

## REPORT

# Recursive construction of perfect DNA molecules from imperfect oligonucleotides

Gregory Linshiz<sup>1,2,3</sup>, Tuval Ben Yehezkel<sup>2,3</sup>, Shai Kaplan<sup>1</sup>, Ilan Gronau<sup>1</sup>, Sivan Ravid<sup>1</sup>, Rivka Adar<sup>2</sup> and Ehud Shapiro<sup>1,2,\*</sup>

<sup>1</sup> Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel and <sup>2</sup> Department of Biological Chemistry, Weizmann Institute of Science, Rehovot, Israel

<sup>3</sup> These authors contributed equally to this work

\* Corresponding author. Department of Computer Science and Applied Mathematics, and Department of Biological Chemistry, Weizmann Institute of Science, Rehovot 76100, Israel. Tel.: + 972 8 9344506; Fax: + 972 8 947 1746; E-mail: ehud.shapiro@weizmann.ac.il

Received 8.1.08; accepted 13.3.08

**Making faultless complex objects from potentially faulty building blocks is a fundamental challenge in computer engineering, nanotechnology and synthetic biology. Here, we show for the first time how recursion can be used to address this challenge and demonstrate a recursive procedure that constructs error-free DNA molecules and their libraries from error-prone oligonucleotides. Divide and Conquer (D&C), the quintessential recursive problem-solving technique, is applied *in silico* to divide the target DNA sequence into overlapping oligonucleotides short enough to be synthesized directly, albeit with errors; error-prone oligonucleotides are recursively combined *in vitro*, forming error-prone DNA molecules; error-free fragments of these molecules are then identified, extracted and used as new, typically longer and more accurate, inputs to another iteration of the recursive construction procedure; the entire process repeats until an error-free target molecule is formed. Our recursive construction procedure surpasses existing methods for *de novo* DNA synthesis in speed, precision, amenability to automation, ease of combining synthetic and natural DNA fragments, and ability to construct designer DNA libraries. It thus provides a novel and robust foundation for the design and construction of synthetic biological molecules and organisms.**

*Molecular Systems Biology* 6 May 2008; doi:10.1038/msb.2008.26

*Subject Categories:* synthetic biology; computational methods

*Keywords:* automation; DNA synthesis; error correction; recursion; synthetic biology

This is an open-access article distributed under the terms of the Creative Commons Attribution Licence, which permits distribution and reproduction in any medium, provided the original author and source are credited. This licence does not permit commercial exploitation or the creation of derivative works without specific permission.

## Introduction

Making faultless complex objects from potentially faulty building blocks is a fundamental challenge in computer engineering (John Von Neumann, 1952), nanotechnology (Drexler, 1992; Merkle, 1997), and synthetic biology (Carr *et al.* 2004; Forster and Church 2006). Complex mathematical objects such as functions (Rogers, 1967), fractals (Mandelbrot, 1982), natural and formal languages (Chomsky, 1964; Hopcroft and Ullman, 1979), and computer data structures (Aho *et al.*, 1983) are typically described using recursion. Although the promise of recursion to physical construction has been recognized (Merkle, 1997), its application in engineering has been scarce (Knight, 2003; <http://www.sloning.de/>). Here, we present a recursive procedure for constructing faultless DNA molecules and libraries from faulty short synthetic oligonucleotides.

Long DNA molecules encoding novel genetic elements are in broad demand (Ryu and Nam, 2000; Tian *et al.*, 2004; Forster and Church, 2006; Heinemann and Panke, 2006); however, only short oligonucleotides (<100 nt) are made quickly and cheaply by machines (Caruthers, 1985). Such oligonucleotides are used as building blocks to construct longer DNA molecules using one of two basic construction strategies, namely polymerase cycling assembly (PCA) of multiple overlapping synthetic oligonucleotides (Stemmer *et al.*, 1995) and ligation of synthetic oligonucleotides (Au *et al.*, 1998).

The utility of synthetic DNA constructs in biology depends on their being free of sequence errors (Carr *et al.*, 2004; Tian *et al.*, 2004; Forster and Church, 2006), yet the synthetic oligonucleotides serving as their building blocks are error prone (about one sequence error per 160 nt) (Tian *et al.*, 2004; Forster and Church, 2006). Therefore, all DNA construction protocols struggle with the labor-intensive time-consuming task of cloning and sequencing synthetic DNA fragments,

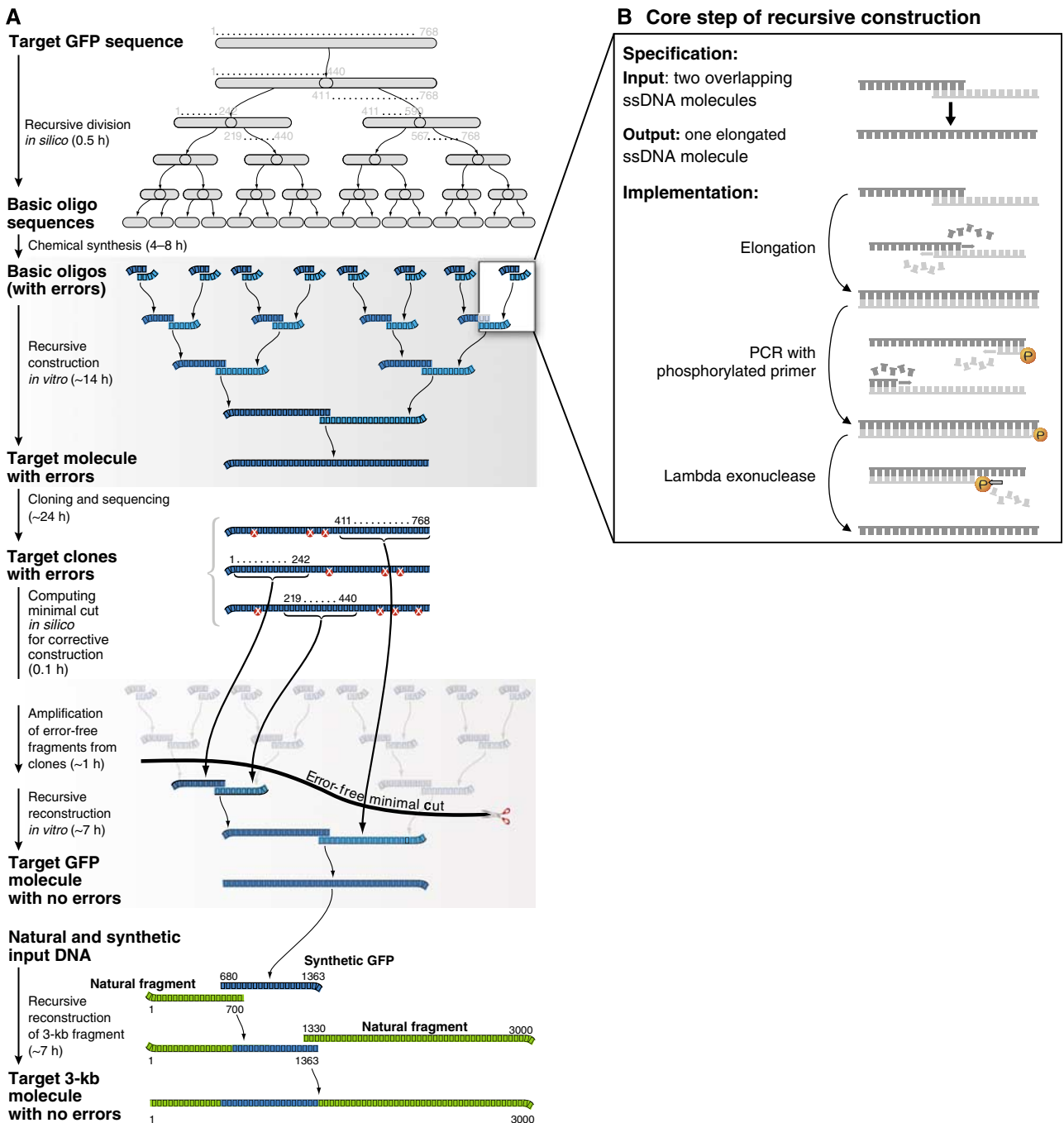
seeking an error-free one. If none is found, a clone with sufficiently few errors that can be patched without undue effort using site-directed mutagenesis (Hutchison *et al*, 1978) is used.

The problem is exacerbated for longer synthetic DNA since the probability of a molecule, and hence of a clone, to be error free decreases exponentially with its length. To partially address this problem, a two-step assembly process is commonly applied in which 300- to 500-bp fragments are constructed, cloned, sequence-validated and then assembled into the desired target molecule (Xiong *et al*, 2004). Other methods enrich error-free DNA molecules with the use of special mismatch-binding proteins (Tian *et al*, 2004; Forster

and Church, 2006) or improve site-directed mutagenesis (Xiong *et al*, 2006) to address this fundamental problem in *de novo* DNA construction.

Our procedure for constructing error-free DNA molecules integrates recursive construction and error correction. It uses Divide and Conquer (D&C) (Aho *et al*, 1983; Alsuwailay, 1999), the quintessential recursive problem-solving technique, to construct long DNA molecules from short oligonucleotides and then to error-correct the resulting molecules, until an error-free molecule is obtained.

D&C solves a problem (in our case, the construction of a particular ssDNA molecule) by dividing it *in silico* into two



smaller subproblems (in our case, the construction of two shorter ssDNA molecules, as shown in Figure 1 top); solving each subproblem recursively, using D&C; and combining *in vitro* the solutions to the subproblems into a solution to the original problem (in our case, combining the two ssDNA molecules into the desired longer ssDNA molecule, as shown in Figure 1). If the problem is small enough (in our case, the ssDNA molecule is short enough), it is not divided further but is solved directly (in our case, synthesized as an oligo).

Solving problems with D&C is naturally implemented using recursive procedures.

A fundamental prerequisite of a recursive procedure is that its output be of the same type as its inputs. Examples of DNA composition procedures that do not comply with this input-output compatibility requirement include overlap extension, which takes two ssDNA fragments that overlap at their 3' as input and produces the corresponding elongated dsDNA molecule as output, and PCA, mentioned above, which takes two or more overlapping DNA molecules as input and produces a mixture of the input molecules and some elongated dsDNA molecules as output. Our construction procedure (shown in Figure 1B and Supplementary Figures 1 and 2) is thus designed so that it accepts two overlapping ssDNA molecules as input and produces an elongated ssDNA molecule as its output (Figure 1B), utilizing three known enzymatic reactions: overlap extension between ssDNAs, PCR with 5' phosphate labeling and Lambda exonuclease-mediated ssDNA generation. It can be applied recursively since its input and output are of the same type (ssDNA). In principle, a recursive construction procedure that uses dsDNA as its input and output can also be devised. We chose ssDNA rather than dsDNA because the extension of overlapping ssDNA molecules can be performed in quasi-equilibrium (i.e. denaturation and then very slow cooling to annealing temperature), thereby greatly improving control, yield and specificity (see Results for CE fragment analysis of composition reactions) of elongation products. This is in contrast to the rapid thermal cycling conditions commonly used when elongating two or more dsDNA molecules, which often result in low elongation yield and in nonspecific elongated products (see Supplementary Figure 3).

The D&C recursive algorithm receives a user-specified target sequence as its input and returns as output a list of oligos to be

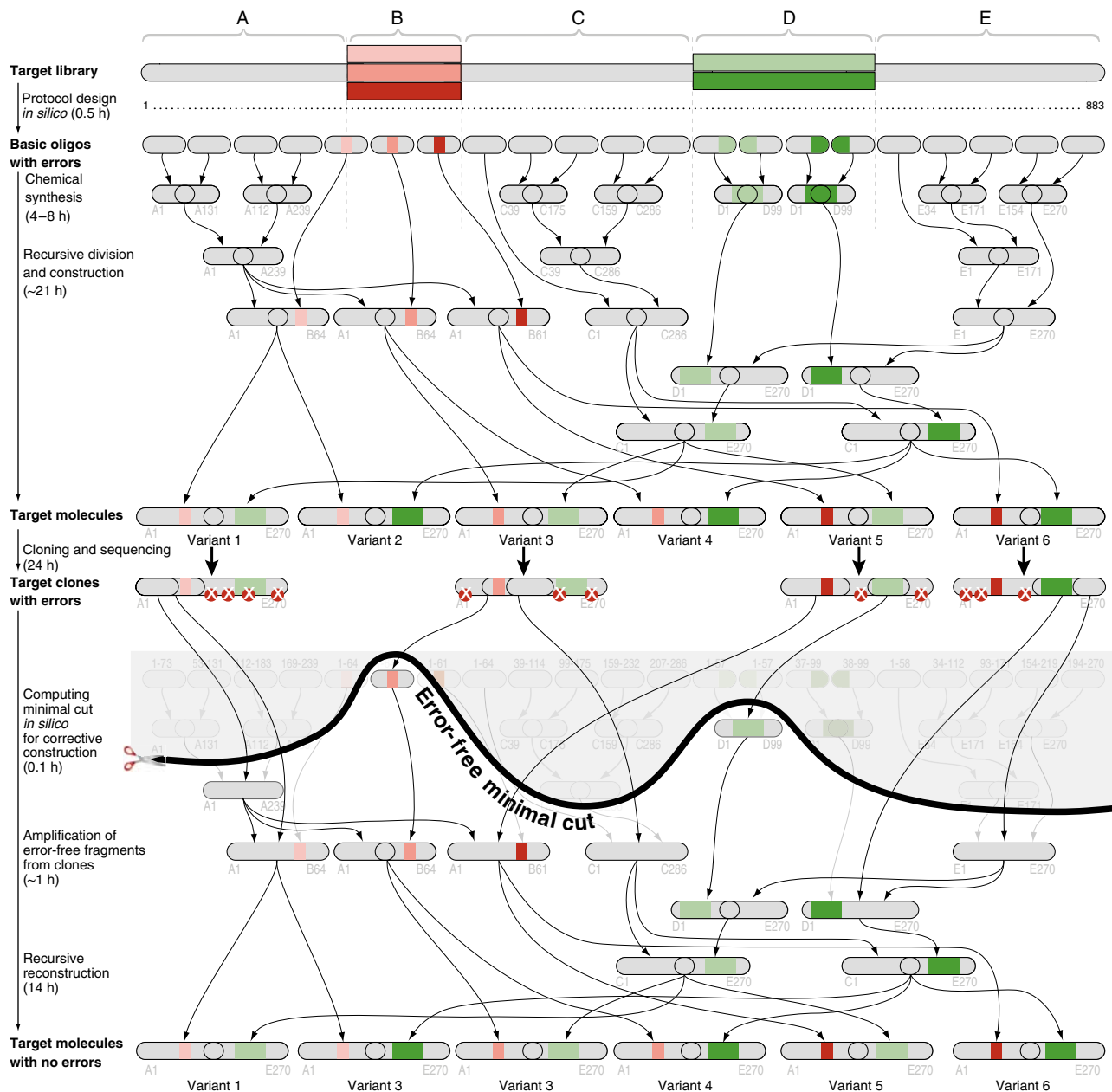
synthesized and a protocol in the form of a robot control program that can be used to construct the desired DNA molecule using the specified set of oligos.

The basic recursive subroutine of the algorithm takes as input the sequence of a target molecule and returns as output a recursive construction protocol and its associated cost. This subroutine divides the target sequence into two overlapping sequences and calls itself recursively with these subtarget sequences as new input. The cost of constructing the target molecule by this protocol is computed by adding the cost of assembling the two overlapping subfragments to the cost of constructing these two individual subfragments. The computed cost accounts for the various features of the construction process, including the number and length of oligos, number of reactions and the total number of levels in the protocol (see Supplementary information). The recursive division ends if the subroutine's target is short enough to be synthesized directly as an oligonucleotide.

Division points are not chosen so that oligos are of equal length, as usually practiced in PCA methods (Smith *et al*, 2003). Instead, division points are selected to minimize the cost of constructing the target and to respect a set of constraints, including whether good PCR primers exist for each of the subtargets and whether the two subtargets can be elongated together efficiently and specifically in the elongation reaction described in Figure 1B. Validation of specificity and affinity of elongation overlaps and PCR primers is performed using sequence alignment algorithms and  $T_m$  calculations, respectively (see Supplementary information). The optimized recursive protocol is then transformed into a robot control program that instructs the robot to construct the molecule bottom-up. It starts with the leaves of the recursive construction tree and iteratively executes the basic chemical step (Figure 1B) all the way up to the root of the tree until the target molecule is constructed.

The hierarchical structure of our procedure, induced by the use of recursion, enables DNA construction by pairwise composition reactions that are performed independently of each other and in equilibrium, which greatly increases the predictability (and hence amenability to automation) of the core biochemical reactions of our procedure. The hierarchical structure of the recursive construction tree is also at the foundation of our error correction procedure.

**Figure 1** Recursive construction of error-free DNA molecules from error-prone oligonucleotides. **(A)** Recursive construction of the *GFP* DNA. The Divide & Conquer procedure, as applied to the construction of the 768-nt *GFP*, is illustrated from top to bottom. The target sequence is recursively divided *in silico* into overlapping oligonucleotide sequences (16 oligos of average size 75 bp for the synthesis of *GFP*). The specified oligos are synthesized by conventional means and serve as inputs (in blue) for recursive construction, performed *in vitro*. Construction proceeds by recursively combining pairs of overlapping ssDNA molecules into ever longer ssDNA molecules, as described in **(B)** until the target molecule is formed. Target molecules thus produced typically have the same error rate as their source oligos, and hence are subject to recursive error correction as follows. A certain number of target molecules are cloned and sequenced (this number is optimized as described in the text, seven in the case of *GFP*). Errors (marked in red) are identified. Error-free segments found in the clones are then amplified from the clones and used as inputs to another recursive reconstruction of the target molecule (one half molecule and two quarter molecules in this case). The error-free segments are chosen to correspond to nodes in the recursive construction tree, so that they can be amplified using the same primers used in the initial procedure and are further optimized, using the mathematical notion of minimal cut in a graph (explained in the text) so as to minimize the number of reactions needed for reconstruction (only 2 reactions out of the total of 15 in this case). This second iteration of the procedure typically (as in this case and all our experiments to date) results in an error-free clone. However, if errors remain another error-correcting iteration of the procedure can be performed. The figure further demonstrates the construction of a 3-kb DNA fragment by combining, using the same construction procedure, the synthetically produced *GFP* molecule and DNA from a natural source as input (bacterial plasmid, in green), which yielded an error-free molecule. Expected optimal times for each step using state-of-the-art standard equipment are shown on the left. The cloning step could potentially be replaced by single molecule PCR. **(B)** The core step of recursive construction receives two overlapping ssDNA molecules as inputs and produces the elongated ssDNA molecule as output, as follows: the overlapping ssDNA molecules hybridize and prime each other for an overlap extension elongation reaction to form a dsDNA molecule (elongation), which is then amplified by PCR with one of the two primers phosphorylated at its 5' end (PCR with phosphorylated primers). The phosphate-labeled PCR strand is then degraded with Lambda exonuclease, yielding an elongated ssDNA molecule as output (Lambda exonuclease).



**Figure 2** Recursive construction and error correction of a simple combinatorial library. The recursive construction of six *p53* variants is illustrated top to bottom: a diagram describes the shared (A, C, E in gray) and unique (B, D, colored) components of the target *p53* combinatorial DNA library. A library construction protocol is computed, where target library sequences are recursively divided into shared and unique components and then further divided into basic oligonucleotide sequences, which are then synthesized conventionally. Gray oligos are shared by all library variants, and colored segments are used by variants with the corresponding colors. Oligos are recursively combined *in vitro* as shown to form the six target *p53* variants. These variants were cloned and sequenced, and errors were identified (marked in red on top of clones). An error-free minimal cut (the non-faded part of the graph below the minimal cut black line) of the library construction graph was computed from only four error-prone clones (variants 1, 3, 5 and 6). Error-free segments out of these clones (delimited) were used as inputs for another iteration of the recursive reconstruction protocol, this time producing error-free clones of all six target library members. Expected execution times for each step using standard equipment are shown on the left.

The molecules produced by the first iteration of our recursive construction procedure are error prone (see Supplementary Table 1) and have the same error rate as the oligos used to produce them. Our recursive construction procedure enables a novel error-correction strategy that employs the very same construction methodology and reagents to produce error-free molecules. Like previous DNA construction proto-

cols (Tian *et al*, 2004), our error-correction procedure uses cloning and sequencing to identify faults, but unlike previous protocols it does not require additional or external methods or reagents to turn the error-prone DNA into error-free DNA. The overall strategy is described in Figure 1: short oligos are used as error-prone basic components and composed as described above till the target DNA molecule is constructed. However,

unlike other methods, if no error-free molecules are found by cloning and sequencing, then error-free parts of the erroneous target DNA molecules are identified and used as new, typically longer, inputs to the same recursive construction procedure. Since this construction starts from typically larger DNA-building blocks that are error free, the number of errors in the resulting reconstructed DNA is expected to decrease, possibly down to zero, eschewing additional screening of clones.

Specifically, the error-prone clones from the initial construction are analyzed to find a minimal cut in the recursive construction tree, defined as follows (see also mathematical definitions in Supplementary information). A node in the tree is said to be covered by a set of clones if its sequence occurs error free in at least one of the clones. A set of clones induce a minimal cut on the tree, defined to be the set of the most shallow (closest to the root) nodes in the tree that are covered by the clones. If some leaf is not covered it means that the oligo is erroneous in all clones. In such a case, we can either analyze additional clones in the hope to find that leaf error free and re-compute the minimal cut or, if we reason that a systematic error has occurred in the synthesis of an oligo (i.e. the same error is represented uniformly in all clones), then there is no reason to analyze additional clones and we simply re-synthesize that oligo and try again. Mathematically, we simply assume that the newly ordered oligo would cover the leaf node and proceed with the computation of the minimal cut. Since the boundaries of the error-free DNA fragments that constitute the minimal cut coincide with boundaries of fragments of the initial recursive construction tree they can be extracted from their respective clones using PCR and the same primers used in their corresponding composition step (Figure 1B). As a result, no additional methods or reagents are needed to obtain error-free molecules beyond those used in the initial construction.

Moreover, based on the known rate and distribution of errors we can predict the number of times error-free components will occur in a given number of constructed objects. Furthermore, we can calculate the probability that a certain number of error-free components would collectively span the entire target object. Conversely (and more importantly), we can calculate the number of object copies (clones) required so that their error-free components span the entire target object with a desired probability (chosen to be 95% in this work, see Supplementary information).

Indeed, in all our experiments, a single re-application of the recursive construction procedure, using as input error-free components copied using PCR from molecules produced during the first application of the procedure, yielded error-free synthetic DNA molecules out of almost every clone.

## Results and discussion

We constructed the gene for *GFP* using the process shown in Figure 1. The construction-protocol-generating algorithm (Figure 1, top and Supplementary information) recursively divided the target sequence into basic overlapping oligos according to multiple criteria (see Supplementary information) using D&C (Aho *et al*, 1983; Sloning, BioTechnology GmbH, 2006) (Figure 1 top). The oligos were ordered from a

commercial provider (see Supplementary information) with standard desalting. The algorithm also generated a liquid-handling robot control program, using a robot programming language developed by one of us (see <http://www.weizmann.ac.il/udi/papers/rpl.pdf> for detailed description) that controlled the execution of the construction protocol by the robot using only off-the-shelf reagents (as shown in Figure 1). While the protocol can be executed fully automatically using standard commercially available reagents and robotic peripheral equipment, in the protocols used for the construction of *GFP* and in the other constructions reported here some procedures specified by the robot program were performed manually (see Supplementary information) due to lack of the relevant robotic peripheral equipment (robotic centrifuge for plates). This resulted in construction times longer than those specified in the fully automated timeline accompanying Figures 1 and 2. We also integrated automated quality control monitoring at all stages of the recursive construction and error-correction procedure including capillary electrophoresis fragment analysis of all fragments that occur during construction to a single base-pair resolution, gel electrophoresis, real-time PCR and DNA sequencing (see Supplementary figures for all these controls). The robot control program instructed the robot to recursively construct the *GFP* DNA molecule. DNA molecules produced from the first iteration of the automated recursive construction process described in Figure 1 (and Supplementary Figure 4) were cloned, sequenced and their errors reflected an error rate of  $\sim 1/160$ , as expected, reflecting the error rate of unpurified desalted synthetic oligonucleotides (Hecker and Rill, 1998; Tian *et al*, 2004). Given that errors are distributed randomly and with a known rate, we computed the minimal number of clones required to obtain an error-free minimal cut with a maximum depth four (Supplementary information). Practically, this means that we could expect to be able to 'lift' from these clones error-free molecules that can be used as input for re-application of the recursive construction procedure, but this time with a recursive construction tree of depth at most four.

The actual minimal cut of depth two for the *GFP* sequence, shown in Figure 1, was computed using three clones (see also Supplementary information). The error-free fragments constituting this minimal cut were used as input for a re-application of the recursive construction procedure (see Supplementary Figure 5), which resulted in an error-free clone. From the clones produced in the first iteration, we could have computed a minimal cut of depth one using only a pair of clones for reconstruction (see Supplementary Figure 6), one for each half of the target molecule. Instead, we chose to show a minimal cut consisting of three clones, one contributing about a half and two contributing about a quarter each of the target molecule, for illustrative purpose. The clones produced in the corrective construction show an error rate of  $< 1/5000$ , reflecting a  $> 30$ -fold decrease in error rate compared to the starting material and approaches the error rate of the DNA polymerase used in the construction process. This might be further improved in the future by using polymerases with higher fidelity. The entire process of automated *de novo* construction and error correction of the *GFP* molecule according to our method was repeated by an external student. Capillary fragment analysis and gel electrophoresis of each

step in the construction and reconstruction process reproduced our results. Sequencing results also reproduced our results with respect to construction and reconstruction robustness and error rates, resulting in similar construction times and minimal cuts.

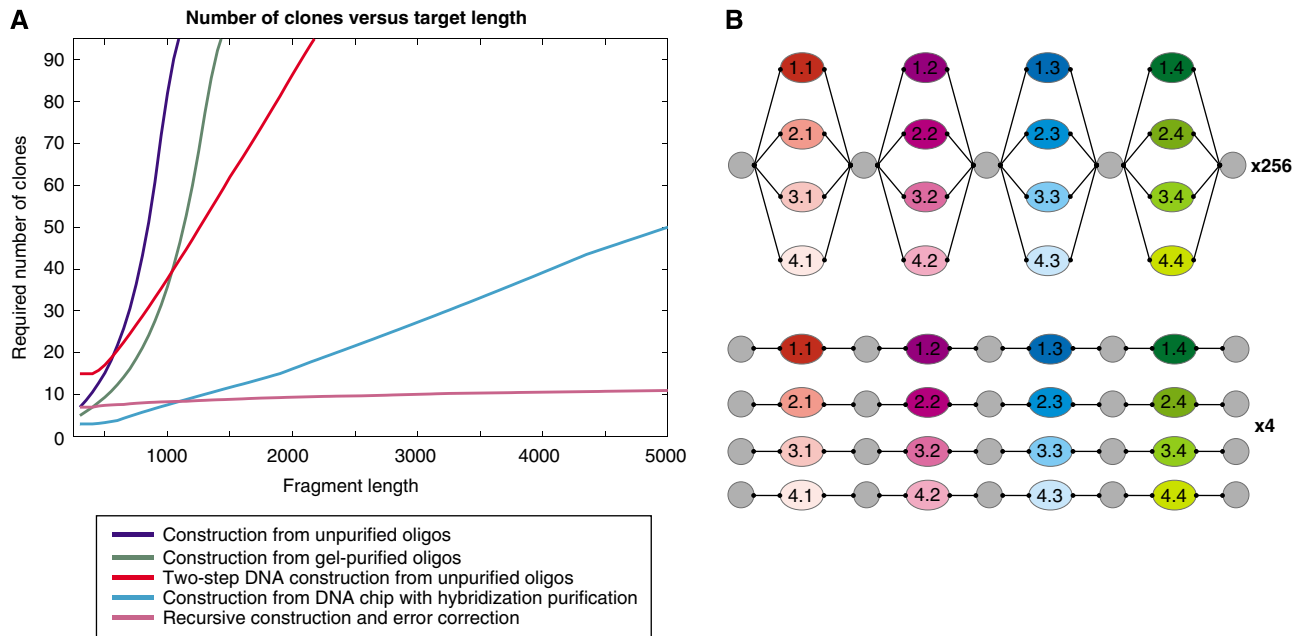
If any fragment of the target sequence is already available as existing DNA (say in a plasmid or in previously constructed DNA), the algorithm can take this information into account and use these fragments as input to the construction process instead of synthesizing it from basic oligos (Figure 1). To illustrate this and that recursive construction can also be used to construct longer fragments, we recursively constructed a 3 kb-long molecule by composing the previously constructed synthetic *GFP* molecule with two more sequences present on a plasmid, 700 nt and 1700 nt long (Figure 1 and see Supplementary Figures 7 and 8). This was executed using the same principles used for constructing shorter sequences only this time using the synthetic *GFP* molecule and a plasmid as input instead of synthetic oligos.

To further test the robustness of our protocol, we used it to recursively construct the *Escherichia coli* codon usage optimized 823-bp-long *TachylectinII* gene. Low complexity genes, like the *TachylectinII* (which utilizes a minimal set of codons (20) and consists of five nearly identical subunit repeats), pose a potential challenge to DNA synthesis methods that perform elongation reactions during construction (Tian *et al*, 2004). This is due to its repetitive sequence elements which, if positioned at the 3' of oligos or any other fragment that occurs in the recursive construction tree (and therefore in real construction), may lead to miss-priming and to subsequent formation of nonspecific products. Since our method is hierarchical we can spot the elements that are repetitive and separate them into different reactions. Also, our algorithm designs the oligos and all other fragments that occur in the recursive construction tree to have unique 3' termini that promote specific elongation reactions. This is crucial condition full automation, which is hindered by nonspecific products. We were able to recursively construct the low complexity *TachylectinII* gene in a single automated application of the recursive construction procedure (see Supplementary Figures 9 and 10 for detailed account of results). A visualization of the fragments that occurred in the recursive construction tree is presented on top of a dot plot revealing the repetitive elements in the *TachylectinII* gene (see Supplementary Figure 11). It shows how our algorithm breaks down the DNA sequence into fragments that minimize miss-priming during construction by positioning the repetitive elements away from parts that can lead to miss-priming (i.e. 3' termini of fragments that occur in the recursive construction tree). The sequences of all oligos, primers, construction intermediates and full lengths reported in this work are available online (see Supplementary information).

The basic principles used to construct DNA molecules can also be applied to construct DNA libraries. DNA libraries are an important source for selecting molecules encoding novel genetic sequences for use in medicine, research and industry (Heinemann and Panke, 2006). Numerous methods for constructing large DNA libraries, mostly by random recombining (Coco *et al*, 2001) and mutagenesis (Cadwell and Joyce, 1992) have been developed for directed evolution (Matsuura

and Yomo, 2006). On the other hand, in the computation-intensive practice of rational design and study of polymers only a small number of specified constructs, typically generated by site-directed mutagenesis (Caruthers, 1985) are investigated experimentally (Cedrone *et al*, 2000). Recursive construction can be extended to produce error-free combinatorial DNA libraries with pre-specified and/or randomized members. Most construction methods deliver combinatorial libraries in 'one pot', which poses a limitation on the methods that can be used for their screening. Our library construction protocol can deliver each library member separately, say in a separate well of a plate, which may facilitate a richer set of screening methods. In addition, the starting material for the libraries can be either natural or synthetic DNA. We demonstrate the feasibility of building user-specified combinatorial DNA libraries by constructing a small library containing six variants of the *p53* gene, specified in Figure 2. The mutants of the library were user-specified (i.e. site-directed) and were chosen arbitrarily, to demonstrate the creation of libraries of mutants with our method. First, target library DNA sequences are analyzed *in silico* identifying segments that are unique and shared between library members, so that shared segments are only produced once and not separately for each variant. These segments are further divided into overlapping oligos. The recursive division algorithm searches for an optimal library construction protocol based on chemical constraints and a cost function, to minimize the number of components and reactions needed to construct the entire library (Supplementary information). All six different *p53* genes were recursively constructed in an automated manner from basic unpurified oligos (Figure 2 top and Supplementary Figure 12), and the resulting molecules were cloned and sequenced (Figure 2 center). In this application of library construction, our error-correction method becomes even more efficient since we only need to find one error-corrected instance of fragments that are shared between several library members. An error-free minimal cut of the entire library was computed in this way from only four clones, and a corrective construction process using the specified error-free fragments from these four clones produced error-free clones of all six full-length library members (Supplementary Figure 13), as predicted (see Supplementary information). The error rate of the uncorrected clones was, as in previous constructions, 1/160, and a total of 1000 nt of error-free fragments taken from these four faulty clones were sufficient to generate (in one error-correcting procedure) a complete library of six members which contain together more than 5200 nt error-free nucleotides (see Supplementary Figure 14).

The clones produced from the corrective construction show an error rate better than 1/5700, computed over 86 000 nt of sequenced clones (see Supplementary Table 1). Moreover, in the future error correction of larger libraries can be further economized. For example, in the construction of a library with 256 members (Figure 3B top), a subset of only four clones containing all library components (Figure 3B bottom) should be initially constructed and error corrected. Only then, should all 256 members of the library be constructed from these four error-free corrected clones. In hindsight, we could have used the same principle to the *p53* library and could have reconstructed it from only three clones instead of four. This



**Figure 3** Comparative analysis of error-correction methodologies. **(A)** Error correction of a single molecule. The required number of clones that have to be sequenced to obtain an error-free synthetic DNA molecule as a function of its length is shown for different methods of construction: naïve construction from synthetic oligos with no error correction (blue); construction from gel-purified oligos (green); a two-step DNA construction, where in the first step molecules of length 500 are constructed, cloned, sequenced, and in the second step these error-free molecules are used as building blocks for larger molecules (red); a two-step construction from oligos purified by hybridization (Tian *et al*, 2004) (cyan); and recursive construction with iterative error correction (purple) (see Supplementary information for mathematical analysis). **(B)** Error correction of libraries: a graph representing a DNA library with four variable sites, each containing four variants, totaling 256 possible library members (top). Using recursive construction one can first construct and error correct a representative set of only four library members, which constitute a minimal cut through the construction graph of the entire library. A subsequent iteration of the protocol can use error-free fragments obtained from these four library members to efficiently construct the entire 256-strong library. This dramatically economizes the error correction of libraries compared to the correction of each library member separately, as presented in (A).

principle, of first constructing and error correcting a minimal kernel from which the entire library can later on be generated, improves on the efficiency of our error correction for libraries compared to error correction for single sequences (shown in Figure 3A). By applying the principles outlined above, we are currently constructing larger pre-specified DNA libraries and believe this may become a routine molecular biology procedure in the future.

A major outcome of our work is that it provides a platform with which combinatorial libraries can be constructed where each library member is provided separately (e.g. in a separate plate well). This would allow screening each library member independently and, once a successful member is found its sequence can be known immediately. Naturally, some parts of any library member can be randomized, as in standard combinatorial libraries. In this case, we would not need to apply error correction to the randomized positions since they are designed to be variable.

Complex human-made objects are usually constructed hierarchically: buildings (floor, apartment, room, wall, brick), airplanes (body, wing, flap, screw) and of course computers. Hierarchical construction requires a different procedure at each level: the procedure for assembling an engine is different from that for assembling a flap, and both are different from the procedure for assembling a wing. This is necessary since the input objects (e.g. engine, flap) and the output objects (assembled wing) of each hierarchical construction procedure are of a different type. In contrast, in a recursive procedure in

general, and in a recursive construction procedure in particular, the inputs and outputs are of the same type. The immediately apparent benefit of recursive construction is that the same procedure is used at all levels of the hierarchy, which makes the entire process efficient and scalable. A less apparent benefit is the ability to employ our error-correcting procedure, which seeks error-free subcomponents in previously constructed objects and reuses them in another recursive construction attempt. The uniformity of recursive construction enables mixing such subcomponents from various levels of the hierarchy without any difficulty.

*In vitro* pairwise composition, as reported here, compared to ‘one-pot’ PCA of multiple overlapping DNA fragments, enables finer control over reaction conditions and the interactions between the DNA-building blocks, thus reducing the formation of by-products. On the other hand, pairwise construction requires a larger number of reactions than one-pot assembly. Therefore, up to a certain length (of  $\sim 500$  bp) one-pot assembly may sometimes, but not always, be less expensive and/or time consuming. However, whether PCA would work or not cannot be reliably predicted, and unpredictable failures often hinder the assembly process. Furthermore, in one-pot construction of fragments longer than  $\sim 500$  bp, traditional PCA methods often suffer from faulty construction attempts and the need to separate correct from incorrect products. Such separation is typically done by extracting accurately sized fragments out of a gel, hindering automation. In addition, predicting by computational methods the potential interac-

tions between reaction components is easier for pairwise reactions, as in the recursive composition procedure, than in reactions with multiple components such as PCA.

Regarding the error rate of the synthetic oligo-building blocks, we have taken into consideration the nonlinear relationship between oligo length and mutation rate. Nonetheless, we have chosen to optimize for longer construction oligos since shorter oligos come with the cost of performing more reactions, the cost of which is integrated into our cost function. The reduction in error rate due to shortening of oligos is small ( $\sim 2$ -fold) compared to the reduction achieved with our directed error correction ( $\sim 30$ -fold); therefore, the saving in the number of reactions due to longer oligos is cost effective. More importantly, our method incurs only a small addition in cost due to the higher error rate in longer oligos compared to shorter ones, since the number of clones we need to construct an error-free molecule increases only linearly with the error rate of the oligos, and not exponentially as in other methods, see Figure 3A.

An important feature of our error-correction procedure is that it bypasses a major obstacle in constructing synthetic DNA, namely the exponential decrease in the fraction of error-free molecules with the length of the molecule, as seen in naïve approaches to DNA synthesis (Figure 3A, blue plot). This is possible since our error-correction procedure avoids the difficult task of finding complete error-free molecules. Instead, it efficiently utilizes small error-free parts and combines them back into an error-free target molecule. The probability of finding an error-free fragment of a fixed small size is high and (more importantly) fixed regardless of the overall length of the target molecule. Hence the small linear increase in the number of clones needed to construct increasingly larger error-free target molecules (Figure 3A, purple plot) compared to the exponential increase in the number of clones needed when constructing DNA without any error correction (Figure 3A, blue plot). Even if some sort of building block (oligo) purification is applied, e.g. PAGE purification (Figure 3A, green plot), the number of clones still becomes overwhelming in the construction of DNA several kilobase pairs long.

Other methods for DNA synthesis also employ a hierarchical strategy in construction and error correction. For example, fragments of  $\sim 500$  bp are constructed by PCA, cloned and screened for error-free molecules, which are then combined into larger fragments by different methodologies (Xiong *et al*, 2004). Such a two-step construction strategy is compared to ours in Figure 3A (red plot). Although we are not aware of evidence that PCA works with automation level robustness at  $\sim 500$  bp, for this plot we assumed it does and that cloning of PCA products occur uniformly at this length. The purification of initial building blocks by PAGE (Figure 3A, green plot) or even an improved building block purification technology (Tian *et al*, 2004) combined with a two-step assembly process (Figure 3A, cyan plot) still do not avoid the large number of molecules that need to be screened to construct molecules several kilobase pairs long.

Other error-correction methods not presented in Figure 3A include those which enrich error-free DNA molecules with the use of special mismatch binding or cleaving proteins (Carr

*et al*, 2004; Forster and Church, 2006; Bang and Church, 2008) or improve site-directed mutagenesis (Xiong *et al*, 2006). The former requires the use of special mismatch-binding proteins and is limited to relatively short fragments with only a few errors. The latter performs corrective PCR with corrective primers for each error, which requires both the retrospective synthesis of new PCR primers for each such error and that the newly corrected PCR fragments be combined back into the target sequence. The fact that the identity of the new PCR fragments and the resulting structure of the construction protocol are dictated by the random distribution of errors and not by engineering considerations impairs robustness and hence amenability to automation. This is also why we do not choose any error-free fragments from our clones or design new primers which span them, but only the ones that coincide with fragments from our construction plan.

We cannot provide actual dollar costs of executing the protocol at this stage, however, a framework for designing and selecting construction protocols that minimize the cost of the process (as described in the paper and in Supplementary information) has been established. In general, the major costs that require reduction in DNA synthesis are the costs associated with (the typically manual labor intensive) production of clones and the cost of sequencing their DNA. The magnitude of these tasks is dramatically reduced using our method, as shown in Figure 3.

We have demonstrated recursive construction and error correction of DNA several kilobase pairs long, accounting for most genes, on producing longer molecules using our methods is a subject of current work. We expect to be able to use our method up to the limit of long-range PCR (about 20–30 kb). Going beyond that limit would probably require shifting from the *in vitro* system reported here to *in vivo* systems capable of copying and maintaining DNA fragments of this length.

Recursive construction improves on previous approaches to DNA synthesis (Stemmer *et al*, 1995; Au *et al*, 1998; Gao *et al*, 2003; Smith *et al*, 2003; Tian *et al*, 2004; Xiong *et al*, 2004) by enabling rapid, fully automated construction of long error-free synthetic DNA molecules. It performs construction *in vitro* and therefore requires no *in vivo* selection steps inherent to some methods (Knight, 2003; Kodumal *et al*, 2004) and has no constraints regarding avoidance or inclusion of restriction sites; it reduces the error rate  $\sim 30$ -fold compared to construction from standard oligos (see Supplementary Table 1) and dramatically decreases the number of clones that have to be sequenced to make an error-free molecule (Figure 3); it easily combines synthetic and natural DNA fragments; and it enables efficient design and accurate synthesis of exactly pre-specified combinatorial DNA libraries with shared and variable components.

We demonstrated recursive construction and error correction of long DNA molecules and libraries employing standard available technology. Additionally, our recursive construction and error-correction method can take full advantage of other improvements in biochemical methods for DNA error correction (Carr *et al*, 2004), of advances in oligo synthesis, including synthesis on a chip (Tian *et al*, 2004) and of improvements in liquid handling such as microfluidic ‘lab on a chip’ technology (Whitesides, 2006).

## Materials and methods

The core recursive construction step (Figure 1B) requires four basic enzymatic reactions: phosphorylation, elongation, PCR and Lambda exonuclease. They are described in the order of execution by our protocol:

### Phosphorylation of all PCR primers used by the recursive construction protocol is performed beforehand simultaneously, according to the following protocol

5' DNA termini (300 pmol) in a 50  $\mu$ l reaction containing 70 mM Tris-HCl, 10 mM MgCl<sub>2</sub>, 7 mM dithiothreitol, pH 7.6 at 37°C, 1 mM ATP, 10 U T4 polynucleotide kinase (NEB). Incubation is at 37°C for 30 min and inactivation is at 65°C for 20 min.

### Overlap extension elongation between two ssDNA fragments

5' DNA termini (1–5 pmol) of each progenitor in a reaction containing 25 mM TAPS pH 9.3. at 25°C, 2 mM MgCl<sub>2</sub>, 50 mM KCl, 1 mM  $\beta$ -mercaptoethanol, 200  $\mu$ M each of dNTP, 4 U Thermo-Start DNA Polymerase (ABgene). Thermal cycling program is as follows: enzyme activation at 95°C for 15 min, slow annealing at 0.1°C/s from 95 to 62°C, elongation at 72°C for 10 min.

### PCR amplification of the above elongation product with two primers, one of which is phosphorylated

Template (1–0.1 fmol), 10 pmol of each primer in a 25  $\mu$ l reaction containing 25 mM TAPS pH 9.3 at 25°C, 2 mM MgCl<sub>2</sub>, 50 mM KCl, 1 mM  $\beta$ -mercaptoethanol, 200  $\mu$ M each of dNTP, 1.9 U AccuSure DNA Polymerase (BioLINE). Thermal Cycler program is: enzyme activation at 95°C for 10 min, denaturation at 95°C, annealing at  $T_m$  of primers, extension at 72°C for 1.5 min per kb to be amplified 20 cycles.

### Lambda exonuclease digestion of the above PCR product to re-generate ssDNA

5' Phosphorylated DNA termini (1–5 pmol) in a reaction containing 25 mM TAPS pH 9.3 at 25°C, 2 mM MgCl<sub>2</sub>, 50 mM KCl, 1 mM  $\beta$ -mercaptoethanol, 5 mM 1,4-dithiothreitol, 5 U Lambda Exonuclease (Epicentre). Thermal Cycler program is: 37°C for 15 min, 42°C for 2 min and enzyme inactivation at 70°C for 10 min.

### Chemical oligonucleotide synthesis

Oligonucleotides for all experiments were ordered by commercial providers (Sigma Genosys and IDT) with standard desalting.

### Automated DNA purification

Automated DNA purification was performed with Qiagen's QIAquick 96-well PCR purification kit using standard protocols adapted to work with Tecan Freedom 200 and a vacuum manifold.

### Preparation of reactions

The preparation of all construction reactions listed above including QC sampling for capillary and gel electrophoresis were done automatically by a Tecan Freedom 200 liquid handling robot controlled with in-house developed software.

## Protocol automation

Parts of protocol that were executed automatically were performed by a Tecan Freedom 200 robot mounted with a Biometra T-Robot PCR block controlled with in-house developed software. Some parts were not performed robotically due to lack of automation-related equipment in our lab. Some DNA purifications were done manually using a tabletop microcentrifuge due to the lack of an automated plate centrifuge in our lab. Transfer of capillary electrophoresis and RT-PCR plates from the robot to their slots in the corresponding machinery was also done manually due to lack of a robotic arm that does so in our lab.

## DNA purification

Manual DNA purification was performed with Qiagen's MinElute PCR purification kit using standard procedures.

## Cloning

Fragments were cloned into the pGEM T easy Vector System1 from Promega. Vectors containing cloned fragments were transformed into JM109 competent cells from Promega1 and sequenced.

## Supplementary information

Supplementary information is available at the *Molecular Systems Biology* website ([www.nature.com/msb](http://www.nature.com/msb)).

## Acknowledgements

This research was supported by the Yeshaya Horowitz Association through the Center for Complexity Science, research grant from Dr Mordecai Roshwald, grant from Kenneth and Sally Leafman Appelbaum Discovery Fund, the Estate of Karl Felix Jakubskind, the Estate of Funnie Sherr, the Clore Center for Biological Physics and The Louis Chor Memorial Trust. Ehud Shapiro is the Incumbent of The Harry Weinrebe Professorial Chair of Computer Science and Biology and of The France Telecom—Orange Excellence Chair for Interdisciplinary Studies of the Paris 'Centre de Recherche Interdisciplinaire' (FTO/CRI).

## References

- Aho AV, Hopcroft JE, Ullman JD (1983) *Data Structures and Algorithms*. Reading, MA/London: Addison-Wesley
- Alsuwaiyel MH (1999) *Algorithms: Design Techniques and Analysis*. Singapore/New Jersey: World Scientific
- Au LC, Yang FY, Yang WJ, Lo SH, Kao CF (1998) Gene synthesis by a LCR-based approach: high-level production of leptin-L54 using synthetic gene in *Escherichia coli*. *Biochem Biophys Res Commun* **248**: 200–203
- Bang D, Church GM (2008) Gene synthesis by circular assembly amplification. *Nat Methods* **5**: 37–39
- Cadwell RC, Joyce GF (1992) Randomization of genes by PCR mutagenesis. *PCR Methods Appl* **2**: 28–33
- Carr PA, Park JS, Lee YJ, Yu T, Zhang S, Jacobson JM (2004) Protein-mediated error correction for *de novo* DNA synthesis. *Nucleic Acids Res* **32**: e162
- Caruthers MH (1985) Gene synthesis machines: DNA chemistry and its uses. *Science* **230**: 281–285
- Cedrone F, Menez A, Quemeneur E (2000) Tailoring new enzyme functions by rational redesign. *Curr Opin Struct Biol* **10**: 405–410
- Chomsky N (1964) *Syntactic Structures*. The Hague: Mouton
- Coco WM, Levinson WE, Crist MJ, Hektor HJ, Darzins A, Pienkos PT, Squires CH, Monticello DJ (2001) DNA shuffling method for

- generating highly recombined genes and evolved enzymes. *Nat Biotechnol* **19**: 354–359
- Drexler KE (1992) *Nanosystems: Molecular Machinery, Manufacturing, and Computation*. New York: Wiley
- Forster AC, Church GM (2006) Towards synthesis of a minimal cell. *Mol Syst Biol* **2**: 45
- Gao X, Yo P, Keith A, Ragan TJ, Harris TK (2003) Thermodynamically balanced inside-out (TBIO) PCR-based gene synthesis: a novel method of primer design for high-fidelity assembly of longer gene sequences. *Nucleic Acids Res* **31**: e143
- Hecker KH, Rill RL (1998) Error analysis of chemically synthesized polynucleotides. *Biotechniques* **24**: 256–260
- Heinemann M, Panke S (2006) Synthetic biology—putting engineering into biology. *Bioinformatics* **22**: 2790–2799
- Hopcroft JE, Ullman JD (1979) *Introduction to Automata Theory, Languages, and Computation*. Reading, MA: Addison-Wesley
- Hutchison III CA, Phillips S, Edgell MH, Gillam S, Jahnke P, Smith M (1978) Mutagenesis at a specific position in a DNA sequence. *J Biol Chem* **253**: 6551–6560
- John Von Neumann RSP (1952) *Lectures on Probabilistic Logics and the Synthesis of Reliable Organisms from Unreliable Components*. Pasadena: California Institute of Technology
- Knight T (2003) *Idempotent Vector Design for Standard Assembly of Biobricks*. Boston: MIT Synthetic Biology Working Group
- Kodumal SJ, Patel KG, Reid R, Menzella HG, Welch M, Santi DV (2004) Total synthesis of long DNA sequences: synthesis of a contiguous 32-kb polyketide synthase gene cluster. *Proc Natl Acad Sci USA* **101**: 15573–15578
- Mandelbrot BB (1982) The fractals book. *Observatory* **102**: 151
- Matsuura T, Yomo T (2006) *In vitro* evolution of proteins. *J Biosci Bioeng* **101**: 449–456
- Merkle RC (1997) Convergent assembly. *Nanotechnology* **8**: 18–22
- Rogers H (1967) *Theory of Recursive Functions and Effective Computability*. New York: McGraw-Hill
- Ryu DD, Nam DH (2000) Recent progress in biomolecular engineering. *Biotechnol Prog* **16**: 2–16
- Sloning, BioTechnology and GmbH (2006) *De Novo Enzymatic Production of Nucleic Acid Molecules*. Munich: Sloning
- Smith HO, Hutchison III CA, Pfannkoch C, Venter JC (2003) Generating a synthetic genome by whole genome assembly: phiX174 bacteriophage from synthetic oligonucleotides. *Proc Natl Acad Sci USA* **100**: 15440–15445
- Stemmer WP, Cramer A, Ha KD, Brennan TM, Heyneker HL (1995) Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. *Gene* **164**: 49–53
- Tian J, Gong H, Sheng N, Zhou X, Gulari E, Gao X, Church G (2004) Accurate multiplex gene synthesis from programmable DNA microchips. *Nature* **432**: 1050–1054
- Whitesides GM (2006) The origins and the future of microfluidics. *Nature* **442**: 368–373
- Xiong A-S, Yao Q-H, Peng R-H, Duan H, Li X, Fan H-Q, Cheng Z-M, Li Y (2006) PCR-based accurate synthesis of long DNA sequences. *Nat Protoc* **1**: 791–797
- Xiong AS, Yao QH, Peng RH, Li X, Fan HQ, Cheng ZM, Li Y (2004) A simple, rapid, high-fidelity and cost-effective PCR-based two-step DNA synthesis method for long gene sequences. *Nucleic Acids Res* **32**: e98



*Molecular Systems Biology* is an open-access journal published by *European Molecular Biology Organization* and *Nature Publishing Group*.

This article is licensed under a Creative Commons Attribution-NonCommercial-No Derivative Works 3.0 Licence.