewed in (108)), the use of computational docking in identifying enzyme substrates has been validated in retrospective studies on several different superfamilies (109–111). *In silico* screening of a collection of ground-state metabolites has also predicted substrates for members of the enolase superfamily in studies that were validated in parallel by *in vitro* screening of the same set of possible substrates (110, 112). For enzymes that undergo large conformational changes upon ligand binding, these ground-state metabolite docking methods can adapted for use with homology models generated using ligand-bound structures of homologs as templates even in the absence of an experimentally determined crystal structure (110).

Docking methods can also take advantage of the fact that many enzyme active sites show preferential binding to the transition state of the reactions that they catalyze rather than the ground-state substrate. In exploring the substrate specificity of cryptic members of the amidohydrolase superfamily, Hermann and coworkers reduced the scale of their docking study from the complete set of metabolites found in the KEGG database to only those that would provide reactive groups for the range of reactions catalyzed by this superfamily. The members of this reduced library of metabolites were then converted to their respective high-energy intermediate forms for docking studies, which performed as well or better than analogous studies using ground-state metabolites (109). The high-energy intermediate approach was subsequently used to predict the function of an amidohydrolase superfamily member with a novel *S*-inosylhomocysteine deaminase activity using a docking screen of over 4,000 potential substrates (113). Many other studies have used docking of high-energy intermediates to assist in functional annotation of amidohydrolase superfamily members (114–116) and the method holds promise for other superfamilies and the possible annotation of new enzymes and metabolic pathways.

High-throughput experimental approaches for gene discovery and functional annotation

Although these methods provide insight into the reactivity of characterized families of enzymes, there are also many open reading frames for which homology cannot be used to infer function. The availability of experimental information derived from high-throughput studies presents an opportunity to overcome these obstacles by facilitating annotation based on biological rather than biochemical function (Figure 4C). High-throughput protein-protein interaction screens (117–120) have provided vast amounts of data on the physical association of proteins, allowing enzymes to be grouped in terms of their interaction partners. Accordingly, several databases have been developed to catalog this information (121–123). Microarray and RNA sequencing datasets have likewise allowed the construction of co-expression profiles (124–127), which examine gene expression across various conditions or species, allowing inference of functional linkage between genes whose expression is co-regulated.

In several model organisms, it is also possible to use high-throughput approaches to examine genetic interaction by comparing the fitness or phenotypes of single mutants in the presence of an additional perturbation such as a deletion, overexpression, treatment with chemical inhibitors, or gene knockdown by RNA interference (128). The resulting information can then be used to construct a genetic interaction profile for the gene of interest, providing information about biological function. These methods are best developed for *Saccharomyces cerevisiae*, for which many quantitative single-gene genetic interaction maps (129–133) as well as a genome-scale genetic interaction map (134) have been constructed. While most high-throughput genetic interaction mapping has been done in model organisms, the method has recently been expanded to mammalian cell culture (135, 136) and has great potential for use in other types of systems.

The increasing availability of knockout collections provides a useful tool for evaluating the output of high-throughput functional genomic screens. Deletion strains for genes identified through high-throughput screening can be evaluated for the phenotype of interest. In one recent example, microarray analysis (137) combined with targeted knockout collection screening identified a key cellodextrin transporter in *Neurospora crassa*, which was then characterized and used to engineer cellodextrin transport in *S. cerevisiae* for improved ethanol production (138). One can also take advantage of knockout collections to run highly parallel screens to monitor physiological traits. The construction of knockout libraries in a number of yeast and bacterial species containing molecular ‘barcode’ identifiers has facilitated pooled screening in which fitness can be assessed by the use of oligonucleotide microarrays complementary to the barcodes (139). The benefit of these approaches is that uncharacterized enzyme families or new linkages across pathways can be revealed in this process.

ENGINEERING NEW OR ALTERED ENZYME FUNCTION

One of the major challenges for synthetic biology is not only to find enzymes that catalyze the reaction of interest with high specificity, but also to identify enzymes with expression characteristics that are amenable for synthetic pathway construction because *in vivo* productivity is a function of both enzyme concentration and rate of turnover. As the functional expression of an enzyme relies on complex pathways for protein synthesis, folding, and maturation, it can often be difficult to overcome the bottleneck of protein solubility. Thus, advances in protein engineering and design can help to alter the characteristics of a specific enzyme by utilizing a well-behaved scaffold to engineer new activity or change substrate specificity in ways that allow the particular enzyme component to meet the metabolic demands of the pathway of interest. If the reaction of interest can be made essential for cell survival, then a genetic selection can also be used in order to search through greater library diversity but most enzymes typically require more intensive screening methods to find sufficiently robust hits for synthetic biology applications. In these cases, protein engineering efforts are typically more successful when the mechanism or source of substrate specificity is well understood, which allows the design of focused libraries that reduce the number of members to be screened or aid in training design algorithms (104–106, 140, 141). Growing insight into physiological mechanisms of evolution can also help to inform the design of laboratory-based evolution while allowing us to explore basic mechanisms of evolution and the contribution that neutral drift and promiscuous function may play in generation of new activities.

In vitro evolution of new and altered enzyme characteristics

While many of the methods described above are useful for isolating the genes responsible for a particular biological activity, in some cases an enzyme is desired to catalyze a chemical transformation not known to occur in nature. *In vitro* laboratory evolution, using both combinatorial and rational approaches, has been successful at producing new enzyme activities that are useful for synthetic biology applications (59, 142–144). In some cases, the evolution of new activities allows an alternative route to a biosynthetic intermediate (145). In other cases, the improvement in an enzyme that serves as a bottleneck in the overall pathway can lead to large amplifications in yield (146, 147). Furthermore, there are times when simple changes, such as the conversion of an NADPH-dependent enzyme to an NADH-dependent one, can have large implications in overall cellular processes like that observed for C_5 sugar assimilation in the production of biofuels (148–151). As the number of protein sequences increases and their annotation improves, bioinformatics approaches can be used for protein design to focus on residues of interest (152–154) or even for prediction of precise amino acid changes that would alter specificity.

The mechanistic diversity of certain enzyme superfamilies can be utilized for the design of new reactivities from a single scaffold. Mechanistically diverse superfamiles are often considered to be particularly ‘evolvable’ because they share mechanistic features and a common fold while catalyzing a wide variety of different reactions (reviewed in (155)). The enolase superfamily was the first to be systematically explored in this regard and members were found to share a common Mg^{2+}-stabilized enolate intermediate formed upon abstraction of the a-proton of a carboxylic acid (156). From this intermediate, a wide range of reaction paths ensue depending on the details of the active site (157) (Figure 3A). This understanding has been used to rationally design enolase superfamily members to catalyze new reactions and can also be used to identify markers of function on a sequence level (158). In other mechanistically diverse superfamilies, such as the Asp/Glu racemase superfamily and the enoyl-CoA hydratase superfamily, there are also several examples in which an understanding of mechanism has facilitated engineering of large changes in activity based on a small number of amino acid changes, which may indicate the generality of this approach (159).

Enzyme promiscuity and neutral drift

While many enzymes have evolved as proficient catalysts with high efficiency and specificity, many others have been found that adventitiously catalyze secondary reactions (160–164). The efficiency of catalysis for these reactions is usually quite low, but can represent a rate acceleration of many orders of magnitude above the uncatalyzed reaction. Mutations conferring this catalytic promiscuity are believed to accumulate as a result of neutral drift (165, 166), in which enzymes evolve under a selective pressure to maintain their original activities (Figure 5A). Enzyme promiscuity has been proposed to play an important role in the evolution of new enzyme functions, requiring that when a selective pressure arises that makes a secondary activity beneficial, it can be improved with just a few mutations to enhance the selective advantages (Figure 5B). This plasticity is believed to account for the apparent rapid evolution of enzymes that proficiently degrade anthropogenic toxins that only appeared in the 20th century, including tetrachlorohydroquinone dehalogenase (167), which is proposed to have evolved from maleylacetoacetate isomerase (Figure 5C), and phosphotriesterase (164, 168), which is evolutionarily related to the phophotriesterase-like lactonases (Figure 5D). Recent experiments have shown that beyond providing a platform for evolution of single new activities, promiscuous reactions can also result in the appearance of serendipitous metabolic pathways (169). Because of their high evolutionary potential, promiscuous enzymes are an excellent starting point for enzyme redesign. Indeed, some enzyme folds are thought to be more permissive and flexible towards evolving new functions (155, 170–172)

Terpene synthases and oxidosqualene cyclases have been extensively explored with respect to their catalytic promiscuity because of the relatively small number of enzymes known compared to the quantity of high value products that they produce. Furthermore, their product distribution can also quantitatively report on the alternative reaction pathways taken by the carbocation intermediates as a result of rearrangements, quenching with water, or deprotonation at the incorrect position (173). The primary determinants of reaction outcome are thought to be the shape of the active site and the placement of reactive functional groups; accordingly, mutation to introduce or remove a hydroxyl group can lead to large changes in product distribution. For example, mutation of Ile to Thr in the *ent*-kaurene synthases converts them to pimaradiene synthases (174), while an Ala to Ser mutation in abietadiene synthase produces a pimaradiene cyclase (175). Similary, removal of a hydroxyl group through a Thr to Ile mutation in *syn*-pimara-7,15-diene synthase produces an aphidicolene-specific synthase (176). The latter case is notable because rather than short-circuiting the natural reaction to produce a simpler diterpene, the mutation increases reaction complexity and confers a specificity not observed in nature. Steric bulk at the active site also plays a role in determining the reaction path. In cycloartenol synthase, a decrease in bulk at the active site through a Tyr to Thr mutation produced a lanosterol synthase (177). Because all terpene synthases share a conserved fold, it has also been possible to reciprocally interconvert the activities of pairs of enzymes using analogous mutations (178, 179). Fold conservation also facilitated the identification of key plasticity residues at the active site of γ-humulene synthase, allowing the ‘designed evolution’ of seven distinct terpene synthases from a small saturation mutagenesis library (154). Taken together, this work demonstrates the potential for application of our understanding of the origins of catalytic promiscuity for the synthesis of new and potentially bioactive molecules.

OPTIMIZING FLUX THROUGH SYNTHETIC METABOLIC PATHWAYS

The biosynthesis of complex molecules through synthetic metabolic pathways involves the expression or overexpression of several enzymatic components, often from a variety of different organisms. Because these genes are expressed outside their natural contexts and reconnected in unique ways, native regulatory mechanisms are often missing or compromised, which can decrease target molecule output. Without regulatory mechanisms to control expression of the pathway and flux, the host organism may exhibit slow growth owing to the metabolic burden or toxicity of protein overexpression (180), the depletion of host-derived precursors to such an extent that its needs for growth are not met (181–184), or the accumulation of toxic intermediates (185). Metabolic flux may also be compromised by bottlenecks caused by enzyme activity levels that are not commensurate with the activity of more efficient enzymes in the pathway or by the expression of the subunits of multienzyme complexes at suboptimal stoichiometric ratios. Synthetic metabolic pathways must therefore include engineered regulatory mechanisms to balance protein expression and to direct intermediates down the target pathway to maximize target molecule production (Figure 6).

Identifying and overcoming pathway bottlenecks

Owing to variation in specific activity, reaction equilibrium, and solubility between the enzymes in synthetic pathways, bottlenecks often arise that limit flux and product titers. A major challenge following the construction of pathway interactions is to find these limiting steps and to develop approaches to manage them. In the course of these efforts, new bottlenecks often appear each time previous ones are alleviated. The origin of these bottlenecks does not necessarily result from only poor solubility or expression but can also have a root in the kinetic behavior of a pathway. For example, various bottlenecks have been identified during the development of a high-yielding engineered pathway for the production of an antimalarial drug precursor and were found to arise from kinetic mismatching for the clearance of a toxic intermediate (185–187), loss of diffusible intermediates from the cell, or limitations in rate of an irreversible step to drive the pathway flux in the forward direction (188). Thus, the elucidation of the biochemical basis for the performance of a synthetic pathway and its dependence on individual steps often aids in the development of strategies to increase titers of the target small molecule. In general, improvement of product yields can be achieved by titrating the expression of a particular protein (29, 188) or by *in vitro* evolution of improved kinetic parameters of the enzyme in question (154, 189).

Working toward understanding the underlying physiological source of a particular bottleneck can also lead to changes in pathway design itself and provide insight into the function of native pathways. During the course of engineering *E. coli* for the production of *n*-butanol, a second-generation biofuel, several groups had identified a similar bottleneck in the enoyl-CoA reduction step that is required to produce butyryl-CoA, a key intermediate, from crotonyl-CoA (29, 42, 171). Cellular studies demonstrated that butyryl-CoA is not dissipated through native cellular pathways, suggesting that if an effectively irreversible reduction of crotonyl-CoA could be achieved, it would commit carbon to the synthetic *n*-butanol pathway (29). Replacement of butyryl-CoA dehydrogenase (Bcd), the native clostridial system for crotonyl-CoA reduction, with a member of the more unusual and mechanistically distinct *trans*-enoyl-CoA reductase (Ter) family led to an order of magnitude increase in *n*-butanol titers to 4.7 g L^{-1} (29). Further analysis of different strains led to the conclusion that Ter is capable of acting as a kinetic trap because of its chemical mechanism of direct hydride transfer from NADH. In contrast, the native Bcd enzyme utilizes a flavin cofactor, which decreases the kinetic barrier to the reverse reaction via a more energetically accessible intermediate, allowing butyryl-CoA to revert to crotonyl-CoA, which can be lost to competing cellular pathways (29). The use of *in vivo* product titers to assay enzyme activity in this case allows us to build strains that produce commercially viable levels of a target compound as well as to assess the function of an enzyme inside the cell and to begin identifying the basis for the parallel evolution of mechanististically distinct enzymes that catalyze the same reaction. In addition to experimental approaches for synthetic pathway construction, the continual development of more robust methods for modeling pathways and identifying bottlenecks through computational approaches should greatly facilitate this process (190–192). Advances in rapid multiplexed *in vitro* evolution at the genome level also hold much promise as a combinatorial method to examine relationships between enzymes in a native or engineered pathway and to identify bottlenecks or points of regulatory control (21).

Engineering pathway balance

Adjusting promoter strength has long been a widely used method for controlling and optimizing protein expression levels (Figure 6D). In addition, promoter titrations are often the fastest approach to identifying key bottlenecks. A wide variety of promoters are available that provide a means for regulation of expression levels with a small molecule inducer. Some promoters, like that controlled by arabinose (*araC*-P_{BAD}) (193) and propionate (*prpR*-P_{prpB}) (194), offer both tunable expression in the presence of the inducer and tight control of expression in the absence of inducer and have been developed further for use in synthetic biology. Given that yields depend on individual cellular behavior, these expression systems sometimes require genomic modifications in order to enforce cell-to-cell tunability rather then the formation of mixed sub-populations (195). The use of promoters as the primary means of controlling gene expression remains problematic, however, in that genes that must be expressed at different levels to achieve pathway balance must each be expressed under a different promoter requiring a different inducer.

More recent work has focused on uncoupling transcriptional control of gene expression from translational control, allowing several genes to be encoded in a single operon but to be expressed at different levels (Figure 6A). An early approach utilized the incorporation of RNase E recognition sites between coding regions that would be endonucleolytically cleaved following transcription, generating two independent secondary transcripts (196). The stability of these transcripts can modulated by adding engineered secondary structural elements, allowing control of protein expression through mRNA stabilization or destabilization. This idea was subsequently expanded through the incorporation of a library of tunable intergenic regions (TIGRs) encoding two variable hairpins flanking variable RNase E sites (186). Using this library, relative expression of the two coding regions could be varied over a 100-fold range. Application of TIGRs to the heterologous mevalonate pathway in *E. coli* attenuated the effects of the accumulation of the toxic intermediate 3-hydroxymethylglutaryl-CoA, which limits target molecule production, leading to a seven-fold enhancement in mevalonate output. Interestingly, the mechanisms by which the improvement was achieved were counterintuitive and involved differential mRNA processing, transcription termination, and sequestration of ribosome binding sites (Figure 6B). Ribozymes and RNA aptamers have also been applied to modulate gene expression. RNA aptamers have been engineered that bind to transcriptional repressors, activating gene expression. Combination ribozyme-aptamer (‘aptazyme’) RNAs have been designed in which ligand binding to the aptamer mediates ribozyme self-cleavage, altering mRNA stability (197) or ribosome binding site accessibility (198). These systems show great potential for application in the construction of feedback loops to couple gene expression to changes in the concentration of a small molecule of interest.

In many cases, the metabolic demands of protein expression interfere with cellular requirements for growth, leading to lower yields of the target molecule. Linking gene transcription to the metabolic state of the cell could alleviate this problem by only allowing the transcription and expression of protein when cellular resources become available. One approach utilized a reengineered Ntr regulon to couple glucose availability to transcription by sensing acetyl phosphate (199). Placement of a gene whose expression is normally toxic to *E. coli* under the promoter of the Ntr regulon relieved growth inhibition by repressing protein expression until stationary phase, when glucose became abundant. Incorporation of this system into a synthetic pathway for the production of lycopene increased yields ~20-fold compared to use of standard inducible promoters, demonstrating the utility of dynamic control of protein expression in increasing yields.

As a complement to transcriptional regulation, gene expression can also be tuned at the translational level (Figure 6BC). In bacteria, ribosome binding sites (RBSs) are used to control translation initiation, which in most cases is the rate-limiting step in translation. The creation of libraries of RBSs with different rates of translation initiation has been studied for the optimization of synthetic pathways (200). Because the library size required for optimization increases combinatorially with the number of genes in a pathway, a computational RBS optimization algorithm was developed to predict the translation initiation rate for a particular sequence, decreasing the time and resources required for RBS optimization (201). This model correctly predicted that the rate of translation initiation varies depending on RBS context and was experimentally verified for over 100 genes, demonstrating that RBS sequence can vary gene expression over a range of 100,000-fold.

Maximizing pathway flux through engineered spatial organization

Rather than acting in isolation, natural enzymes often participate in multienzyme complexes, are localized to specific cellular compartments, or are found as fusion proteins in which a single polypeptide catalyzes two or more activities (Figure 7). Based on whole-genome affinity purification-mass spectrometry studies in *Saccharomyces cerevisiae*, it was estimated that there are at least 491 protein complexes in yeast (117, 118). A later study using a protein-fragment complementation assay suggested that at least 1124 yeast proteins participate in 2270 multiprotein complexes (119), demonstrating the ubiquity of relatively stable protein-protein interactions in cellular pathways. Such spatial organization often confers kinetic benefits by increasing local concentration of intermediates, by avoiding loss of reactive intermediates, and by preventing intermediates from entering other pathways. Further, such metabolic channels can help to avoid toxicity of poisonous metabolites by keeping local concentrations high while keeping overall cellular concentrations low. In addition to function of the pathway itself, approaches to increasing the efficiency of metabolic pathways without increasing intracellular enzyme concentration are important as the metabolic burden that protein synthesis places on the cell can be substantial and detrimental to product titers.

Many different mechanisms have evolved for achieving metabolic channeling through spatial organization. Some enzymes, such as tryptophan synthase (202) and carbamoyl phosphate synthetase (203), use physical channels to deliver reactive intermediates from one active site to another within multisubunit complexes (Figure 7A). Other pathways utilize fusion proteins that carry out several reactions, such as synthesis of 5-enolpyruvylshikimate 3-phosphate in the shikimate pathway (204) (Figure 7B). Many examples exist in which spatial organization is achieved through the formation of multienzyme, multiple activity complexes, including the tricarboxylic acid cycle (205), the Calvin cycle (206), glycolysis (207), fatty acid oxidation (208), and protein degradation (209), that allow the cell to rapidly control both flux and selectivity of multistep reaction pathways. For example, the assembly and disassembly of the enzymes of purine biosynthesis, which involves intermediates unstable in the cellular milieu, is regulated by cellular conditions and can be rapidly controlled to turn biosynthesis on or off in low and high purine levels, respectively (22) (Figure 7D). Cellulosomes similarly comprise multisubunit complexes, but complex formation is triggered by scaffold proteins that specify the stoichiometry of each enzyme involved in cellulose breakdown depending on the carbon source available to the organism (23) (Figure 7C). Enzymes can also be compartmentalized by protein organelles, such as the carboxysome structure found in cyanobacteria that traps CO_2 within in order to manage the poor selectivity and kinetic behavior of ribulose-1,6-bisphosphate carboxylase/oxygenase (RuBisCO) in carbon fixation (25).

While multienzyme complexes often mediate metabolic channeling, they also play roles in the regulation of metabolic pathways by using stable and transient protein-protein interactions or enzyme localization to control the output and product distribution of promiscuous enzymes and pathways. In the biosynthesis of dhurrin, for example, the formation of transient, low-affinity complexes both prevents the diffusion of reactive and toxic intermediates and enforces the need for expression of a specific glucosyltransferase despite the promiscuity and ubiquity of these enzymes (210, 211). For enzymes that participate in multiple pathways, such as glyceraldehyde-3-phosphate dehydrogenase of glycolysis and the Calvin cycle, differential regulation mechanisms can be used to control its behavior when it is complexed with different enzyme partners (206). Although the precise organization of complexes involved in plant phenylpropanoid biosynthesis remains to be established, the distribution of the diverse set of possible products, including lignin, sinapate esters, stilbenes, and flavonoids, is coordinated by the localization of several different enzymes in the pathway (212).

Engineered spatial organization has long been considered as a means to enhance productivity of enzymes by mimicking metabolic channeling. Early studies focused on enforcing the proximity of enzymes by physically immobilizing them on polymer beads (213). While modest enhancements in product formation were achieved, enzyme immobilization is problematic in terms of scaling up for industrial applications. More desirable would be to mimic nature by designing self-replicating, genetically encoded systems. The use of fusion proteins for this application was explored in the incipient days of recombinant DNA technology (214, 215). While some fusion proteins were successful, in other cases, fusions led to decreased activity due to interference with the formation of other multiprotein complexes that were required for activity, highlighting the limited scope of this strategy.

Recent efforts to engineer spatial organization have focused on using scaffold proteins to template enzyme colocalization. In an example based on natural cellulosomes, two cellulases were fused to scaffold binding domains known as dockerins, allowing their precise assembly onto the scaffold protein scaffoldin (216). The resulting ‘designer cellulosomes’ showed synergistic behavior that conferred a two-fold enhancement in their ability to degrade cellulose. A second approach utilized scaffolds with interaction domains derived from the metazoan signaling machinery. By tagging the enzymes of the mevalonate pathway with the cognate peptide ligands for the interaction domains, the stoichiometry of the mevalonate-producting enzymes was optimized, increasing product titers by 77-fold (187). The same scaffold strategy was applied to enhance production of glucaric acid from an engineered pathway by 5-fold, demonstrating the system’s modularity (187, 217). By colocalizing pathway proteins, the scaffold decreases the metabolic burden of overexpression, suggesting that it could be useful for difficult to express or poorly soluble proteins. In addition to product yield enhancement, this approach could also be used to control the product distribution of promiscuous enzymes, especially when considering that substrate flexibility can often arise as a result of protein engineering efforts.

CONCLUSIONS

The remarkable diversity of enzyme-catalyzed transformations observed in nature makes biological systems ideal for addressing a wide variety of problems in chemical synthesis. The availability of thousands of genome sequences and improving technologies for the low-cost assembly of large synthetic DNAs means that we are limited by our ability to design rather than to construct *de novo* synthetic pathways that are sufficiently robust to displace existing processes for small molecule production. Beyond applications to low-cost chemical production, synthetic biology also offers an alternative and complementary platform to more traditional reductionist approaches to study and elucidate how enzymes and other biochemical components work inside living cells to produce organism-level phenotypes. In this regard, synthetic pathway construction can also serve as a powerful tool for enzymology by providing an interesting intermediate level of study between *in vitro* studies where we can carefully measure physicochemical properties of enzymes and genetic studies where we can assess and validate physiological function. In comparison to these approaches, product titers from synthetic pathways can be used both as a measure of enzyme activity when filtered through the context of thousands of other highly regulated and potentially opposing chemical reactions within a living cell as well as a genetic phenotype with high dynamic range to score the fitness of individual components or a pathway as a whole. Furthermore, the perturbation to host metabolism caused by introduction of an exogenous metabolic pathway has the potential to uncover the organizational and regulatory principles that control the complex metabolic network of the cell. Thus, the synergy between deconstruction and *de novo* construction can advance both our understanding of the complex behavior of metabolic pathways and networks and our ability to engineer microbes capable of producing commercially viable titers of an expanding repertoire of target molecules.

ABBREVIATIONS

EST	expressed sequence tag
EC	Enzyme Commission
OMPDC	orotidine 5′-monophosphate decarboxylase
NAL	N-acetylneuraminate lyase
TIGR	tunable intergenic region
RBS	ribosome binding site

References

1. Behr A. Carbon-Dioxide as an Alternative C_1 Synthetic Unit - Activation by Transition-Metal Complexes. Angew Chem, Int Ed Engl. 1988; 27:661–678.
2. Sakakura T, Choi JC, Yasuda H. Transformation of carbon dioxide. Chem Rev. 2007; 107:2365–2387. [PubMed: 17564481]
3. Benson EE, Kubiak CP, Sathrum AJ, Smieja JM. Electrocatalytic and homogeneous approaches to conversion of CO2 to liquid fuels. Chem Soc Rev. 2009; 38:89–99. [PubMed: 19088968]
4. Feig AL, Lippard SJ. Reactions of non-heme iron(II) centers with dioxygen in biology and chemistry. Chem Rev. 1994; 94:759–805.
5. Groves JT. High-valent iron in chemical and biological oxidations. J Inorg Biochem. 2006; 100:434–447. [PubMed: 16516297]
6. Que L, Tolman WB. Biologically inspired oxidation catalysis. Nature. 2008; 455:333–340. [PubMed: 18800132]

7. Chang MCY, Eachus RA, Trieu W, Ro DK, Keasling JD. Engineering *Escherichia coli* for production of functionalized terpenoids using plant P450s. Nat Chem Biol. 2007; 3:274–277. [PubMed: 17438551]

8. Chen MS, White MC. A predictably selective aliphatic C-H oxidation reaction for complex molecule synthesis. Science. 2007; 318:783–787. [PubMed: 17975062]

9. Glazer, AN.; Nikaido, H. Microbial biotechnology: Fundamentals of applied microbiology. W. H. Freeman and Co; New York: 1995.

10. Elander RP. Industrial production of β-lactam antibiotics. Appl Microbiol Biotechnol. 2003; 61:385–392. [PubMed: 12679848]

11. Knop DR, Draths KM, Chandran SS, Barker JL, von Daeniken R, Weber W, Frost JW. Hydroaromatic equilibration during biosynthesis of shikimic acid. J Am Chem Soc. 2001; 123:10173–10182. [PubMed: 11603966]

12. Sánchez AM, Bennett GN, San KY. Novel pathway engineering design of the anaerobic central metabolic pathway in *Escherichia coli* to increase succinate yield and productivity. Metab Eng. 2005; 7:229–239. [PubMed: 15885621]

13. Stephanopoulos G, Sinskey AJ. Metabolic engineering - Methodologies and future prospects. Trends Biotechnol. 1993; 11:392–396. [PubMed: 7764086]

14. Rohlin L, Oh MK, Liao JC. Microbial pathway engineering for industrial processes: Evolution, combinatorial biosynthesis and rational design. Curr Opin Microbiol. 2001; 4:330–335. [PubMed: 11378488]

15. Prather KL, Martin CH. *De novo* biosynthetic pathways: Rational design of microbial chemical factories. Curr Opin Biotechnol. 2008; 19:468–474. [PubMed: 18725289]

16. Keasling JD. Synthetic biology for synthetic chemistry. ACS Chem Biol. 2008; 3:64–76. [PubMed: 18205292]

17. Mardis ER. Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet. 2008; 9:387–402. [PubMed: 18576944]

18. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005; 437:376–380. [PubMed: 16056220]

19. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008; 456:53–59. [PubMed: 18987734]

20. Gibson DG, Glass JI, Lartigue C, Noskov VN, Chuang RY, Algire MA, Benders GA, Montague MG, Ma L, Moodie MM, et al. Creation of a bacterial cell controlled by a chemically synthesized genome. Science. 2010; 329:52–56. [PubMed: 20488990]

21. Wang HH, Isaacs FJ, Carr PA, Sun ZZ, Xu G, Forest CR, Church GM. Programming cells by multiplex genome engineering and accelerated evolution. Nature. 2009; 460:894–U133. [PubMed: 19633652]

22. An S, Kumar R, Sheets ED, Benkovic SJ. Reversible compartmentalization of *de novo* purine biosynthetic complexes in living cells. Science. 2008; 320:103–106. [PubMed: 18388293]

23. Shoham Y, Lamed R, Bayer EA. The cellulosome concept as an efficient microbial strategy for the degradation of insoluble polysaccharides. Trends Microbiol. 1999; 7:275–281. [PubMed: 10390637]

24. Tanaka S, Sawaya MR, Yeates TO. Structure and mechanisms of a protein-based organelle in *Escherichia coli*. Science. 2010; 327:81–84. [PubMed: 20044574]

25. Yeates TO, Kerfeld CA, Heinhorst S, Cannon GC, Shively JM. Protein-based organelles in bacteria: Carboxysomes and related microcompartments. Nat Rev Microbiol. 2008; 6:681–691. [PubMed: 18679172]

26. Martin VJ, Pitera DJ, Withers ST, Newman JD, Keasling JD. Engineering a mevalonate pathway in *Escherichia coli* for production of terpenoids. Nat Biotechnol. 2003; 21:796–802. [PubMed: 12778056]

27. Atsumi S, Hanai T, Liao JC. Non-fermentative pathways for synthesis of branched-chain higher alcohols as biofuels. Nature. 2008; 451:86–89. [PubMed: 18172501]

28. Tang X, Tan Y, Zhu H, Zhao K, Shen W. Microbial conversion of glycerol to 1,3-propanediol by an engineered strain of *Escherichia coli*. Appl Environ Microbiol. 2009; 75:1628–1634. [PubMed: 19139229]

29. Bond-Watts B, Bellerose R, Chang M. Enzyme mechanism as a kinetic control element for the design of synthetic biofuel pathways. Nat Chem Biol. 2011; 7:222–227. [PubMed: 21358636]

30. Ro DK, Paradise EM, Ouellet M, Fisher KJ, Newman KL, Ndungu JM, Ho KA, Eachus RA, Ham TS, Kirby J, et al. Production of the antimalarial drug precursor artemisinic acid in engineered yeast. Nature. 2006; 440:940–943. [PubMed: 16612385]

31. Wender PA, Lechleiter JC. Total synthesis of (+/−) isabelin. J Am Chem Soc. 1980; 102:6340–6341.

32. Holton RA, Somoza C, Kim HB, Liang F, Biediger RJ, Boatman PD, Shindo M, Smith CC, Kim SC, Nadizadeh H, et al. First total synthesis of Taxol.1 Functionalization of the B-Ring. J Am Chem Soc. 1994; 116:1597–1598.

33. Furstner A, Weintritt H. Total synthesis of roseophilin. J Am Chem Soc. 1998; 120:2817–2825.

34. Winkler JD, Rouse MB, Greaney MF, Harrison SJ, Jeon YT. The first total synthesis of (+/−)-ingenol. J Am Chem Soc. 2002; 124:9726–9728. [PubMed: 12175229]

35. Richter JM, Ishihara Y, Masuda T, Whitefield BW, Llamas T, Pohjakallio A, Baran PS. Enantiospecific total synthesis of the hapalindoles, fischerindoles, and welwitindolinones via a redox economic approach. J Am Chem Soc. 2008; 130:17938–17954. [PubMed: 19035635]

36. Martin CL, Overman LE, Rohde JM. Total synthesis of (+/−)- and (−)-actinophyllic acid. J Am Chem Soc. 2010; 132:4894–4906. [PubMed: 20218696]

37. O'Hagan D, Schaffrath C, Cobb SL, Hamilton JT, Murphy CD. Biochemistry: Biosynthesis of an organofluorine molecule. Nature. 2002; 416:279. [PubMed: 11907567]

38. Lee J, Simurdiak M, Zhao H. Reconstitution and characterization of aminopyrrolnitrin oxygenase, a Rieske *N*-oxygenase that catalyzes unusual arylamine oxidation. J Biol Chem. 2005; 280:36719–36727. [PubMed: 16150698]

39. Lee J, Zhao HM. Mechanistic studies on the conversion of arylamines into arylnitro compounds by aminopyrrolnitrin oxygenase: Identification of intermediates and kinetic studies. Angew Chem, Int Ed Engl. 2006; 45:622–625. [PubMed: 16342311]

40. Zhang Y, Zhu X, Torelli AT, Lee M, Dzikovski B, Koralewski RM, Wang E, Freed J, Krebs C, Ealick SE, et al. Diphthamide biosynthesis requires an organic radical generated by an iron-sulphur enzyme. Nature. 2010; 465:891–896. [PubMed: 20559380]

41. Zhang Q, Li Y, Chen D, Yu Y, Duan L, Shen B, Liu W. Radical-mediated enzymatic carbon chain fragmentation-recombination. Nat Chem Biol. 2011; 7:154–160. [PubMed: 21240261]

42. Atsumi S, Cann AF, Connor MR, Shen CR, Smith KM, Brynildsen MP, Chou KJ, Hanai T, Liao JC. Metabolic engineering of *Escherichia coli* for 1-butanol production. Metab Eng. 2008; 10:305–311. [PubMed: 17942358]

43. Steen EJ, Kang Y, Bokinsky G, Hu Z, Schirmer A, McClure A, Del Cardayre SB, Keasling JD. Microbial production of fatty-acid-derived fuels and chemicals from plant biomass. Nature. 2010; 463:559–562. [PubMed: 20111002]

44. Wallace KK, Bao ZY, Dai H, Digate R, Schuler G, Speedie MK, Reynolds KA. Purification of crotonyl-CoA reductase from *Streptomyces collinus* and cloning, sequencing and expression of the corresponding gene in Escherichia coli. Eur J Biochem. 1995; 233:954–962. [PubMed: 8521864]

45. Erb TJ, Brecht V, Fuchs G, Muller M, Alber BE. Carboxylation mechanism and stereochemistry of crotonyl-CoA carboxylase/reductase, a carboxylating enoyl-thioester reductase. Proc Natl Acad Sci U S A. 2009; 106:8871–8876. [PubMed: 19458256]

46. Demerec M, Hartman P. Complex loci in microorganisms. Annu Rev Microbiol. 1959; 13:377–406.

47. Banik JJ, Brady SF. Recent application of metagenomic approaches toward the discovery of antimicrobials and other bioactive small molecules. Curr Opin Microbiol. 2010; 13:603–609. [PubMed: 20884282]

48. Corre C, Challis GL. New natural product biosynthetic chemistry discovered by genome mining. Nat Prod Rep. 2009; 26:977–986. [PubMed: 19636446]

49. Vaillancourt FH, Yin J, Walsh CT. SyrB2 in syringomycin E biosynthesis is a nonheme Fe^{II} α-ketoglutarate- and O_2-dependent halogenase. Proc Natl Acad Sci U S A. 2005; 102:10111–10116. [PubMed: 16002467]

50. Balskus EP, Walsh CT. The genetic and molecular basis for sunscreen biosynthesis in cyanobacteria. Science. 2010; 329:1653–1656. [PubMed: 20813918]

51. Vaillancourt FH, Yeh E, Vosburg DA, O'Connor SE, Walsh CT. Cryptic chlorination by a non-haem iron enzyme during cyclopropyl amino acid biosynthesis. Nature. 2005; 436:1191–1194. [PubMed: 16121186]

52. Liu W, Christenson SD, Standage S, Shen B. Biosynthesis of the enediyne antitumor antibiotic C-1027. Science. 2002; 297:1170–1173. [PubMed: 12183628]

53. Thibodeaux CJ, Melancon CE, Liu HW. Unusual sugar biosynthesis and natural product glycodiversification. Nature. 2007; 446:1008–1016. [PubMed: 17460661]

54. Winkler R, Hertweck C. Biosynthesis of nitro compounds. Chembiochem. 2007; 8:973–977. [PubMed: 17477464]

55. Metcalf WW, van der Donk WA. Biosynthesis of phosphonic and phosphinic acid natural products. Annu Rev Biochem. 2009; 78:65–94. [PubMed: 19489722]

56. Lu Y, Yeung N, Sieracki N, Marshall NM. Design of functional metalloproteins. Nature. 2009; 460:855–862. [PubMed: 19675646]

57. Fasan R, Chen MM, Crook NC, Arnold FH. Engineered alkane-hydroxylating cytochrome $P450_{BM3}$ exhibiting nativelike catalytic properties. Angew Chem, Int Ed Engl. 2007; 46:8414–8418. [PubMed: 17886313]

58. Lewis JC, Bastian S, Bennett CS, Fu Y, Mitsuda Y, Chen MM, Greenberg WA, Wong CH, Arnold FH. Chemoenzymatic elaboration of monosaccharides using engineered cytochrome P450BM3 demethylases. Proc Natl Acad Sci U S A. 2009; 106:16550–16555. [PubMed: 19805336]

59. Lewis JC, Mantovani SM, Fu Y, Snow CD, Komor RS, Wong CH, Arnold FH. Combinatorial alanine substitution enables rapid optimization of cytochrome $P450_{BM3}$ for selective hydroxylation of large substrates. Chembiochem. 2010; 11:2502–2505. [PubMed: 21108271]

60. Höhne M, Schatzle S, Jochens H, Robins K, Bornscheuer UT. Rational assignment of key motifs for function guides *in silico* enzyme identification. Nat Chem Biol. 2010; 6:807–813. [PubMed: 20871599]

61. Schirmer A, Rude MA, Li X, Popova E, del Cardayre SB. Microbial biosynthesis of alkanes. Science. 2010; 329:559–562. [PubMed: 20671186]

62. Dennis MW, Kolattukudy PE. Alkane biosynthesis by decarbonylation of aldehyde catalyzed by a microsomal preparation from *Botryococcus braunii*. Arch Biochem Biophys. 1991; 287:268–275. [PubMed: 1898004]

63. Aarts MG, Keijzer CJ, Stiekema WJ, Pereira A. Molecular characterization of the CER1 gene of Arabidopsis involved in epicuticular wax biosynthesis and pollen fertility. Plant Cell. 1995; 7:2115–2127. [PubMed: 8718622]

64. Janga SC, Moreno-Hagelsieb G. Conservation of adjacency as evidence of paralogous operons. Nucleic Acids Res. 2004; 32:5392–5397. [PubMed: 15477389]

65. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. The use of gene clusters to infer functional coupling. Proc Natl Acad Sci U S A. 1999; 96:2896–2901. [PubMed: 10077608]

66. Dandekar T, Snel B, Huynen M, Bork P. Conservation of gene order: A fingerprint of proteins that physically interact. Trends Biochem Sci. 1998; 23:324–328. [PubMed: 9787636]

67. Pignatelli M, Serras F, Moya A, Guigo R, Corominas M. CROC: Finding chromosomal clusters in eukaryotic genomes. Bioinformatics. 2009; 25:1552–1553. [PubMed: 19389737]

68. Liu X, Han B. Evolutionary conservation of neighbouring gene pairs in plants. Gene. 2009; 437:71–79. [PubMed: 19264115]

69. Dávila López M, Martinez Guerra JJ, Samuelsson T. Analysis of gene order conservation in eukaryotes identifies transcriptionally and functionally linked genes. PLoS One. 2010; 5:e10654. [PubMed: 20498846]

70. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. Proc Natl Acad Sci U S A. 1999; 96:4285–4288. [PubMed: 10200254]

71. Gonzalez O, Zimmer R. Assigning functional linkages to proteins using phylogenetic profiles and continuous phenotypes. Bioinformatics. 2008; 24:1257–1263. [PubMed: 18381403]

72. Harvey A. Strategies for discovering drugs from previously unexplored natural products. Drug Discovery Today. 2000; 5:294–300. [PubMed: 10856912]

73. Croteau R, Ketchum RE, Long RM, Kaspera R, Wildung MR. Taxol biosynthesis and molecular genetics. Phytochem Rev. 2006; 5:75–97. [PubMed: 20622989]

74. Barkovich R, Liao JC. Metabolic engineering of isoprenoids. Metab Eng. 2001; 3:27–39. [PubMed: 11162230]

75. Chang MC, Keasling JD. Production of isoprenoid pharmaceuticals by engineered microbes. Nat Chem Biol. 2006; 2:674–681. [PubMed: 17108985]

76. O'Connor SE, Maresh JJ. Chemistry and biology of monoterpene indole alkaloid biosynthesis. Nat Prod Rep. 2006; 23:532–547. [PubMed: 16874388]

77. Liscombe DK, Facchini PJ. Evolutionary and cellular webs in benzylisoquinoline alkaloid biosynthesis. Curr Opin Biotechnol. 2008; 19:173–180. [PubMed: 18396034]

78. Liscombe DK, Usera AR, O'Connor SE. Homolog of tocopherol C methyltransferases catalyzes N methylation in anticancer alkaloid biosynthesis. Proc Natl Acad Sci U S A. 2010; 107:18793–18798. [PubMed: 20956330]

79. Hagel JM, Facchini PJ. Dioxygenases catalyze the O-demethylation steps of morphine biosynthesis in opium poppy. Nat Chem Biol. 2010; 6:273–275. [PubMed: 20228795]

80. Jennewein S, Wildung MR, Chau M, Walker K, Croteau R. Random sequencing of an induced Taxus cell cDNA library for identification of clones involved in Taxol biosynthesis. Proc Natl Acad Sci U S A. 2004; 101:9149–9154. [PubMed: 15178753]

81. Murata J, Bienzle D, Brandle JE, Sensen CW, De Luca V. Expressed sequence tags from Madagascar periwinkle (*Catharanthus roseus*). FEBS Lett. 2006; 580:4501–4507. [PubMed: 16870181]

82. Jennewein S, Rithner CD, Williams RM, Croteau RB. Taxol biosynthesis: taxane 13 alpha-hydroxylase is a cytochrome P450-dependent monooxygenase. Proc Natl Acad Sci U S A. 2001; 98:13595–13600. [PubMed: 11707604]

83. Yonekura-Sakakibara K, Saito K. Functional genomics for plant natural product biosynthesis. Nat Prod Rep. 2009; 26:1466–1487. [PubMed: 19844641]

84. Hall DE, Robert JA, Keeling CI, Domanski D, Quesada AL, Jancsik S, Kuzyk MA, Hamberger B, Borchers CH, Bohlmann J. An integrated genomic, proteomic and biochemical analysis of (+)-3-carene biosynthesis in Sitka spruce (Picea sitchensis) genotypes that are resistant or susceptible to white pine weevil. Plant J. 2011; 65:936–948. [PubMed: 21323772]

85. Desgagne-Penix I, Khan MF, Schriemer DC, Cram D, Nowak J, Facchini PJ. Integration of deep transcriptome and proteome analyses reveals the components of alkaloid metabolism in opium poppy cell cultures. BMC Plant Biol. 2010; 10:252. [PubMed: 21083930]

86. Liscombe DK, Ziegler J, Schmidt J, Ammer C, Facchini PJ. Targeted metabolite and transcript profiling for elucidating enzyme function: isolation of novel N-methyltransferases from three benzylisoquinoline alkaloid-producing species. Plant J. 2009; 60:729–743. [PubMed: 19624470]

87. Haarmann T, Machado C, Lubbe Y, Correia T, Schardl CL, Panaccione DG, Tudzynski P. The ergot alkaloid gene cluster in *Claviceps purpurea*: Extension of the cluster sequence and intra species evolution. Phytochemistry. 2005; 66:1312–1320. [PubMed: 15904941]

88. Ketchum RE, Rithner CD, Qiu D, Kim YS, Williams RM, Croteau RB. Taxus metabolomics: methyl jasmonate preferentially induces production of taxoids oxygenated at C-13 in Taxus x media cell cultures. Phytochemistry. 2003; 62:901–909. [PubMed: 12590117]

89. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990; 215:403–410. [PubMed: 2231712]

90. Blair Hedges S, Kumar S. Genomic clocks and evolutionary timescales. Trends Genet. 2003; 19:200–206. [PubMed: 12683973]

91. Chen FC, Chen CJ, Li WH, Chuang TJ. Gene family size conservation is a good indicator of evolutionary rates. Mol Biol Evol. 2010; 27:1750–1758. [PubMed: 20194423]

92. Rost B. Enzyme function less conserved than anticipated. J Mol Biol. 2002; 318:595–608. [PubMed: 12051862]

93. Tian W, Skolnick J. How well is enzyme function conserved as a function of pairwise sequence identity? J Mol Biol. 2003; 333:863–882. [PubMed: 14568541]

94. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res. 1997; 25:3389–3402. [PubMed: 9254694]

95. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL, et al. The Pfam protein families database. Nucleic Acids Res. 2008; 36:D281–288. [PubMed: 18039703]

96. Wilson D, Pethica R, Zhou Y, Talbot C, Vogel C, Madera M, Chothia C, Gough J. SUPERFAMILY--sophisticated comparative genomics, data mining, visualization and phylogeny. Nucleic Acids Res. 2009; 37:D380–386. [PubMed: 19036790]

97. Orengo CA, Jones DT, Thornton JM. Protein superfamilies and domain superfolds. Nature. 1994; 372:631–634. [PubMed: 7990952]

98. Pegg SC, Brown SD, Ojha S, Seffernick J, Meng EC, Morris JH, Chang PJ, Huang CC, Ferrin TE, Babbitt PC. Leveraging enzyme structure-function relationships for functional inference and experimental design: The structure-function linkage database. Biochemistry. 2006; 45:2545–2555. [PubMed: 16489747]

99. Lisewski AM, Lichtarge O. Rapid detection of similarity in protein structure and function through contact metric distances. Nucleic Acids Res. 2006; 34:e152. [PubMed: 17130161]

100. Barker JA, Thornton JM. An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. Bioinformatics. 2003; 19:1644–1649. [PubMed: 12967960]

101. Torrance JW, Bartlett GJ, Porter CT, Thornton JM. Using a library of structural templates to recognise catalytic sites and explore their evolution in homologous families. J Mol Biol. 2005; 347:565–581. [PubMed: 15755451]

102. Porter CT, Bartlett GJ, Thornton JM. The Catalytic Site Atlas: A resource of catalytic sites and residues identified in enzymes using structural data. Nucleic Acids Res. 2004; 32:D129–133. [PubMed: 14681376]

103. Kinoshita K, Nakamura H. eF-site and PDBjViewer: Database and viewer for protein functional sites. Bioinformatics. 2004; 20:1329–1330. [PubMed: 14871866]

104. Chen F, Gaucher EA, Leal NA, Hutter D, Havemann SA, Govindarajan S, Ortlund EA, Benner SA. Reconstructed evolutionary adaptive paths give polymerases accepting reversible terminators for sequencing and SNP detection. Proc Natl Acad Sci U S A. 2010; 107:1948–1953. [PubMed: 20080675]

105. Voigt CA, Martinez C, Wang ZG, Mayo SL, Arnold FH. Protein building blocks preserved by recombination. Nat Struct Biol. 2002; 9:553–558. [PubMed: 12042875]

106. Heinzelman P, Snow CD, Wu I, Nguyen C, Villalobos A, Govindarajan S, Minshull J, Arnold FH. A family of thermostable fungal cellulases created by structure-guided recombination. Proc Natl Acad Sci U S A. 2009; 106:5610–5615. [PubMed: 19307582]

107. Li Y, Drummond DA, Sawayama AM, Snow CD, Bloom JD, Arnold FH. A diverse family of thermostable cytochrome P450s created by recombination of stabilizing fragments. Nat Biotechnol. 2007; 25:1051–1056. [PubMed: 17721510]

108. Kolb P, Ferreira RS, Irwin JJ, Shoichet BK. Docking and chemoinformatic screens for new ligands and targets. Curr Opin Biotechnol. 2009; 20:429–436. [PubMed: 19733475]

109. Hermann JC, Ghanem E, Li Y, Raushel FM, Irwin JJ, Shoichet BK. Predicting substrates by docking high-energy intermediates to enzyme structures. J Am Chem Soc. 2006; 128:15882–15891. [PubMed: 17147401]

110. Kalyanaraman C, Imker HJ, Fedorov AA, Fedorov EV, Glasner ME, Babbitt PC, Almo SC, Gerlt JA, Jacobson MP. Discovery of a dipeptide epimerase enzymatic function guided by homology modeling and virtual screening. Structure. 2008; 16:1668–1677. [PubMed: 19000819]

111. Favia AD, Nobeli I, Glaser F, Thornton JM. Molecular docking for substrate identification: The short-chain dehydrogenases/reductases. J Mol Biol. 2008; 375:855–874. [PubMed: 18036612]

112. Song L, Kalyanaraman C, Fedorov AA, Fedorov EV, Glasner ME, Brown S, Imker HJ, Babbitt PC, Almo SC, Jacobson MP, et al. Prediction and assignment of function for a divergent *N*-succinyl amino acid racemase. Nat Chem Biol. 2007; 3:486–491. [PubMed: 17603539]

113. Hermann JC, Marti-Arbona R, Fedorov AA, Fedorov E, Almo SC, Shoichet BK, Raushel FM. Structure-based activity prediction for an enzyme of unknown function. Nature. 2007; 448:775–779. [PubMed: 17603473]

114. Cummings JA, Nguyen TT, Fedorov AA, Kolb P, Xu C, Fedorov EV, Shoichet BK, Barondeau DP, Almo SC, Raushel FM. Structure, mechanism, and substrate profile for Sco3058: The closest bacterial homologue to human renal dipeptidase. Biochemistry. 2010; 49:611–622. [PubMed: 20000809]

115. Hall RS, Fedorov AA, Marti-Arbona R, Fedorov EV, Kolb P, Sauder JM, Burley SK, Shoichet BK, Almo SC, Raushel FM. The hunt for 8-oxoguanine deaminase. J Am Chem Soc. 2010; 132:1762–1763. [PubMed: 20088583]

116. Xiang DF, Kolb P, Fedorov AA, Meier MM, Fedorov LV, Nguyen TT, Sterner R, Almo SC, Shoichet BK, Raushel FM. Functional annotation and three-dimensional structure of Dr0930 from *Deinococcus radiodurans*, a close relative of phosphotriesterase in the amidohydrolase superfamily. Biochemistry. 2009; 48:2237–2247. [PubMed: 19159332]

117. Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dumpelfeld B, et al. Proteome survey reveals modularity of the yeast cell machinery. Nature. 2006; 440:631–636. [PubMed: 16429126]

118. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, et al. Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. Nature. 2006; 440:637–643. [PubMed: 16554755]

119. Tarassov K, Messier V, Landry CR, Radinovic S, Serna Molina MM, Shames I, Malitskaya Y, Vogel J, Bussey H, Michnick SW. An *in vivo* map of the yeast protein interactome. Science. 2008; 320:1465–1470. [PubMed: 18467557]

120. Yu H, Braun P, Yildirim MA, Lemmens I, Venkatesan K, Sahalie J, Hirozane-Kishikawa T, Gebreab F, Li N, Simonis N, et al. High-quality binary protein interaction map of the yeast interactome network. Science. 2008; 322:104–110. [PubMed: 18719252]

121. Bader GD, Betel D, Hogue CW. BIND: The Biomolecular Interaction Network Database. Nucleic Acids Res. 2003; 31:248–250. [PubMed: 12519993]

122. Breitkreutz BJ, Stark C, Reguly T, Boucher L, Breitkreutz A, Livstone M, Oughtred R, Lackner DH, Bahler J, Wood V, et al. The BioGRID Interaction Database: 2008 update. Nucleic Acids Res. 2008; 36:D637–640. [PubMed: 18000002]

123. Aranda B, Achuthan P, Alam-Faruque Y, Armean I, Bridge A, Derow C, Feuermann M, Ghanbarian AT, Kerrien S, Khadake J, et al. The IntAct molecular interaction database in 2010. Nucleic Acids Res. 2010; 38:D525–531. [PubMed: 19850723]

124. Luo F, Yang Y, Zhong J, Gao H, Khan L, Thompson DK, Zhou J. Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory. BMC Bioinformatics. 2007; 8:299. [PubMed: 17697349]

125. Wang K, Narayanan M, Zhong H, Tompa M, Schadt EE, Zhu J. Meta-analysis of inter-species liver co-expression networks elucidates traits associated with common human diseases. PLoS Comput Biol. 2009; 5:e1000616. [PubMed: 20019805]

126. Ruan J, Dean AK, Zhang W. A general co-expression network-based approach to gene expression analysis: Comparison and applications. BMC Syst Biol. 2010; 4:8. [PubMed: 20122284]

127. Wang Z, Gerstein M, Snyder M. RNA-Seq: A revolutionary tool for transcriptomics. Nat Rev Genet. 2009; 10:57–63. [PubMed: 19015660]

128. Collins SR, Roguev A, Krogan NJ. Quantitative Genetic Interaction Mapping Using the E-Map Approach. Methods Enzymol. 2010; 470:205–231. [PubMed: 20946812]

129. Collins SR, Miller KM, Maas NL, Roguev A, Fillingham J, Chu CS, Schuldiner M, Gebbia M, Recht J, Shales M, et al. Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. Nature. 2007; 446:806–810. [PubMed: 17314980]

130. Schuldiner M, Collins SR, Thompson NJ, Denic V, Bhamidipati A, Punna T, Ihmels J, Andrews B, Boone C, Greenblatt JF, et al. Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. Cell. 2005; 123:507–519. [PubMed: 16269340]

131. Aguilar PS, Frohlich F, Rehman M, Shales M, Ulitsky I, Olivera-Couto A, Braberg H, Shamir R, Walter P, Mann M, et al. A plasma-membrane E-MAP reveals links of the eisosome with sphingolipid metabolism and endosomal trafficking. Nat Struct Mol Biol. 2010; 17:901–908. [PubMed: 20526336]

132. Fiedler D, Braberg H, Mehta M, Chechik G, Cagney G, Mukherjee P, Silva AC, Shales M, Collins SR, van Wageningen S, et al. Functional organization of the *S. cerevisiae* phosphorylation network. Cell. 2009; 136:952–963. [PubMed: 19269370]

133. Breslow DK, Collins SR, Bodenmiller B, Aebersold R, Simons K, Shevchenko A, Ejsing CS, Weissman JS. Orm family proteins mediate sphingolipid homeostasis. Nature. 2010; 463:1048–1053. [PubMed: 20182505]

134. Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, Sevier CS, Ding H, Koh JL, Toufighi K, Mostafavi S, et al. The genetic landscape of a cell. Science. 2010; 327:425–431. [PubMed: 20093466]

135. Schlabach MR, Luo J, Solimini NL, Hu G, Xu Q, Li MZ, Zhao Z, Smogorzewska A, Sowa ME, Ang XL, et al. Cancer proliferation gene discovery through functional genomics. Science. 2008; 319:620–624. [PubMed: 18239126]

136. Bassik MC, Lebbink RJ, Churchman LS, Ingolia NT, Patena W, LeProust EM, Schuldiner M, Weissman JS, McManus MT. Rapid creation and quantitative monitoring of high coverage shRNA libraries. Nat Methods. 2009; 6:443–445. [PubMed: 19448642]

137. Tian C, Beeson WT, Iavarone AT, Sun J, Marletta MA, Cate JH, Glass NL. Systems analysis of plant cell wall degradation by the model filamentous fungus Neurospora crassa. Proc Natl Acad Sci U S A. 2009; 106:22157–22162. [PubMed: 20018766]

138. Galazka JM, Tian C, Beeson WT, Martinez B, Glass NL, Cate JH. Cellodextrin transport in yeast for improved biofuel production. Science. 2010; 330:84–86. [PubMed: 20829451]

139. Mazurkiewicz P, Tang CM, Boone C, Holden DW. Signature-tagged mutagenesis: barcoding mutants for genome-wide screens. Nat Rev Genet. 2006; 7:929–939. [PubMed: 17139324]

140. Ehren J, Govindarajan S, Moron B, Minshull J, Khosla C. Protein engineering of improved prolyl endopeptidases for celiac sprue therapy. Protein Eng Des Sel. 2008; 21:699–707. [PubMed: 18836204]

141. Liao J, Warmuth MK, Govindarajan S, Ness JE, Wang RP, Gustafsson C, Minshull J. Engineering proteinase K using machine learning and synthetic genes. BMC Biotechnol. 2007; 7:16. [PubMed: 17386103]

142. Wen F, Nair NU, Zhao H. Protein engineering in designing tailored enzymes and microorganisms for biofuels production. Curr Opin Biotechnol. 2009; 20:412–419. [PubMed: 19660930]

143. Ang EL, Obbard JP, Zhao H. Probing the molecular determinants of aniline dioxygenase substrate specificity by saturation mutagenesis. FEBS J. 2007; 274:928–939. [PubMed: 17269935]

144. Savile CK, Janey JM, Mundorff EC, Moore JC, Tam S, Jarvis WR, Colbeck JC, Krebber A, Fleitz FJ, Brands J, et al. Biocatalytic asymmetric synthesis of chiral amines from ketones applied to sitagliptin manufacture. Science. 2010; 329:305–309. [PubMed: 20558668]

145. Dietrich JA, Yoshikuni Y, Fisher KJ, Woolard FX, Ockey D, McPhee DJ, Renninger NS, Chang MC, Baker D, Keasling JD. A novel semi-biosynthetic route for artemisinin production using engineered substrate-promiscuous $P450_{BM3}$. ACS Chem Biol. 2009; 4:261–267. [PubMed: 19271725]

146. Yoshikuni Y, Dietrich JA, Nowroozi FF, Babbitt PC, Keasling JD. Redesigning enzymes based on adaptive evolution for optimal function in synthetic metabolic pathways. Chem Biol. 2008; 15:607–618. [PubMed: 18559271]

147. Leonard E, Ajikumar PK, Thayer K, Xiao WH, Mo JD, Tidor B, Stephanopoulos G, Prather KL. Combining metabolic and protein engineering of a terpenoid biosynthetic pathway for overproduction and selectivity control. Proc Natl Acad Sci U S A. 2010; 107:13654–13659. [PubMed: 20643967]

148. Jeppsson M, Johansson B, Hahn-Hagerdal B, Gorwa-Grauslund MF. Reduced oxidative pentose phosphate pathway flux in recombinant xylose-utilizing *Saccharomyces cerevisiae* strains improves the ethanol yield from xylose. Appl Environ Microbiol. 2002; 68:1604–1609. [PubMed: 11916674]

149. Grotkjaer T, Christakopoulos P, Nielsen J, Olsson L. Comparative metabolic network analysis of two xylose fermenting recombinant *Saccharomyces cerevisiae* strains. Metab Eng. 2005; 7:437–444. [PubMed: 16140032]

150. Watanabe S, Abu Saleh A, Pack SP, Annaluru N, Kodaki T, Makino K. Ethanol production from xylose by recombinant Saccharomyces cerevisiae expressing protein-engineered NADH-preferring xylose reductase from *Pichia stipitis*. Microbiology. 2007; 153:3044–3054. [PubMed: 17768247]

151. Ha SJ, Galazka JM, Rin Kim S, Choi JH, Yang X, Seo JH, Louise Glass N, Cate JHD, Jin YS. Engineered *Saccharomyces cerevisiae* capable of simultaneous cellobiose and xylose fermentation. Proc Natl Acad Sci U S A. 2011; 108:504–509. [PubMed: 21187422]

152. Hiraga K, Arnold FH. General method for sequence-independent site-directed chimeragenesis. J Mol Biol. 2003; 330:287–296. [PubMed: 12823968]

153. Park HS, Nam SH, Lee JK, Yoon CN, Mannervik B, Benkovic SJ, Kim HS. Design and evolution of new catalytic activity with an existing protein scaffold. Science. 2006; 311:535–538. [PubMed: 16439663]

154. Yoshikuni Y, Ferrin TE, Keasling JD. Designed divergent evolution of enzyme function. Nature. 2006; 440:1078–1082. [PubMed: 16495946]

155. Gerlt JA, Babbitt PC. Divergent evolution of enzymatic function: Mechanistically diverse superfamilies and functionally distinct suprafamilies. Annu Rev Biochem. 2001; 70:209–246. [PubMed: 11395407]

156. Babbitt PC, Mrachko GT, Hasson MS, Huisman GW, Kolter R, Ringe D, Petsko GA, Kenyon GL, Gerlt JA. A functionally diverse enzyme superfamily that abstracts the alpha protons of carboxylic acids. Science. 1995; 267:1159–1161. [PubMed: 7855594]

157. Gerlt JA, Babbitt PC, Rayment I. Divergent evolution in the enolase superfamily: The interplay of mechanism and specificity. Arch Biochem Biophys. 2005; 433:59–70. [PubMed: 15581566]

158. Schmidt DM, Mundorff EC, Dojka M, Bermudez E, Ness JE, Govindarajan S, Babbitt PC, Minshull J, Gerlt JA. Evolutionary potential of (β/α)8-barrels: functional promiscuity produced by single substitutions in the enolase superfamily. Biochemistry. 2003; 42:8387–8393. [PubMed: 12859183]

159. Gerlt JA, Babbitt PC. Enzyme (re)design: Lessons from natural evolution and computation. Curr Opin Chem Biol. 2009; 13:10–18. [PubMed: 19237310]

160. O'Brien P, Herschlag D. Sulfatase activity of *E. coli* alkaline phosphatase demonstrates a functional link to arylsulfatases, an evolutionarily related enzyme family. J Am Chem Soc. 1998; 120:12369–12370.

161. O'Brien PJ, Herschlag D. Functional interrelationships in the alkaline phosphatase superfamily: Phosphodiesterase activity of *Escherichia coli* alkaline phosphatase. Biochemistry. 2001; 40:5691–5699. [PubMed: 11341834]

162. Taylor Ringia EA, Garrett JB, Thoden JB, Holden HM, Rayment I, Gerlt JA. Evolution of enzymatic activity in the enolase superfamily: Functional studies of the promiscuous *o*-succinylbenzoate synthase from Amycolatopsis. Biochemistry. 2004; 43:224–229. [PubMed: 14705949]

163. Wang SC, Johnson WH Jr, Whitman CP. The 4-oxalocrotonate tautomerase- and YwhB-catalyzed hydration of 3E-haloacrylates: Implications for the evolution of new enzymatic activities. J Am Chem Soc. 2003; 125:14282–14283. [PubMed: 14624569]

164. Afriat L, Roodveldt C, Manco G, Tawfik DS. The latent promiscuity of newly identified microbial lactonases is linked to a recently diverged phosphotriesterase. Biochemistry. 2006; 45:13677–13686. [PubMed: 17105187]

165. Bershtein S, Goldin K, Tawfik DS. Intense neutral drifts yield robust and evolvable consensus proteins. J Mol Biol. 2008; 379:1029–1044. [PubMed: 18495157]

166. Kimura, M. The Neutral Theory of Molecular Evolution. Cambridge University Press; Cambridge: 1983.

167. Anandarajah K, Kiefer PM Jr, Donohoe BS, Copley SD. Recruitment of a double bond isomerase to serve as a reductive dehalogenase during biodegradation of pentachlorophenol. Biochemistry. 2000; 39:5303–5311. [PubMed: 10820000]

168. Roodveldt C, Tawfik DS. Shared promiscuous activities and evolutionary features in various members of the amidohydrolase superfamily. Biochemistry. 2005; 44:12728–12736. [PubMed: 16171387]

169. Kim J, Kershner JP, Novikov Y, Shoemaker RK, Copley SD. Three serendipitous pathways in *E. coli* can bypass a block in pyridoxal-5′-phosphate synthesis. Mol Syst Biol. 2010; 6:436. [PubMed: 21119630]

170. Wise E, Yew WS, Babbitt PC, Gerlt JA, Rayment I. Homologous (β/α) 8-barrel enzymes that catalyze unrelated reactions: Orotidine 5′-monophosphate decarboxylase and 3-keto-L-gulonate 6-phosphate decarboxylase. Biochemistry. 2002; 41:3861–3869. [PubMed: 11900527]

171. Fischer CR, Tseng HC, Tai M, Prather KL, Stephanopoulos G. Assessment of heterologous butyrate and butanol pathway activity by measurement of intracellular pathway intermediates in recombinant *Escherichia coli*. Appl Microbiol Biotechnol. 2010; 88:265–275. [PubMed: 20625717]

172. Yew WS, Akana J, Wise EL, Rayment I, Gerlt JA. Evolution of enzymatic activities in the orotidine 5′-monophosphate decarboxylase suprafamily: Enhancing the promiscuous D-arabino-hex-3-ulose 6-phosphate synthase reaction catalyzed by 3-keto-L-gulonate 6-phosphate decarboxylase. Biochemistry. 2005; 44:1807–1815. [PubMed: 15697206]

173. Lodeiro S, Xiong Q, Wilson WK, Kolesnikova MD, Onak CS, Matsuda SP. An oxidosqualene cyclase makes numerous products by diverse mechanisms: A challenge to prevailing concepts of triterpene biosynthesis. J Am Chem Soc. 2007; 129:11213–11222. [PubMed: 17705488]

174. Xu M, Wilderman PR, Peters RJ. Following evolution’s lead to a single residue switch for diterpene synthase product outcome. Proc Natl Acad Sci U S A. 2007; 104:7397–7401. [PubMed: 17456599]

175. Wilderman PR, Peters RJ. A single residue switch converts abietadiene synthase into a pimaradiene specific cyclase. J Am Chem Soc. 2007; 129:15736–15737. [PubMed: 18052062]

176. Morrone D, Xu M, Fulton DB, Determan MK, Peters RJ. Increasing complexity of a diterpene synthase reaction with a single residue switch. J Am Chem Soc. 2008; 130:5400–5401. [PubMed: 18366162]

177. Lodeiro S, Schulz-Gasch T, Matsuda SP. Enzyme redesign:Two mutations cooperate to convert cycloartenol synthase into an accurate lanosterol synthase. J Am Chem Soc. 2005; 127:14132–14133. [PubMed: 16218577]

178. Greenhagen BT, O’Maille PE, Noel JP, Chappell J. Identifying and manipulating structural determinates linking catalytic specificities in terpene synthases. Proc Natl Acad Sci U S A. 2006; 103:9826–9831. [PubMed: 16785438]

179. Kampranis SC, Ioannidis D, Purvis A, Mahrez W, Ninga E, Katerelos NA, Anssour S, Dunwell JM, Degenhardt J, Makris AM, et al. Rational conversion of substrate and product specificity in a Salvia monoterpene synthase: Structural insights into the evolution of terpene synthase function. Plant Cell. 2007; 19:1994–2005. [PubMed: 17557809]

180. Jones KL, Kim SW, Keasling JD. Low-copy plasmids can perform as well as or better than high-copy plasmids for metabolic engineering of bacteria. Metab Eng. 2000; 2:328–338. [PubMed: 11120644]

181. Glick BR. Metabolic load and heterologous gene expression. Biotechnol Adv. 1995; 13:247–261. [PubMed: 14537822]

182. Lin H, Vadali RV, Bennett GN, San KY. Increasing the acetyl-CoA pool in the presence of overexpressed phosphoenolpyruvate carboxylase or pyruvate carboxylase enhances succinate production in *Escherichia coli*. Biotechnol Prog. 2004; 20:1599–1604. [PubMed: 15458351]

183. Singh A, Lynch MD, Gill RT. Genes restoring redox balance in fermentation-deficient *E. coli* NZN111. Metab Eng. 2009; 11:347–354. [PubMed: 19628049]

184. Singh A, Cher Soh K, Hatzimanikatis V, Gill RT. Manipulating redox and ATP balancing for improved production of succinate in *E. coli*. Metab Eng. 2011; 13:76–81. [PubMed: 21040799]

185. Pitera DJ, Paddon CJ, Newman JD, Keasling JD. Balancing a heterologous mevalonate pathway for improved isoprenoid production in *Escherichia coli*. Metab Eng. 2007; 9:193–207. [PubMed: 17239639]

186. Pfleger BF, Pitera DJ, Smolke CD, Keasling JD. Combinatorial engineering of intergenic regions in operons tunes expression of multiple genes. Nat Biotechnol. 2006; 24:1027–1032. [PubMed: 16845378]

187. Dueber JE, Wu GC, Malmirchegini GR, Moon TS, Petzold CJ, Ullal AV, Prather KL, Keasling JD. Synthetic protein scaffolds provide modular control over metabolic flux. Nat Biotechnol. 2009; 27:753–759. [PubMed: 19648908]

188. Anthony JR, Anthony LC, Nowroozi F, Kwon G, Newman JD, Keasling JD. Optimization of the mevalonate-based isoprenoid biosynthetic pathway in *Escherichia coli* for production of the anti-malarial drug precursor amorpha-4,11-diene. Metab Eng. 2009; 11:13–19. [PubMed: 18775787]

189. Atsumi S, Liao JC. Directed evolution of *Methanococcus jannaschii* citramalate synthase for biosynthesis of 1-propanol and 1-butanol by *Escherichia coli*. Appl Environ Microbiol. 2008; 74:7802–7808. [PubMed: 18952866]

190. Contador CA, Rizk ML, Asenjo JA, Liao JC. Ensemble modeling for strain development of L-lysine-producing *Escherichia coli*. Metab Eng. 2009; 11:221–233. [PubMed: 19379820]

191. Rizk ML, Liao JC. Ensemble modeling for aromatic production in *Escherichia coli*. PLoS One. 2009; 4:e6903. [PubMed: 19730732]

192. Tan Y, Rivera JG, Contador CA, Asenjo JA, Liao JC. Reducing the allowable kinetic space by constructing ensemble of dynamic models with the same steady-state flux. Metab Eng. 2011; 13:60–75. [PubMed: 21075211]

193. Guzman LM, Belin D, Carson MJ, Beckwith J. Tight regulation, modulation, and high-level expression by vectors containing the arabinose P_{BAD} promoter. J Bacteriol. 1995; 177:4121–4130. [PubMed: 7608087]

194. Lee SK, Keasling JD. Propionate-regulated high-yield protein production in *Escherichia coli*. Biotechnol Bioeng. 2006; 93:912–918. [PubMed: 16333863]

195. Khlebnikov A, Datsenko KA, Skaug T, Wanner BL, Keasling JD. Homogeneous expression of the P(BAD) promoter in *Escherichia coli* by constitutive expression of the low-affinity high-capacity AraE transporter. Microbiology. 2001; 147:3241–3247. [PubMed: 11739756]

196. Smolke CD, Carrier TA, Keasling JD. Coordinated, differential expression of two genes through directed mRNA cleavage and stabilization by secondary structures. Appl Environ Microbiol. 2000; 66:5399–5405. [PubMed: 11097920]

197. Win MN, Smolke CD. A modular and extensible RNA-based gene-regulatory platform for engineering cellular function. Proc Natl Acad Sci U S A. 2007; 104:14283–14288. [PubMed: 17709748]

198. Ogawa A, Maeda M. An artificial aptazyme-based riboswitch and its cascading system in *E. coli*. Chembiochem. 2008; 9:206–209. [PubMed: 18098257]

199. Farmer WR, Liao JC. Improving lycopene production in *Escherichia coli* by engineering metabolic control. Nat Biotechnol. 2000; 18:533–537. [PubMed: 10802621]

200. Basu S, Gerchman Y, Collins CH, Arnold FH, Weiss R. A synthetic multicellular system for programmed pattern formation. Nature. 2005; 434:1130–1134. [PubMed: 15858574]

201. Salis HM, Mirsky EA, Voigt CA. Automated design of synthetic ribosome binding sites to control protein expression. Nat Biotechnol. 2009; 27:946–950. [PubMed: 19801975]

202. Hyde CC, Ahmed SA, Padlan EA, Miles EW, Davies DR. Three-dimensional structure of the tryptophan synthase $\alpha_2\beta_2$ multienzyme complex from Salmonella *typhimurium*. J Biol Chem. 1988; 263:17857–17871. [PubMed: 3053720]

203. Thoden JB, Holden HM, Wesenberg G, Raushel FM, Rayment I. Structure of carbamoyl phosphate synthetase: A journey of 96 Å from substrate to product. Biochemistry. 1997; 36:6305–6316. [PubMed: 9174345]

204. Hawkins AR, Smith M. Domain structure and interaction within the pentafunctional arom polypeptide. Eur J Biochem. 1991; 196:717–724. [PubMed: 1849480]

205. Morgunov I, Srere PA. Interaction between citrate synthase and malate dehydrogenase - Substrate channeling of oxaloacetate. J Biol Chem. 1998; 273:29540–29544. [PubMed: 9792662]

206. Graciet E, Lebreton S, Camadro JM, Gontero B. Thermodynamic analysis of the emergence of new regulatory properties in a phosphoribulokinase-glyceraldehyde 3-phosphate dehydrogenase complex. J Biol Chem. 2002; 277:12697–12702. [PubMed: 11815615]

207. Campanella ME, Chu HY, Low PS. Assembly and regulation of a glycolytic enzyme complex on the human erythrocyte membrane. Proc Natl Acad Sci U S A. 2005; 102:2402–2407. [PubMed: 15701694]

208. Binstock JF, Pramanik A, Schulz H. Isolation of a multienzyme complex of fatty-acid oxidation from *Escherichia coli*. Proc Natl Acad Sci U S A. 1977; 74:492–495. [PubMed: 322129]

209. Gallastegui N, Groll M. The 26S proteasome: Assembly and function of a destructive machine. Trends Biochem Sci. 2010; 35:634–642. [PubMed: 20541423]

210. Moller BL, Conn EE. The biosynthesis of cyanogenic glucosides in higher plants. Channeling of intermediates in dhurrin biosynthesis by a microsomal system from *Sorghum bicolor* (Linn) Moench. J Biol Chem. 1980; 255:3049–3056. [PubMed: 7358727]

211. Tattersall DB, Bak S, Jones PR, Olsen CE, Nielsen JK, Hansen ML, Hoj PB, Moller BL. Resistance to an herbivore through engineered cyanogenic glucoside synthesis. Science. 2001; 293:1826–1828. [PubMed: 11474068]

212. Vogt T. Phenylpropanoid biosynthesis. Molecular Plant. 2010; 3:2–20. [PubMed: 20035037]

213. Mosbach K, Mattiasson B. Matrix-bound enzymes. II Studies on a matrix-bound two-enzyme-system. Acta Chem Scand. 1970; 24:2093–2100. [PubMed: 4394961]

214. Bulow L, Ljungcrantz P, Mosbach K. Preparation of a soluble bifunctional enzyme by gene fusion. Bio/Technology. 1985; 3:821.

215. Bulow L. Characterization of an artificial bifunctional enzyme, β-galactosidase/galactokinase, prepared by gene fusion. Eur J Biochem. 1987; 163:443–448. [PubMed: 3104037]

216. Fierobe HP, Mechaly A, Tardif C, Belaich A, Lamed R, Shoham Y, Belaich JP, Bayer EA. Design and production of active cellulosome chimeras. Selective incorporation of dockerin-containing enzymes into defined functional complexes. J Biol Chem. 2001; 276:21257–21261. [PubMed: 11290750]

217. Moon TS, Dueber JE, Shiue E, Prather KL. Use of modular, synthetic scaffolds for improved production of glucaric acid in engineered *E. coli*. Metab Eng. 2010; 12:298–305. [PubMed: 20117231]

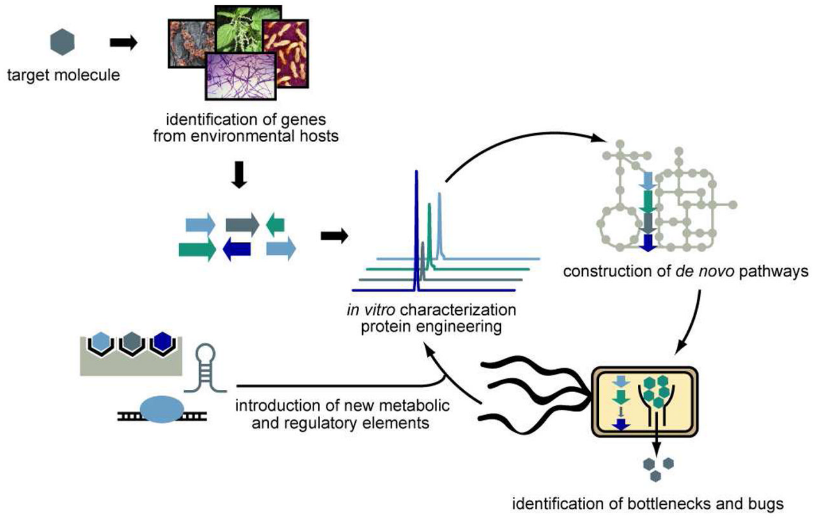
target molecule
identification of genes
from environmental hosts
in vitro characterization
protein engineering
construction of *de novo* pathways
introduction of new metabolic
and regulatory elements
identification of bottlenecks and bugs

Figure 1.

Pipeline for construction of a *de novo* metabolic pathway. Enzymes from environmental hosts are identified, assembled, and transplanted into a heterologous host for target molecule production. Bottlenecks that decrease product titer are identified by a combination of *in vivo* and *in vitro* characterization to insight into their source. Incorporation of additional metabolic and regulatory elements are used to alleviate these bottlenecks, which reveals new factors that limit production yields for further optimization.

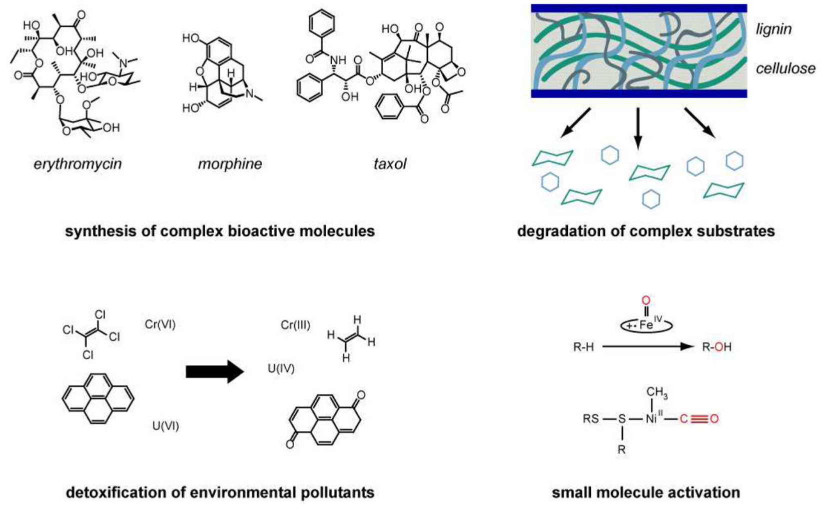
erythromycin
morphine
taxol
lignin
cellulose
synthesis of complex bioactive molecules
degradation of complex substrates
Cr(VI)
U(VI)
Cr(III)
U(IV)
detoxification of environmental pollutants
Fe IV
R-H
R-OH
RS
S
Ni II
C
O
CH3
R
small molecule activation

Figure 2.

Chemical phenotypes of interest for *de novo* metabolic pathway construction.

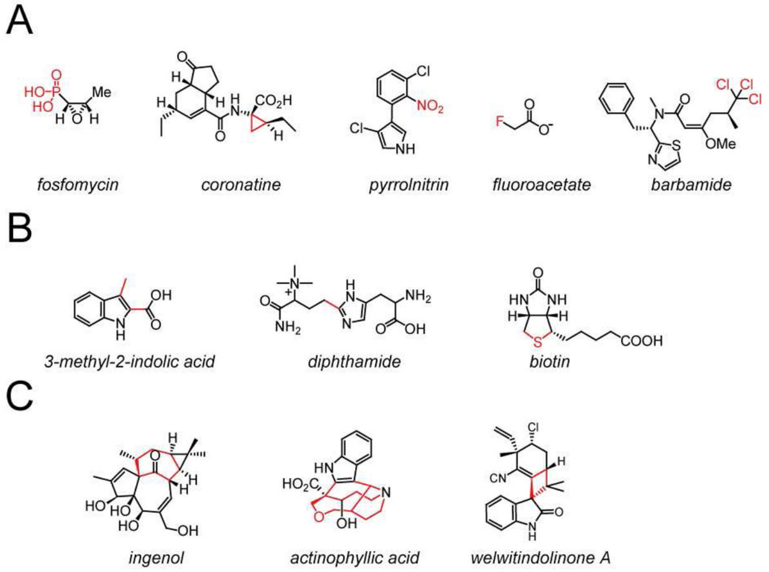

Figure 3.

Specialized structural motifs and unusual functional groups in natural products. Structural motifs and functional groups of interested are highlighted in red. (A) Unique functional groups. (B) Unusual bond couplings. (C) Strained ring structures.

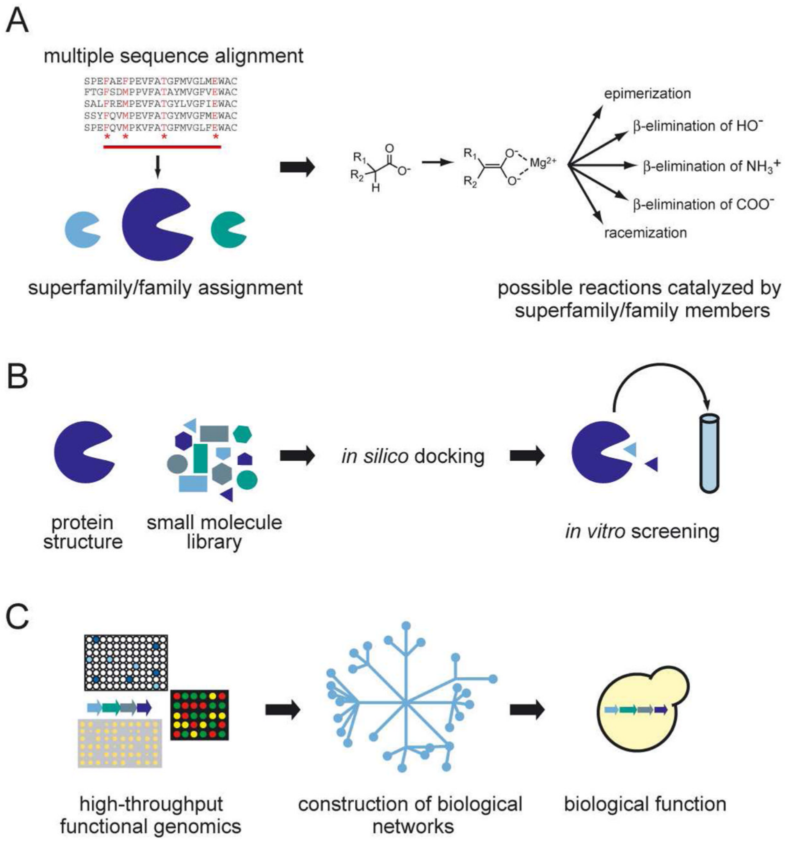

Figure 4.

Methods for functional gene annotation. (A) Multiple sequence alignments are often sufficient to classify an uncharacterized protein into a superfamily or family, limiting the scope of possible reactions. (B) *In silico* docking utilizes structural information about the protein of interest for docking potential substrates and ranking them according to favorability of binding. The results limit the size of libraries that must be screened *in vitro* to determine enzyme function. (C) Functional genomic approaches including protein-protein interaction screens, genetic interaction mapping, and microarray analysis facilitate gene annotation based on biological rather than biochemical function.

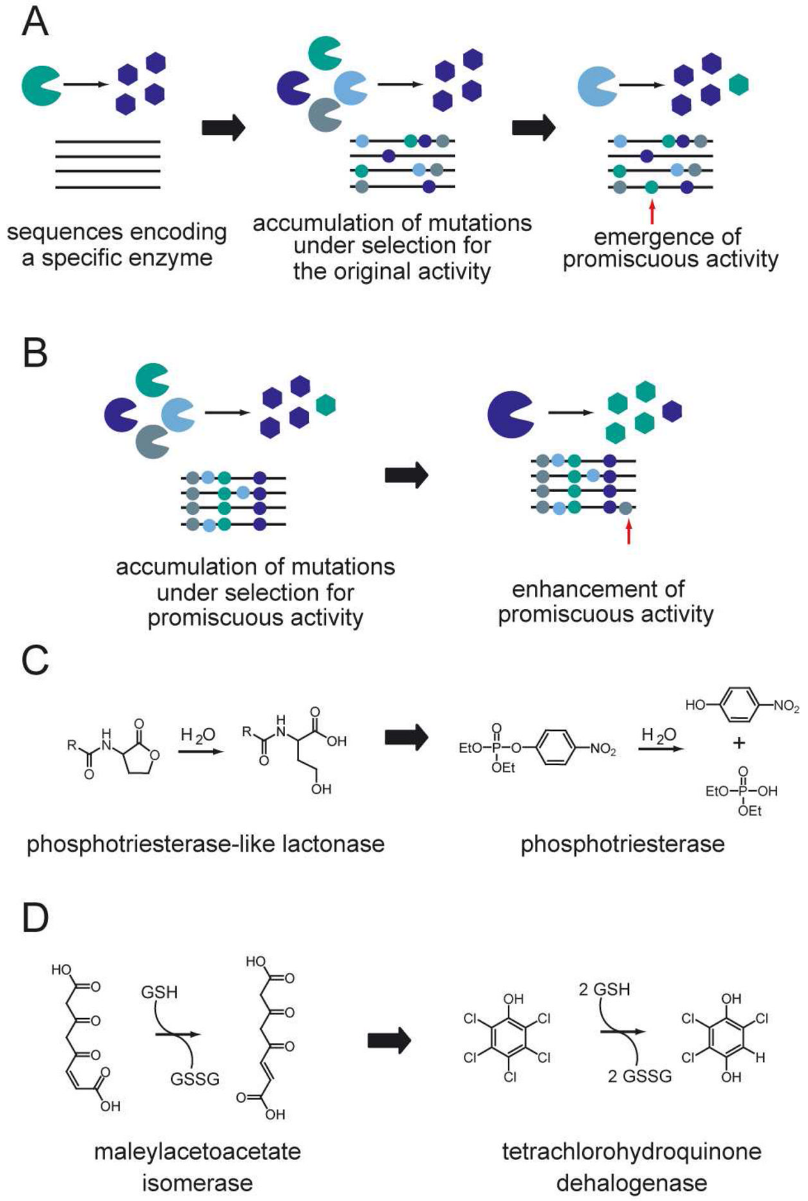

Figure 5.

The neutral drift mechanism of enzyme evolution. (A) A sequence encoding an enzyme with a specific substrate accumulates mutations under selective pressure to maintain the original activity, resulting in the emergence of promiscuous activities and the maintenance of the initial activity. (B) When a selective pressure favoring the promiscuous activity arises, the accumulation of a small number of mutations enhances the promiscuous activity. (C) Phosphotriesterases are proposed to have evolved recently from lactonases. (D) Tetrachlorohydroquinone dehalogenase is proposed to have evolved from maleylacetoacetate isomerase (GSH, glutathione; GSSG, glutathione disulfide).

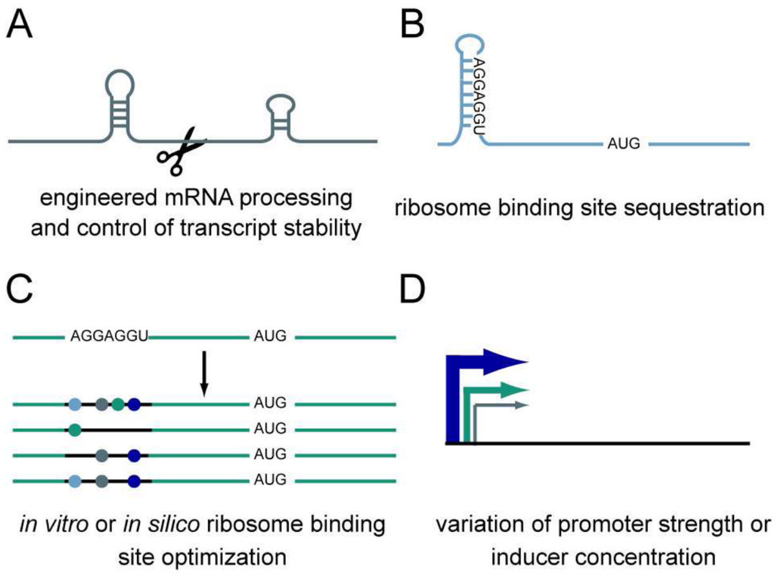
A
engineered mRNA processing and control of transcript stability
B
AGGAGGU
AUG
ribosome binding site sequestration
C
AGGAGGU
AUG
AUG
AUG
AUG
AUG
in vitro or *in silico* ribosome binding site optimization
D
variation of promoter strength or inducer concentration

Figure 6.

Engineering pathway balance. (A) Expression of pathways genes appropriate levels can be achieved by adding RNA regulatory elements. (B) Control of ribosome binding site accessibility or (C) ribosome binding site optimization can be used to tune protein expression at the translational level. (D) Variation of promoter strength or inducer concentration can be used to tune protein expression at the transcriptional level.

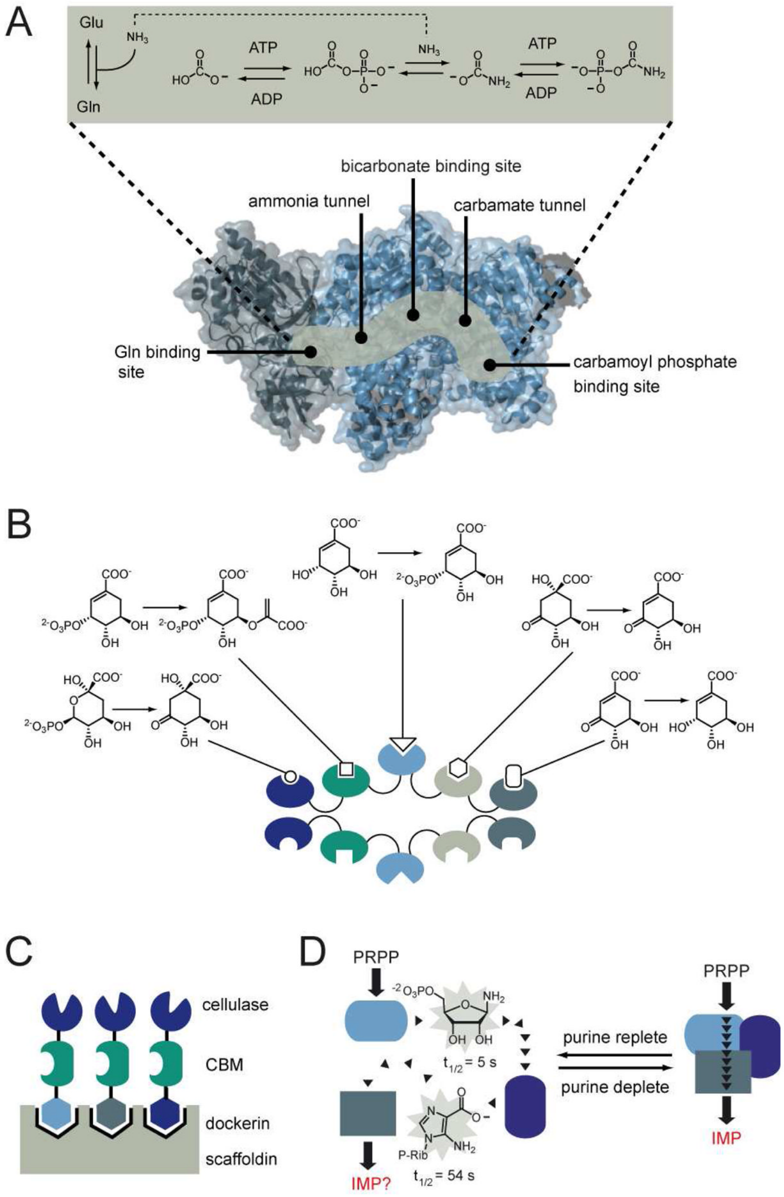

Figure 7.

Spatial organization in natural enzyme systems. (A) Carbamoyl phosphate synthetase utilizes a physical channel that protects the labile intermediates ammonia, carboxy phosphate, and carbamate from the cellular environment. (B) In the fungal shikimate pathway, a pentafunctional polypeptide is utilized to catalyze the five reactions required to convert 3-deoxy-D-arabino-heptulosonate 7-phosphate to 5-enolpyruvylshikimate 3-phosphate. (C) The breakdown of cellulose requires multiple enzymes of the glycosyl hydrolase superfamily, including reducing and non-reducing end exoglucanases as well as endoglucanases. The spatial organization of cellulosomes allows the modular docking and exchange of different enzymes in a single scaffold for synergistic and tunable degradation of the sugar polymer (CBM, celluolose binding module). (D) Formation of the purinosome complex between the enzymes that catalyze *de novo* purine biosynthesis is believed to protect the short-lived intermediates, phosphoribosylamine (PRA) and 4-carboxyaminoimidazole ribonucleotide (CAIR), in the mammalian pathway between 5-

phosphoribosyl-α-pyrophosphate (PRPP) and inosine monophosphate (IMP). In most bacteria, yeasts, and plants, the precursor to CAIR, N^5-carboxy-4-aminoimidazole ribonucleotide (N^5-CAIR, $t_{½}$ = 15 s), also exists as a short-lived intermediate but is channeled by a multifunctional enzyme in mammals.