Rewriting DNA Synthesis

Emily Leproust
Twist Bioscience

Traditional methods of DNA synthesis are slow and costly, and hinder the design-build-test cycle for creating optimal gene sequences and protein variants. This article presents a novel approach that will allow researchers to explore synthetic biology's full potential.

Synthetic biology is one of the most important technological advances of the 21st century. Redesigning organisms offers the potential for new breakthroughs, such as novel medicines and therapies, sustainable energy from biofuels, plastics and fibers made from sugar rather than unsustainable oil, and food crops that can fertilize themselves. The possible benefits are almost endless.

DNA reading and writing: Two faces of the same coin

In the past decade, the development of next-generation sequencing (NGS) — a fast, reliable, and affordable method of reading a DNA sequence — revolutionized synthetic biology. Using this massively parallel high-throughput method, genomes of hundreds of organisms — from bacteria and viruses to plants and animals — were sequenced, providing an unprecedented amount of genetic information with single-base resolution.

Through large-scale data curation and sharing initiatives such as GenBank and ClinGen, end-users can access genetic data, which fuels research and enables a better understanding of sequence-to-function relationships. This facilitates hypothesis-driven redesign of specific genomic loci or even whole genomes, which scientists and engineers can then test to assess the impact on, and ultimately optimize, the processes that generate useful, high-value products.

The design-build-test cycle: Accelerating evolutionary processes

The cornerstone of synthetic biology is the design, build, and test cycle (Figure 1), an iterative process that requires DNA for rapid and affordable generation and optimization of custom pathways and organisms.

In the design phase, the adenine (A), cytosine (C), thymine (T), and guanine (G) nucleotides that constitute DNA are formulated into various gene sequences that comprise the locus or pathway of interest. Each hypothesis that will be tested requires a variant gene sequence. These variant gene sequences represent subsets of sequence space, a concept that originated in evolutionary biology and pertains to the totality of sequences that make up genes, genomes, transcriptome, and proteome.

Bioengineers compile many different variants for each design-build-test cycle to enable adequate sampling of sequence space and maximize the probability of finding an optimized design. These sequences, made accessible by DNA reading, are then manufactured through conventional synthesis methods. Once the various designs are assembled, they are tested for desired function within a chosen model system, which creates nanoscale biofactories out of cellular systems.

This cycle is repeated many times, testing permutations of the sequences and varying nucleotides at multiple

positions, until an optimal candidate is identified. Although straightforward in concept, process bottlenecks related to speed, throughput, and quality slow the pace, extending development time. Despite recent advances in technologies that support the overall design-build-test process, the ability to build each hypothesis for testing remains the rate-limiting step. Bioengineers are unable to sufficiently explore sequence space due to the high cost of highly accurate DNA and the limited throughput of current synthesis technologies.

Beginning with the build phase, two processes underlie success: oligo synthesis and gene synthesis. In the past, scientists synthesized different gene variants through molecular cloning. In this process, a gene or sequence of interest is extracted from a given organism. The gene is inserted into a vector system that replicates and produces large amounts of the gene for further study. While this approach is robust, it is not scalable.

A successful design-build-test process relies heavily on the ability to iterate on the design and its underlying sequences. Supporting this cyclical process requires large amounts of DNA that consist of hundreds or thousands of sequence variants. Simply put, cycle time depends on how quickly and how broadly this DNA can be accessed for the test phase.

In William Stemmer's classic 1995 paper dealing with the synthesis of long DNA sequences, it took approximately 50 to 60 40-mer oligonucleotides (oligos) to construct a 1-kilobase (kb) gene using assembly polymerase chain reaction (PCR) *(1)*. Synthesis of oligonucleotides was carried out using the classic phosphoramidite chemistry developed by Marvin Caruthers *(2, 3)*. This robust process has stood the test of time.

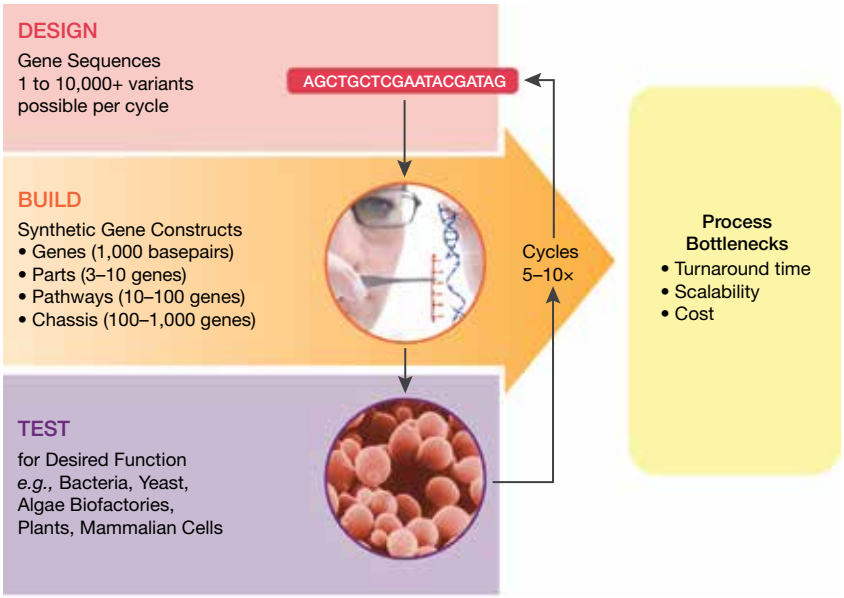
Most DNA synthesis platforms that exist today leverage phosphoramidite chemistry. Coupling phosphoramidite chemistry with gene assembly methods like PCR is what enables synthetic biology. New innovations centered on synthetic biology techniques improve quality, throughput, scalability, and turnaround time of bioprocesses. Technological advances in DNA writing, like those in DNA reading, drive the development of new applications in synthetic biology by increasing access to sequence space for further testing. However, unlike sequencing, DNA writing has not advanced enough to keep up with the rate at which DNA can be read, hindering the exploration of synthetic biology's full potential.

The evolution of gene synthesis

Early chemical gene synthesis efforts focused on producing a large number of oligos with overlapping sequence homology *(1)*. These were then pooled and subjected to multiple rounds of PCR, which concatenates the overlapping oligos into a full-length double-stranded gene. This method of gene synthesis requires equimolar amounts of each oligonucleotide, normalization of their concentrations, and pooling of the oligonucleotides prior to gene assembly. While effective, these requirements make this method both time- and labor-intensive, and limit its scalability.

This method requires high volumes of phosphoramidites, an expensive raw material, and other ancillary reagents. And, because it produces nanomolar amounts of the final product, significantly more than required for downstream steps, the process is unnecessarily expensive. In addition, the large number of separate oligos requires one 96-well plate for the synthesis of one gene. This is the same roadblock that hinders traditional cloning — the synthesis method does not scale to meet the synthetic biology market need for throughput and cost efficiency.

The use of microarrays significantly increases the throughput of gene synthesis *(4)*. A large number of oligos can be synthesized on the microarray surface, then cleaved off and pooled together. Each oligo — destined for a specific gene — contains a unique barcode sequence that enables that specific subpopulation of oligos to be depooled and assembled into the gene of interest. In this phase of the process, each subpool is transferred into one well of a 96-well plate, increasing throughput to 96 genes. Throughput is nearly two



▲ **Figure 1.** The design, build, and test cycle that underlies each DNA application development process requires DNA for rapid generation and optimization of custom pathways and organisms.

orders of magnitude higher, but it still does not adequately support the design-build-test cycles that require thousands of sequences at one time, due to a lack of cost efficiency and slow turnaround times.

A paradigm shift is needed

The field of synthetic biology impacts many different market segments and reaches into various biological disciplines. A recurring theme in research and development is the need for scalability. With highly resolved next-generation sequencing data underpinning the tested hypotheses, the ability to scale is critical, particularly when going from a small proof-of-concept with tens to hundreds of genes or sequences, to larger-scale efforts that reconstruct entire pathways. Furthermore, using a consistent synthesis platform minimizes variability and increases the ability to draw relevant and appropriate conclusions.

Today's DNA synthesis technologies are limited either at the low end, where a few sequences are produced at high-nanomolar amounts, or at the high end, where the synthesis platforms are only able to offer economy of scale if thousands of megabases worth of DNA are ordered at the same time. To support the application development process, it is apparent that a paradigm shift toward a rapid, scalable method of DNA synthesis is needed. A technology that scales will provide researchers access to DNA that will enable them to test all their ideas, simple or complex.

In the past, similar paradigm shifts have taken place in tech industries, spurred by the need to accelerate and increase access to information. For example, in the telecommunications industry, the mid-2000s brought about a shift from dial-up service to broadband internet. That paradigm shift enabled us to jump from a single-channel connection over a telephone line — transmitting 56 kb of data per second — to a technology able to transmit 25 to 100 Mb per second over multiple channels in multimedia data formats.

The field of DNA sequencing is another example, moving from the one-sample/one-lane, capillary-based sequencing platform used to sequence the human genome to NGS, a massively parallel platform enabled by miniaturized reactions on a flow-cell surface. Able to support both DNA and RNA, NGS brought about a multidimensional perspective on biological structure and function, which now drives discovery and development in synthetic biology.

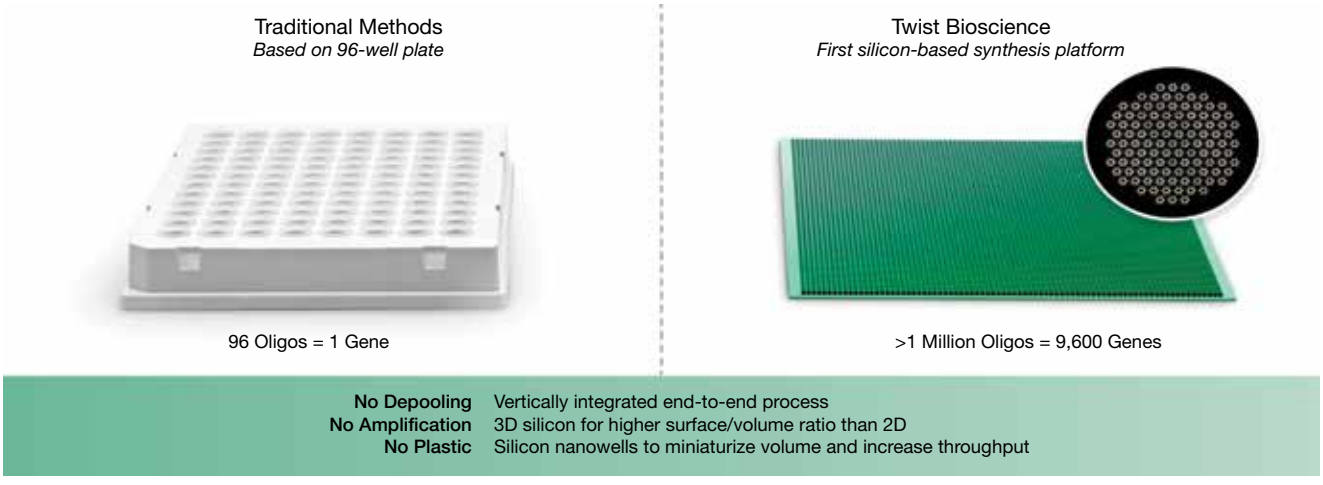
Revolutionizing a 50-year-old technology

Twist Bioscience addresses the challenges of throughput and speed in DNA synthesis with a revolutionary platform that combines miniaturization, parallelization, and vertical integration of the end-to-end process from oligo synthesis to gene assembly within nanowells on silicon.

Rather than the standard 96-well plate, Twist Bioscience's method of synthesis uses a silicon platform with 9,600 nanowells. This enables reaction volumes to be reduced by a factor of 1,000, while increasing reaction efficiency through the use of silicon, an excellent conductor of heat. Improved reaction efficiency at low reaction volumes in turn increases the accuracy of synthesis and significantly reduces the cost of production.

With the same footprint as a 96-well plate (Figure 2), the Twist Bioscience silicon synthesis platform has a throughput that is 100 to 1,000 times more than that of traditional synthesis methods. It can produce up to approximately 1,000,000 oligonucleotides, or about 10,000 genes, in a single highly parallelized run.

This scalable approach to gene synthesis enables unprecedented access to DNA for both low- and high-volume



▲ **Figure 2.** This silicon-based synthesis platform has the same footprint as a traditional 96-well plate. It increases DNA synthesis throughput through miniaturization, enabling scalability from one to 10,000 genes per run.

users. This facilitates the democratization of DNA, making it highly accessible to all types of users and applications, primed and ready to fuel advances in synthetic biology.

Twist Bioscience's silicon-based DNA synthesis platform gives researchers unprecedented access to high-quality DNA, enabling sufficient sampling of sequence space, faster screening times, lower costs, shorter cycles, and fewer iterations. For researchers in academia or industry, the impact of this advancement is tremendous — reducing the overall cost of experimentation while accelerating the time to publication of a novel technique, or minimizing time to market for synthetic-biology-derived products.

Powering gene editing and drug discovery through DNA synthesis

DNA is the raw material that fuels biological research and development. It is the biological scaffold that allows the redesign and testing of simple to highly complex pathways, even entire organisms. Like any building block, DNA can be used in many different ways and within many different workflows to create larger systems. Bioengineers can create optimized systems for production of high-value products.

With the advent of next-generation sequencing, high-resolution genomic data have become the lifeblood of studies that delve into the biological roles of various genes in both normal biology and disease pathogenesis. At the core of this research is the central dogma of molecular biology and the concept of "residue-by-residue transfer of sequential information" *(5)*: genomic information encoded in the DNA is transcribed into messenger RNA (mRNA) that is then translated into the protein that is the active product within a given biological pathway. Testing specific hypotheses through these studies requires access to the other dimensions of sequence space — RNA and protein.

Development of highly targeted therapeutics has been the holy grail for translational researchers in academia as well as for those in the pharmaceutical industry. To that end, highly specific techniques are extremely valuable to manipulate dysregulated pathways — biological pathways containing genetic perturbations that interfere with normal regulatory processes and result in the disease state. Specific therapeutic approaches involve identification of the underlying mechanism responsible for the disease, engineering a correction to the genetic perturbation, and therefore repairing the disease. Targeted therapeutics are possible through technologies such as specific editing of the gene coding regions to alter the mRNA sequence and enabling changes to the protein, or designing variants to the protein to optimize its binding properties.

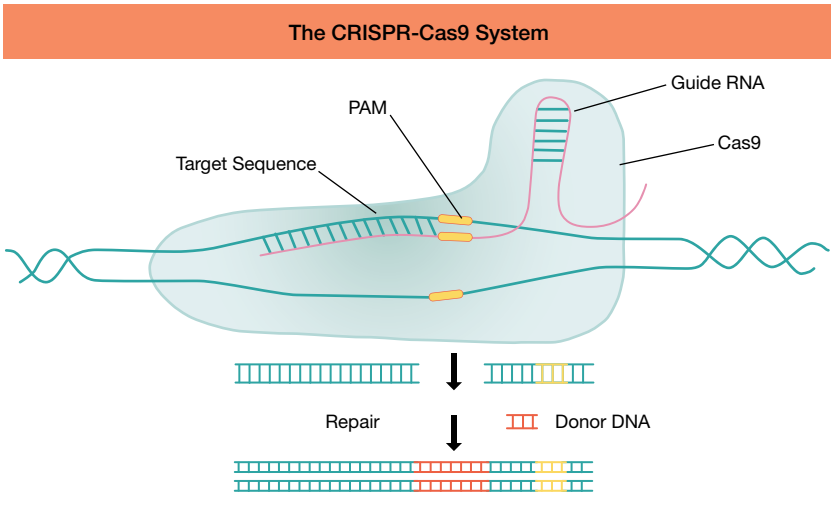
Genome editing for targeted gene therapies

The discovery of gene editing using the clustered regularly interspaced short palindromic repeats (CRISPR) system *(6)* has transformed molecular biology and empowered synthetic biology. CRISPR makes highly accurate editing of a precise location in any gene within any genome possible. CRISPR greatly accelerates production of gene variants that can then be used to understand gene function, produce modified organisms, further enhance beneficial properties, or correct the gene's function. Taking advantage of the power of CRISPR requires an affordable source of oligo pools with highly accurate, full-length sequences.

In the CRISPR-Cas9 system, an RNA sequence called a guide RNA, with a sequence complementary to its target, directs the Cas9 enzyme to the exact location in a gene where a change is to be made. Cas9's endonuclease activity then creates a double-stranded break at the specific site, which inactivates the gene. New DNA can be inserted at the location of the break to modify functionality (Figure 3).

Cellular assays can be used to further assess the impact of genome editing with the CRISPR-Cas9 system. More than one locus may be targeted simultaneously, and each locus may have multiple guides designed against it. For example, a genome-wide CRISPR library may contain 100,000 unique oligo sequences, and five guides that target each of the 20,000 genes in the human genome. Casting a wide net simplifies the process and increases the efficiency of screening, thereby reducing overall cost and time to optimization.

Article continues on next page



▲ **Figure 3.** CRISPR-Cas9 enables highly specific gene editing with a simple three-component system: guide RNA, Cas9 enzyme, and donor DNA. The protospacer adjacent motif (PAM) is a DNA sequence that immediately follows the sequence targeted by the Cas9 enzyme.

Critical to genome editing are precision and 100% guide representation. Precision (brought about by sequence accuracy of the guides) and effective guide representation are possible only through highly uniform synthesis.

In the context of gene editing, the ability to dial in from a genome-scale targeting approach down to single genes within a minimal number of design iterations is important. This may yield targeted gene therapies, which would be a major accomplishment.

Finding the needle in the haystack

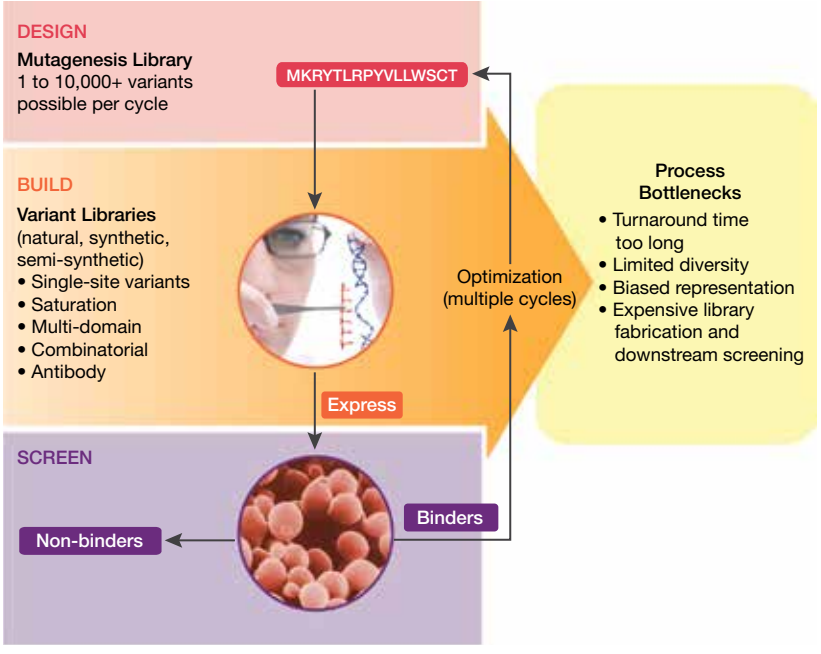
Another exciting area of study is the discovery, development, and manufacture of therapeutic molecules focused on a highly specific cellular target. Diverse mutagenesis libraries that consist of gene mutants are at the core of development pipelines for targeted therapeutics. Mutagenesis libraries are highly diverse collections of gene variants, which in turn can produce variant proteins. The design-build-screen protein engineering cycle ideally culminates in an optimized gene for increased expression of a protein with high affinity for its therapeutic target (Figure 4).

These libraries are often generated using slow, laborious, and inefficient methods, such as random mutagenesis, recombinant mutagenesis, site-directed mutagenesis, and combinatorial mutagenesis. The degenerate methods, which include random and recombinant mutagenesis, provide no precise control over amino acid composition and can generate stop codons that truncate proteins, thereby limiting the sampling of the total sequence space in the targeted gene. Although these current approaches may serve their purpose today, the need to provide a better technology that delivers on speed and cost without sacrificing quality and diversity remains unmet.

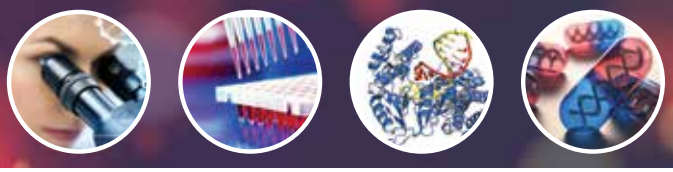
Drug development using the protein engineering cycle can take more than 10 years and may cost $1 billion *(7)*. As with other DNA applications, access to large amounts of DNA to build highly complex libraries remains a major bottleneck. As an example, consider the binding pocket of a receptor, which is typically the region within the protein that confers function or activity. These types of domains are of particular interest in drug development because this is the region that needs to be altered or modified to modulate function for a drug to have a particular treatment effect. Binding pockets can be folded into three-dimensional conformations, meaning that the sequences that form the binding pocket and those that come into contact with the ligand (target) are not necessarily adjacent to one another within the primary sequence, but are often from another part of the protein. If the crystal structure of a receptor-ligand complex has been determined, then all of the amino acid residues that come into contact with each other during binding can be identified. These residues can then be specifically and systematically altered to modify the binding properties. The ability to test all sequence permutations of all residues within the binding pocket simultaneously will allow for a thorough exploration of the best possible binding pocket structure with respect to the novel target, increasing chances of success.

A saturation mutagenesis library, which attempts to generate all possible mutations at a specific site within the receptor, is one approach to this development challenge. Although costly and time- and labor-intensive, it enables each variant to be introduced into each position. In contrast, combinatorial mutagenesis, in which a few selected positions or a short stretch of DNA may be modified extensively, generates an incomplete repertoire of variants with biased representation.

While a common misconception in drug development holds that the higher the diversity, the better the library, in actuality, balancing diversity with screening capabilities is more efficient. A precision library, made up of fewer variants but containing only variants that are directly relevant to the desired outcome, has lower costs and enables shorter screening times because unnecessary and unwanted variants do not need to be



▲ **Figure 4.** The design-build-screen cycle in protein engineering can be used to create targeted therapeutics starting from diverse mutagenesis libraries.



generated and screened. Within the drug development pipeline, though library production itself is very costly (typically hundreds of thousands of dollars per library), screening represents the bulk of this overall development cost.

As an example, a saturation library of about 250 positions would typically cost around $125,000 to $250,000. This type of library might be composed of about 5,000 variants that are readied for screening at a cost of $25 to $50 per variant. For each of these variants, functional assays are conducted to screen each variant, with each assay costing approximately $100 per variant screened. Therefore, the total cost of screening alone is about double the cost of the library. This highlights the need for an efficient library in order to reduce the number of variants that move downstream for screening and thus reduce associated costs.

With the capability for massively parallel DNA synthesis, Twist Bioscience's synthetic mutagenesis approach enables explicit and precise introduction of each intended variant at the desired frequency. To the end-user, this translates into the ability to not only thoroughly sample sequence space, but also to query these hypotheses in an efficient manner, reducing cost and screening time.

The challenge to reimagine

The advent of a technology that enables synthesis of large amounts of DNA at low cost opens up avenues to reimagine workflows. Workflows that today are encumbered by common bottlenecks associated with DNA — such as cost, throughput, and quality — can be optimized or redesigned to accommodate increased explorations of sequence space within the same budget.

In the future, engineers may be able to reconstruct entire pathways and genomes to re-engineer biological systems that can be accessed by end-users. With the rapid advances in DNA sequencing and synthesis, the challenges of synthetic biology are being resolved, accelerating development of applications that have potential to greatly benefit human existence and the environment that supports it.

Perhaps the most significant impact of DNA writing will be on discovery and development pipelines, as molecular cloning is replaced by inexpensive synthetic genes that can be obtained in large numbers in days, rather than weeks. The advancement of scientific knowledge around gene structure and function, disease progression, and the function of organisms will increase exponentially. Moving downstream, further testing and optimization for robust production and manufacturing processes can be accelerated, reducing time for development and ultimately lowering overall costs and reducing time to market for high-value products.

Recently, novel applications of DNA have come to fruition. DNA has been suggested as a means of stable long-term storage of data, with a gram of DNA storing as much as a zettabyte (10^{21} bytes) of data. This is an excellent example of an application of synthetic DNA with high growth potential that would not have been imagined only a few years ago. This is only the beginning. With open access to larger amounts of high-quality DNA at low cost, the challenge to reimagine biology can now be tackled. The next steps are yours. CEP

Literature Cited

1. **Stemmer, W. P. C.,** "Single-Step Assembly of a Gene and Entire Plasmid from Large Numbers of Oligodeoxyribonucleotides," *Gene,* **164** (1), pp. 49–53 (1995).
2. **Matteucci, M. D., and M. H. Caruthers,** "Synthesis of Deoxyoligonucleotides on a Polymer Support," *Journal of the American Chemical Society,* **103** (11), pp. 3185–3191 (1981).
3. **Beaucage, S. L., and M. H. Caruthers,** "Deoxynucleoside Phosphoramidites — A New Class of Key Intermediates for Deoxypolynucleotide Synthesis," *Tetrahedron Left*, **22** (20), pp. 1859–1862 (1981).
4. **Kosuri, S., *et al.***, "Scalable Gene Synthesis by Selective Amplification of DNA Pools from High-Fidelity Microchips," *Nature Biotechnology,* **28** (12), pp. 1295–1299 (2010).
5. **Crick, F. H. C.,** "On Protein Synthesis." *Symposia of the Society for Experimental Biology XII,* pp. 139–163 (1958).
6. **Doudna, J. A., and E. Charpentier,** "Genome Editing. The New Frontier of Genome Engineering with CRISPR-Cas9," *Science*, **346** (6213), pp. 1258096-1–1258096-9 (2014).
7. **Hughes, J. P., *et al.*,** "Principles of Early Drug Discovery," *British Journal of Pharmacology,* **162** (2), pp. 1239–1249 (2011).

EMILY LEPROUST, PhD, is the Chief Executive Officer of Twist Bioscience. As an early pioneer in the high-throughput synthesis and sequencing of DNA, she is disrupting the process of gene synthesis to enable the exponential growth of synthetic biology applications in multiple fields, including medicine, DNA data storage, agricultural biology, and industrial chemicals. In 2015, she was named one of *Foreign Policy's* 100 Leading Global Thinkers of 2015 for fast-tracking the building blocks of life, and *Fast Company* named her one of the most creative people in business for synthesizing DNA faster than ever. Prior to Twist Bioscience, she held positions at Agilent Technologies, where she architected the successful SureSelect product line that lowered the cost of sequencing and elucidated dozens of Mendelian diseases. She also developed the Oligo Library Synthesis technology, where she initiated and led product and business development activities for the team. Leproust designed and developed multiple commercial synthesis platforms to streamline microarray manufacturing and fabrication. Prior to Agilent, she worked with X. Gao at the Univ. of Houston developing DNA and RNA parallel synthesis processes on solid support, a project developed commercially by Xeotron Corp. She has published more than 30 peer-reviewed papers, many on applications of synthetic DNA, and is the author of numerous patents. She earned her PhD in organic chemistry from the Univ. of Houston and her M.Sc. in industrial chemistry from the Lyon School of Industrial Chemistry in France.