INEXTRICABLY BOUND: MEASUREMENT AND THE BIOECONOMY

Synthetic biology has been described as the design and construction of biological devices and systems for useful purposes [1]. The synthesis of DNA is a critical part of this construction. Advanced measurements have been both enabling and motivating for advances in DNA synthesis chemistry. Building on decades of development of chemical synthesis of DNA [2] and the development of DNA microarrays [3], additional careful attention to minimizing rare side reactions and very small non-idealities in reaction yields has enabled unprecedented levels of synthesis perfection and throughput [4]. The industrialization of this advanced chemistry has been shown to serve as a robust and economical basis for highly sensitive and specific hybridization assays [5]. It has also been shown to serve as a robust and economical source of user defined DNA oligonucleotides of sufficient quality to be used for synthetic biology [6]. The availability of high quality DNA oligonucleotides, coupled with analogously industrialized processes for combining them into larger constructs, opens up the possibility of widespread adoption of synthetic biology methods. New measurement modalities are being developed as a consequence. These examples, along with others elaborated elsewhere in this journal, illustrate the close and sometimes unpredictable interplay amongst measurement, science, and biotechnology, and the foundational role of measurement in advancing the bio-economy.

By **Emily M. LEPROUST*, Derek LINDSTROM*, Maurice SANCIAUME** and Stephen LADERMAN***

* Agilent Technologies, Inc., Santa Clara, California 95051, Etats-Unis.

** Agilent Technologies France, S.A.S., Diegem, BC B1813, Belgique.

MEASUREMENT MOTIVATES AND ENABLES HIGH QUALITY DNA OLIGONUCLEOTIDE SYNTHESIS

The function of synthetic DNA in the laboratory, like naturally occurring DNA, is determined by its sequence, or order in the polymer chain, of the bases (A, C, G or T). Over the years, a number of chemistries have been developed for the serial addition of bases to synthesize DNA, most notably the method developed by Marvin Caruthers [7]. In this approach, the synthesis of DNA is performed by the serial addition of bases (A, C, G or T) onto an attachment linker on a solid support. The incoming base is created to be a highly reactive moiety, which couples onto a particular chemical site, a hydroxyl group, on the growing DNA chain. The bond created between the two bases is an unnatural phosphite (P_{III} oxidation state), which is subsequently oxidized into a natural phosphate (P_{IV} oxidation state) bond. Following oxidation, the terminal end of the newly added base, previously protected by a Di O-Methyl Trytil (DMT) group, is de-protected to create a hydroxyl end. This step, called de-block, creates the attachment point for the next base addition. The combination of the coupling, oxidation, de-block steps is called DNA synthesis cycle, and as many cycles as necessary are performed to synthesize a DNA strand of a given length. This chemistry was advanced to an unprecedented level in the course of pursuing high performance DNA microarray assays [4].

By enabling complete flexibility of the synthesized sequences, chemical synthesis of DNA is also essential for synthetic biology applications. At the same time, many synthetic biology applications require longer lengths of DNA than can yet be achieved with chemical synthesis. These demands require a balance between chemical synthesis and subsequently combining the chemically synthesized oligonucleotides into longer constructs with enzymes. The longer and more perfect the initially chemically synthesized materials, the more universal and practical is the entire process.

The overall length of the synthetic DNA strand is limited by the chemical efficiency of the DNA synthesis cycle, as any error in any step will result in a failed DNA synthesis. For each cycle, there is, even in the best chemistries, a small probability of failure. The overall yield of the synthesis is determined by compounding the errors in each cycle. The overall yield, then, is the yield of each cycle raised to the power of the length. Small effects become compounded very quickly. As an example, a cycle yield of 99% results in an overall yield of only 36% after 100 cycles, or 13% after 200 cycles. Typically, cycle yields of 99.8% and better are necessary to obtain synthetic DNA suitable for widespread application. To achieve high yields after two-hundred or three-hundred cycles required the development of creative chemistry and imaginative engineering translatable to high precision, robust manufacturing processes, quality control, and quality assurance.

As with the manufacturing of DNA microarrays, the DNA synthesis cycle may be controlled spatially to enable the simultaneous synthesis of tens of thousands of different DNA sequences on a flat solid support, typically glass or silicon, enabling high throughput downstream applications [3]. One approach for such spatial control is the use of inkjet printing technology to print fluids containing coupling phosphoramidite derivatives of the four DNA bases: A, C, G and T; analogous to the four standard color inks: black, magenta, cyan, and yellow. Since only one step of the synthesis cycle needs to be spatially controlled, the other steps, oxidation and de-block, may be performed in a non-spatially controlled manner in a flow cell. In addition to a printing module and flow cell module, a DNA synthesis instrument would also contain an automation module to transfer the solid support to and from the printing and flow cell modules, as well as an inspection module to image the printed drops and detect any misprinting.

Engineering controls during DNA synthesis on an automated instrument are essential to achieve the high quality and robustness required for routine commercial operation. For example, during printing, the solid support is maintained at sub-millimeter distances from the print heads and displaced at speeds of about one meter per second (m/s) over more than a meter during printing of drops at millions of discreet locations and densities up to 1,000 features per mm^2. Since the droplets of all four bases must always be printed at the same location, relative alignment of the print heads is essential. This is achieved by three-dimensional, piezo-assisted precision movements of each print head, along with precision timing of the signals initiating delivery of ink from the print head nozzles. Similarly, in the flow cell, the thickness must be tightly controlled over the dimensions of the solid support to ensure uniform flow of the reagents.

The fluid or “ink” formulation is a key aspect of this technology. The ink composition must be optimized to ensure efficient 1. jetting, i.e. ejection of droplets from the print heads of uniform volume, speed and trajectory, 2. printing, i.e. landing of the droplets on the solid support without splashing or spreading, and 3. coupling, i.e. the chemical attachment of the incoming base onto the growing DNA chain. The surface tension and viscosity of the ink are the two examples of parameters affecting the droplet formation at the print head and the absorption of the droplet kinetic energy at impact, and are among those that must be tightly controlled. The chemical composition of the ink must also be controlled to ensure high reaction rates, low side reactions and the physical integrity of the reaction area despite the large hydrodynamic forces accompanying the process.

The solid support used during the synthesis also has key characteristics that must be tightly controlled, such as its dimensions, to ensure reproducible and effective movement by transfer robots and to avoid even slight bowing in flow cells. The active surface must have an optimum and reproducible loading of DNA attachment points where DNA synthesis is initiated. To that end, glass, or a silicon oxide layer on silicon, is treated to achieve a durable attachment of a monolayer of a material with optimized physical and chemical properties. Its quality must be carefully measured to insure utility.

High quality DNA synthesis writers must be fully automated. An efficient and automated machine that can process twelve wafers simultaneously has a printing capacity of 1Giga bases per day. To balance printing and flow chemistry, such a machine is composed of one printing module linked to twelve flow cell modules so that there is always one wafer ready to print as soon as printing of a previous wafer is completed. The instrument can be run asynchronously where a wafer can be added or removed from the synthesis instrument, and scheduler software can be used to optimize the transfer of wafers from module to module in order to minimize waiting times, similarly to logistics software optimizing the inventory of goods between a warehouse and stores. Finally, each module can be driven by an independent logic center, which interfaces with the scheduler to exchange status, and with a central data storage to capture log information for future traceability. After synthesis, the wafers are de-protected in basic solution, to finalize the synthetic DNA into its natural structure, and washed. Each wafer can then be diced, for example, into approximately 25 mm by 75 mm slides using non-contact laser dicing, after each slide has been individually labeled.

Multiple layers of quality control of the synthetic DNA are critical to ensure accurate and reproducible results for downstream applications. First, during incoming inspection, the DNA synthesis reagents are tested analytically and functionally prior to use in manufacturing to ensure that they contain the intended active ingredients at the appropriate concentration, and that they do not contain any known detrimental contaminants. Second, during in-process quality control of the synthesis on the instrument, the instrument checks the correct execution of the process by measuring reaction times, reagent pressures and temperatures, as well as by imaging every printed droplet to ensure its correct location and volume. If any deviation is observed, the instrument is stopped and corrected immediately, ensuring that a minimal, if any, quantity of product has to be discarded. Third, during destructive testing, a portion of each wafer is analyzed in a way similar to downstream use to ensure that the synthesized DNA meets specification. This test usually involves the hybridization of the synthetic DNA with a mixture of simple and complex labeled targets to measure background signal, linearity of binding, lower level of detection and the extent of any side reaction such as de-purination. Fourth and finally, all the manufacturing and QC parameters are studied using statistical process control methodologies to detect any drift from standard behaviors. This ensures ongoing and continuously improving control of the processes and provides early warning of any quality control issues. All data acquired for each reagent lot, wafer and slide can be stored to provide a device history file that can be queried later, if needed, in the course of routine or special troubleshooting.

SYNTHETIC BIOLOGY MOTIVATES AND ENABLES NEW MEASUREMENTS

Single celled organisms such as bacteria have long been the focus for synthetic biology because of their simplicity: They are easy to manipulate genetically and environmentally. By growing large populations of identical cells under defined conditions, a wealth of RNA, protein, and metabolite data has been collected and analyzed using *in vitro* measurements. Interestingly, measuring the number of RNA or protein molecules in a single cell often shows a population average which is an incomplete description of an individual cell. Part of the discrepancy is due to how RNA is produced from DNA: RNA messages are not always churning out constantly from a DNA template like a smooth-running printing press. Instead, many are produced in discreet bursts of activity [8]. Thus, the concentration of a given message will spike and decline over time due to bursts of synthesis followed by degradation and dilution. This dissonant production schedule is further confounded by the fact that many RNAs exist at very low copy number. A burst may produce 10 copies of RNA, which can decline to zero before the next burst of synthesis. Low copy numbers lead to high stochastic variation, especially during cell division. If a cell contains only a handful of molecules of a particular RNA, when it divides it can often give birth to one cell with the entire handful and a second cell with none. These variations in RNA levels can also be compounded during their translation into protein, which can exhibit the same burst-like mode of synthesis. The end result is wide cell-to-cell variation of these molecules within a population. These variations in molecular components translate into differences in phenotype. The fact that genetically identical cells can behave so differently can profoundly affect our ability to predict the outcomes of modifications such as introducing new metabolic systems or genetic circuits. Measuring the phenotypic probability distributions of individual RNAs, proteins and metabolites in single cells has become an important source of data needed to accurately model gene expression, systems-level interactions, and even whole cell behavior [9]. Indeed, the understanding of basic biology has been

changed by the results of single-cell measurements on populations of genetically identical cells grown in the laboratory. Dramatic differences are seen between the DNA instructions (the genotype, identical in each cell) and the products of those instructions (the phenotype, or resulting behaviors) [10]. These differences have not only changed the implications of genetic inheritance, they have also contributed significantly to the refinement and success of synthetic biology. The realization that phenotype is not a simple value derived from the product of genetics and environment, but is better described as a probability function derived from these inputs, has led to more robust predictive modeling of systems and even whole cells [11].

If a given phenotype is derived from a probability distribution of certain RNAs and proteins, what set of measurements should be made to determine phenotype at the single-cell level? Predictive approaches first identify all of the RNAs and proteins that contribute to a particular phenotype and then measure them all to predict the phenotypic outcome. While becoming more tractable over time as measurement and computation advance, both making and fully interpreting more than a few measurements on a single cell remains slow and expensive.

At the other extreme, if the phenotype can be defined by a single molecule, a single measurement would suffice. If that single molecule is an RNA or protein, it is already known how to genetically encode fluorescent reporters for most such molecules. But if the molecular phenotype of interest is best represented by a metabolite, there is not yet a systematic method for generating reporters for these chemically diverse molecules. Here, using synthetic biology as a means for generating new tools shows great promise for developing precise measurement devices, or biosensors, to measure metabolite levels in single cells. Biosensors are genetically encoded molecules that provide two functions: An input function accomplished by binding to the molecule of interest and an output function comprising the generation of a measureable signal. Many modular outputs have been developed for biosensors, but a predictable way to generate a desired input for any molecule of interest is still lacking. Inputs have been appropriated on an *ad hoc* basis from natural sources, for example a protein known to bind to a particular metabolite. With the advent of high throughput sequencing, many DNA sequences encoding potential binding proteins have never actually been physically isolated. Synthetic biology now enables the testing of such a large number of candidate inputs from natural sources by enabling their construction and testing. Moreover, synthetic biology approaches can be used to generate large libraries of variations on a theme: to create new candidates based on variations of a single natural source that can be screened for new functions. Ultimately, libraries of completely novel protein sequences can be screened for new functions [12]. In this way, advanced synthesis provides an opportunity to overcome a lack of prior knowledge, accelerating the development of compelling new measurements.

BIBLIOGRAPHY

[1] ENDY (D.), Foundations for engineering biology, *Nature*, 438, 449-453, 2005.

[2] CARUTHERS (M.H.), Chemical Synthesis of DNA and DNA Analogues, *Acc. Chem. Res.*, 24, 278-284, 1991.

[3] FRIEND (S.H.) and STOUGHTON (R.B.), The Magic of Microarrays, *Scientific American*, 286, 44-53, 2002.

[4] LEPROUST (E.M.), PECK (B.J.), SPIRIN (K.), MCCUEN (H.B.), MOORE (B.), NAMSARAEV (E.) and CARUTHERS (M.H.), Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process, *Nucleic Acids Research,* 38, 2522-2540, 2010.

[5] BARRETT (M.T.), SCHEFFER (A.), BENDOR (A.), SAMPAS (N.), LIPSON (D.), KINCAID (R.), TSANG (P.), CURRY (B.), K BAIRD (K.), MELTZER (P.S), YAKHINI (Z.), BRUHN (L.) and LADERMAN (S.): Comparative genomic hybridization using oligonucleotide microarrays and total genomic DNA, *PNAS*, 101, 17765-17770, 2004.

[6] KOSURI (S.), EROSHENKO (N.), LEPROUST (E.M), SUPER (M.), WAY (J.), LI (J.B) and CHURCH (G.M.), Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips, *Nat Biotechnol*, 28, 1295-1299, 2010.

[7] CARUTHERS (M.), Gene Synthesis Machines: DNA Chemistry and Its Uses, *Science*, 230, 281-285, 1985.

[8] SEGAL (E.) and WIDOM (J.), From DNA sequence to transcriptional behavior: a quantitative approach, *Nat Rev Genet*, 10, 443-456, 2009.

[9] KARR (J.R), SANGHVI (J.C.), MACKLIN (D.N.), GUTSCHOW (M.V.), JACOBS (J.M.), JR BOLIVAL (B.), ASSAD-GARCIA (N.), GLASS (J.I.), and COVERT (M.W.), A whole-cell computational model predicts phenotype from genotype, *Cell*, 150, 389-401, 2012.

[10] NANDAGOPAL (N.) and ELOWITZ (M.B.), Synthetic biology: integrated gene circuits, *Science*, 333, 1244-1248, 2011.

[11] CARDINALE (S.) and ARKIN (A.P.), Contextualizing context for synthetic biology - identifying causes of failure of synthetic biological systems, *Biotechnol J*, 7, 856-866, 2012.

[12] CHERNY (I.), KOROLEV (M.), KOEHLER (A.N.) and HECHT (M.H.), Proteins from an Unevolved Library of *de novo* Designed Sequences Bind a Range of Small Molecules, *ACS Synth Biol*, 1, 130-138, 2012.