# Designing Living Matter. Can We Do Better than Evolution?

*Synthetic Biology is Sharpening its Profile*

*Thanks to substantial improvements in the theory of metabolic fluxes and the application of $^{13}C$ isotope markers in experimental flux studies, Pareto efficiency of bacterial metabolism can now be determined and direct answers to the long standing questions of optimization according to multiple criteria in nature can be given. Cells or organisms operate close to Pareto optima but the performance with respect to every single criterion is almost always improvable. Rational design and evolutionary methods are routinely used for the production of biomolecules with optimized properties. Examples are proteins for technical applications, for example in detergents, and optimally binding nucleic acid molecules called aptamers. Among the various perspectives of synthetic biology, the usage of DNA for information storage is particularly promising: In a pilot experiment, an entire book including figures and a Java script, in total more than 5 megabit, were stored on a single DNA molecule. © 2013 Wiley Periodicals, Inc. Complexity 18: 21–31, 2013*

**Key Words:** evolutionary design; flux balance; metabolism; optimality; Pareto surface; rational design

**PETER SCHUSTER**

*Editor-in-chief of Complexity is at the Institut für Theoretische Chemie der Universität Wien, Währingerstraße 17, Wien, 1090, Austria*
*(E-mail: pks@tbi.univie.ac.at)*

## 1. INTRODUCTION

Less than 1 year ago, the *Jena Life Science Forum* organized a meeting with precisely the same title as this essay. After having created this title, we were shocked first by our own hubris. Do we really believe that we can do better than nature? Thinking more deeply, however, we realized that whether or not man can outperform evolution depends entirely on the interpretation of better. It is not difficult to beat nature in case one picks out a single criterion and improves the performance related to it.

In this essay, we shall be dealing first with the question of optimality and multiple criteria in nature. Then, the essay will focus on the design of molecules with predefined functions. Rational design will be confronted with evolutionary design sometimes also characterized as "irrational design." At the end, we shall try to review the perspectives of synthetic biology.

 Vol. 18 No. 6
DOI 10.1002/cplx.21461
Published online 30 July 2013 in Wiley Online Library
(wileyonlinelibrary.com)

## 2. OPTIMALITY IN NATURE

Optimization for multiple criteria, which—as usual—are not completely independent, is anything but trivial and requires the determination of a so-called Pareto[1] surface (for an essay dealing with this problem, see the recent column "Simply Complex"[1]). What about optimality in nature? Ever since the Neolithic Period, man is redesigning nature through manipulation of organisms for his purposes. In this process, the genetics of species has been changed, first by trial and error then specifically, in particular since the second half of the 20th century—otherwise there were neither field crops nor vegetables nor fruits and no domestic animals. The history of mankind is tantamount to the history of modifying nature to the benefit of human society. Until the 20th century, the manipulation of living beings was done without any knowledge of the fundamental mechanisms, which govern the change of the properties of organisms and species. The spectacular development of molecular biology and the explosion of biological knowledge opened entirely novel avenues toward targeted modifications of biological entities from molecules to whole organisms. It turned out that it is not difficult, in essence, to "improve" natural properties in the sense of more, faster, larger, smaller, more specific, more stable, and so forth. Proteins were not only modified in this sense, they were also adapted to non-natural conditions such as nonaqueous media and they were redesigned to catalyze reactions that do not occur in nature (see the section on "rational design"). Nature in contrast to the human designer is not able to afford optimization of single features—selection operates on entire organisms in ecosystems and what counts is fertile progeny and nothing else. At the same time, evolution has to build with already existing blocks and *de novo* design as well as engineering with new materials is impossible. Nature follows the principle of "evolutionary tinkering" or "bricolage"[2,3], and constructs by combing units that are around. The only property on which success in evolution is built is functionality. It is easy to find suboptimal solutions of problems in higher organisms, which came about as the result of evolutionary tinkering and historical contingency. We mention two examples representative for many others: (i) The blind spot in the eye of vertebrates originates from the fact the nerve fibers leave the light sensitive cells of the retina on the wrong side—the side at which the light arrives. The fibers must pass the retina in bundled form as optical nerve creating a light insensitive spot. In contrast, the evolution got the design of the cephalopod eye right—the nerve fibers are bundled on the opposite side of the incoming light and hence need not pass the retina. (ii) The mammalian pharynx where trachea and esophagus are crossing is a misconstruction with the danger of choking on food that can be harmless but also fatal when large pieces fall into the windpipe.

Obvious questions concern cellular metabolism: Is metabolism optimal or is it just functional? If it appears to be optimal what were the criteria of optimization? Biochemists were successfully exploring the metabolism of cells since more than a century and are still working on it. All important reaction paths as well as their stoichiometry and the catalyzing enzymes are known and comprised in the famous metabolic map of Boehringer-Mannheim[4] but nevertheless, fundamental questions concerning the dominant components of the metabolic flux as well as the optimality criteria are still unanswered. Early works dealing with the analysis of biochemical reaction networks in 1970 recognized already the need of a joint approach by suitable experiments and mathematical modeling [5–7]. About 25 years ago, a new mathematical model has been conceived with the aim to develop a simplified analysis of the reaction fluxes in metabolic networks that, in essence, takes care only of the governing constraints, which have to be fulfilled by all expressible phenotypes, and implements some objective function for optimization, for example maximal growth [8]. The new approach has been called flux balance analysis (FBA), and it became an indispensible tool for genome-based models of microbial metabolism [9–11]. FBA is a computer-based approach that needs to be combined with appropriate experiments to allow for quantitative predictions. The networks of metabolic fluxes, which combine results from genomics and proteomics with biochemical knowledge, are essentially unconstrained and make no restrictions concerning the accessibility of flux combinations in a Cartesian space (Figure 1). Various constraints define the regions in solution space that are accessible in reality. These constraints come in several categories: (i) constraints arising from basic physical chemistry, in particular chemical kinetics and thermodynamics, (ii) spatial and topological constraints, (iii) environmental constraints depending on the experimental conditions, and (iv) internal regulatory restraints. Taken all together, the different constraints define a space of allowed solutions, which lie in a cone-like subspace of the Cartesian space—a five-sided pyramid in Figure 1—with the apex in the origin of the coordinate system were all fluxes vanish. The

[1] *Pareto optimality is named after the Italian economist Vilfredo Pareto (1848–1923) and it characterizes a kind of pseudo-optimality in the sense that the performance with respect to one criterion cannot be improved without making the efficiency worth with respect to another criterion.*

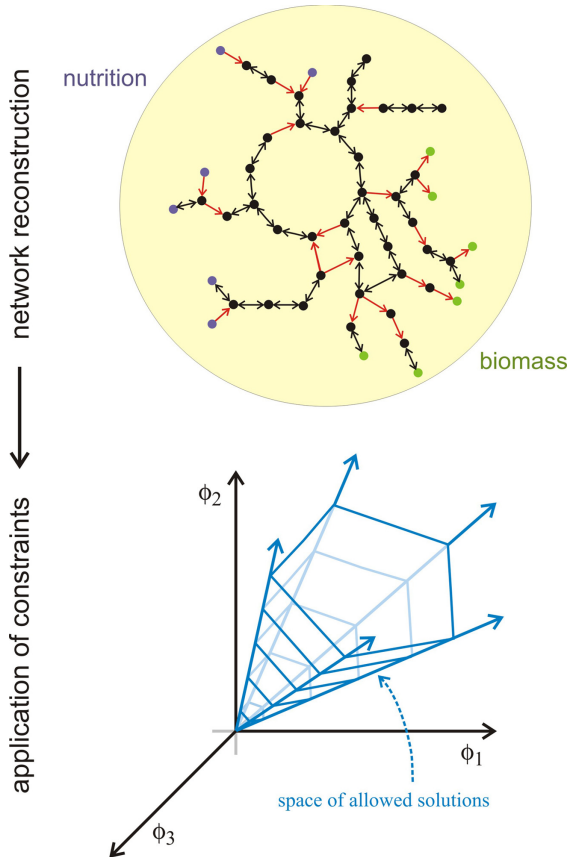


DOI 10.1002/cplx

allowed subspace is commonly confined from above by one of the conservation relations—in the case reported in the next paragraph the Pareto surface provides this limitation. The individual steps of FBA are described in detail in the review by Palsson and coworkers [11]. The combination of flux balance and energy balance analysis provides a basis for the definition of cost-functions with multiple criteria and allows for the calculation of Pareto optimal curves and surfaces [12,13]. Despite full internal consistency of the theory of metabolic fluxes, FBA, as said above, can only make qualitative predictions on the distribution of fluxes in the networks and requires experimental data for completion.

To illustrate the complexity of microbial metabolism, we present a few numbers from the bacterium *Escherichia coli* [14]: The genome is $4.6 \times 10^6$ base pairs long and contains about 4500 genes, which are coding for some 5000 transcripts. Transcription, translation, and post translational modifications yield 6000–10,000 proteins.[2] Molecular diversity of proteins is supplemented by around 2000 metabolites, mainly low-molecular-weight compounds. The entire ensemble of a bacterial cell is thus composed of some 20,000 molecular species. Inevitably, global analysis based on the entire genome is possible only when the number of dimensions is drastically reduced, and the model is limited to core metabolism. Despite such drastic reductions in complexity, it took until last year before an experimentally supported analysis of *E. coli* metabolism was reported. A corresponding paper by Sauer and coworkers at the ETH Zürich was published in *Science* magazine [15]. The food source or the starting product of the bacterial metabolism is labeled with the stable isotope $^{13}$C—glucose in position 1 in the experiment described—and the metabolic fluxes within the cells are determined by a combination of mass spectrometric analysis and computer-based calculation of the distributions of the label in the metabolites. The kinetic basis of the investigation is a stoichiometric model of the core metabolism in *E. coli*, which consists of 79 individual reaction steps and an

**FIGURE 1**

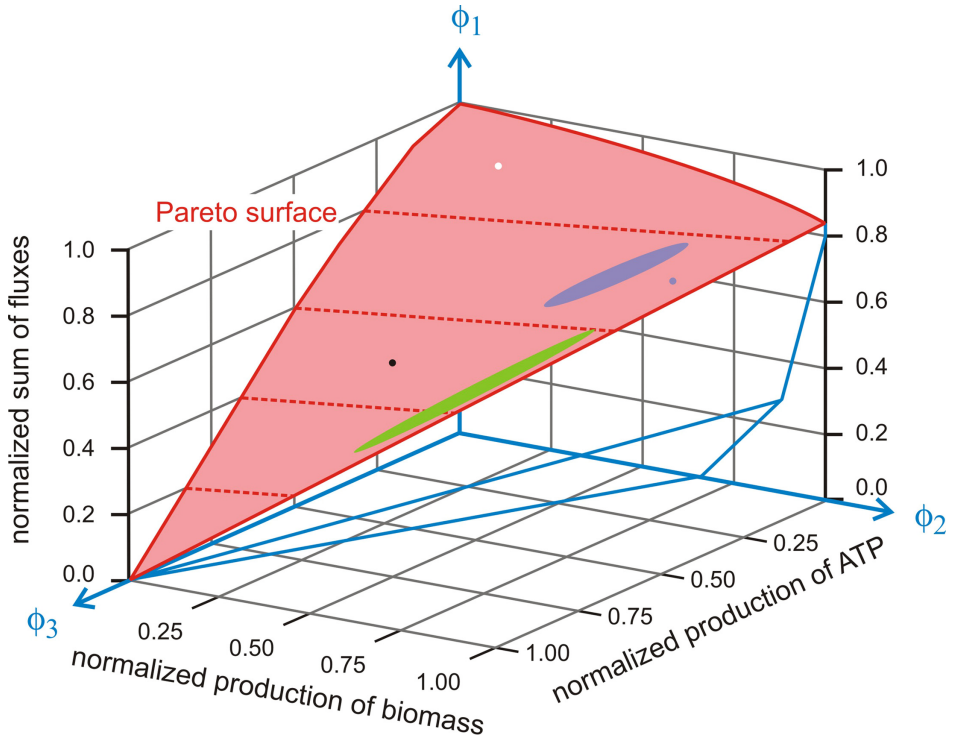



The metabolic flux space and its limitations through chemical stoichiometry and other constraints. "*Flux-balance analysis*" (FBA) considers the chemical reaction fluxes ($\phi$) in a metabolic reaction network that has be reconstructed from genome wide biochemical data (upper part of the figure). The global flux in the metabolic network leads from food source to biomass production through cell growth and cell division. Every instantaneous state of the fluxes in the network corresponds to a point in the space of fluxes but not every point in the Cartesian coordinate system of fluxes represents a realizable state, because the system has to meet several constraints. As sketched in the lower part of the figure, the constraints cut out the space of allowed solution, which has the shape of a coneÅa five-sided pyramid in the figure, from the Cartesian coordinate system, and the search for acceptable solutions can be confined to the inner part of the cone. Such constraints are the stoichiometry of individual reaction steps, thermodynamic constraints like balance of free energy, and mass conservation (Source: Refs. [9] and [11]).

[2]*The number of proteins accounts for differences in protein folding and size and covers covalent post translational modifications. Bacteria do not splice transcripts regularly on "splicosomes," but some messenger-RNAs are nevertheless processed by "self-splicing." These mechanisms and others result in numbers of protein molecules that exceed the numbers of genes from which they are expressed.*


DOI 10.1002/cplx

**FIGURE 2**




The Pareto surface of the three dominant fluxes in *Escherichia coli* metabolism. The surface of the Pareto optimal fluxes that describe best the metabolism of *E. coli*, the sum of all individual fluxes ($\phi_1$), the production of biomass ($\phi_2$), and the production of energy in form of the ATP yield ($\phi_3$), was determined by means of $^{13}$C-based flux analysis (red surface corresponding to minimal $\phi_1$, maximal $\phi_2$, and maximal $\phi_3$).[14,15] The points and areas near the red surface were calculated from measured $^{13}$C distributions in metabolites of *E. coli* under different growth conditions(aerobe cultures: excess glucose = blue, glucose deficiency = green, nitrogen deficiency = black, and, for comparison, anaerobe culture: white). All points lie systematically somewhat below the Pareto surface (explanation see text). The Cartesian coordinate system differs from Figure 1 through normalization of the fluxes: "1" means best possible fulfillment of the optimization criterion - minimum of the sum of fluxes, maximum biomass, and maximum ATP yield—and "0" characterizes the solution at maximal distance from the optimum. The origin, accordingly, symbolizes the poorest possible solution (Source: Ref. 15; the figure is redrawn form Figure 1A in this publication); reproduced with permission of the American Association of the Advancement of Science.

*E. coli* specific gross balance equation for biomass production. To identify the relevant gross cost-functions spanning a representative coordinate system in this high-dimensional space, the distributions of $^{13}$C-labels were calculated for 54 cost-functions and the deviations from the 44 *in vivo* measured flux distributions were determined. Five cost-functions were found to be consistent with the *in vivo* fluxes: (i) production of adenosine tri-phosphate (ATP), (ii) biomass production, (iii) acetate yield, (iv) carbon dioxide ($CO_2$) production, and (v) minimal sum of absolute fluxes in the sense of most efficient utilization of resources. No single cost-function was able to provide an adequate description of the measured fluxes and also no pair of cost-functions doing this job was found. Out of all triples, the combination (i), (ii), and (v)—production of energy as ATP yield, biomass production, and optimal allocation of resources corresponding to minimal sum of fluxes—was found to be most suitable to model the metabolism of the *E. coli* cell and provided the best basis for an analysis of optimality. The indication for suitability is seen in the nearness to the Pareto surface of the positions of all 44 fluxes.

Figure 2 shows the Pareto surface of *E. coli* metabolism in a coordinate system spanned by the three cost-functions chosen (see next paragraph): All aerobic cultures are situated at positions close to the Pareto surface: (i) cultures grown at excess glucose are close together and marked blue, (ii) carbon-limited—glucose-deficient—cultures are lying in the green zone, (iii) one nitrogen-limited culture occupies the black dot. An anaerobic culture is shown for comparison (white dot).



DOI 10.1002/cplx

Three results of the work of Sauer and coworkers [15] at the ETH Zürich and recent works of the same institute [16] are of general importance: (i) the points for metabolic fluxes in other bacterial species—*Bacillus subtilis* (several strains), *Zymomonas mobilis*, *Pseudomonas fluorescence*, *Rhodobacter sphaeroides*, *Pseudomonas putida*, *Agrobacterium tumefaciens, Sinorhizobium maliloti*, and *Paracoccus versutus*—are situated very close to the Pareto surface of *E. coli,* (ii) bacteria have a biochemical repertoire to measure and regulate metabolic fluxes as shown in case of the glycolytic flux in *E. coli* [16], and (iii) precise inspection of the positions of fluxes reveals that all points lie significantly below the Pareto surface. The calculations of metabolic fluxes for different food sources provide the explanation: combinations of fluxes (precisely) at the Pareto surface for one particular food source are relatively far away from the Pareto optimal combinations for another food source, and consequently the adaptation from one optimal condition to the other takes fairly long time. At some distance away from the Pareto surface points can be found, where changing the fluxes from one condition to the other requires a small effort only and bacteria operating at these suboptimal states can switch quickly from one food source to the other. Evolution does not only drive the organisms toward efficiency of metabolism at the present conditions, it takes into account also the necessity of flexible response to environmental changes. The given explanation of variability of metabolism at the expense of metabolic efficiency in variable environments—properly characterized as minimal cost for flux adjustment [15]—provides a more plausible alternative explanation to the previously favored interpretation as an adaptative memory of microorganisms to historical sequences of changes in nutrition [17].

## 3. RATIONAL DESIGN

Two alternative strategies for the creation of molecules and organisms with predefined properties and functions are available: (i) rational design that makes use of the entire knowledge of structures and functions of biomolecules and (ii) evolutionary design applying the principles of biological evolution to the selection of entities with predetermined properties [18]. Both methods have their advantages and deficiencies and we shall mention here a few typical examples only, because the literature on molecular design is enormous.

Rational design is built on the paradigm of conventional theoretical structural biology

Sequence ⇒ structure ⇒ function:

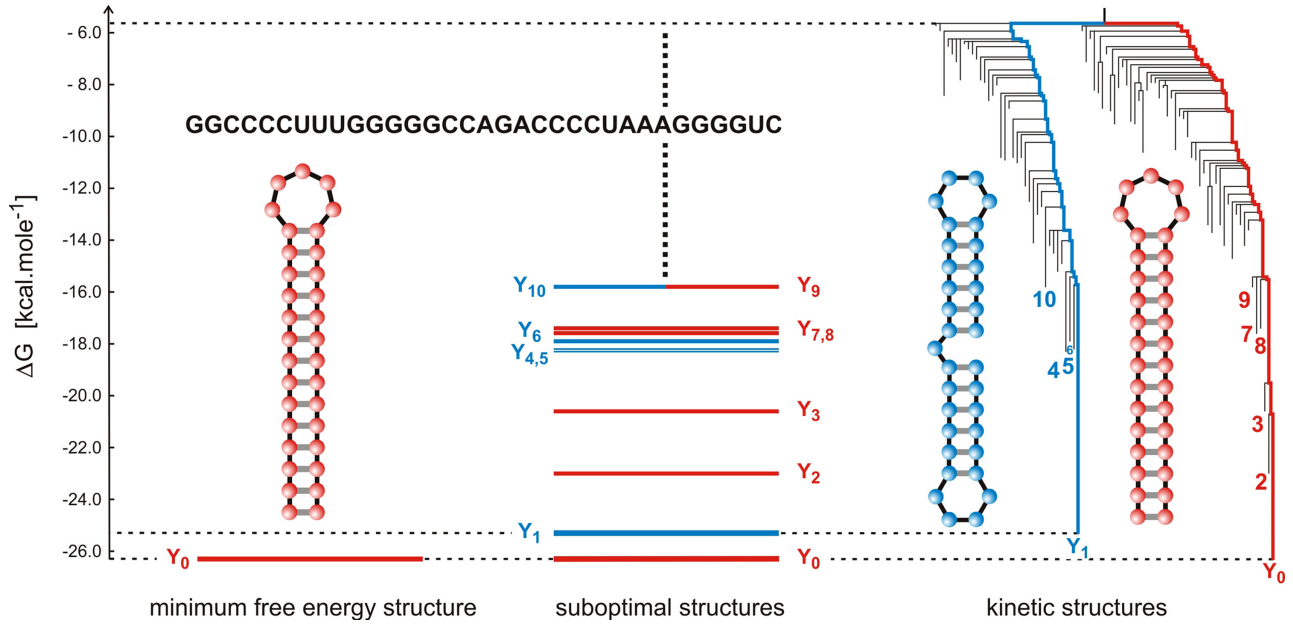
In essence, the paradigm says that the structure of the biopolymer molecule can be predicted provided the sequence and the conditions of structure formation through folding are known. In absence of special conditions, the conventional assumptions are constant temperature *T*, constant pressure *p*, neutral pH = 7, and an ideal solution[3]—the search will aim at the structure with minimal free energy, $\Delta G^0_T$. The known structure of a molecule, according to the paradigm of structural biology, allows for inferences about function. The two steps, sequence ⇒ structure and structure ⇒ function, yield reliable predictions from first principles only in exceptionally simple cases. Input of empirical data is essential for successful algorithms. Rational design provides the advantage of a direct or targeted search strategy and is economic with respect to search time and required material. Its deficits are the still unsatisfactory methods for the prediction of biopolymer structures and functions. Evolutionary design has the advantage to operate independently of the knowledge of structure provided a technique to select for the desired function is at hand but it is expensive and requires large quantities for experiments. Here, we shall discuss first the rational design of enzyme molecules, which is known also under the name of "protein engineering" and add some information on the state-of-the-art in rational design of ribonucleic acids (RNAs).

The prerequisite for rational protein design was the development and the implementation of techniques for "site-directed mutagenesis" that allows for targeted replacement of every amino acid by each of the other nineteen amino acids at every position of the protein sequence [19]. In the beginning, protein design was used mainly for the analysis of sequence–structure relations with the goals to understand the principles of protein folding and to predict reliably the structures of minimal free energy [20]. In the late 1990, the combination of computer calculations and large-scale empirical data opened a new avenue for the design of protein structures [21–23]. This new approach eventually became one of the most popular methods for protein structure prediction [24]. An important property of designed proteins is thermodynamic stability, in particular stability at higher temperatures, because almost all natural proteins are not sufficiently stable for technical usage, for example in detergents. Based on theoretical computations and empirical data, enzymes were made thermodynamically more stable through targeted mutations without changing the

[3] *The application of standard conditions, 1 k bar pressure, diluted solution, and so forth, is expressed by the superscript zero, '0'. A special condition would be, for example, in situ folding of the biomolecule during synthesis. Then not the entire sequence is folded but only the already synthesized strand in statu nascendi beginning from one end.*


DOI 10.1002/cplx

**FIGURE 3**




Structures of minimal free energy and suboptimal conformations of RNA molecules. A small RNA molecule with a chain length of $n = 33$ nucleotides and the sequence shown in the figure yields a hairpin structure with a stacking region of 14 base pairs as its structure of minimal free energy (mfe-structure $Y_0$, red; left-hand drawing). The mfe-structure is accompanied by a spectrum of suboptimal conformations (drawing in the middle). To be able to calculate the kinetics of conformational transitions, we need to know also the activation energies (right-hand drawing), $\Delta E_a$, which are related to the rates parameter through the Arrhenius equation: $k_r = A.\ \exp(\Delta E_a/RT)$ with A being a proportionality factor. Structures together with their mean life times, $\tau = k_r^{-1}$, are called kinetic structuresÅhere the double hairpin structure $Y_1$ (blue; source: Ref. [30]).

enzymatic activity [25]. An impressive extreme example is the human protein procarboxypeptidase, where the designed molecule is more than $|\Delta\Delta G^0_T| > 10$ kcal mole$^{-1}$ more stable than the natural protein [26]. Recent developments allowed for the design of protein catalysts for chemical reactions that have no counterpart in nature [27]. In general, basic design yields only poor catalytic activities compared to natural enzymes but the modest activities can be improved substantially with respect to thermal stability, binding affinity, specificity, and turnover rates by molecular dynamics simulations and evolutionary methods (see next section).

As an example of technical usage of natural and artificial—designed—enzymes, we describe here the application of proteins to make laundry detergent more efficient [28,29]. The idea to use enzymes in washing agents is fairly old: the German pharmacist, chemist, and businessman Otto Röhm isolated in 1913 a protease-containing extract from animal pancreas, used it in prewash, and made a successful patent application. Because of impurities and high-production costs, the new detergent was no real success. It took until 1959 before the first washing powder containing a bacterial protease has been launched in Switzerland. It took 10 more years before the usage of enzymes in detergents became standard. Today proteins, which are produced by genetic engineering in essentially two bacterial species, *Bacillus licheniformis* and *Bacillus amyloliquefacies*, are indispensible for the detergent industry. Enzymes used in detergents amount to approximately two thirds of the total enzyme production for technical use. A modern detergent for use in the washing machine or the dishwasher contains enzymes from up to four different classes: (i) proteases for degradation of protein impurities, (ii) amylases for the degradation of starch, (iii) lipases for the cleavage of fat, and (iv) cellulases for surface degradation of cotton fibers to soften cotton fabric after washing. Protein design is used primarily to make the enzymes more stable and to maintain their activities at higher temperatures and alkaline pH-values.



DOI 10.1002/cplx

RNA molecules are particularly well-suited for rational design, because there is a well-defined simplified notion of RNA structure, the so-called secondary structure, which allows for rigorous mathematical analysis [30]. The secondary structure is a graph that is equivalent to a listing of base pairs in the double-helical stretches or "stacks" of the RNA structure.[4] Figure 3 shows three different concepts of RNA structures:

1. Conventional structural biology assigns one structure to every sequence in the sense of the "sequence $\Rightarrow$ structure" paradigm (Figure 3, left drawing: $Y_0$ is a hairpin with 14-base pairs). This structure is the thermodynamically most stable structure or the conformation of minimal free energy (mfe).[6] Given the sequence, the structure of the molecule can be calculated by means of algorithms making use of linear programming [31–33]. These algorithms, however, are not suitable for RNA design, because for this goal the inverse problem, sequence $\Leftarrow$ structure, has to be solved. Algorithms have also been developed for the inverse problem, which predicts the sequences folding into given structures [30,34–36]. A major results that is important for RNA design is the nonuniqueness of the inverse folding problem: Many sequences form the same mfe-structure [37].
2. RNA molecules like all polymers form many conformations corresponding to local minima of the free energy. A whole spectrum of metastable[5] suboptimal structures belongs to a given sequence as shown in the middle drawing of Figure 3. These suboptimal conformations differ with respect to their lifetime. In case of the example in the figure, there is a structure $Y_1$ consisting of two hairpins with six-base pairs each, which has a free energy that lies only about 1 kcal mole$^{-1}$ above the global free energy minimum of $Y_0$. The metastable conformation $Y_1$ corresponds to a local minimum of the free energy and provides a thermodynamic alternative to the most stable structure $Y_0$. The free energy of a conformation determines the statistical weight in the partition function, which is tantamount to its presence at the thermodynamic equilibrium. It tells nothing, however, about the timescale on which the conformational change takes place.
3. In addition to the existence of suboptimal conformations and their presence at equilibrium, the rate parameters for the transition between conformations are required for a full characterization of suboptimal states. According to the Arrhenius equation (Figure 3, right drawing), the rate of a conformational change is determined by the height of the barrier separating the valleys of the two local minima. The barrier height in Figure 3 is about 20 kcal mole$^{-1}$ and the two structures $Y_0$ and $Y_1$ live long enough to be observed.

In the case of RNA molecules, it is possible to develop algorithms that allow for the design of bistable and multistable molecules with predefined barrier heights [38]. Making use of methods in discrete mathematics, it is possible to proof that for each pair of arbitrarily chosen structures there exists at least one sequence, which can form the two structures [37]. The result is called intersection theorem, an extension of the proof to three structures is not possible and counterexamples are readily constructed. Schultes and Bartel were choosing two completely unrelated RNA molecules of the same length of 88 nucleotides; one is an artificially designed RNA ligase,[6] and the second one is a natural RNA cleaving enzyme from hepatitis δ-virus. Sequences, which can fold into both structures were designed and indeed are able to perform both catalytic functions [39]. RNA molecules, which form two active conformations with different functions, are well-known in nature. The so-called "riboswitches" regulate the synthesis of enzymes for metabolic reactions [40].

## 4. EVOLUTIONARY DESIGN

Darwin's principle of natural selection is built on three prerequisites: (i) multiplication through reproduction, (ii) variation, and (iii) selection as a consequence of limited resources. None of the three processes is bound to the existence of cellular life and there is no reason why Darwinian evolution should not be observable in cell-free laboratory assays. Spiegelman recognized this fact already in the 1960th and designed and implemented the first successful experiments evolving molecules in the test tube[41–43]: All three conditions were fulfilled by the properties of the RNA molecules

[4] *For the sake of simplicity, we shall use the word ,structure "as a synonym for secondary structure." The notion of "conformation" is used when two or more structures are assigned to a single sequence (Figure 3; middle and right-hand drawing).*

[5] *A conformation is characterized as "metastable," if it corresponds to a minimum of the free energy that is higher than the free energy of the mfe-structure.*

[6] *A ligase is an enzyme that ligates or concatenates two molecules, here two RNAs.*


DOI 10.1002/cplx

and the experimental setup: (i) the RNA had been isolated from *E. coli* cells that were infected by the bacteriophage Qβ and is multiplied by an RNA dependent RNA polymerase—Qβ-replicase isolated from infected *E. coli* cells as well,[7] (ii) variation is provided by imperfect copying, and (iii) the conditions for selection are determined by the setup of the experiment. Frequently the "serial transfer" technique is applied: stock solution, which contains the building blocks for RNA synthesis and Qβ-replicase, is inoculated by a small sample of the RNA that is suitable for replication.[8] RNA synthesis starts instantaneously and after a certain time span a small sample is transferred to the next test tube, which contains fresh stock solution and replaces the consumed materials. RNA replication sets in again and the whole procedure is repeated about 100 times. Spiegelman observed that the rate of RNA synthesis increases during the serial transfers and he gave the correct interpretation of the result: faster replicating variants that originated from replication as copying errors replaced the original, less efficient replicating molecules until the velocity of replication reached a value, which can be increased no more under the conditions of the experiment. In a way such serial transfer experiments with molecules under laboratory conditions represent evolution in time lapse mode as an entire series can be carried out in a few days. The consumption of material during RNA synthesis within a single-test tube changes the kinetics of RNA replication [44] and special automata have constructed to be able to perform serial transfer under precisely controllable conditions [45]. The fundament of an understanding of the evolution of molecules on the basis of chemical reaction kinetics has been laid down by Eigen in form of a kinetic theory of molecular evolution [46,47].

The observation that molecules can be evolved by Darwinian selection in the test tube initiated the development of a novel branch of biotechnology in the sense of evolutionary design of biomolecules with predetermined properties and functions [48,49]. In contrast to rational design, the evolutionary approach does neither require knowledge of molecular structures nor is it necessary to know the relation between structure and function. What is necessary is only a test system for the desired molecular property and a technique that allows for selection of molecules, which meet best the prescribed conditions, from an ensemble with other properties. The principles of evolutionary "breeding" of molecules[9] are sketched in Figure 4. Initially, a population of molecules with sufficient genetic diversity has to be prepared. Several methods are currently available to meet this goal. One can use, for example, a DNA synthesizer to produce an ensemble of random DNA sequences, transcribe it into RNA, and eventually translate the RNA into protein. The best suited variants are picked by means of a suitable selection method and used to create a new population by means of amplification and diversification, for example through error-prone replication. The new population is subjected to the next selection-cycle and, in general, some 20–30 cycles are sufficient to produce optimal molecules. Evolutionary techniques were used for many different purposes—we mention here two examples: (i) the production of "aptamers" [50]—being optimally binding RNA molecules, and (ii) targeted evolution of proteins [51].

RNA molecules are especially well-suited for the application of evolutionary methods, because they can be directly replicated and mutated [49]. Here, we illustrate the technique by sketching the **S**ystematic **E**volution of **L**igands by **EX**ponential Enrichment (SELEX) method (Fig. 5), which is nowadays routinely applied for the evolution of molecules that bind optimally to predefined targets [52,53]. The basic protocol is readily described: initially a "pool" of random RNA sequences is created and transferred into a suitable solvent. A column is prepared for affinity chromatography, on which the target molecules are covalently bound to the stationary phase of the column. Then, the solution containing the RNA molecules is poured over the column and the molecules that bind strongest to the stationary phase are retained. Using a different solvent, these currently best binders are eluted and subjected to the next selection-cycle consisting of amplification, diversification, and selection. The solvent is changed from cycle to cycle to make it more and more difficult for the binders to be retained and after a sufficient number of cycles—commonly 20–30—RNA molecules called aptamers binding optimally to the target are obtained. SELEX allows for the preparation of aptamers that have binding constants almost as high as the strongest binding constants found in nature.

Targeted evolution of proteins was and is pursuing two different goals: (i) an improvement in the understanding

[7] *Qβ denotes a barteriophage, a virus that infects bacteria, which lives in bacteria of the species Escherichia coli. Qβ-replicase is an RNA replicating enzyme that can be isolated from infected bacteria.*

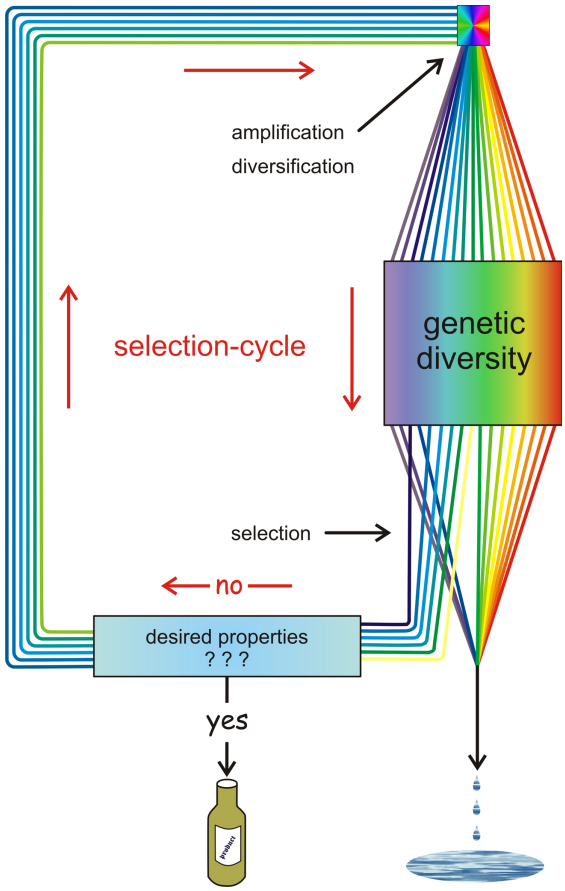
[8] *To be suitable for replication, the RNA has to contain a binding site for Qβ-replicase.*

[9] *"Breeding" is a good characteristic for the techniques of evolutionary biotechnology as the experimenter like an animal breeder has to pick out molecules, which meet the properties of the predefined target as closely as possible, from a diverse mixture.*



DOI 10.1002/cplx

**FIGURE 4**




The strategy of evolutionary biotechnology. Molecular functions are optimized in selection-cycles. Initially, an ensemble of biomolecules - RNAs, DNA,s or proteins - is prepared, which has sufficiently high-genetic diversity - for example, random sequences through synthesis that is not controlled by a program in the form of a prescribed biopolymer sequence. Variants that carry the desired feature are picked out from the ensemble by means of a suitable selection procedure. The selected sample is subjected to the next selection-cycle, which consists of the steps (i) amplification, (ii) diversification, and again (iii) selection. Amplification and diversification can be combined in a single step in case biomolecules are copied by a process with a suitably high-mutation rate. Selection-cycles are repeated until the desired result has been obtained or no more improvements are recorded (Source: Refs. 48 and 49).

of thermodynamic stabilities and functions of proteins, which in natural evolution are commonly blurred by the complex superposition of multiple optimization criteria, and (ii) the production of non-natural proteins, which reveal physical and chemical properties free of the historical contingencies of biological evolution [54]. In addition, *in vitro* evolution allows for isolation and analysis of all intermediate states of an evolutionary process and in this way one gets otherwise not achievable insights into the paths along which populations of molecules are optimized evolutionarily [55]. Finally, we mention a recent publication, in which it has been shown by *in vitro* evolution that a ancient class of proteins—a primordial "fold"—has the capability to change structure through a few evolutionary mutations and to adopt the function of a specific RNA ligase [56].

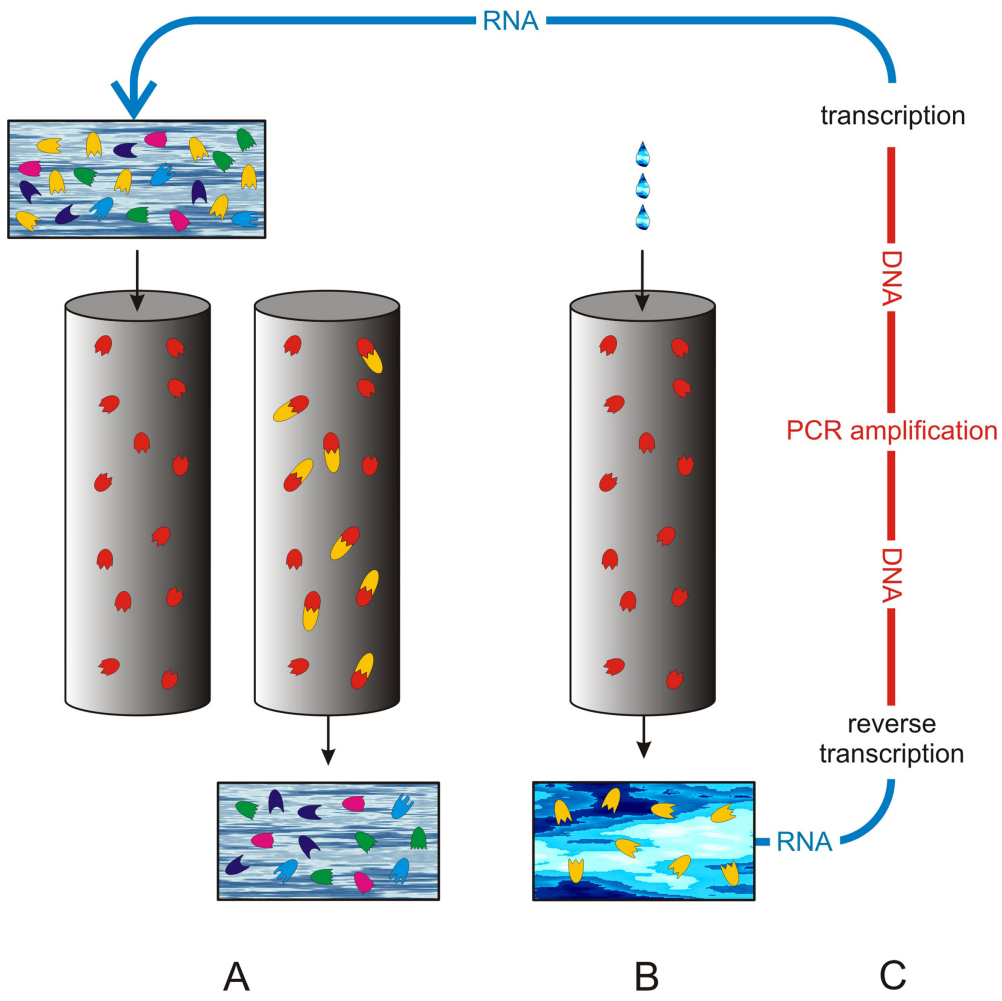
## 5. SYNTHETIC BIOLOGY "QUO VADIS"?

The notion "synthetic biology" appears the first time in the year 1913 in a *Letter to Nature* with the heading "Synthetic biology and the mechanism of life" [57] and refers to a book by Leduc with the title "*La biologie synthétique*" [58] that had been published in Paris in the year before. In this monograph, the author, a French biologist, tries to reduce processes in living organisms to the physics and chemistry of diffusion in liquid solutions. Synthetic biology as we know it nowadays—still not yet a homogeneous field with a well-defined methodology—has been shaped during the second half of the 20th century as it was growing out of molecular biology and molecular genetics. As important milestones in the development of present day synthetic biology, we mention:

i. Watson and Crick's model of DNA structure—Nobel prize for physiology 1962,
ii. restriction nucleases for highly specific cleavage of DNA by Werner Arber, Daniel Nathans, and Hamilton Othanel Smith—Nobel prize for physiology 1978—and their application in molecular genetics through recombinant cloning by Paul Berg [59]—Nobel prize for chemistry 1980,
iii. the development of novel DNA sequencing techniques that allow for the sequencing off whole genomes by Walter Gilbert [60] and Frederick Sanger [61,62]—Nobel prize for chemistry 1980,
iv. the synthesis of the gene regulation network for an artificial oscillator *in vivo* called 'repressilator' through the incorporation of three repressor genes into *E. coli* cells [63], and the incorporation *in vivo* of a reversible


DOI 10.1002/cplx

**FIGURE 5**




The SELEX method in the production of RNA molecules binding optimally to target structures. SELEX provides a simple experimental example for the selection of RNA molecules that bind strongly to predefined target structures. Molecules with high binding affinities called "aptamers" are selected from a population of RNA molecules with different binding constants. The separation process applied is called affinity chromatography: a target structure is bound covalently to the stationary phase of a chromatographic column, a solution containing the molecules to be separated is poured over the column, and the separation is achieved through retention of the strongly binding aptamers on the column (**A**). The bound molecules are eluted by means of another solvent (**B**) and prepared for the next separation step on the column by means of amplification and diversification (**C**). In case of RNA molecules, step **C** is commonly achieved through reverse transcription RNA $\Rightarrow$ DNA by means of a reverse transcriptase, PCR amplification of the DNA, and transcription, DNA $\Rightarrow$ RNA. Often, reverse transcription and PCR amplification are sufficiently inaccurate for the creation of a sufficiently broad mutant spectrum. If necessary, error rates can be artificially enhanced. To increase the affinity the solvent for the retention is varied from selection-cycle to selection-cycle in the sense that stronger and stronger binding parameters are required to be retained. In general 20–30 selection-cycles are sufficient for the production of optimally binding aptamers. The binding parameters often are as high as with naturally evolved binders (Source: Refs. 52 and 53).

artificial genetic switch in the same bacterium [64], both in the year 2000, and

v. the chemical total synthesis of a genome and its implementation in a cell of the same bacterial species [65] in the year 2010.

The last few years saw a new although not far to seek application for synthetic DNA: the usage for storage of digital information [66]. Church and his coworkers at Harvard University synthesized a DNA that contains an entire book with 53,426 words, 11



DOI 10.1002/cplx

figures as jpeg-files, and one Java script stored in 5.27 megabit of its sequence. This storage facility holds the current record with respect to density of information with almost $10^{16}$ bits per $mm^3$, and beats thereby all physical storage devices including quantum holograms. DNA after all has been designed by biological evolution to be a storage device for information, and making copies by PCR is efficient and cheap compared to other memories with high density storage.

The key toward a novel DNA-based information technology is DNA synthesis in sufficiently high quantities at low price. The "next generation technology" in DNA-synthesis seems to meet this requirement, indeed [67] Scientists and engineers at the company *Gen9* in Cambridge (MA) announced that they are in a position to synthesize as much DNA as the rest of the world. It remains to be seen whether this promise can be fulfilled or not.

Recently—precisely on April 9, 2013—the American Chemical Society published an exposé with the title: "Engineering for the 21st Century: Synthetic Biology" [68]. The closing remarks of the essay will be taken from this report: "For years, scientists have hoped that biology would find its engineering counterpart—a series of principles that could be used as reliably as chemical engineering is for chemistry. Thanks to major advances in synthetic biology, those hopes may soon be realized." The production of DNA-constructs could well be such a biological core technology, says the author of the exposé. It will find numerous applications in diverse disciplines such as DNA-nanotechnology, targeted ribosomal protein synthesis with natural and artificial amino acids, and genetic variation of entire organisms. Similarly, as chemical core technology integrates a rich repertoire of surrounding processes into the well-defined discipline of chemical engineering, a new kind of biotechnology could serve as the common denominator for the large number of subareas in present day synthetic biology that look to us like a smorgasbord.

## REFERENCES


1. Schuster, P. Optimization of multiple criteria: Pareto efficiency and fast heuristics should be more popular than they are. Complexity 2012, 18(2), 5–7.
2. (a) Jacob, F. Evolution and tinkering. *Science* 1977, 196, 1191–1196; (b) The possible and the actual, Pantheon Books: New York, 1982.
3. Duboule, D.; Wilkins, A.S. The evolution of 'bricolage'. Trends Genet 1998, 14, 55–59.
4. Biochemical Pathway Maps. ExPASy: web.expasy.org/pathways/, Swiss Institute of Bioinformatics and Roche-Boehringer Mannheim GmbH Biochemie, 1993.
5. (a) Savageau, M.A. Biochemical systems analysis. I. Some mathematical properties of the rate law for the component enzymatic reactions J Theor Biol 1969, 25, 365–369; (b) II. The steady-state solutions for an *n*-pool system using a power-law approximation J Theor Biol 1969, 25, 370–379; (c) III. Dynamic solutions using a power-law approximation J Theor Biol 1970, 26, 215–226.
6. Garfinkel, D.; Garfinkel, L.; Pring, M.; Green, S.B.; Chance, B. Computer applications to biochemical kinetics. Annu Rev Biochem 1970, 39, 473–498.
7. Heinrich, R.; Rapoport, S.M.; Rapoport, T.A. Metabolic regulation and mathematical models. Prog Biophys Mol Biol 1977, 32, 1–81.
8. Joshi, A.; Palsson, B.Ø. Metabolic dynamics in the human red blood cell. Part I. A comprehensive kinetic model. J Theor Biol 1989, 141, 515–528.
9. Varma, A.; Palsson, B.Ø. Metabolic flux balancing: Basic concepts, scientific and practical use. (Nature) BioTechnology 1994, 12, 994–998.
10. Edwards, J.S.; Palsson, B.Ø. How will bioinformatics influence metabolic engineering? Biotechnol Bioeng 1998, 58, 162–169.
11. Price, N.D.; Reed, J.L.; Palsson, B.Ø. Genome-scale models of microbial cells: Evaluating the consequences of constraints. Nat Rev Microbiol 2004, 2, 886–897.
12. Nagrath, D.; Avila-Elchiver, M.; Berthiaume, F.; Tilles, A.W.; Messac, A.; Yarmush, M.L. Integrated energy and flux balance based multiobjective framework for large-scale metabolic networks. Ann Biomed Eng 2007, 35, 863–885.
13. Messac, A.; Ismail-Yahaya, A.; Mattson, C.A. The normalized normal constraint method for generating the Pareto frontier. Struct Multidiscip Opt 2003, 25, 86–98.
14. Sauer, U. Metabolic networks in motion: 13C-based flux analysis. Mol Syst Biol 2006, 2,:e62.
15. Schuetz, R.; Zamboni, N.; Zampieri, M.; Heinemann, M.; Sauer, U. Multidimensional optimality of microbial metabolism. Science 2012, 366, 601–604.
16. Kochanowski, K.; Volkmer, B.; Gerosa, L.; van Rijsewijk, B.R.H.; Schmidt, A.; Heinemann, M. Functioning of a metabolic flux sensor in *Escherichia coli*. Proc Natl Acad Sci USA 2013, 110, 1130–1135.
17. Mitchell, A. Romano, G.H.; Groisman, B.; Yona, A.; Dekel, E.; Kupiec, M.; Dahan, O.; Pilpel, Y. Adaptive prediction of environmental changes by microorganisms. Nature 2009, 460, 220–224.
18. Nixon, A.E.; Firestine, S.M. Rational and "irrational" design of proteins and their use in biotechnology. IUBMB Life 2000, 49, 181–187.
19. Knowles, J.R. Tinkering with enzymes: What are we learning? Science 1987, 236, 1252–1258.



DOI 10.1002/cplx

20. DeGrado, W.F.; Summa, C.M.; Pavone, V.; Nastri, F.; Lombardi, A. *De novo* design and structural characterization of proteins and metalloproteins. Annu Rev Biochem 1999, 68, 779–819.
21. Malakauskas, S.M.; Mayo, S.L. Design, structure and stability of a hyperthermophilic protein variant. Nat Struct Biol 1998, 5, 470–475.
22. Hellinga, H.W. Rational protein design: Combining theory and experiment. Proc Natl Acad Sci USA 1997, 94, 10015–10017.
23. Hellinga, H.W. Computational protein engineering. Nat Struct Biol 1998, 5, 525–527.
24. Guerois, R.; López de la Paz, M. Protein design: Methods and applications. In: *Methods in Molecular Biology*, Vol. 340; Humana Press: Totowa, NJ, 2007.
25. Korkegian, A.; Black, M.E.; Baker, D.; Stoddard, B.L. Computational thermostabilization of an enzyme. Science 2005, 308, 857–860.
26. Dantas, G.; Corrent, C.; Reichov, S.L.; Havranek, J.J.; Eletr, Z.M.; Isern, N.G.; Kuhlman, B.; Varani, G.; Merritt, E.A.; Baker, D. High-resolution and thermodynamic analysis of extreme stabilization of human precarboxypeptidase by computational protein design. J Mol Biol 2007, 366, 1209–1221.
27. (a) Kiss, G.; Çelebi-Ölçüm, N.; Moretti, R.; Baker, D.; Houk, K.N. Computerbasiertes enzyme design. Angew Chem 2013, 125, 5810–5836; (b) Computational enzyme design. Angew Chem Internat Ed 2013, 52, 5700–5725.
28. Jain, M. Enzymes used in detergents. Amity Institute of Biotechnology: Noida, India, 2008, available at: http://www.scribd.com/doc/11523933/Enzymes-in-Detergents (accessed on May 26, 2013).
29. Hasan, F.; Ali Shah, A.; Javed, S.; Hammed, A. Enzymes used in detergents: Lipases. Afr J Biotechnol 2010, 9, 4836–4844.
30. Schuster, P. Prediction of RNA secondary structures: From theory to models and real molecules. Rep Prog Phys 2006, 69, 1419–1477.
31. Waterman, M.S.; Smith, T.F. RNA secondary structure: Complete mathematical-analysis. Math Biosci 1978, 42, 257–266.
32. Zuker, M.; Stiegler, P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res 1981, 9, 133–148.
33. Hofacker, I.L.; Fontana, W.; Stadler, P.F.; Bonhoeffer, L.S.; Tacker, M.; Schuster, P. Fast folding and comparison of RNA secondary structures. Mh. Chem. 1994, 125, 167–188.
34. Andronescu, M.; Fejes, A.P.; Hutter, F.; Hoos, H.H.; Condon, A. A new algorithm for RNA secondary structure design. J Mol Biol 2004, 336, 607–624.
35. Busch, A.; Backofen, R. INFO-RNA: A fast approach to inverse folding. Bioinformatics 2006, 22, 1823–1831.
36. Zadeh, J.N.; Wolfe, B.R.; Pierce, N.A. Nucleic acid sequence design via efficient ensemble defect optimization. J Comput Chem 2011, 32, 439–452.
37. Reidys, C.; Stadler, P.F.; Schuster, P. Generic properties of combinatory maps. Neutral Networks of RNA secondary structures. Bull Math Biol 1997, 59, 339–397.
38. Flamm, C.; Hofacker, I.L.; Maurer-Stroh, S.; Stadler, P.F.; Zehl, M. Design of multistable RNA molecules. RNA 2001, 7, 254–265.
39. Schultes, E.A.; Bartel, D.P. One sequence, two ribozymes: Implications for the emergence of new ribozyme folds. Science 2000, 289, 448–452.
40. Mandal, M.; Boese, B.; Barrick, J.E.; Winkler, W.C.; Breaker, R.R. Riboswitches control fundamental biochemical pathways in *Bacillus subtilis* and other bacteria. Cell 2003, 113, 577–587.
41. Mills, D.R.; Peterson, R.L.; Spiegelman, S. An extracellular Darwinian experiment with a self-duplicating nucleic acid molecule. Proc Natl Acad Sci USA 1967, 58, 217–224.
42. Spiegelman, S. An approach to the experimental analysis of precellular evolution. Q Rev Biophys 1971, 4, 213–253.
43. Joyce, G.F. Forty years of *in vitro* evolution. Angew Chem Int Ed 2007, 46, 6420–6436.
44. Biebricher, C.; Eigen, M.; Gardiner, W.C., Jr. Kinetics of ribonucleic acid replication. Biochemistry 1983, 22, 2544–2559.
45. Strunk, G.; Ederhof, T. Machines for automated evolution experiments in vitro based on the serial transfer concept. Biophys Chem 1997, 66, 193–202.
46. Eigen, M. Selforganization of matter and the evolution of biological macromolecules. Die Naturwissenschaften 1971, 58, 465–532.
47. (a) Eigen, M.; Schuster, P. The hypercycle: A principle of natural self-organization. Part A: The emergence of the hypercycle. Die Naturwissenschaften 1977, 64, 541–565; (b) Part B: The abstract hypercycle. Die Naturwissenschaften, 1978, 65, 7–41; (c) Part C: The realistic hypercycle. Die Naturwissenschaften 1978, 65, 341–369.
48. Schuster, P. Evolutionary Biotechnology: Theory, facts and perspectives. Acta Biotechnol. 1996, 16(1), 3–17.
49. Schuster, P. How to search for RNA structures. Theoretical concepts in evolutionary biotechnology. J Biotechnol 1995, 41, 239–257.
50. Klussmann, S., Ed. The aptamer handbook. Functional oligdonucleotides and their applications. Wiley-VCh Verlag: Weinheim, 2006.
51. Brakmann, S.; Johnson, K.; Eds. Directed molecular evolution of proteins or how to improve enzymes for Biocatalysis. Wiley-VCh Verlag: Weinheim, 2002.
52. Tuerk, C.; Gold, L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science 1990, 249, 505–510.
53. Ellington, A.D.; Szostak, J.W. In vitro selection of RNA molecules that bind specific ligands. Nature 1990, 346, 818–822.
54. Arnold, F.H.; Wintrode, P.L.; Miyazaki, K.; Gershenson, A. How enzymes adapt: Lessons from directed evolution. Trends Biochem Sci 2001, 26, 100–106.
55. Romero, P.A.; Arnold, F.H. Exploring protein fitness landscapes by directed evolution. Nat Rev Mol Cell Biol 2009, 10, 866–876.



DOI 10.1002/cplx

56. Chao, F.-A.; Morelli, A.; Haugner, J.C., III; Churchfield, L.; Hagmann, L.H.; Shi, L.; Masterson, L.R.; Sarangi, R.; Veglia, G.; Seelig, B. Structure and dynamics of a primordial catalytic fold generated by in vitro evolution. Nat Chem Biol 2013, 9, 81–83.
57. W.A.D. Synthetic biology and the mechanism of life. Nature 1913, 91, 270–272.
58. Leduc, S. *La biologie synthétique*. A. Poinat, Editeur, Paris, 1912.
59. Jackson, D.A.; Symons, R.H.; Berg, P. Biochemical method for inserting new genetic information into DNA of simian virus 40: Circular SV40 DNA molecules containing λ phage genes and the galactose operon of Escherichia coli. Proc Natl Acad Sci USA 1972, 69, 2904–2909.
60. Maxam, A.M.; Gilbert, W. A new method of sequencing DNA. Proc Natl Acad Sci USA 1977, 74, 560–564.
61. Sanger, F.; Coulson, A.R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol 1975, 94, 441–448.
62. Sanger, F.; Nicklen, S.; Coulson, A.R. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 1977, 74, 5463–5467.
63. Elowitz, M.B.; Leibler, S. A synthetic oscillatory network of transcriptional regulators. Nature 2000, 403, 335–339.
64. Gardner, T.S.; Cantor, C.R.; Collins, J.J. Construction of a genetic toggle switch *in Escherichia coli*. Nature 2000, 403, 339–342.
65. Gibson, D.G.; Glass, J.I.; Lartigue, C.; Noskov, V.N.; Chuang, R.-Y.; Algire, M.A.; Benders, G.A.; Montague, M.G.; Ma, L.; Moodie, M.M.; Merryman, C.; Vashee, S.; Krishnakumar, R.; Assad-Garcia, N.; Andrews-Pfannkoch, C.; Denisova, E.A.; Young, L.; Qi, Z.-Q.; Segall-Shapiro, T.H.; Calvey, C.H.; Parmar, P.P.; Hutchinson, C.A., III; Smith, H.O.; Venter, J.C. Creation of a bacterial cell controlled by a chemically synthesized genome. Science 2010, 329, 52–56.
66. Church, G.M.; Gao, Y.; Kosuri, S. Next-generation digital information storage in DNA. Science 2012, 337, 1628.
67. Goldberg, M. BioFab: Applying Moore's Law to DNA Synthesis. Ind Biotechnol 2013, 9(2), 10–12.
68. Munnelly, K. Engineering fort the 21st Century: Synthetic Biology. ACS Synth Biol 2013, 2, 213–215.


DOI 10.1002/cplx