Computational tools for the synthetic design of biochemical pathways

Marnix H. Medema[1,2], *Renske van Raaphorst*[1,2], *Eriko Takano*[1] *and Rainer Breitling*[2,3]

Abstract | As the field of synthetic biology is developing, the prospects for *de novo* design of biosynthetic pathways are becoming more and more realistic. Hence, there is an increasing need for computational tools that can support these efforts. A range of algorithms has been developed that can be used to identify all possible metabolic pathways and their corresponding enzymatic parts. These can then be ranked according to various properties and modelled in an organism-specific context. Finally, design software can aid the biologist in the integration of a selected pathway into smartly regulated transcriptional units. Here, we review key existing tools and offer suggestions for how informatics can help to shape the future of synthetic microbiology.

Parts
Basic building blocks that can be incorporated into a design in synthetic biology; for example, a ribosome binding site, promoter or enzyme coding sequence.

Biosynthetic pathway
Sequence of enzymatically catalysed reactions that convert one or more source metabolites into a product compound.

[1]Department of Microbial Physiology and [2]Groningen Bioinformatics Centre, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Nijenborgh 7, 9747 AG, Groningen, The Netherlands.
[3]Institute of Molecular, Cell and Systems Biology, College of Medical, Veterinary and Life Sciences, Joseph Black Building, University of Glasgow, Glasgow G12 8QQ, UK.
Correspondence to E.T.
e-mail: e.takano@rug.nl

doi:10.1038/nrmicro2717
Published online
23 January 2012

A key promise of synthetic biology is the possibility to customize the metabolic system of microorganisms for the commercial production of a wide range of high-value biofuels[1,2] or natural products[3–5]. Pathways for the production of alcohols, biodiesels, polyketides and terpenoids have successfully been constructed by introducing combinations of parts from various origins into a bacterial host that is easy to cultivate[1,6–10]. Potentially, entire metabolic pathways can be (re) designed *in silico* and implemented in specialized host organisms[11–15]. Successes obtained in pioneering work on the antimalarial drug artemisinin[16–18] suggest that such approaches can be very fruitful. A biosynthetic pathway towards this compound was successfully engineered in *Saccharomyces cerevisiae* and *Escherichia coli* (BOX 1), and this pathway has the potential to enable much more cost-effective production of this important drug compared to the costly and laborious process of harvesting it from the source plant *Artemisia annua*.

The experimental work involved in engineering a synthetic pathway is considerable, and even systematically planned experiments are usually accompanied by much trial and error. When conceiving the design of a novel biosynthetic pathway (FIG. 1), the synthetic biologist has to find optimal solutions for selecting pathways, enzymes or host organisms from an abundance of possibilities. In this Review, we explore how the use of powerful computational tools (TABLE 1) can lead to better-informed and more rapid design and implementation of novel pathways, and we propose ways in which tools from different fields of computation can be linked together effectively. We discuss the different methodologies for identifying all possible metabolic pathways that can lead to the synthesis of a compound of choice, and how to rank these pathways based on various criteria. Subsequently, we consider how flux balance analysis of pathways can be applied to identify the most suitable candidate host organisms. We also examine how to effectively search sequence databases to obtain a list of candidate parts (such as genes and operons) for the execution of each step in the proposed pathway. Finally, we discuss how computational methods can aid in refactoring these parts and integrating them into well-designed transcriptional units that are optimized for a specific host organism.

For specific case studies and more detailed explanations on the inner workings of each of the computational methods, we refer the reader to a range of excellent specialist reviews that have been published recently[15,19–21].

Prediction and prioritization of possible pathways

For compounds of biotechnological value, often only a single specific biosynthetic pathway has been characterized. The key promise of the synthetic biology approach to pathway design is, however, that one does not remain limited to biosynthetic routes that already exist in nature. Instead, realistic biosynthetic pathways can, for instance, be constructed from first principles to optimize their thermodynamic efficiency.

During the past decade, a range of computational pathway prediction algorithms has been generated that can aid in pathway (re)design. Some predictors focus on changing existing pathways through making

Box 1 | Experimental successes in synthetic pathway engineering

Several pioneering experimental efforts in the construction of synthetic pathways have highlighted the potential of the field. Arguably the most famous example is that of artemisinin, a potent antimalaria drug that is naturally produced by the plant *Artemisia annua*. As large-scale production of the compound from plant biomass is very difficult, synthetic biologists instigated a project to engineer its biosynthetic pathway in the bacterium *Escherichia coli*. In 2003, researchers succeeded in introducing a yeast-derived pathway for the production of the isoprenoid precursors of artemisinin in *E. coli*[16]. Later, they also succeeded in developing a synthetic pathway consisting of plant- and microorganism-derived enzymes that was capable of producing artemisinic acid (which can be converted into artemisinin in just two chemical steps) at high titres in *E. coli* and *Saccharomyces cerevisiae*[17,18,100]. In a largely similar fashion, others have successfully introduced a plant-derived pathway to produce taxadiene, the first committed intermediate for the anticancer drug taxol, in *E. coli*[10]. After carefully balancing the expression of the heterologous pathway and the native pathway producing the necessary isoprenoid precursors, production levels were increased by more than 10,000-fold.

Another elegant early example of synthetic engineering of biosynthetic pathways is displayed in the work of Müller *et al.*[101], who engineered a pathway for the biosynthesis of D-hydroxyphenylglycine, an important building block for the side chain of semi-synthetic penicillins and cephalosporins. They combined a hydroxymandelate synthase and a hydroxymandelate oxidase from *Streptomyces coelicolor* and *Amycolatopsis orientalis* with a stereo-inverting hydroxyphenylglycine aminotransferase from *Pseudomonas putida*. Although the yields that were obtained initially were not very high, the results highlighted the potential of combining enzymes from various biological sources into a novel pathway.

Regarding biofuel production, synthetic pathways for re-routing bacterial native metabolism towards the production of isopropanol and higher alcohols were introduced into *E. coli* in a similar fashion by testing enzymes from a range of different organisms (including engineered versions of native enzymes) and finally expressing the combination that had been tested to result in the highest yields[6,7]. More elaborate synthetic approaches, which also entailed the redesign of specific transcriptional units and simple regulatory circuits in combination with introducing enzymes from other microorganisms, later led to the production of biodiesels and waxes in *E. coli* directly from simple sugars[1].

Perhaps even more intriguingly, Bayer *et al.*[102] engineered an efficient pathway for the synthesis of methyl halides in a fully fledged synthetic manner. They selected all 89 putative homologues of the enzyme methyl halide transferase from bacteria, plants, fungi and archaea that were identified by a BLAST search on the entire NCBI sequence database. Subsequently, they designed codon-optimized versions of all of them using Gene Designer. Finally, they used these to chemically synthesize a synthetic gene library that could be tested to find the enzyme that performed the desired function most effectively in the host strain, which resulted in production titres of up to 190 mg l^{-1} h^{-1}.

Dunlop *et al.*[103] engineered microbial biofuel export and tolerance by creating a similar synthetic library of hydrophobe/amphiphile efflux transporters. In addition to a simple BLAST homology search, they used substrate specificity predictions based on the specificity-determining regions of the transporters to generate a subset of 43 homologues that represented a uniform distribution across all candidates. In this study, the genes in the synthetic library were not codon-optimized, and this could be the reason that the native *E. coli* gene still ranked highest in a large number of assays performed. An additional codon optimization step could have led to even more impressive results.

Generally, as the engineering aims become more ambitious in the more recent of these examples, a trend towards more prominent application of computational tools is noticeable; as this trend continues, it is likely to lead to a rapid increase in the ease and speed with which efficient synthetic pathways of unprecedented complexity are designed and constructed.

Flux balance analysis
Computational method for analysing a metabolic system under the assumption of metabolic steady state.

KEGG
The Kyoto Encyclopedia of Genes and Genomes, a portal of web databases on genomes, metabolic pathways and enzymes. The KEGG PATHWAY database contains biochemical pathways in the context of the rest of the metabolic network, and the KEGG LIGAND database contains chemical substances and the biochemical reactions that interconvert them.

Enzyme Commission classification
Hierarchical numerical classification of enzymes standardized by the Enzyme Commission; the complete Enzyme Commission number of an enzyme consists of four numbers, separated by dots, that define with increasing detail the enzyme class and subclasses that it belongs to.

knockouts or adding novel enzymes[22]. Other predictors have been built to identify possible metabolic pathways from first principles[23,24] on the basis of possible biotransformations between chemical structures. More recently, several algorithms have been constructed that use more complex search heuristics to find and rank all possible pathways that lead to a desired end compound (FIG. 1; TABLE 1).

Software for metabolic pathway identification and ranking. One accessible and user-friendly system for pathway identification is From Metabolite to Metabolite (FMM), a freely available web service through which one can search possible pathways between known input and output compounds[25]. It combines the KEGG maps and KEGG LIGAND information to form combined pathway maps, identifies the corresponding genes and organisms and gives an output in which different pathways can be compared. The drawbacks of this system are that it is limited to characterized pathways that are present in the (often incomplete) KEGG framework and that it does not give further insight into the practical or thermodynamic feasibility of the pathway. However, the fact that it can quickly give a clear overview of different possible metabolic routes towards a product of interest can make it a convenient starting point for many investigations.

A more advanced method, BNICE[26], predicts novel pathways on the basis of the somewhat broader reaction rules of the Enzyme Commission classification system. Because BNICE is not restricted to entries from a specific database, it can also predict unknown pathways that are potentially chemically feasible. In its search for pathways, it takes into account the starting compound and/or product, the requested length of the pathway and the range of reactions searched over. The last criterion means that one can choose to only search for a pathway using enzyme reactions out of one known pathway, a combination of multiple pathways or the whole metabolic network. This makes it easy to perform a more targeted search for shorter, more efficient pathways. BNICE can be a good first step in finding possible pathways, but a lot of subsequent analysis of the results is needed to obtain a useful outcome. This is a

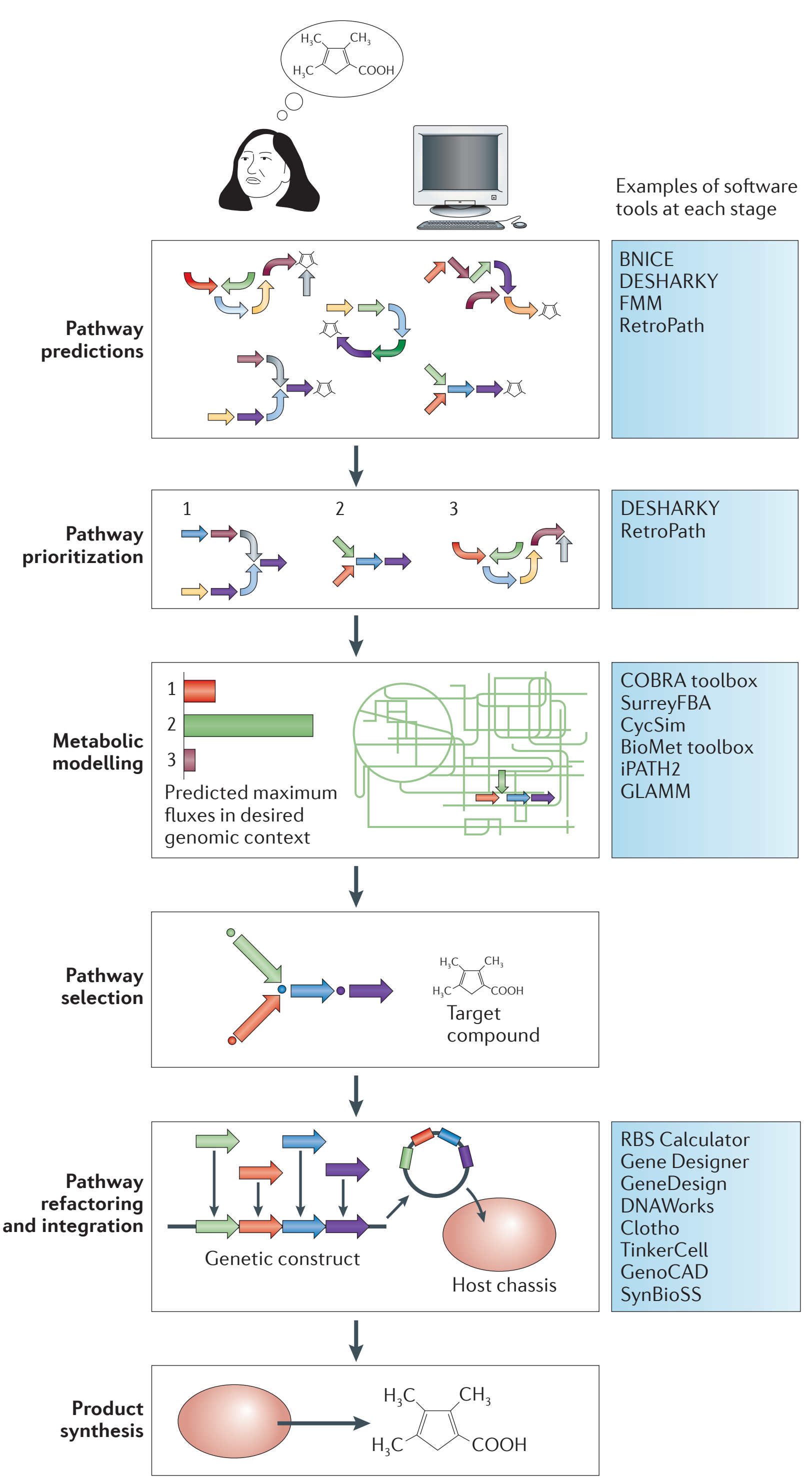


Figure 1 | **Generalized workflow for *de novo* engineering of biosynthetic pathways, from initial idea to final product.** First, a large number of possible pathways is predicted based on chemical reaction rules and/or metabolic maps. Subsequently, the resulting pathways are prioritized based on a number of criteria. Comparative modelling is then performed to predict theoretical production capacities of candidate host organisms for the compounds using the different pathways in the context of the topology of their metabolic network. Finally, one or a few suitable pathways are selected for which synthetic expression constructs can be designed. A diversity of computational tools is essential for all steps, as indicated schematically at the right side of the figure. The order of the steps is not necessarily linear; various iterations and feedback loops between the steps may be necessary for optimization, as sometimes information obtained in a 'later' stage suggests revision of an 'earlier' decision.

conscious choice of the developers: because the BNICE framework is restricted to the first analysis steps, it can be applied to various different kinds of investigations, including not only the engineering of novel pathways but also the analysis and retrosynthesis of metabolic pathways[27].

In some searches, BNICE predicts more than 10,000 different pathways for the biosynthesis or degradation of a certain compound owing to the few criteria the system relies on. Therefore, it is of paramount importance to not only predict possible pathways but also to rank them based on discriminative criteria. Recently, Henry *et al.*[28] have pioneered a prioritization approach in the BNICE framework by ranking novel 3-hydroxypropanoate biosynthesis pathways by thermodynamic feasibility, pathway length, maximum achievable yield and maximum achievable activity. In this manner, they could obtain an informative ranking of the otherwise randomly ordered list of thousands of predicted pathways. Interestingly, the currently commercially used pathway was among the top four pathways in the ranked list, but it was matched (and in some aspects exceeded) by three novel pathways that could provide interesting alternative designs for industrial implementation.

Another prediction system, DESHARKY, is based on enzymatic reactions, but approaches the search for novel pathways quite differently[29]. The major difference concerns the choice of host organism for the pathway, which for DESHARKY is the first step of the pathway prediction after compound design. The algorithm searches for all possible pathways that connect the metabolic network of the organism to a target compound, after which the thermodynamic favourability and the energy loss in transcription and translation are calculated. The tool is most useful if the host organism of choice has already been determined and one needs to search for the pathway that will work most efficiently in that organism for generating a certain compound.

Cho *et al.*[30] have constructed a unified system that both predicts and ranks pathways. Starting with a database of reactions categorized by type and a database of reaction rules describing different reactions, the system first predicts a wide range of possible pathways. A ranking algorithm then prioritizes the pathways based on binding site covalence (similarity of reactions in terms of chemical structure changes), chemical similarity, thermodynamic favourability, pathway distance and organism specificity.

However, rankings only address the symptoms, not the cause of the explosion of possibilities — devising a method to flexibly constrain the search space would be a major breakthrough. The recently published web server RetroPath[31] offers such a principled way to manually determine the strictness of the initial search for reactions. It searches based on molecular signatures of the compounds and reactions involved. Each signature has different 'heights' that correspond to levels of structural detail. By varying the height, one can retrieve numbers of reactions varying from the large numbers of reactions found using BNICE to the small number

Table 1 | **Key computational tools currently available for pathway construction**

System	Description	Refs
Pathway prediction		
BNICE	Biochemical Network Integrated Computational Explorer; framework for identification and thermodynamic assessment of all possible pathways for the degradation or production of a given compound	26
System of Cho *et al.*	Framework for identification and prioritization of biosynthetic pathways for the synthesis of a user-specified chemical	30
DESHARKY	Pathway identification algorithm. Identifies pathways that best match up to the native metabolic network of a specific host and provides the user with amino acid sequences of corresponding enzymes from phylogenetically closely related organisms	29
RetroPath	Web server hosting a unified framework for retrosynthetic pathway design, integrating pathway prediction and ranking, prediction of compatibility with host genes, toxicity prediction and metabolic modelling	31
FMM	From Metabolite to Metabolite; web server that finds biosynthetic routes between two metabolites within the KEGG database	25
CarbonSearch	Algorithm that identifies pathways within existing metabolic networks by tracking the conservation of atoms moving through them	109
OptStrain	Computational framework that advises on optimization of the host's metabolic network to add a particular metabolic pathway by adding or deleting reactions	22
Parts identification		
Registry of Standard Biological Parts	Massachusetts Institute of Technology parts registry, containing various types of biological parts such as promoters, RBSs, transcriptional terminators and plasmids; the registry mostly contains parts collected during iGEM competitions	
Standard Biological Parts knowledgebase	Knowledgebase with parts (including all parts from the Registry of Standard Biological Parts) that have been transformed into Synthetic Biology Open Language to make the information computable	60
IMG	Integrated Microbial Genomes; environment for the comparative and evolutionary analysis of microbial genomes, including gene neighbourhood orthology searches	71
antiSMASH	Identification, annotation and comparative analysis of secondary metabolite biosynthesis gene clusters	72
KEGG	Key collection of metabolite and metabolic pathway databases; includes organism-specific and general maps of metabolic pathways and networks, gene–enzyme associations, orthology information and more	110
ASC	Active Site Classification; uses a protein structure to find residues near the active site of enzymes, which it uses to construct support vector machines that classify subclasses (for example, substrate specificities) of enzymes within an enzyme family	69
Parts refactoring and synthesis		
RBS Calculator	Automated design of RBSs based on a thermodynamic model of transcription initiation	81
RBSDesigner	Algorithm for prediction of mRNA translation efficiencies, as well as design of RBSs for a desired protein expression level	83
Gene Designer 2.0	Software package for gene, operon and vector design, codon optimization and primer design	75
GeneDesign	Web server with algorithms ranging from codon optimization and codon bias graphing to insertion of restriction sites into a protein-coding nucleotide sequence and designing building blocks based on restriction site overlaps	74
Gene Composer	Commercial software suite for genetic construct design, codon optimization and gene assembly	
Optimizer	Web server that performs codon optimization on an input protein-coding DNA sequence using a codon usage table	73
DNAWorks	Web server for oligonucleotide design for PCR-based gene synthesis, with integrated codon optimization	84
TmPrime	Web server for oligonucleotide design for PCR-based gene synthesis, with integrated codon optimization	85
CloneQC	Web application for controlling the quality of sequenced clones by detecting errors in DNA synthesis	111

Table 1 (cont.) | **Key computational tools currently available for pathway construction**

System	Description	Refs
Pathway and circuit design software packages		
Biojade	Software tool for design and simulation of genetic circuits	
Clotho	Flexible interface for synthetic biological systems design; within the interface, a range of apps/plugins can be utilized to import, view, edit and share DNA parts and system designs	
TinkerCell	CAD software that allows drag-and-drop drawing and simulation of biological systems	112
Asmparts	Computational tool that generates models of biological systems by assembling models of parts	113
GenoCAD	CAD software for design of multigene DNA sequences, with the option to assist the user through interactive 'grammar checking' of the design drafts	76
WebGEC	Web simulator from Microsoft for genetic circuit design and testing	
SynBioSS	Software suite for designing, modelling and simulating synthetic genetic constructs; the SynBioSS Designer can be used to transform a sequence of BioBricks (from the Registry of Standard Biological Parts) or other parts into a model that can be simulated in the SynBioSS Desktop Simulator	79
CellDesigner	Editor for graphical drawing of regulatory and biochemical networks that can be stored in Systems Biology Markup Language (SBML)	
BioNetCAD	CellDesigner plug-in for CAD and simulation of biochemical networks	
Metabolic modelling/flux balance analysis		
COBRA Toolbox	Standard toolbox for metabolic modelling and FBA	51
SurreyFBA	Command-line tool and graphical user interface for constraint-based modelling of genome-scale networks	52
CycSim	Web server for analysing genome-scale metabolic models; includes enzyme knockout simulations	53
BioMet Toolbox	Web toolbox for analysing genome-scale metabolic models; includes gene knockout analysis, flux optimization and more	54
iPATH2	Interactive visualization of data on metabolic pathways; items on KEGG-based metabolic maps can be coloured based on the user's preferences	58
GLAMM	Interactive visualization of data on metabolic pathways; can use host-specific metabolic networks and allows detection of pathways within a network	59

CAD, computer-aided design; RBS, ribosome binding site.

of original reactions that are present in the KEGG database. RetroPath also hosts several interesting additional features for filtering and ranking, such as predictions of promiscuous activity of enzymes, predictions of the compatibility between host and heterologous genes, and compound toxicity predictions.

Once a pathway has been selected for introduction into a specific host bacterium, the consequences of this manipulation have to be predicted in the new metabolic context. One system that aims to monitor the effects of the new pathway on the host is OptStrain[22] — a method that uses flux analysis to give advice on how production could be optimized by altering the host's gene expression. After constructing a hypothetical biosynthetic pathway towards the target compound, the OptStrain system changes the pathway in such a way that as many enzymes from the pathway as possible are native to the host organism. With the use of a purely stoichiometric model of the host's metabolic network, OptStrain then predicts the effect of novel enzymes in the pathway, as well as which host genes should be up- or downregulated in order to increase the production yield.

Criteria for ranking pathways. In *de novo* pathway engineering, it may sometimes be desirable to search for pathways before choosing a suitable host. Therefore, the best pathways have to be chosen on the basis of a few theoretical criteria. Not everyone agrees on the criteria to be used for this process of prioritization (FIG. 1). As mentioned above, Cho *et al.*[30] use five criteria, including organism specificity and pathway distance. According to extreme pathway analysis by Papin *et al.*[32], the length of the pathway does not influence production rate. Even so, the energetic costs of producing more enzymes should be taken into account when considering longer pathways, which still makes pathway distance a relevant parameter. Organism specificity can only be an applicable criterion if a host has already been chosen. It can be disputed whether it is more desirable that the pathway consists of enzymes specific for one particular organism; testing multiple combinations of enzymes from different organisms selected for experimentally tested activities is actually likely to yield more effective compound production, as the most catalytically efficient combination of enzymes that is available can then be identified from these. Other criteria can be the theoretically achievable yield and the achievable activity, which are both difficult to predict without analysis of the metabolic network of the host.

One of the few factors that can be determined independently is the theoretical thermodynamic favourability. In most methods, the thermodynamic favourability is measured by a group contribution method, which measures the Gibbs free energy of formation of groups of atoms of the products and intermediates[33]. These groups are added up to a total Gibbs energy of every reaction in the pathway. When the Gibbs energy of formation adds up to a negative value, the reaction is defined as thermodynamically favourable. Cho *et al.*[30] went beyond only using Gibbs free energy of formation by also taking into account the fluctuation of Gibbs energy between the reactions in the pathway: the less fluctuation, the more thermodynamically favourable the pathway, as a product of each reaction in the pathway is a reactant for the next. Less fluctuation of Gibbs free energy along the pathway also reduces the accumulation of intermediates, thus avoiding potential adverse effects on the host organism.

All in all, the algorithms and tools described provide a useful toolbox for the synthetic biologist. By smartly combining the advantages of several tools, the predictions have the potential to make possible the exploration of pathways that are chemically more versatile and hopefully at least as effective as those found in nature.

Metabolic modelling in candidate host organisms

One cannot predict, construct and investigate a new metabolic pathway without taking the host organism into account, as every new pathway has to take its place in the overall topology of a large native metabolic network[34]. Competition with native pathways and metabolites, unpredicted side products and feedback loops are only some of the possible effects that the context of the host organism can have on the new pathway.

One approach for finding a suitable host organism for a pathway is to look for an organism that already has most of the enzymes from the designed pathway present in its native metabolic network. In this way, fewer enzymes would have to be introduced into the organism and thereby the metabolic network would be disturbed less. This idea has been incorporated into a ranking system designed by Cho *et al.*[30] that prefers pathways that consist of many enzymes originating from the same organism. However, there are also reasons why using a host with many usable native enzymes may not be the best option. For example, it would then also be more likely that in the given organism, more pathways exist that compete with the pathway that is to be introduced. Partially knocking out such native pathways can be a solution to this problem if they are not essential to the host. Algorithms such as OptKnock and OptFlux[35,36] are available to supply the researcher with suggestions on which pathway to knock out on the basis of metabolic flux simulations. More ambitiously, orthogonal synthetic systems — for example, for transcription or translation — can be used to insulate the synthetic pathways even more. Unnatural DNA and amino acids are now available to keep the production machinery of the desired compound separate from the host at many levels[37–41], although the fundamental metabolic link will still need to be maintained.

Genome-scale metabolic modelling. One of the most important computational developments that is likely to facilitate true *de novo* pathway engineering in the future is the development of genome-scale metabolic models[19,42–44]. Such models allow *in silico* prediction of the behaviour of a pathway in a candidate host organism using constraint-based flux balance analysis[45,46]. In this approach, a steady-state flux distribution of the metabolic network is predicted based on the stoichiometry of each reaction, mass–balance constraints and an objective function that specifies for which goal the fluxes are optimized. In traditional studies, the objective function is often the maximization of biomass production, whereas in the analysis of genetically engineered microorganisms, a common alternative objective is the ‘minimization of metabolic adjustments’ (MOMA): that is, finding the flux distribution in the engineered strain that is closest to the wild-type situation, but still feasible for engineering[47]. Flux balance analysis is often used effectively to increase product titres. For example, vanillin production in baker’s yeast was increased twofold by selecting genes for knockout constructs on the basis of flux balance analysis simulations[48]. In a similar fashion, Asadollahi *et al.*[49] achieved 85% titre improvement after constructing model-guided knockouts. The more recently developed Thermodynamic Metabolic Flux Analysis (TMFA)[50] adds further thermodynamic constraints — which are based on the Gibbs free energy change of each reaction and the concentrations of metabolites — to identify only thermodynamically feasible flux distributions and thereby increase the predictive power.

The MATLAB-based COBRA Toolbox[51], which has become a near-standard in the field, can be used to perform TMFA, but it can also be performed with user-friendly web servers or graphical user interfaces (GUIs), such as SurreyFBA[52], CycSim[53] and the BioMet Toolbox[54]. The advantage of CycSim and BioMet is that they are web-based and therefore require no installation, but they are limited to the analysis of a small number of model organisms. The stand-alone tool SurreyFBA may have a somewhat steeper learning curve, but it is more complete and can be accessed both through a GUI and through (scripting from) the command line. One of the major bottlenecks in the use of metabolic models is the low level of standardization of the SBML files that are used to store genome-scale models, sometimes making models incompatible between tools.

As essential as it might be, computational prediction of the effects of introducing a pathway into a host is a procedure that is still in the pioneering stage. In one early example, the previously discussed pathway prediction system BNICE was used to find novel pathways for biodegradation of the pollutant 1,2,4–trichlorobenzene by *Pseudomonas putida*, which is a known pollutant degrader[55]. To predict which pathway was most effective

Metabolic flux
Flow of metabolites through a metabolic system.

Genome-scale metabolic models
Models of all of the enzymatic reactions encoded in a genome, based on the genome annotation.

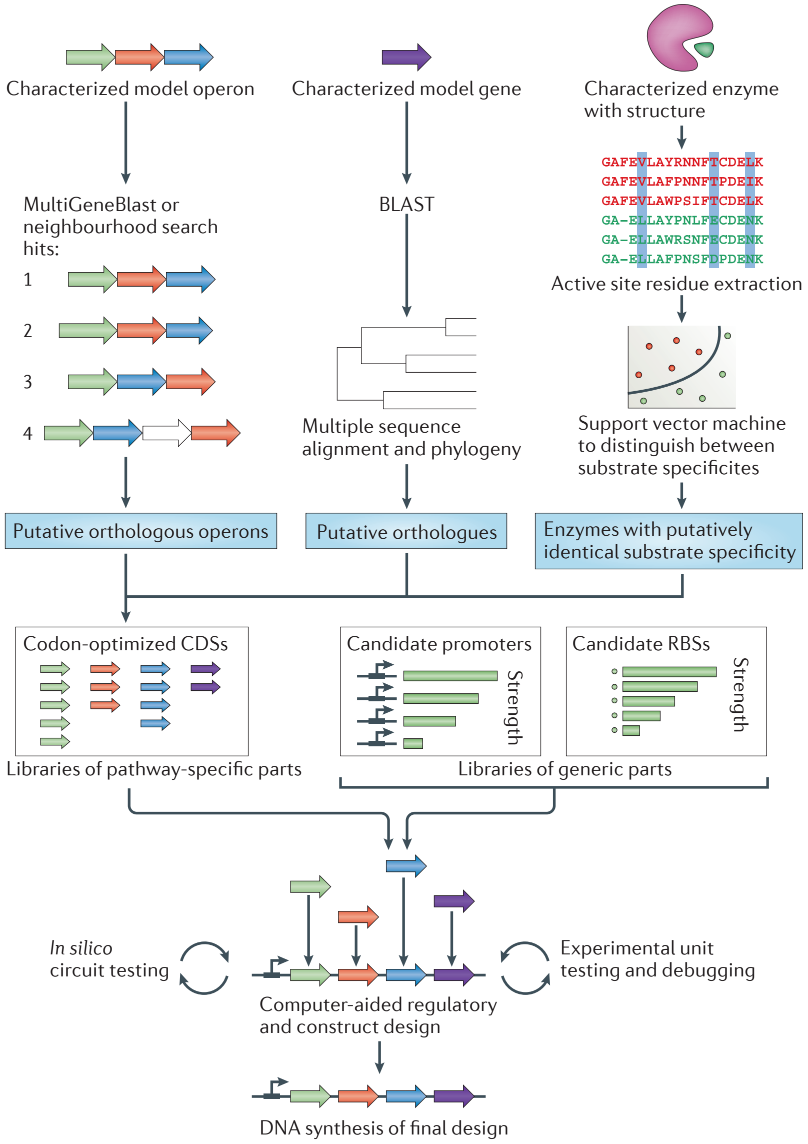


Figure 2 | **Scheme showing the steps involved in the identification of various parts, their refactoring and their integration into transcriptional units.** At the top, three strategies are shown for the identification of libraries of variants of pathway-specific genetic parts on the basis of genetic and biochemical knowledge. The left panel shows how operons or gene clusters that are homologous to a characterized operon or gene cluster can be detected using neighbourhood orthology analysis (as in Integrated Microbial Genomes (IMG)[71]) or MultiGeneBlast analysis (as in antiSMASH[72]). The middle panel shows the well-accepted procedure for the identification of orthologues of a characterized gene by homology search, multiple sequence alignment and phylogenetic tree construction. The right panel shows the method recently pioneered by Röttig *et al.*[69] for identifying enzymes with identical substrate specificity to a model enzyme if this enzyme is part of an enzyme family that contains multiple specificities. The method proceeds by automatic extraction of active site residues on the basis of a crystal structure, training of a support vector machine to distinguish between the different substrate specificity variants in the enzyme family, and classification of all homologues to identify those enzymes that have the desired specificity. Coding sequences (CDSs) of the identified pathway-specific parts are codon-optimized using, for example, Optimizer[73], GeneDesign[74] or Gene Designer[75]. Libraries of generic parts are then acquired. Host-specific promoters are collected, and ribosome binding sites (RBSs) are obtained with the aid of, for example, RBS Calculator[81,82] and/or RBSDesigner[83]. Both generic and pathway-specific parts are then used in computer-aided design of genetic constructs encoding the target biochemical pathway. After extensive *in silico* and *in vivo* testing and debugging, oligonucleotides are designed (for example, using TmPrime[85] or DNAWorks[84]) to synthesize the final design.

and what would be the growth rate of the host with the implemented pathway, TMFA was applied to a genome-scale model of the metabolic network of the host. By implementing thermodynamic constraints in the modelling, the number of candidate pathways was reduced by around 200-fold. This approach seems to be most useful when testing not-yet-existing pathways generated by pathway prediction algorithms.

Around 40 manually curated genome-scale metabolic reconstructions of different bacteria have been published so far[19]. Interestingly, a new approach has recently been published that can automatically generate metabolic models on the basis of genome sequences by annotating them in a uniform fashion, linking predicted enzymes to reactions and filling in gaps[56]. Although the models require subsequent manual curation, this methodology now provides the intriguing option of reconstructing multiple models in parallel and performing comparative analyses between them[57]. The fact that the reconstruction method is uniform removes the large annotator bias that otherwise makes manually constructed models difficult to compare. Such models can be used to predict the suitability of the metabolic network topologies of multiple candidate hosts for the production of a specified compound *in silico* by introducing the pathway into these models, performing TMFA with a dual objective consisting of the biomass and the production of the compound, and comparing the predicted maximized fluxes to the target compound.

The outputs of *in silico* TMFA experiments are often complex and difficult to interpret in the context of the whole metabolic map of an organism. Fortunately, powerful visualization tools have recently been made available that can colour pathways according to predicted fluxes. For example, iPATH2 (REF. 58) can generate coloured KEGG-based metabolic maps on the basis of simple tables of Enzyme Commission numbers and corresponding colours. Interestingly, GLAMM[59], a similar web service, offers the additional possibility to automatically highlight routes between two compounds in the metabolic map. Such visualization tools are indispensable for gaining a rapid understanding of the large output tables that are usually obtained through modelling. Combined with the user-friendly TMFA simulation tools mentioned above, they make this complex technique accessible to the non-specialist experimental microbiologist.

Strategies for parts identification

In order to construct a pathway, one will of course need to find the parts — in the form of gene domains, genes and operons — that can carry out the proposed enzymatic and regulatory steps (FIG. 2). A range of parts is already available in online parts registries[60], and several major efforts are under way to systematically characterize and standardize large numbers of biological parts in a consistent manner[61]. However, these parts registries are currently much more suitable for finding regulatory elements than for finding coding sequences of biosynthetic enzymes, as these parts are much more unique and specified. Therefore, effective genome mining of enzymatic parts is a crucial step in the construction and optimization of biosynthetic pathways. To optimize metabolic fluxes through the candidate pathways, one needs to find the optimal combination of enzymatic parts in terms of individual catalytic efficiency as well as overall pathway reaction stoichiometry. Importantly, these parts do not necessarily have to originate from one existing natural system in one particular organism; instead, one could select a range of candidate parts from many different sources, characterize them (if possible, using some high-throughput assay) and integrate the codon-optimized versions of the most optimal combination of parts into a new pathway (BOX 1). To find combinations of enzymes that cooperate effectively, full exploitation of the genomic databases to make an optimal selection of all available parts is probably necessary.

The optimal computational strategies for harvesting potential parts from the genomic databases differ according to the nature of the parts: that is, whether they are domains, genes or operons.

Identification of genes and domains. Usually, gene domains can most easily be identified by traditional searching for the encoded protein domains with curated-profile hidden Markov models from the PFAM[62], SMART[63] or TIGRFAM databases. This can be done manually using the HMMer package[64] or using the SMART or PFAM architecture searches available on the Internet. After detection of the total set of domains, a high-priority subset can be identified by generating a phylogenetic tree and identifying the branches that specifically contain curated entries of domains with the desired enzymatic activity. If certain active site residues are known to be essential for this activity, their presence can be verified for the entries in multiple-sequence alignments of the selected branches. Alternatively, active sites can first be predicted *ab initio* on the basis of sequence conservation and/or structural information[65,66]. If the desired parts are individual genes, candidates can be identified by a simple BLAST search and analysed according to their phylogeny and active site residues in the same fashion as described above. As an additional criterion, the genomic context of candidate parts may be studied to verify the likelihood of the gene having the desired function in the context of the whole operon in which it is located. Finally, databases such as KEGG can be used to find isoenzymes that might catalyse the same reaction, even if they have no substantial sequence homology. To increase the number of potential parts that can be tested even further, the set of identified genes can be further supplemented by predicted evolutionary intermediates or ancestor genes or domains[67,68] of different taxonomic branches that have a high potential of encoding the desired parts.

In some cases, a manual inspection of active site residues of candidate gene or domain parts may not be sufficient to reliably predict their enzymatic activity. This is especially the case if the desired enzyme is a member of a broader enzyme family that encompasses multiple substrate specificities. More elaborate *in silico* approaches for finding enzymes with the right substrate specificity or for identifying the right mutations to get those enzymes will then be needed. Recently, an automated method was developed for classifying active sites from enzymes throughout an enzyme family using support vector machines trained on sets of residues around the active site of enzymes with known substrate specificities from within the family[69]. The approach has already been successfully implemented to predict the extremely variable substrate specificities of non-ribosomal peptide synthetases[70].

Identification of multigene modules. Often, the desired parts are not single domains or genes but complete operons, such as the biosynthetic operon for a precursor needed to produce the compound of choice. In these cases, one will want to avoid having to combine the results of a large number of individual BLAST searches manually to find the homologous genomic regions. One currently available means of identifying genomic regions that are homologous to a certain model operon is the 'identify homologous regions' tool in the Integrated Microbial Genomes (IMG) system of the Joint Genome Institute[71]. However, an important disadvantage of this tool is that it does not cover all information stored in genomic databases such as GenBank. Additionally, it is difficult to specifically look for overall homologues of genetic elements at the operon or gene cluster levels, as the search is always performed on the neighbourhoods of a single gene. If the biosynthetic operon is always part of a specific type of secondary metabolite biosynthetic gene cluster, the comparative gene cluster analysis module from antiSMASH[72] can be used to find gene clusters that have the same operon. Additionally, to be able to very specifically look for all genomic regions homologous to a given query operon or gene cluster, a new tool, MultiGeneBlast, is currently under development in our group. This tool effectively combines the individual BLAST results of the various genes in the query gene cluster to rank all genomic regions from GenBank on the basis of the number of BLAST hits from the query gene cluster, the conservation of gene order and the cumulative BLAST bit score.

Designing transcriptional units

When raw candidate parts have been screened and an optimal combination has been selected, one will still need to optimize the sequences for the targeted host organisms and combine them into transcriptional units with a well-designed regulatory circuitry (FIG. 2). Several algorithms have been developed to aid in these processes. Additionally, the development of drag-and-drop computer-aided design (CAD) approaches[21] makes the life of the biological designer much easier.

Computer-aided design software. When trying to heterologously express a genetic construct that consists of non-native parts, a crucial factor in obtaining high protein expression rates is optimizing the codon usage in the coding regions by matching it to the more

Support vector machines
Machine learning classification algorithms based on hyperplanes that separate two or more classes in a multidimensional parameter space.

Computer-aided design
(CAD). The use of computers for the design process in engineering, including, for example, functionalities for drafting and simulation.

Codon usage
Genome-specific frequency of occurrence of the different codons that can encode each amino acid.

abundant tRNA species in the host. A range of tools has been developed that can use a codon frequency table of the target host organism to optimize the codon usage of a given protein coding sequence (CDS). Arguably the most straightforward to use is Optimizer[73], a simple web tool that does exactly this. Another web server, GeneDesign[74], offers several further options such as the addition of restriction sites. For advanced users, there is Gene Designer 2.0 (REF. 75), which offers a comprehensive drag-and-drop user interface for the construction of genetic constructs that also include regulatory parts such as promoters, ribosome binding sites (RBSs) and transcriptional terminators. Another CAD tool is GenoCAD[76], which is less comprehensive but generally easier to handle and provides an additional interesting feature: it assists the user in correctly designing genetic constructs by interactively checking them against a user-specified set of grammatical rules (for example, a transcription unit is always composed according to the following pattern: promoter–[RBS–CDS]$_n$–terminator). For example, to design a simple *E. coli* three-gene operon construct with GenoCAD, one would first select the *E. coli* grammar with the parts libraries that contain parts matching this grammar. Then one would arrange the desired architecture in a way that complies with the grammar (promoter–RBS–CDS–RBS–CDS–RBS–CDS–terminator), and finally select the specific parts for each entry from the libraries.

Regulatory circuits are crucial for designing well-balanced genetically engineered machines, and many synthetic biology efforts have focused on developing these[77,78]. Several CAD tools are available that have integrated simulation capacities that can be used to predict whether the envisioned regulatory mechanisms will function as intended. SynBioSS[79], for example, simulates regulatory circuits by creating network models of reactions that represent transcription, translation, *cis* and *trans* regulatory effects, and degradation. Of specific interest for pathway design is BioNetCad[80], which allows a similar kind of simulation of the logic inherent in biochemical networks consisting of enzymes and metabolites.

Designing regulatory parts. Optimized regulatory parts for use in genetic circuits and constructs are unlikely to be already available and will usually have to be designed for the intended usage. Synthesizing and characterizing host-specific libraries of characterized RBSs and sigma-factor binding sites (SFBSs) will be useful in many cases, as regulatory parts with the correct binding strength can then be selected from these for incorporation into the genetic construct. However, even if this is done, the strength of an RBS (and probably of an SFBS as well) partially depends on its DNA sequence context[81]. Therefore, some RBSs may not function as expected in their new context. To find a suitable RBS in such cases — or in cases when no RBS library is available — the online tools RBS Calculator[81,82] or RBSDesigner[83] can be used to suggest an RBS sequence for a given desired translation initiation rate. These algorithms are based on thermodynamic models of the molecular interactions between the ribosome complex and mRNA transcripts, and use these to predict a translation initiation rate. Although RBS Calculator does not necessarily succeed in reaching a global optimum, it enables rapid design of RBS sequences that are sufficient for most purposes. It would be helpful if a similar tool could be generated for SFBS design in the near future.

DNA synthesis and integration. When an adequate design has been obtained, the next step will be the synthesis of the DNA parts. Online algorithms — notably DNAWorks[84] and TmPrime[85] — are available for designing oligonucleotides for PCR-based gene synthesis, and such algorithms also include integrated codon optimization functionalities. Optimization can be important for the efficient heterologous expression of a gene; for example, in terpenoid-producing *E. coli* strains, Chang *et al.*[18] reported a 2.5-fold increase in production titres after codon optimization. Several recent developments in DNA synthesis[86–88] now allow DNA synthesis in a throughput that is sufficient even to construct thousands of variants of any DNA parts, which can then be tested to pinpoint the ones that function most efficiently. These developments provide new opportunities for computational tools: for example, software could be designed to construct libraries containing part variants that optimally cover the relevant areas of sequence space.

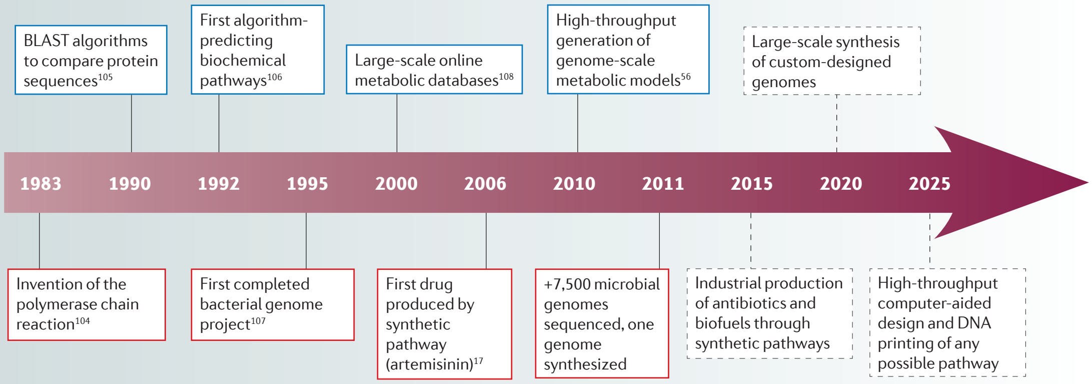
In order to arrive at a functional design for the DNA construct encoding a biosynthetic pathway, important lessons can be learned from programming. For example, 'unit testing' — the functional testing and debugging of every individual component before putting everything together — is crucial for keeping the complexity of the debugging process manageable. Fluorescent or other biological signals may be introduced into the construct to mimic the 'print statement' that is often used in computer software debugging to test the successful execution of a programmatic unit. When one has arrived at a functional design, it can be inserted into the chromosome of a specified plug-and-play host[4] or in a multigene expression plasmid[89].

Future perspectives

In general, the ability to computationally predict pathways, identify variants of the necessary parts and model them in genome-scale metabolic networks of candidate host organisms offers great promise for accelerating the developing field of synthetic pathway engineering. In this area, systems biology can inform and complement synthetic biology approaches, which could lead to important breakthroughs in the near future (TIMELINE).

The ambition of synthetic biology is to design biological systems on the basis of first principles, irrespective of which combinations of parts happen to be used in nature. The algorithms that have been developed in recent years are likely to facilitate reaching this goal. However, owing to the immense challenge of biological

Timeline | **Important events in the development of *de novo* pathway engineering**



1983 – Invention of the polymerase chain reaction[104]

1990 – BLAST algorithms to compare protein sequences[105]

1992 – First algorithm-predicting biochemical pathways[106]

1995 – First completed bacterial genome project[107]

2000 – Large-scale online metabolic databases[108]

2006 – First drug produced by synthetic pathway (artemisinin)[17]

2010 – High-throughput generation of genome-scale metabolic models[56]

2011 – +7,500 microbial genomes sequenced, one genome synthesized

2015 – Industrial production of antibiotics and biofuels through synthetic pathways

2020 – Large-scale synthesis of custom-designed genomes

2025 – High-throughput computer-aided design and DNA printing of any possible pathway

Computational milestones are highlighted in blue, experimental milestones in red, and all future milestones have a dashed box.

complexity[90], the development of additional algorithms specifically focused on the needs of synthetic biology projects will be crucial to allow the field to mature.

In all key aspects of synthetic microbiology, the speed of progress will largely depend on the headway made in software development and data systematization. Fortunately, there are great opportunities for computational breakthroughs all around.

In the field of metabolic modelling, the advances in high-throughput generation of genome-scale models[56] are likely to inaugurate a whole new era. A field of comparative modelling[57] will develop in which the metabolic network topologies from a range of genome-scale models can be compared rapidly. After algorithms have been developed to automatically add one or a few enzymes to each model to simulate growth on a certain growth medium, the approach will be a useful way to compare organisms across a range of medium types in terms of the suitability of their network topologies for growth and productivity. Also, when a specific host organism has already been chosen, the same algorithm could be implemented to aid in medium design. Ideally, the design of synthetic pathways and growth media would be an integrated process in which the pathways are supplemented with the enzymes necessary to arrive at an optimal combination of medium and pathway. The field of metabolomics, which has its own crucial branch of software development[91], will be key in coupling the predictions to actual experimental measurements.

The available pathway prediction tools will need to be made more user-friendly (with GUIs or web servers) and linked to well-curated databases of experimentally characterized enzymatic parts in an integrated framework. Over the past decades, an enormous number of enzymes have been characterized that are involved in the synthesis or tailoring of small molecule scaffolds[5], but little systematic data archiving has been performed thus far. For such purposes, databases such as KEGG provide sparse information, being largely focused on primary metabolism. Besides enzymes, it would be helpful if transporters for small molecules and resistance genes against toxins or antibiotics were also categorized systematically in a database and linked to the chemical structure of the corresponding small molecule. Algorithms for chemical (sub)structure similarity searching[92,93] could then be used to rapidly search for enzymes or transporters that are likely to synthesize or transport a compound of choice, or that would otherwise have a functionality that is related closely enough to be modified successfully using, for example, directed evolution.

It is also crucial that computational developments go hand-in-hand with developing experimental approaches. High-throughput transcription factor binding affinity characterization using protein binding microarrays[94] or microfluidics[95] could, for example, be perfectly linked up to algorithms for the design of transcription factor binding sites and SFBSs[96]. Alternatively, when a close homologue of a desired transcription factor has already been characterized in a related species, phylogenetic footprinting approaches[97] could be used to automatically predict species-specific binding motifs of the orthologue that one wants to use. Yet when a synthetic transcription factor is heterologously expressed, algorithms would probably still be necessary to correct for the difference in GC content between the species, which is likely to influence the heterologous binding dynamics on a chromosome that is foreign to those for which experimental data have been obtained. Finally, with the expected advances in synthetic genomics, the development of algorithms for the optimal integration of genetic constructs into a chromosomal design will be important for arranging both operon-level genetic organization[98] and higher-level organization[99] to achieve optimal gene expression levels.

When computational and experimental breakthroughs thus go hand-in-hand in an integrated manner, synthetic pathway engineering may well become one of the major drivers of applied synthetic biology.

Orthologue
Orthologues are genes with similar functions that derive from a common ancestor through vertical descent.

1. Steen, E. J. *et al.* Microbial production of fatty-acid-derived fuels and chemicals from plant biomass. *Nature* **463**, 559–562 (2010).
2. Bond-Watts, B. B., Bellerose, R. J. & Chang, M. C. Enzyme mechanism as a kinetic control element for designing synthetic biofuel pathways. *Nature Chem. Biol.* **7**, 222–227 (2011).
This article reports the engineering of biosynthetic pathways that convert simple sugars into biofuels in *E. coli*.
3. Khalil, A. S. & Collins, J. J. Synthetic biology: applications come of age. *Nature Rev. Genet.* **11**, 367–379 (2010).
4. Medema, M. H., Breitling, R., Bovenberg, R. & Takano, E. Exploiting plug-and-play synthetic biology for drug discovery and production in microorganisms. *Nature Rev. Microbiol.* **9**, 131–137 (2011).
References 3 and 4 describe the ways in which synthetic biology can be applied practically for drug discovery, biofuel production and biomaterial production.
5. Walsh, C. T. & Fischbach, M. A. Natural products version 2.0: connecting genes to molecules. *J. Am. Chem. Soc.* **132**, 2469–2493 (2010).
6. Hanai, T., Atsumi, S. & Liao, J. C. Engineered synthetic pathway for isopropanol production in *Escherichia coli*. *Appl. Environ. Microbiol.* **73**, 7814–7818 (2007).
7. Atsumi, S., Hanai, T. & Liao, J. C. Non-fermentative pathways for synthesis of branched-chain higher alcohols as biofuels. *Nature* **451**, 86–89 (2008).
8. Zhang, W., Li, Y. & Tang, Y. Engineered biosynthesis of bacterial aromatic polyketides in *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **105**, 20683–20688 (2008).
9. Menzella, H. G. *et al.* Redesign, synthesis and functional expression of the 6-deoxyerythronolide B polyketide synthase gene cluster. *J. Ind. Microbiol. Biotechnol.* **33**, 22–28 (2006).
10. Ajikumar, P. K. *et al.* Isoprenoid pathway optimization for Taxol precursor overproduction in *Escherichia coli*. *Science* **330**, 70–74 (2010).
11. Prather, K. L. J. & Martin, C. H. *De novo* biosynthetic pathways: rational design of microbial chemical factories. *Curr. Opin. Biotechnol.* **19**, 468–474 (2008).
12. Martin, C. H., Nielsen, D. R., Solomon, K. V. & Prather, K. L. Synthetic metabolism: engineering biology at the protein and pathway scales. *Chem. Biol.* **16**, 277–286 (2009).
References 11 and 12 review the principles and ideas that are key to the engineering of synthetic pathways.
13. Tyo, K. E., Kocharin, K. & Nielsen, J. Toward design-based engineering of industrial microbes. *Curr. Opin. Microbiol.* **13**, 255–262 (2010).
14. Weeks, A. M. & Chang, M. C. Constructing *de novo* biosynthetic pathways for chemical synthesis inside living cells. *Biochemistry* **50**, 5404–5418 (2011).
15. Soh, K. C. & Hatzimanikatis, V. DREAMS of metabolism. *Trends Biotechnol.* **28**, 501–508 (2010).
16. Martin, V. J. J., Pitera, D. J., Withers, S. T., Newman, J. D. & Keasling, J. D. Engineering a mevalonate pathway in *Escherichia coli* for production of terpenoids. *Nature Biotech.* **21**, 796–802 (2003).
17. Ro, D. K. *et al.* Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature* **440**, 940–943 (2006).
18. Chang, M. C., Eachus, R. A., Trieu, W., Ro, D. K. & Keasling, J. D. Engineering *Escherichia coli* for production of functionalized terpenoids using plant P450s. *Nature Chem. Biol.* **3**, 274–277 (2007).
19. Feist, A. M., Herrgard, M. J., Thiele, I., Reed, J. L. & Palsson, B. Ø. Reconstruction of biochemical networks in microorganisms. *Nature Rev. Microbiol.* **7**, 129–143 (2009).
20. Xu, D. Computational methods for protein sequence comparison and search. *Curr. Protoc. Protein Sci.* Ch. 2, Unit 2.1 (2009).
21. Marchisio, M. A. & Stelling, J. Computational design tools for synthetic biology. *Curr. Opin. Biotechnol.* **20**, 479–485 (2009).
22. Pharkya, P., Burgard, A. P. & Maranas, C. D. OptStrain: a computational framework for redesign of microbial production systems. *Genome Res.* **14**, 2367–2376 (2004).
23. McShan, D. C., Rao, S. & Shah, I. PathMiner: predicting metabolic pathways by heuristic search. *Bioinformatics* **19**, 1692–1698 (2003).
24. Hou, B. K., Wackett, L. P. & Ellis, L. B. M. Microbial pathway prediction: a functional group approach. *J. Chem. Inf. Comput. Sci.* **43**, 1051–1057 (2003).
25. Chou, C., Chang, W., Chiu, C., Huang, C. & Huang, H. FMM: a web server for metabolic pathway reconstruction and comparative analysis. *Nucleic Acids Res.* **37**, W129–W134 (2009).
26. Hatzimanikatis, V. *et al.* Exploring the diversity of complex metabolic networks. *Bioinformatics* **21**, 1603–1609 (2005).
Description of the original BNICE framework for metabolic pathway identification.
27. Bachmann, B. O. Biosynthesis: is it time to go retro? *Nature Chem. Biol.* **6**, 390–393 (2010).
28. Henry, C. S., Broadbelt, L. J. & Hatzimanikatis, V. Discovery and analysis of novel metabolic pathways for the biosynthesis of industrial chemicals: 3-hydroxypropanoate. *Biotechnol. Bioeng.* **106**, 462–473 (2010).
29. Rodrigo, G., Carrera, J., Prather, K. J. & Jaramillo, A. DESHARKY: automatic design of metabolic pathways for optimal cell growth. *Bioinformatics* **24**, 2554–2556 (2008).
30. Cho, A., Yun, H., Park, J., Lee, S. & Park, S. Prediction of novel synthetic pathways for the production of desired chemicals. *BMC Systems Biol.* **4**, 35 (2010).
31. Carbonell, P., Planson, A. G., Fichera, D. & Faulon, J. L. A retrosynthetic biology approach to metabolic pathway design for therapeutic production. *BMC Syst. Biol.* **5**, 122 (2011).
32. Papin, J. A., Price, N. D. & Palsson, B. Ø. Extreme pathway lengths and reaction participation in genome-scale metabolic networks. *Genome Res.* **12**, 1889–1900 (2002).
33. Jankowski, M. D., Henry, C. S., Broadbelt, L. J. & Hatzimanikatis, V. Group contribution method for thermodynamic analysis of complex metabolic networks. *Biophys. J.* **95**, 1487–1499 (2008).
34. Breitling, R., Vitkup, D. & Barrett, M. P. New surveyor tools for charting microbial metabolic maps. *Nature Rev. Microbiol.* **6**, 156–161 (2008).
35. Burgard, A. P., Pharkya, P. & Maranas, C. D. Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol. Bioeng.* **84**, 647–657 (2003).
36. Rocha, I. *et al.* OptFlux: an open-source software platform for *in silico* metabolic engineering. *BMC Syst. Biol.* **4**, 45 (2010).
37. Neumann, H., Wang, K., Davis, L., Garcia-Alai, M. & Chin, J. W. Encoding multiple unnatural amino acids via evolution of a quadruplet-decoding ribosome. *Nature* **464**, 441–444 (2010).
38. Dixon, N. *et al.* Reengineering orthogonally selective riboswitches. *Proc. Natl Acad. Sci. USA* **107**, 2830–2835 (2010).
39. Neumann, H., Slusarczyk, A. L. & Chin, J. W. *De novo* generation of mutually orthogonal aminoacyl-tRNA synthetase/tRNA pairs. *J. Am. Chem. Soc.* **132**, 2142–2144 (2010).
40. An, W. & Chin, J. W. Synthesis of orthogonal transcription-translation networks. *Proc. Natl Acad. Sci. USA* **106**, 8477–8482 (2009).
41. Wang, K., Neumann, H., Peak-Chew, S. Y. & Chin, J. W. Evolved orthogonal ribosomes enhance the efficiency of synthetic genetic code expansion. *Nature Biotech.* **25**, 770–777 (2007).
42. Durot, M., Bourguignon, P. Y. & Schachter, V. Genome-scale models of bacterial metabolism: reconstruction and applications. *FEMS Microbiol. Rev.* **33**, 164–190 (2009).
43. Price, N. D., Reed, J. L. & Palsson, B. Ø. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nature Rev. Microbiol.* **2**, 886–897 (2004).
44. Oberhardt, M. A., Palsson, B. Ø. & Papin, J. A. Applications of genome-scale metabolic reconstructions. *Mol. Syst. Biol.* **5**, 320 (2009).
45. Edwards, J. S. & Palsson B. Ø. Towards metabolic phenomics: analysis of genomic data using flux balances. *Biotechnol. Prog.* **15**, 288–295 (1999).
46. Orth, J. D., Thiele, I. & Palsson, B. Ø. What is flux balance analysis? *Nature Biotech.* **28**, 245–248 (2010).
47. Segre, D., Vitkup, D. & Church, G. M. Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl Acad. Sci. USA* **99**, 15112–15117 (2002).
48. Brochado, A. R. *et al.* Improved vanillin production in baker's yeast through *in silico* design. *Microb. Cell. Fact.* **9**, 84 (2010).
49. Asadollahi, M. A. *et al.* Enhancing sesquiterpene production in *Saccharomyces cerevisiae* through *in silico* driven metabolic engineering. *Metab. Eng.* **11**, 328–334 (2009).
50. Henry, C. S., Broadbelt, L. J. & Hatzimanikatis, V. Thermodynamics-based metabolic flux analysis. *Biophys. J.* **92**, 1792–1805 (2007).
51. Becker, S. A. *et al.* Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. *Nature Protoc.* **2**, 727–738 (2007).
52. Gevorgyan, A., Bushell, M. E., Avignone-Rossa, C. & Kierzek, A. M. SurreyFBA: a command line tool and graphics user interface for constraint-based modeling of genome-scale metabolic reaction networks. *Bioinformatics* **27**, 433–434 (2011).
53. Le Fevre, F. *et al.* CycSim — an online tool for exploring and experimenting with genome-scale metabolic models. *Bioinformatics* **25**, 1987–1988 (2009).
54. Cvijovic, M. *et al.* BioMet Toolbox: genome-wide analysis of metabolism. *Nucleic Acids Res.* **38**, W144–W149 (2010).
55. Finley, S., Broadbelt, L. & Hatzimanikatis, V. *In silico* feasibility of novel biodegradation pathways for 1,2,4-trichlorobenzene. *BMC Systems Biol.* **4**, 7 (2010).
56. Henry, C. S. *et al.* High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature Biotech.* **28**, 977–982 (2010).
A seminal paper describing a pipeline for the high-throughput generation of genome-scale metabolic models from annotated genomes.
57. Alam, M. T., Medema, M. H., Takano, E. & Breitling, R. Comparative genome-scale metabolic modeling of actinomycetes: the topology of essential core metabolism. *FEBS Lett.* **585**, 2389–2394 (2011).
This is one of the first studies to use a large number of genome-scale metabolic models to perform comparative metabolic model analysis.
58. Yamada, T. Letunic, I., Okuda, S., Kanehisa, M. & Bork, P. iPath2.0: interactive pathway explorer. *Nucleic Acids Res.* **26**, 787–793 (2011).
59. Bates, J. T., Chivian, D. & Arkin, A. P. GLAMM: Genome-Linked Application for Metabolic Maps. *Nucleic Acids Res.* **39**, W400–W405 (2011).
60. Galdzicki, M., Rodriguez, C., Chandran, D., Sauro, H. M. & Gennari, J. H. Standard biological parts knowledgebase. *PLoS ONE* **6**, e17005 (2011).
61. Canton, B., Labno, A. & Endy, D. Refinement and standardization of synthetic biological parts and devices. *Nature Biotech.* **26**, 787–793 (2008).
This perspective article proposes ways in which biological parts and devices can be standardized to allow for their modular and universal use.
62. Finn, R. D. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **38**, D211–D222 (2010).
63. Letunic, I., Doerks, T. & Bork, P. SMART 6: recent updates and new developments. *Nucleic Acids Res.* **37**, D229–D232 (2009).
64. Eddy, S. R. A new generation of homology search tools based on probabilistic inference. *Genome Inform.* **23**, 205–211 (2009).
65. Goyal, K., Mohanty, D. & Mande, S. C. PAR-3D: a server to predict protein active site residues. *Nucleic Acids Res.* **35**, W503–W505 (2007).
66. Bray, T. *et al.* SitesIdentify: a protein functional site prediction tool. *BMC Bioinformat.* **10**, 379 (2009).
67. Thornton, J. W. Resurrecting ancient genes: experimental analysis of extinct molecules. *Nature Rev. Genet.* **5**, 366–375 (2004).
68. Hall, B. G. Simple and accurate estimation of ancestral protein sequences. *Proc. Natl Acad. Sci. USA* **103**, 5431–5436 (2006).
69. Röttig, M., Rausch, C. & Kohlbacher, O. Combining structure and sequence information allows automated prediction of substrate specificities within enzyme families. *PLoS. Comput. Biol.* **6**, e1000636 (2010).
This paper presents an automated method for predicting substrate specificities within enzyme families using the amino acids extracted from the area around the protein active site.
70. Röttig, M. *et al.* NRPSpredictor2 — a web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Res.* **39**, W362–W367 (2011).
71. Mavromatis, K. *et al.* Gene context analysis in the Integrated Microbial Genomes (IMG) data management system. *PLoS ONE* **4**, e7979 (2009).
72. Medema, M. H. *et al.* antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res.* **39**, W339–W346 (2011).
73. Puigbo, P., Guzman, E., Romeu, A. & Garcia-Vallve, S. OPTIMIZER: a web server for optimizing the codon

usage of DNA sequences. *Nucleic Acids Res.* **35**, W126–W131 (2007).

74. Richardson, S. M., Nunley, P. W., Yarrington, R. M., Boeke, J. D. & Bader, J. S. GeneDesign 3.0 is an updated synthetic biology toolkit. *Nucleic Acids Res.* **38**, 2603–2606 (2010).
75. Villalobos, A., Ness, J. E., Gustafsson, C., Minshull, J. & Govindarajan, S. Gene Designer: a synthetic biology tool for constructing artificial DNA segments. *BMC Bioinformat.* **7**, 285 (2006).
76. Czar, M. J., Cai, Y. & Peccoud, J. Writing DNA with GenoCAD. *Nucleic Acids Res.* **37**, W40–W47 (2009).
77. Mukherji, S. & van Oudenaarden, A. Synthetic biology: understanding biological design from synthetic circuits. *Nature Rev. Genet.* **10**, 859–871 (2009).
78. Tamsir, A., Tabor, J. J. & Voigt, C. A. Robust multicellular computing using genetically encoded NOR gates and chemical 'wires'. *Nature* **469**, 212–215 (2011).
79. Weeding, E., Houle, J. & Kaznessis, Y. N. SynBioSS designer: a web-based tool for the automated generation of kinetic models for synthetic biological constructs. *Brief. Bioinform.* **11**, 394–402 (2010).
80. Rialle, S. *et al.* BioNetCAD: design, simulation and experimental validation of synthetic biochemical networks. *Bioinformatics* **26**, 2298–2304 (2010).
81. Salis, H. M., Mirsky, E. A. & Voigt, C. A. Automated design of synthetic ribosome binding sites to control protein expression. *Nature Biotech.* **27**, 946–950 (2009).
    This article describes a predictive method for designing synthetic RBSs that allows the translation of synthetic genes to be tuned according to their desired stoichiometry.
82. Salis, H. M. The ribosome binding site calculator. *Meth. Enzymol.* **498**, 19–42 (2011).
83. Na, D. & Lee, D. RBSDesigner: software for designing synthetic ribosome binding sites that yields a desired level of protein expression. *Bioinformatics* **26**, 2633–2634 (2010).
84. Hoover, D. M. & Lubkowski, J. DNAWorks: an automated method for designing oligonucleotides for PCR-based gene synthesis. *Nucleic Acids Res.* **30**, e43 (2002).
85. Bode, M., Khor, S., Ye, H., Li, M. H. & Ying, J. Y. TmPrime: fast, flexible oligonucleotide design software for gene synthesis. *Nucleic Acids Res.* **37**, W214–W221 (2009).
86. Matzas, M. *et al.* High-fidelity gene synthesis by retrieval of sequence-verified DNA identified using high-throughput pyrosequencing. *Nature Biotech.* **28**, 1291–1294 (2010).
87. Kosuri, S. *et al.* Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips. *Nature Biotech.* **28**, 1295–1299 (2010).
88. Quan, J. *et al.* Parallel on-chip gene synthesis and application to optimization of protein expression. *Nature Biotech.* **29**, 449–452 (2011).
89. Heneghan, M. N. *et al.* First heterologous reconstruction of a complete functional fungal biosynthetic multigene cluster. *Chembiochem* **11**, 1508–1512 (2010).
90. Kwok, R. Five hard truths for synthetic biology. *Nature* **463**, 288–290 (2010).
91. Wishart, D. S. Computational strategies for metabolite identification in metabolomics. *Bioanalysis* **1**, 1579–1596 (2009).
92. Willett, P., Barnard, J. M. & Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **38**, 983–996 (1998).
93. Ridley, D. D. Introduction to structure searching with SciFinder Scholar. *J. Chem. Educ.* **78**, 559 (2001).
94. Berger, M. F. *et al.* Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature Biotech.* **24**, 1429–1435 (2006).
95. Fordyce, P. M. *et al. De novo* identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis. *Nature Biotech.* **28**, 970–975 (2010).
96. van Hijum, S. A., Medema, M. H. & Kuipers, O. P. Mechanisms and evolution of control logic in prokaryotic transcriptional regulation. *Microbiol. Mol. Biol. Rev.* **73**, 481–509 (2009).
97. Francke, C., Kerkhoven, R., Wels, M. & Siezen, R. J. A generic approach to identify transcription factor-specific operator motifs; inferences for LacI-family mediated regulation in *Lactobacillus plantarum* WCFS1. *BMC Genomics* **9**, 145 (2008).
98. Lim, H. N., Lee, Y. & Hussein, R. Fundamental relationship between operon organization and gene expression. *Proc. Natl Acad. Sci. USA* **108**, 10626–10631 (2011).
99. Llopis, P. M. *et al.* Spatial organization of the flow of genetic information in bacteria. *Nature* **466**, 77–81 (2010).
100. Dietrich, J. A. *et al.* A novel semi-biosynthetic route for artemisinin production using engineered substrate-promiscuous $P450_{BM3}$. *ACS Chem. Biol.* **4**, 261–267 (2009).
101. Müller, U. *et al.* Metabolic engineering of the *E. coli* L-phenylalanine pathway for the production of D-phenylglycine (D-Phg). *Metab. Eng.* **8**, 196–208 (2006).
102. Bayer, T. S. *et al.* Synthesis of methyl halides from biomass using engineered microbes. *J. Am. Chem. Soc.* **131**, 6508–6515 (2009).
    A breakthrough in the use of synthetic genes for pathway engineering. The authors generated codon-optimized synthetic genes for all homologues of an enzyme-coding gene, and characterized them in high throughput to find the best-performing enzyme.
103. Dunlop, M. J. *et al.* Engineering microbial biofuel tolerance and export using efflux pumps. *Mol. Syst. Biol.* **7**, 487 (2011).
104. Mullis, K. B. The unusual origin of the polymerase chain reaction. *Sci. Am.* **262**, 56–61, 64–65 (1990).
105. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
106. Mavrovouniotis, M., Stephanopoulos, G. & Stephanopoulos, G. Synthesis of biochemical production routes. *Comput. Chem. Eng.* **16**, 605–619 (1992).
107. Fleischmann, R. *et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512 (1995).
108. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
109. Heath, A. P., Bennett, G. N. & Kavraki, L. E. Finding metabolic pathways using atom tracking. *Bioinformatics* **26**, 1548–1555 (2010).
110. Kanehisa, M., Goto, S., Kawashima, S. & Nakaya, A. The KEGG databases at GenomeNet. *Nucleic Acids Res.* **30**, 42–46 (2002).
111. Lee, P. A. *et al.* CLONEQC: lightweight sequence verification for synthetic biology. *Nucleic Acids Res.* **38**, 2617–2623 (2010).
112. Chandran, D., Bergmann, F. T. & Sauro, H. M. TinkerCell: modular CAD tool for synthetic biology. *J. Biol. Eng.* **3**, 19 (2009).
113. Rodrigo, G., Carrera, J. & Jaramillo, A. Asmparts: assembly of biological model parts. *Syst. Synth. Biol.* **1**, 167–170 (2007).

Competing interests statement

The authors declare no competing financial interests.

FURTHER INFORMATION

Active Site Classification (ASC): http://asc.informatik.uni-tuebingen.de
antiSMASH: http://antismash.secondarymetabolites.org
Asmparts: http://soft.synth-bio.org/asmparts.html
Biojade: http://web.mit.edu/jagoler/www/biojade/
BioMet Toolbox: http://www.sysbio.se/BioMet/
BioNetCAD: http://www.sysdiag.cnrs.fr/BioNetCAD/
CarbonSearch: http://www.kavrakilab.org/atommetanet
CellDesigner: http://www.celldesigner.org
CloneQC: http://cloneqc.thruhere.net
Clotho: http://www.clothocad.org
COBRA Toolbox: http://opencobra.sourceforge.net
CycSim: http://www.genoscope.cns.fr/cycsim/
DESHARKY: http://soft.synth-bio.org/desharky.html
DNAWorks: http://helixweb.nih.gov/dnaworks/
From Metabolite to Metabolite (FMM): http://fmm.mbc.nctu.edu.tw
Gene Composer: http://www.genecomposer.net
Gene Designer 2.0: https://www.dna20.com/genedesigner2/
GeneDesign: http://www.genedesign.org
GenoCAD: http://www.genocad.org
GLAMM: http://glamm.lbl.gov
Integrated Microbial Genomes (IMG): http://img.jgi.doe.gov
iPATH2: http://pathways.embl.de
KEGG: http://www.genome.jp/kegg/
MultiGeneBlast: http://multigeneblast.sourceforge.net/
Optimizer: http://genomes.urv.cat/OPTIMIZER/
RBS Calculator: https://salis.psu.edu/software/
RBSDesigner: http://rbs.kaist.ac.kr
Registry of Standard Biological Parts: http://partsregistry.org
RetroPath: http://www.issb.genopole.fr/~faulon/retropath.php
Standard Biological Parts knowledgebase: http://sbpkb.sbolstandard.org
SurreyFBA: http://sysbio3.fhms.surrey.ac.uk/SurreyFBA.zip
SynBioSS: http://synbioss.sourceforge.net
TinkerCell: http://www.tinkercell.com
WebGEC: http://lepton.research.microsoft.com/webgec/

ALL LINKS ARE ACTIVE IN THE ONLINE PDF