Chemistry & Biology
Review

Putting Synthesis into Biology: A Viral View of Genetic Engineering through De Novo Gene and Genome Synthesis

Steffen Mueller,[1,*] J. Robert Coleman,[1] and Eckard Wimmer[1]
[1]Department of Molecular Genetics and Microbiology, Stony Brook University, Stony Brook, NY 11794-5222, USA
*Correspondence: smueller@ms.cc.sunysb.edu
DOI 10.1016/j.chembiol.2009.03.002

The rapid improvements in DNA synthesis technology hold the potential to revolutionize biosciences in the near future. Traditional genetic engineering methods are template dependent and make extensive but laborious use of site-directed mutagenesis to explore the impact of small variations on an existing sequence "theme." De novo gene and genome synthesis frees the investigator from the restrictions of the pre-existing template and allows for the rational design of any conceivable new sequence theme.

Viruses, being among the simplest replicating entities, have been at the forefront of the advancing biosciences since the dawn of molecular biology. Viral genomes, especially those of RNA viruses, are relatively short, often less than 10,000 bases long, making them amenable to whole genome synthesis with the currently available technology. For this reason viruses are once again poised to lead the way in the budding field of synthetic biology—for better or worse.

A Brief History of DNA Synthesis

The chemical synthesis of nucleotide chains took its first infant steps soon after the discovery of the DNA double helix. The race to elucidate the genetic code was driven by the use of triplet sequences of ribonucleotides synthesized by liquid-phase chemistry. Depending on their sequence these triplets selectively interacted with amino-acylated tRNA (the codon:anticodon recognition) (Nirenberg and Leder, 1964; Soll et al., 1965), which led to the assignment of codons to their respective amino acids, and to a much-deserved Nobel Prize for these heroic efforts in these earliest days of synthetic biology. Khorana's group "raced" to synthesize the first DNA copy of the 75 base pair (bp) tRNA^{Ala} in 1970 (Agarwal et al., 1970), a monumental task requiring 20 man-years of labor, only to be outclassed by himself in 1979 by a 207 bp DNA cassette containing the tyrosine suppressor tRNA gene (Khorana, 1979).

The innovations of synthesizing DNA oligonucleotides ("oligos") on solid supports (Letsinger and Mahadevan, 1965) combined with new activated phosphoramidite nucleosides (Caruthers et al., 1987) led to steady improvements in the availability of quality oligos up to 100 bases long. This resulted in a boost in gene synthesis activity throughout the 1990s that continues unabatedly today. Some of the most notable synthesis achievements are summarized in Figure 1 (Agarwal et al., 1970; Becker et al., 2008; Blight et al., 2000; Cello et al., 2002; Chan et al., 2005; Edge et al., 1981; Ferretti et al., 1986; Gibson et al., 2008; Gupta et al., 1968; Kalman et al., 1990; Khorana, 1979; Kodumal et al., 2004; Nirenberg and Leder, 1964; Pan et al., 1999; Soll et al., 1965; Stemmer et al., 1995; Tian et al., 2004). Significant landmarks include the synthesis of an entire 2.7 kb plasmid sequence by Stemmer et al. (1995), the 4.9 kb MSP-1 gene of *Plasmodium* (Pan et al., 1999), the 7.5 kb of the poliovirus genome as the first synthetic self replicating organism (Cello et al., 2002), and the 32 kb polyketide synthase gene cluster (Kodumal et al., 2004). The trend has culminated in the recent synthesis of 582,970 bp corresponding to the first artificial bacterial genome by the group of Craig Venter (Gibson et al., 2008). Starting with 101 prefabricated segments of 5–7 kb in length (purchased from commercial vendors), Gibson et al. used state-of-the-art methods and brute force to assemble larger and larger DNA pieces, at first by recombination in bacteria, and finally in yeast (Gibson et al., 2008). Alas, the synthetic genome was not, or could not, be "booted" to life, by transplanting the genome into an "empty" chassis as the group has shown previously with a natural genome (Lartigue et al., 2007). Therefore, the first synthetic autonomous life form is still just below the horizon.

Methods for the Assembly of Long Synthetic DNA

It is not yet possible to synthesize entire genes as long continuous strands of DNA from scratch. Rather, all synthetic genes are assembled from short custom-made single-stranded DNA oligonucleotides or "oligos," which are literally strings of a few nucleotides. Oligos are by-and-large still synthesized the same way as they were 15 or 20 years ago. Through incremental improvements in instrumentation and higher throughput, oligos have become a cheap commodity for use in standard recombinant DNA technologies. But, more than anything else, great demand and even greater competition by manufacturers have driven the oligo prices down by about 10-fold over the past 15 years (Figure 2). In comparison, the prices of finished, sequence-confirmed gene synthesis by commercial gene foundries have plummeted 50-fold in only 10 years (Figure 2). As a reference point, at the outset of the poliovirus synthesis project (Cello et al., 2002) in 1999, commercial gene synthesis was simply unheard of. As recently as 2000, after much searching, we found a vendor who agreed to synthesize parts of the genome by special arrangement at a price of $12/bp (Cello et al., 2002).

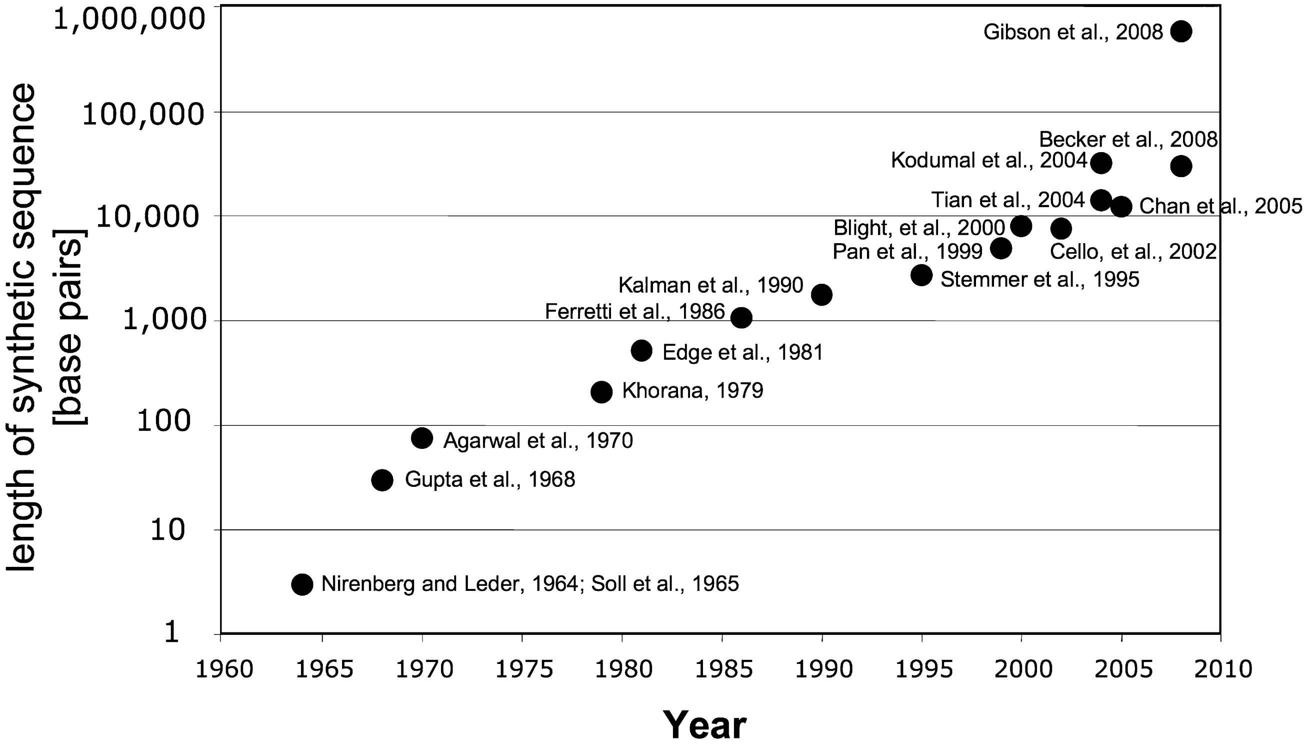

Figure 1. Pushing the Limits: a Historical Progression of Notable Achievements in Gene Synthesis with References
Each point represents a report of an individual gene synthesis accomplishment with respect to the length of the synthetic sequence and the year it was first reported.

In the ideal world, an efficient and economical de novo gene synthesis platform would combine cheap error-free oligo synthesis with accurate assembly methods. Neither one is currently available. There are two dramatically different methods of synthesizing oligos. In the traditional, time-proven method of solid-phase oligo synthesis, each oligo is synthesized individually, on a separate small column or a well on a multiwell plate. The method is high yielding but costly ($0.10–0.20 per nucleotide synthesis cost), which is a critical aspect if the oligos are needed for the assembly of long DNA sequences. The price given above translates into an oligonucleotide cost of approximately $200–400 for a 1 kb DNA sequence—and that's for the raw material only.

The development of optical deprotection chemistries heralded a new era of parallel synthesis methods on micro biochips (Fodor

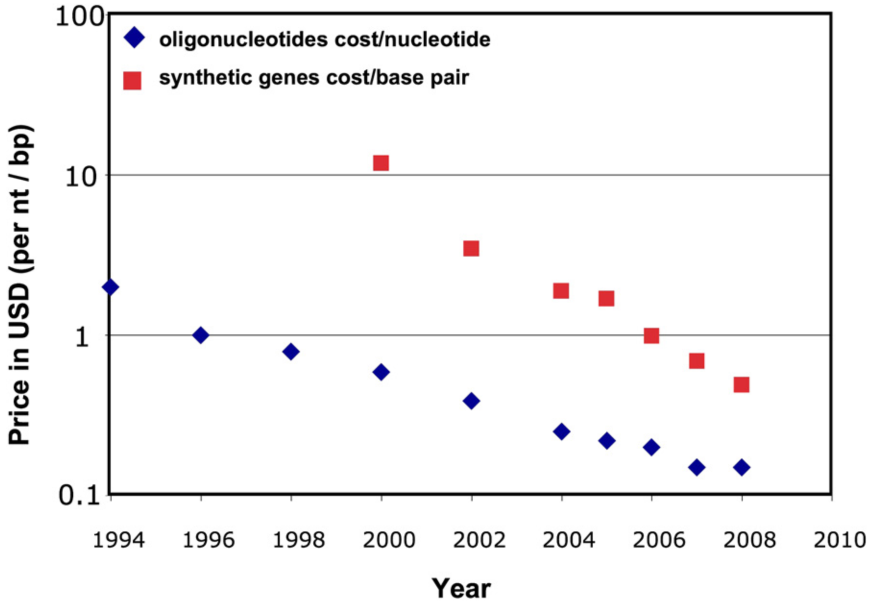

Figure 2. Price Development of Oligonucleotide Synthesis and De Novo Gene Synthesis
Shown are the approximate end user prices per base for oligonucleotides (desalted, nonpurified) or per base pair for synthetic genes (below 3 kb, sequence guaranteed). The data were compiled from a "look back" of vendor invoices, and a survey among colleagues. Although by no means comprehensive, the prices shown here are representative of what the typical research laboratory paid for these services at the time.

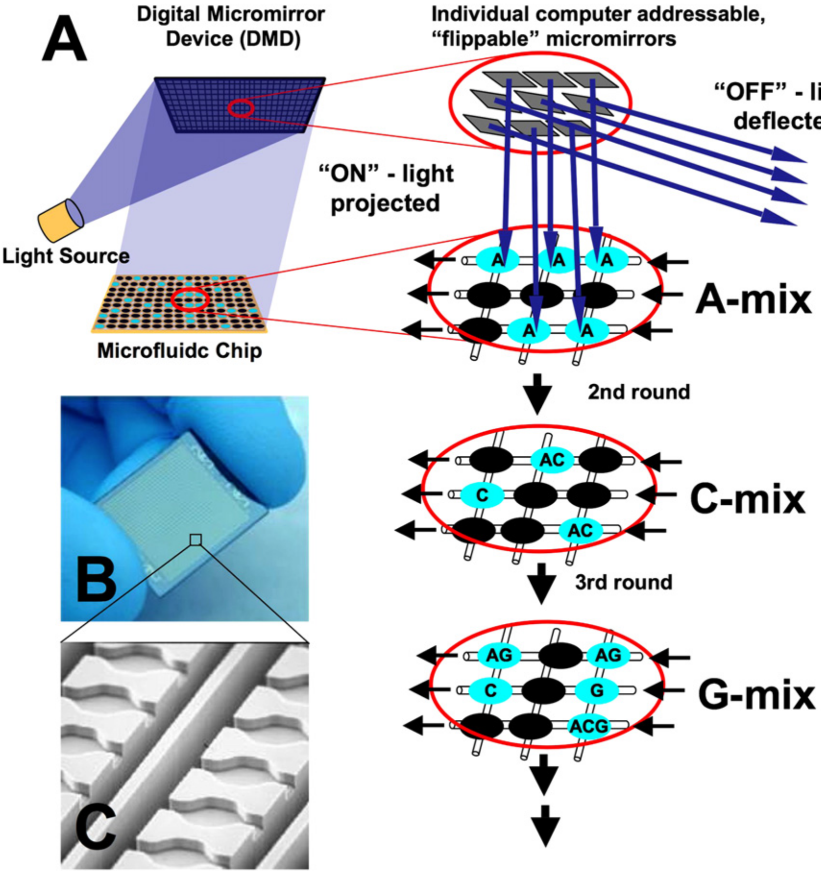

Figure 3. Microfluidic Chip Technology Coupled with Light-Activated Chemistries Hold Great Promise for the Massive Parallel Synthesis of Oligonucleotides

(A, B) On an array of tiny flippable mirrors, each mirror can be separately computer controlled (flipped to an "ON" or "OFF" position). Mirrors in the ON position reflect light onto their corresponding reaction chamber on a microfluidic chip (bright blue spots), leading to the incorporation of the nucleotide currently loaded on the chip (here, A-mix). Although all chambers receive the same nucleotide mix at any one time, no reaction occurs in the dark chambers (black spots). The process is repeated with the next nucleotide mix and a new light pattern, which specifies the chambers to incorporate the new nucleotide. After the last nucleotides are incorporated, the finished oligos are released from the chip and collected as a pool (B) actual size of a microfluidic chip holding 4000 sequence features. Reproduced with permission by LC Sciences, LLC, Houston, Texas. (C) A magnified view of the interconnected microscopic reaction chambers on an Atactic microfluidic chip. Reproduced with permission by LC Sciences, LLC, Houston, Texas.

et al., 1991) that can be used for both oligo or peptide synthesis. Depending on the chip platform being used, several thousands to hundreds of thousands of distinct oligonucleotides can theoretically be synthesized on a single chip.

In an ingenious extension Tian and colleagues (Tian et al., 2004) mated the light-induced deprotection chemistry with microfluidic technology that allows the programmable synthesis of thousands individual oligonucleotides on a tiny chip (Figure 3A). At the heart of this method is the digital light processing technology that was developed for digital projectors and high-definition projection television sets. On a microfluidic chip containing a labyrinth of thousands of connected tiny reaction chambers (Figure 3C), each chamber is computer-addressable by a light beam generated on a digital micromirror device (Singh-Gasson et al., 1999) (akin to the individual color light spots making up the projection-television picture). A DNA synthesis mixture containing the first nucleotide (A, for instance) is pumped through the system. Here, A only "sticks" to the chambers that call for an A at the specific position in their sequence, which are the ones that are being illuminated at that time (Figure 3A). Although all chambers receive the same synthesis mixture at any given time, no reaction occurs in the chambers that are "left in the dark" (in the example above, the ones that need a C, G, or T at their corresponding position). After the first reaction, the A-mix is washed out and the next reaction mix containing the next nucleotide is pumped in and the process is repeated, four times in total. After all four nucleotide reaction mixes have gone through the chip, in each chamber the oligonucleotide chain has now grown by at least one nucleotide of the desired sequence.

At the end of the reaction, the oligonucleotides are eluted from the chambers as a single pool. Each of the oligo sequences is only present in minute quantities. This might present a challenge in further increasing the throughput by increasing the number of reaction chambers per chip, while decreasing their size. Tian et al. (2004) demonstrated the potential power of this technology for the synthesis of large numbers of oligonucleotides to be used in synthetic gene assembly.

Companies already offer parallel on-chip-synthesized custom oligo mixtures that are amenable for gene synthesis (LC Sciences, Houston, TX). Currently the price of a pool of 3912 90-mers is approximately $1000. This technology is still very much in the exploratory stage. One inherent difficulty of the method is that all oligos are released from the chip as a mixture. The low yields of oligos that come off the chip (10^7–10^8 molecules per sequence) are insufficient to drive a gene assembly reaction, which mandates a postsynthesis polymerase chain reaction (PCR) amplification step before oligos can be used. For this purpose each oligo is synthesized with two flanking generic adaptor sequences, which allows amplification of all oligos in parallel in a single PCR reaction using the corresponding adaptor primer pair (Figure 4) (Tian et al., 2004). Using distinct sets of adaptors on distinct subsets of oligos in the same chip-synthesis reaction allows the subsequent selective amplification of a desired subset of oligos, for instance a set necessary for the assembly of one particular gene. Therefore, it is possible that in a separate reaction a different set of oligos can be amplified from the same chip-eluted oligo mix. Thus, fractioning the entire oligo pool into gene-specific subsets will reduce complexity of the mixture, increase concentration of each specific oligo, and reduce potential interference or cross-hybridization from other

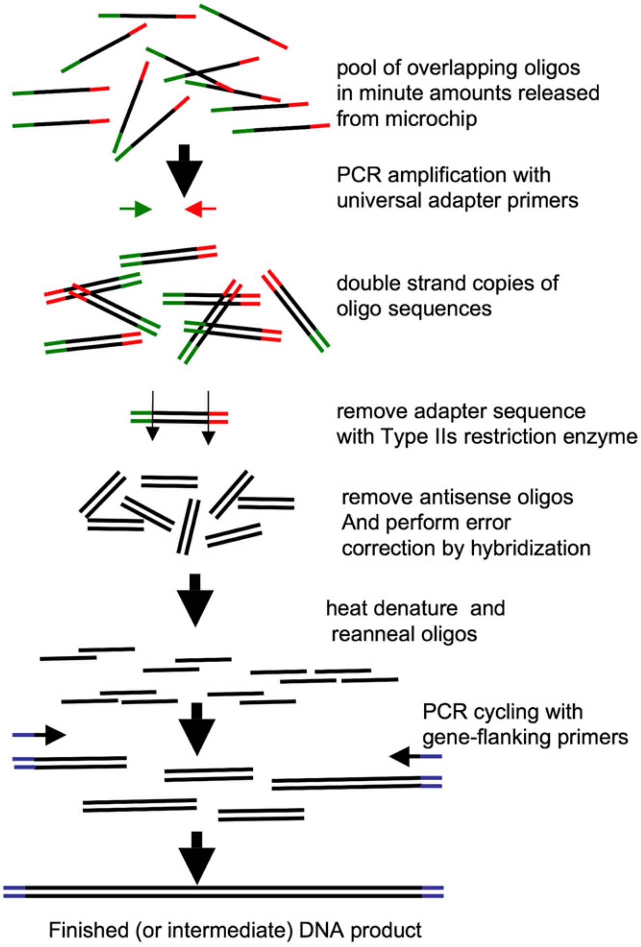

Figure 4. Assembly of Gene Sequences from Chip-Synthesized Oligonucleotides

The pool of overlapping oligos in minute amounts is released from the microchip, followed by PCR amplification with universal adaptor primers. Double-strand copies produced in this way are subjected to type II restriction enzymes to remove the adaptor sequence. Construction oligos are purified by stringent hybridization to immobilized selection oligos. This leads to the elimination of the unwanted antisense oligos and reduces the error frequency in the construction oligos. Next, the eluted construction oligos are heat denatured and reannealed, and subjected to PCR cycling to produce intermediate or final DNA products. The reaction is driven by excess concentration of a gene-flanking primer pair.

oligos in the pool. This will be especially useful as the number of individual sequences synthesized on the chip increases. The higher the number of discrete oligo sequences synthesized per chip, the lower the absolute yield per oligonucleotide (subfemtomolar range) because the total yield of DNA is a direct function of the total reaction surface on the chip. With more distinct oligos the potential for unwanted cross-hybridizations during the gene assembly step also increases.

The second drawback of the chip-based oligo synthesis is that the PCR amplified oligos are now in a double stranded form. The presence of a perfectly matched antisense strand might reduce the efficiency in the subsequent assembly of these oligos into larger genes. The assembly reaction depends on the complementarity of the overlapping "construction" oligos, those designed to build the gene, and the antisense oligos are likely to compete more effectively for the same hybridization partner. To overcome this problem the desired single stranded construction oligos can be selectively enriched by specific hybridization to antisense selection-oligos affixed to a column and subsequent elution (Tian et al., 2004). When done under stringent enough conditions, this procedure also contributes to a significant elimination of error-containing oligos, because they produce mismatches with the selection oligo and consequently elute from the column at a lower temperature. On the downside, this method requires twice the amount of selection oligos than there are construction oligos. In other words, to produce one chip's worth of oligos, one needs two additional chips worth of selection oligos, tripling the cost of synthesis (Tian et al., 2004). This brings the current "rock-bottom" cost of the final construction oligos before the gene assembly to about $0.03/bp.

Although these new multiplex synthesis systems are technically feasible, it is our understanding that the major suppliers of large synthetic DNA for now continue to assemble genes from individually synthesized overlapping oligonucleotides by traditional methods.

The sheer number of different oligonucleotides synthesized on a chip mandates the use of new software programs to handle the complexity of possible interactions of the various oligo sequences in the mix (Czar et al., 2009). Several software programs are freely available to design optimal sets of assembly oligonucleotides. The basic tasks that successful software needs to perform are:

1. Breaking down the target sequences to be synthesized into suitable overlapping oligos.
2. Designing hybridization units, the overlapping portion between two oligos, with the same melting temperature.
3. Ensuring hybridization specificity of each oligo pair to eliminate potential cross-hybridization by choosing the best possible breaking points between oligos for a particular gene, and by altering synonymous codons.

Assembly of Synthetic Genes and Genomes

There are two basic methods available for assembling long DNA sequences, such as virus genomes, from short overlapping synthetic oligonucleotides: direct assembly PCR, and ligase chain reaction (LCR) followed by fusion PCR with flanking primers.

Assembly PCR

Assembly PCR is based on the principle of generating stepwise elongation of the amplicon, a piece of DNA formed in an amplification event, by one oligonucleotide at each end of the growing amplicon with each PCR cycle (Stemmer et al., 1995), and on the possibility of intermediate products to act as overlapping megaprimers to assemble even larger amplicons (Figure 4). Theoretically, the reaction continues until the two outermost oligos are incorporated to give the full-length product. The full-length product is subsequently amplified with an excess of the two flanking PCR primers. Practically, obtaining large DNA fragments in a single assembly reaction is exceedingly difficult. For this reason, and for error-management purposes, it is generally necessary to first synthesize, clone, and verify the sequence of several intermediate-size subfragments (500–1000 bp). These can then be linked by fusion PCR to form larger genes or by standard cloning methods.

LCR followed by Fusion PCR with Flanking Primers

The LCR method is similar in that it uses overlapping oligos. But unlike with PCR assembly, oligos for LCR have to be designed to anneal without gaps between them, head to toe, forming annealed stretches of DNA that are then ligated using a thermostable DNA ligase (Barany, 1991). In contrast to PCR assembly, where a single oligo is added at each end of a synthon in each cycle, during LCR several overlapping oligos can be ligated to one another. Owing to the thermostability of the ligase, LCR can be cycled similar to a PCR reaction, leading to assembly of longer and longer chains, but no net amplification. The desired product is finally amplified by PCR using gene-flanking primers.

Limitations of Current Oligo-Based DNA Synthesis Methods

Regardless of the many variations on the theme of how to assemble a large synthetic DNA, at the core of all current methods are chemically synthesized oligonucleotides. The downward price trend for oligos has slowed significantly over the past 5 years and appears to be bottoming out (currently in the $0.10–0.20/base range). Because the price gap, and therefore the profit margin, between finished synthetic genes and their oligo building blocks is narrowing, it can be expected that oligo-based gene synthesis prices will soon follow. For long DNA synthesis to become economical, radically new technologies need to be developed that either reduce the errors in run-of-the-mill oligos by orders of magnitude, or allow de novo gene synthesis independent of the error-prone oligonucleotide chemistry, perhaps by developing enzyme-based synthesis of long accurate polynucleotides. Barring such breakthrough, the routine synthesis of bacterial or larger genomes will likely remain prohibitively expensive for some time to come. As a case in point, the recent synthesis of the *Mycoplasma* genome (Gibson et al., 2008) cost an estimated $10 million (Herper, 2007). At the research level, however, once gene synthesis hits the $0.10–0.20/bp price range, synthesis will very likely replace the traditional recombinant DNA methods for many smaller scale cloning projects within the next few years.

A major problem with genes assembled from overlapping oligos is the inherent error rate of about 1% during the chemical synthesis of the oligos themselves. The most frequent error is the failure to incorporate bases due to less than perfect deprotection of the reactive groups or incorporation of the incoming nucleotide. It appears that there is a rather hard limit for improving the oligo accuracy during the synthesis step much beyond the 1/100. Therefore, several techniques are being employed, often in combination, to improve the accuracy of oligos and the assembled DNA intermediates.

1. Keeping the oligos and the overlapping regions between them short (40–50 bases) not only reduces the relative error rate per nucleotide in the oligo, but also increases the disruptive effect of mismatches between annealed oligos. Using stringent hybridization conditions thus reduces the chance of incorrect oligos to partake in the assembly reaction (Young and Dong, 2004).
2. A common approach is to gel-purify oligos before the assembly reaction, which helps eliminate many of the shorter aberrant oligo species. This reduces the error rate to about 1 in 500. At this error rate, short (several hundred base pairs long), intermediate assembly products are cloned by traditional recombinant DNA methods and sequence verified. The vetted sequence segments are then either combined by further rounds of cloning, or by assembly PCR. The need for gel purification is another reason to keep oligo length limited, because oligos that are too long can no longer be effectively separated from the most troublesome offender, the (N-1)-mer. If all construction oligos for one specific synthesis project are kept the same length, the gel purification can be done by combining all oligos in one sample, much reducing time and cost (Smith et al., 2003).
3. Another approach relies on the selective hybridization of the construction oligos to a column of immobilized selection oligos (Tian et al., 2004), as noted above.
4. Finally, a second tier of error correction can be implemented after the LCR or PCR assembly of gene fragments. It is based on the enzymatic activity of T7 endonuclease, which recognizes and specifically cleaves dsDNA at mismatched nucleotide pairs (Picksley et al., 1990; Young and Dong, 2004). Following the final PCR amplification, the DNA amplicon is heat denatured and reannealed. Because mutations in the original construction oligo sequences are distributed randomly, the probability of two hybridizing strands carrying a mutation on one and the corresponding compensatory mutation on the other oligo is miniscule. It can therefore be expected that virtually every mutation in every oligo that participates in the assembly reaction will create a mismatch. Similarly, error correction by mismatch binding proteins, such as MutS of *Thermus aquaticus*, can be employed, facilitating the separation of the MutS-bound mismatched DNA from the correct DNA by gel electrophoresis (Carr et al., 2004).

The quality of the oligos critically determines the practical size of the synthesis intermediates that need to be cloned and sequence verified (Carr et al., 2004). If sequence errors follow a normal Gaussian distribution along the length of the DNA, an error rate of 1 in 600 would make it impractical to assemble a DNA longer than 1–2 kb in a single reaction without intermediate sequence verification (Figure 5).

Applications of De Novo Gene and Genome Synthesis

Codon Optimization

In many cases it is desirable to express a gene of interest (often a human gene) in a heterologous, more economical expression system, such as bacteria or yeast. All too often, however, the codon usage within the gene is at odds with the codon usage of the new host species. As a result the gene expresses poorly. Thus, the need for ‘‘codon optimization’’ was born (Itakura et al., 1977). During codon optimization, the codon usage of the gene is altered to reflect that of the host species by replacing suboptimal codons with preferred synonymous codons. Because this often involves many simultaneous sequence changes, it is best done by de novo gene synthesis. Probably the best known example

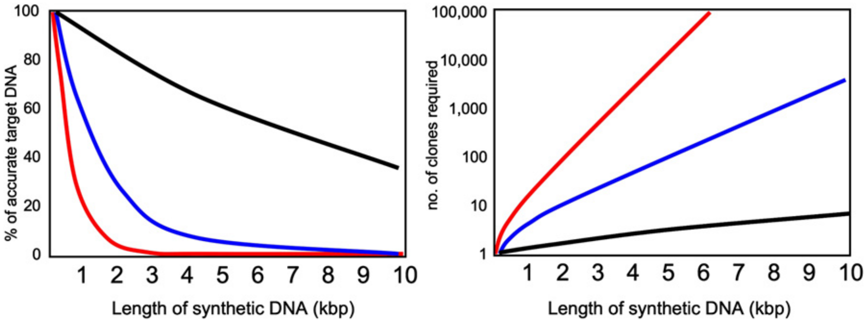

Figure 5. The Impact of Oligonucleotide Error Rate on the Accuracy of Assembled Synthetic Genes

The various curves assume error rates in the construction oligonucleotides typically achieved after different error-correction methods used to assemble a target sequence are 1/600 (red; using gel-purified oligos), 1/1,400 (blue; using hybridization-selected oligos), and 1/10,000 (black; using mismatch-specific endonucleases). Adapted from Carr et al., 2004.

of codon optimization is the ''humanization'' of the green fluorescent protein of the jellyfish *A. victoria* (Zolotukhin et al., 1996). Codon optimization is currently still the most prevalent reason for de novo gene synthesis (Gustafsson et al., 2004).

In some instances gene synthesis has been used to recreate a DNA sequence from a publicly available sequence database in an effort to sidestep licensing, patenting, or material transfer issues.

Creating New Chassis for Protein Engineering

It is theoretically possible to synthesize a bacterial genome in which the redundancy of the genetic code is eliminated, such that each amino acid in every bacterial protein is represented by exactly one codon only. Thus, only 20 codons plus 1 stop codon would be needed to synthesize all the bacteria's own genes. At the same time, the remaining 43 ''orphaned'' codons could be freed up to specify non-natural amino acids. Bacteria with such an expanded genetic code could one day become a powerful chassis for the production of artificial proteins (Carr and Isaacs, 2006; The Economist, 2006).

Viral Gene and Genome Synthesis

Viruses are among the simplest replicating genetic systems. For this reason they have been at the forefront of the advancing biosciences since the dawn of molecular biology. Their small genome sizes (most RNA virus genomes are 10 ± 5 kb) makes them amenable to whole genome synthesis with the currently available technology. For this reason viruses are poised to lead the way in the budding field of synthetic biology.

A significant use for genome synthesis consists in the recreation of viruses or perhaps other organisms in the future, for which no intact natural template is available. The synthesis of the 1918 flu virus was accomplished by piecing together sequence fragments recovered from victims buried in the Alaskan permafrost and archived tissue samples (Tumpey et al., 2005). The creation of bat SARS coronavirus (Becker et al., 2008) and HIV from Chimpanzee feces (Takehisa et al., 2007) also falls into this category. A clever extension of this idea has been the resurrection of live infectious retroviruses assembled from a consensus of ancient remnants that are endogenous to the human genome, and which have perhaps been inactive for millions of years (Dewannieux et al., 2006; Lee and Bieniasz, 2007). Once the stuff of science fiction movies, these ''Jurassic Park*esque*'' projects are likely to be just the teaser trailers of the coming attractions in the budding synthetic technology.

Through the process of natural selection, evolution favors systems that work, especially those that work better than their direct predecessors and competitors. This selection process however does not follow what humans would consider a logical design process. Evolutionary changes are small and incremental following a one-directional ratchet that does not move backward. There is no ''reset'' button that allows evolution to jump back to an earlier version and try again. De novo gene and genome synthesis provides this virtual reset button by allowing the creation of any conceivable genome at will and at once, no matter how different from its predecessor.

One recurring theme in viral genomes is the evolution of overlapping reading frames. This space-saving measure allows a virus to encode portions of two proteins on the same stretch of genome sequence, but in two different reading frames. Studying individual genes and proteins of such a virus genetically and biochemically poses a problem for the experimenter, because manipulating one protein inadvertently changes the other. To simplify these interdependencies in the genome, Chan and colleagues redesigned and synthesized parts of the bacteriophage T7 genome, eliminating the overlapping reading frames (Chan et al., 2005). In the resulting virus, the individual genes could be then manipulated and studied independently, a process they called ''refactoring'' in analogy to the process of redesigning and improving computer code, while retaining its basic function.

Exploiting the Intrinsic Sequence Biases of the Human Genome for the Generation of Synthetic Virus Vaccines

The basic mechanism of mRNA translation is preserved from the simplest virus to the most complex organism. Viruses, just like human cells, need to produce mRNA molecules, which are used to convert their genetic information into proteins. Different viruses have devised different strategies to accomplish this, and have different ways to store this genetic information in their genome. Invariably, however, viruses need to divert the host's cellular machinery for the translation of their proteins, because they themselves cannot execute this function. The degeneracy in the genetic code (several synonymous codons specify the same amino acid) gives an organism the flexibility to encode a given protein sequence in its genome in an unimaginably large number of ways. The poliovirus polyprotein, for instance, could be encoded by a staggering 10^{1100} different mRNA sequences, all of them specifying the same protein sequence (for comparison, the number of atoms in the observable universe is estimated to be on the order of 10^{80}). This raises the question: To what extent is the natural encoding of a gene optimal or

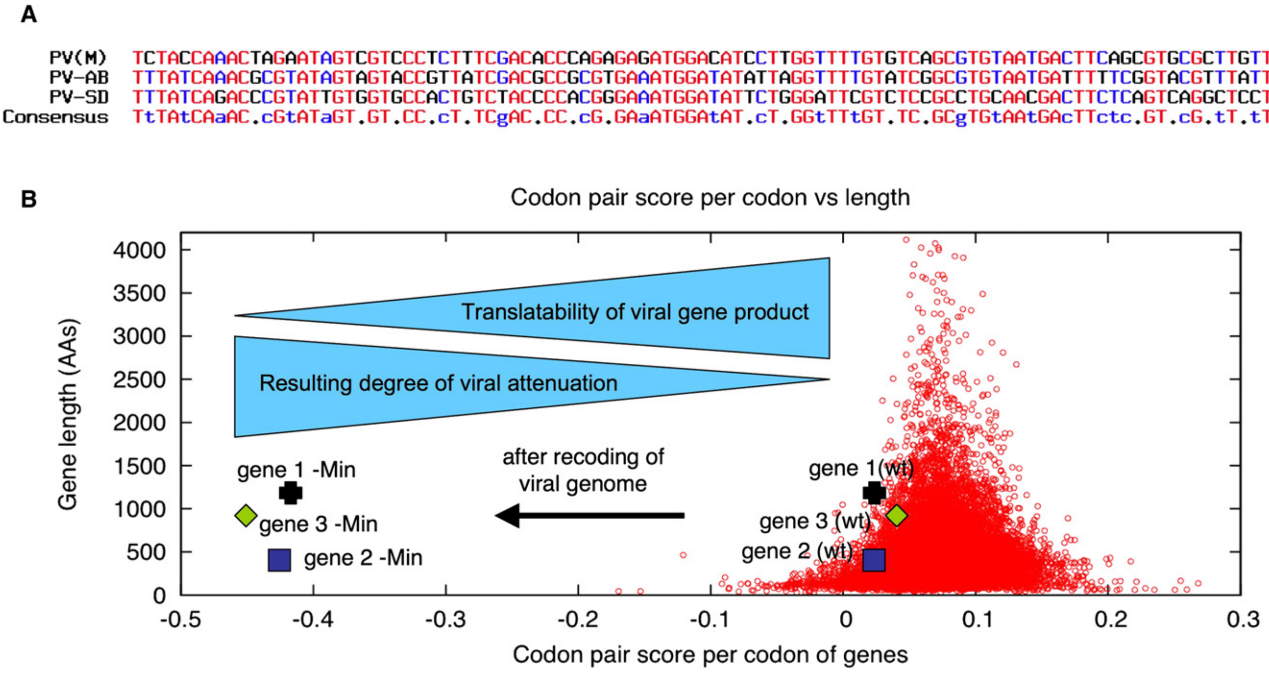

Figure 6. Recoding of Viral Genomes According to the SAVE Method

(A) Example of the level of sequence alteration after codon reassignment of the poliovirus capsid gene (Mueller et al., 2006). PV(M), part of the wild-type capsid coding sequence; PV-AB, the same amino acid sequence encoded by rare codons; PV-SD, the same amino acid sequence encoded by random shuffling of synonymous codons present in the wild-type sequence. Note that the amino acid sequence encoded by all three sequences remains the same.

(B) Codon pair bias after SAVE-mediated codon reassignment of viral genes. The codon pair bias (CPB) score for each of 14,795 confirmed annotated human genes was calculated. Each red dot represents the calculated CPB score of one human gene plotted against its amino acid length. Predominant use of underrepresented codon pairs yields negative CPB scores. The codon-pair scores of three wild-type viral genes fall within the bulk of the human genes. After computer-aided recoding and de novo synthesis of the viral genome according to the *SAVE* algorithm the new genes ("Min" for minimized CPB) have extremely unfavorable CPB, unlike any gene the cellular translation machinery has ever encountered. Note that the amino acid sequence of all proteins remains unchanged during this process. By analogy to other virus systems a decreasing CPB leads to reduced translatability of the mRNA and increased attenuation of the virus. Adapted from Coleman et al., 2008.

special? The cell's preference of one synonymous codon over another to specify the same amino acid is termed "codon bias." It is thought that codon bias is correlated with the abundance of the corresponding cognate tRNAs in the cell. Consequently, rare codons are associated with a suboptimal translation of an mRNA. In addition, the frequencies of which two codons occur next to one another in the genome are not what is statistically expected from the frequencies of the two codons that make up the pair—a phenomenon called the "codon-pair bias." There are codon-pair combinations that are statistically greatly underrepresented whereas others are greatly overrepresented. The significance of codon pair bias has been largely unknown and underappreciated. We have recently shown that it is possible to exploit the codon-pair bias phenomenon for the synthesis of novel live attenuated forms of viruses with incredible properties (Coleman et al., 2008). By using large-scale computer-aided redesign of the viral genome, we engineered hundreds of silent mutations into poliovirus. These mutations were targeted to introduce a maximum number of unfavorable synonymous codon-pairs, without changing codon bias or protein sequence. By forcing a virus to "make do" with this heavily biased synthetic genome, we showed that viral protein translation is greatly reduced. Thus, codon-pair deoptimized viruses cannot reproduce their genetic information as quickly as their wild-type cousins, which puts them at a decisive disadvantage against the host's innate and immune defenses. One of the major benefits of the whole-genome deoptimization strategy is that the resulting attenuated viruses are phenotypically and genotypically extremely stable. The attenuation (*att*) phenotype is dependent on many hundreds, even thousands, of silent mutations, each by themselves virtually inconsequential, or "death by a thousand cuts." Therefore, the fitness gain from reverting individual mutations appears to be too small to drive genetic selection, and thus, reversion apparently does not occur (Coleman et al., 2008). We termed this process of perturbing intrinsic viral genome biases by synthetic genome redesign SAVE (synthetic attenuated virus engineering) (Figure 6).

SAVE attacks a virus at one of the most fundamental processes common to all living systems, the translation of protein, for which viruses depend on the host cell's machinery. Thus, it should be predicted that SAVE will work on most (if not any) viruses.

The rational genetic changes imposed on SAVE-designed viral genomes are completely independent of protein sequence. The viral protein sequences, and therefore their function, remain 100% preserved in the recoding process. Therefore, an understanding of the proteins function is not necessary, sidestepping the need of most of classic virology in order to produce an attenuated vaccine candidate in a very short time with a predictable degree of attenuation in virtually any virus system. Viruses live lives of genetic austerity, and therefore don't usually carry unnecessary genes around. By that rationale, most viral genes product can be considered essential. Depending upon the virus system, interfering just a little bit with the synthesis of several of those genes turns out to pack a great punch against the overall fitness of the virus (Coleman et al., 2008; Mueller et al., 2006).

Using the SAVE method we can profit from these genomic biases that have arisen over evolutionary time-scales and turn them upside down and inside out, undoing eons of viral evolution. If we think of evolution as ''walking'' along a dirt path, SAVE allows us to ''leap'' across the evolutionary universe at warp speed. Because it is evident that many viruses have actively selected against the occurrence of certain sequence features, such as unfavorable codons, codon-pairs, and other sequences motifs, the whole-genome recoding approach by de novo synthesis will very likely have a profound effect on any virus.

General Requirements for the Application of SAVE to a Virus System

Because SAVE targets a virus at the level of protein translation, a function elementary to all viruses, we believe this approach is applicable to many virus systems for which the following basic requirements are met:

1. A target virus has a known genome sequence, preferably available online.
2. The desired deoptimized genome sequence is prepared by computer-aided redesign using the SAVE algorithm.
3. De novo synthesis of the artificial viral genome is performed according to the design specifications, usually outsourced to a commercial vendor.
4. A reverse genetics system is employed to boot the artificial genome to life and make a virus. This is decidedly simple for many human viruses. Often a genome-length copy of the DNA itself or an RNA transcript of that DNA is infectious upon transfection into susceptible cells.
5. A method to screen for viruses of desired phenotype has to be available. An initial screen in susceptible cell culture will yield valuable information as to the viability of various deoptimized virus designs. Clearly the virus still must be able to replicate at least at a low level in order to be useful as a live vaccine.
6. A suitable animal model to test attenuation and immune response is required.

If the above requirements are met, the SAVE strategy can successfully be employed for redesign and synthesis of viruses.

Synthetic virology, i.e., the redesign and synthesis of custom-tailored whole virus genomes, has become economically feasible with recent rapid improvements in DNA synthesis technology. This holds the potential to revolutionize the way virology and vaccinology is done. Viral genomes, especially of RNA viruses and retroviruses, are short enough to make them amenable to whole-genome synthesis with currently available technology. Such freedom of design could provide tremendous power to perform large-scale redesign of DNA/RNA coding sequences, to study the impact of large-scale changes in codon bias, codon-pair bias, dinucleotide biases, GC content, RNA secondary structures, and other sequence signatures, on viral fitness, with the aim to develop a new platform for vaccine design and genetic engineering.

Societal Implications of Synthetic Biology

What is synthetic biology? It is neither a field in its own right, nor a separate science. It is perhaps best described as an improvement of existing enabling technologies that are beginning to penetrate mainstream sciences, as they become more and more economical. This has led to an ''organized'' crossover of different scientific fields (e.g., biology, chemistry, mathematics, engineering) that promises to yield organisms with useful biochemical pathways never seen before.

The new reality of synthetic genes and genomes calls for a fundamental revision of the ways biology is taught to students. The Johns Hopkins University has already embraced these cutting-edge developments, and is now offering an undergraduate course in which students collaboratively work toward synthesizing the yeast genome. Impressively, within only 1 year this unified effort resulted in the synthesis of hundreds of 750 bp cassettes amounting to the 280 kb of the yeast chromosome III (Dymond et al., 2009). An equally imaginative and playful introduction to engineering of biological systems is fostered by the International Genetically Engineered Machine Competition (iGEM; http://www.igem.org) organized by synthetic biologists at MIT. Here undergraduate teams compete in designing and building genetic circuits and systems from an ever expanding toolkit of standard genetic parts, or ''BioBricks™'' (Goodman, 2008).

However, although the excitement about synthetic biology is substantial enough, it faces equally big skepticism and ''fear of the new'' in our society. A disservice to their own science is perhaps the tendency of some researchers in the ''synthetic biology field'' to overvalue its novelty and uniqueness. The most commonly cited public concerns with regard to synthetic biology are probably the ethical implications connected with the creation of ''new life forms'' and the fear of synthetic ''killer viruses.'' These sentiments are often picked up and fuelled by the media, potentiating the perceived fear of the uncertain.

Virtually every organism ever modified in molecular or genetic research is by definition a new life form. This definition could be expanded to all naturally occurring organisms that genetically differ from their parent—in other words: all the living creatures. Why would an organism created by synthetic methods be qualitatively different? The question presents itself: Why do we, as a society, worry more about the possibility of a synthetic designer pathogen, when some of the worst pathogens known to mankind are still raging? Measles virus, as a case in point, is one of the most contagious viruses to humans. As recently as in 2000, approximately 777,000 people died per year from measles, and in third-world countries with poor health care systems the fatality rate can be as high as 28% (Perry and Halsey, 2004). Annually, 250,000–500,000 people die from complications of the flu (WHO, 2003). Additionally, only a few critical mutations in the H5N1 bird flu virus separate us from a virus that can easily spread among humans and lead to an influenza pandemic. The AIDS pandemic, caused by primate viruses that jumped the species barrier to humans, claims approximately 2 million lives annually (http://www.avert.org/worldstats.htm). In 2003, the world barely escaped a pandemic by a SARS-coronavirus now thought to have jumped from bats to humans (Becker et al., 2008 and references therein).

Although in theory at least, we have the capacity to generate any genetic sequence that we can conceive, what we can do with this capacity is in fact quite limited. Although it's easy to think up fantastic and scary scenarios of synthetic killer viruses wiping out humankind, bioterrorists and the brightest scientific thinkers alike would be hard pressed to say what such a designer superpathogen would look like. In reality, all that can be accomplished via synthesis, now and for some time to come, is emulating, copying, and re-creating what mother nature has brought forth and thrown at us incessantly throughout our history on this planet. It is possible to produce variations on an existing theme. It is not possible, as yet, to design from scratch a qualitatively new pathogen that is completely different from any organism that exists now or has existed in the past. The level of abstraction required to ''piece together'' qualitatively new life forms from defined off-the-shelf parts (genes) is far from being realized (Goler et al., 2008). It is probably this misconception, trumpeted by the media, which strikes a cord of fear in the general population. Cases in point:

1. The 2002 poliovirus synthesis (Cello et al., 2002), the first synthesis of a pathogen, caught the world off guard and ignited a heated debate in its aftermath. All we had done was to re-create an exact synthetic copy of the poliovirus genome, except for some genetic ''watermarks'' to prove the authenticity of the synthetic genome. The resulting virus was, at the protein level, 100% identical to the wild-type virus used in countless laboratories around the world, a virus that even now naturally circulates in several countries and that is available for purchase at repositories such as the American Type Culture Collection. Being an exact antigenic match to the currently available poliovirus vaccine, an overwhelming proportion of the world population is immune against this virus. Worldwide vaccine coverage against poliovirus is arguably the greatest of any vaccine-preventable disease. This is hardly a blueprint for an imminent bioterrorist attack. But it was suddenly becoming clear that viruses can never be regarded as extinct, as long as their genome sequence information is preserved, be it on a government-sponsored online database, a 29-year-old *Nature* journal (Kitamura et al., 1980) gathering dust in libraries across the world, or just written down on a smudgy piece of paper forgotten in a desk drawer... It is sufficient to re-create a virus at any point, even long after any traces of its natural presence have vanished. It is this uncomfortable realization that brought about the level of public discussion that the original poliovirus synthesis had. The publication was intended not only to herald a new era in the study of organisms, but also to serve as a ''wake-up call'' for dual use technology.
2. The re-creation of the highly pathogenic 1918 flu virus (Tumpey et al., 2005) out of sequences extracted from influenza victims preserved in the northern permafrost also met with criticism, although no one had maligned the publication of the genome sequence as much as 8 years earlier (Taubenberger et al., 1997). In fact, the synthesis the 1918 virus brought critical new insight into the pathogenesis of the influenza and it is a prerequisite for the production of an adequate vaccine should such a need ever arise. Isn't society in the long run much better off with this knowledge than without it, understanding 1918 flu virus in detail rather than hoping that something like the 1918 flu will never happen again? This sentiment is even more inappropriate with the looming threat of the H5N1 bird flu pandemic.
3. Over 30 years of random, ''unenlightened'' genetic manipulation of viral genomes through recombinant DNA technology by countless laboratories around the world has not shown any evidence that researchers would accidentally and unbeknownst to them create a human supervirus. Whole-genome synthesis will be no different.
4. The adaptation of a human pathogen to an experimental animal species by repeated passaging through that species (a decidedly ''pre-synthetic era'' method) has been employed ever since viruses were discovered. It leads to the increased pathogenicity in the new species compared with the wild-type virus. These host-adapted models have greatly facilitated the study of viruses and the diseases they cause. Equally important is that these experiments resulted in the development of some of the most successful vaccines ever produced (polio, measles, mumps, rubella, and smallpox). As it turned out, passaging these viruses through diverse animal species lead to the mitigation of their disease-causing potential for humans—a process termed ''attenuation.''

All the above considerations notwithstanding, de novo genome synthesis, like many technologies in the past, does hold a potential for dual use. And unlike many technologies before it that require immense resources that cannot escape detection (nuclear proliferation, for instance), the intentional misuse of genome synthesis technologies will become increasingly undetectable. It seems next to impossible that genome synthesis can ever be government-regulated effectively. The technology and its components are too ubiquitous already, and too easy to jury-rig from off-the-shelf parts. The nature of genome synthesis is such that in the very near future pathogens can, and perhaps will, be synthesized in the proverbial hobbyist's basement, high school science lab, or by a bioterrorist organization. These possibilities are not an academic's hyperbole either. In fact, the grassroots ''biohacker'' culture is already flourishing outside the realm of academia, industry, and government oversight (Cowell and Bobe, 2009). When considering these issues, our society would be prudent to shift focus from prevention of such dual-use proliferation to preparing for it. The latter might include the development of new vaccines and/or the stockpiling of available vaccines against the most likely bioterrorist agents.

ACKNOWLEDGMENTS

The work was supported by NIH grant AI075219 (to E.W.).

REFERENCES

Agarwal, K.L., Buchi, H., Caruthers, M.H., Gupta, N., Khorana, H.G., Kleppe, K., Kumar, A., Ohtsuka, E., Rajbhandary, U.L., Van de Sande, J.H., et al. (1970).

Total synthesis of the gene for an alanine transfer ribonucleic acid from yeast. Nature *227*, 27–34.

Barany, F. (1991). Genetic disease detection and DNA amplification using cloned thermostable ligase. Proc. Natl. Acad. Sci. USA *88*, 189–193.

Becker, M.M., Graham, R.L., Donaldson, E.F., Rockx, B., Sims, A.C., Sheahan, T., Pickles, R.J., Corti, D., Johnston, R.E., Baric, R.S., and Denison, M.R. (2008). Synthetic recombinant bat SARS-like coronavirus is infectious in cultured cells and in mice. Proc. Natl. Acad. Sci. USA *105*, 19944–19949.

Blight, K.J., Kolykhalov, A.A., and Rice, C.M. (2000). Efficient initiation of HCV RNA replication in cell culture. Science *290*, 1972–1974.

Carr, P.A., and Isaacs, F. (2006). rE.coli: whole genome engineering. Presented at the Second International Conference on Synthetic Biology (Synthetic Biology 2.0), May 20–22, Berkeley, CA.

Carr, P.A., Park, J.S., Lee, Y.J., Yu, T., Zhang, S., and Jacobson, J.M. (2004). Protein-mediated error correction for de novo DNA synthesis. Nucleic Acids Res. *32*, e162.

Caruthers, M.H., Barone, A.D., Beaucage, S.L., Dodds, D.R., Fisher, E.F., McBride, L.J., Matteucci, M., Stabinsky, Z., and Tang, J.Y. (1987). Chemical synthesis of deoxyoligonucleotides by the phosphoramidite method. Methods Enzymol. *154*, 287–313.

Cello, J., Paul, A.V., and Wimmer, E. (2002). Chemical synthesis of poliovirus cDNA: generation of infectious virus in the absence of natural template. Science *297*, 1016–1018.

Chan, L.Y., Kosuri, S., and Endy, D. (2005). Refactoring bacteriophage T7. Mol. Syst. Biol. *1*, 2005.0018.

Coleman, J.R., Papamichail, D., Skiena, S., Futcher, B., Wimmer, E., and Mueller, S. (2008). Virus attenuation by genome-scale changes in codon pair bias. Science *320*, 1784–1787.

Cowell, M., and Bobe, J. (2009). Straight talk with…Mac Cowell and Jason Bobe. Interview by Prashant Nair. Nat. Med. *15*, 230–231.

Czar, M.J., Anderson, J.C., Bader, J.S., and Peccoud, J. (2009). Gene synthesis demystified. Trends Biotechnol. *27*, 63–72.

Dewannieux, M., Harper, F., Richaud, A., Letzelter, C., Ribet, D., Pierron, G., and Heidmann, T. (2006). Identification of an infectious progenitor for the multiple-copy HERV-K human endogenous retroelements. Genome Res. *16*, 1548–1556.

Dymond, J.S., Scheifele, L.Z., Richardson, S., Lee, P., Chandrasegaran, S., Bader, J.S., and Boeke, J.D. (2009). Teaching synthetic biology, bioinformatics and engineering to undergraduates: the interdisciplinary Build-a-Genome course. Genetics *181*, 13–21.

The Economist. (2006). Life 2.0. The Economist, September 2, 2006, pp. 67–70.

Edge, M.D., Green, A.R., Heathcliffe, G.R., Meacock, P.A., Schuch, W., Scanlon, D.B., Atkinson, T.C., Newton, C.R., and Markham, A.F. (1981). Total synthesis of a human leukocyte interferon gene. Nature *292*, 756–762.

Ferretti, L., Karnik, S.S., Khorana, H.G., Nassal, M., and Oprian, D.D. (1986). Total synthesis of a gene for bovine rhodopsin. Proc. Natl. Acad. Sci. USA *83*, 599–603.

Fodor, S.P., Read, J.L., Pirrung, M.C., Stryer, L., Lu, A.T., and Solas, D. (1991). Light-directed, spatially addressable parallel chemical synthesis. Science *251*, 767–773.

Gibson, D.G., Benders, G.A., Andrews-Pfannkoch, C., Denisova, E.A., Baden-Tillson, H., Zaveri, J., Stockwell, T.B., Brownley, A., Thomas, D.W., Algire, M.A., et al. (2008). Complete chemical synthesis, assembly, and cloning of a Mycoplasma genitalium genome. Science *319*, 1215–1220.

Goler, J.A., Bramlett, B.W., and Peccoud, J. (2008). Genetic design: rising above the sequence. Trends Biotechnol. *26*, 538–544.

Goodman, C. (2008). Engineering ingenuity at iGEM. Nat. Chem. Biol. *4*, 13.

Gupta, N.K., Ohtsuka, E., Sgaramella, V., Buchi, H., Kumar, A., Weber, H., and Khorana, H.G. (1968). Studies on polynucleotides, 88. Enzymatic joining of chemically synthesized segments corresponding to the gene for alanine-tRNA. Proc. Natl. Acad. Sci. USA *60*, 1338–1344.

Gustafsson, C., Govindarajan, S., and Minshull, J. (2004). Codon bias and heterologous protein expression. Trends Biotechnol. *22*, 346–353.

Herper, M. (2007). Venter takes step toward synthetic cells, June 28, 2007. Available at: Forbes http://www.forbes.com/2007/06/28/venter-synthetic-bacteria-tech-science-cx_mh_0628venter.html

Itakura, K., Hirose, T., Crea, R., Riggs, A.D., Heyneker, H.L., Bolivar, F., and Boyer, H.W. (1977). Expression in Escherichia coli of a chemically synthesized gene for the hormone somatostatin. Science *198*, 1056–1063.

Kalman, M., Cserpan, I., Bajszar, G., Dobi, A., Horvath, E., Pazman, C., and Simoncsits, A. (1990). Synthesis of a gene for human serum albumin and its expression in Saccharomyces cerevisiae. Nucleic Acids Res. *18*, 6075–6081.

Khorana, H.G. (1979). Total synthesis of a gene. Science *203*, 614–625.

Kitamura, N., Adler, C., Martinko, J., Nathenson, S., and Wimmer, E. (1980). The genome-linked protein of picornaviruses VII. Genetic mapping of poliovirus VPg by protein and RNA sequence studies. Cell *21*, 295–302.

Kodumal, S.J., Patel, K.G., Reid, R., Menzella, H.G., Welch, M., and Santi, D.V. (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.

Lartigue, C., Glass, J.I., Alperovich, N., Pieper, R., Parmar, P.P., Hutchison, C.A., 3rd, Smith, H.O., and Venter, J.C. (2007). Genome transplantation in bacteria: changing one species to another. Science *317*, 632–638.

Lee, Y.N., and Bieniasz, P.D. (2007). Reconstitution of an infectious human endogenous retrovirus. PLoS Pathog. *3*, e10.

Letsinger, R.L., and Mahadevan, V. (1965). Oligonucleotide synthesis on a polymer support. J. Am. Chem. Soc. *87*, 3526–3527.

Mueller, S., Papamichail, D., Coleman, J.R., Skiena, S., and Wimmer, E. (2006). Reduction of the rate of poliovirus protein synthesis through large-scale codon deoptimization causes attenuation of viral virulence by lowering specific infectivity. J. Virol. *80*, 9687–9696.

Nirenberg, M., and Leder, P. (1964). RNA codewords and protein synthesis. The effect of trinucleotides upon the binding of srna to ribosomes. Science *145*, 1399–1407.

Pan, W., Ravot, E., Tolle, R., Frank, R., Mosbach, R., Turbachova, I., and Bujard, H. (1999). Vaccine candidate MSP-1 from Plasmodium falciparum: a redesigned 4917 bp polynucleotide enables synthesis and isolation of full-length protein from Escherichia coli and mammalian cells. Nucleic Acids Res. *27*, 1094–1103.

Perry, R.T., and Halsey, N.A. (2004). The clinical significance of measles: a review. J. Infect. Dis. *189* (*Suppl 1*), S4–S16.

Picksley, S.M., Parsons, C.A., Kemper, B., and West, S.C. (1990). Cleavage specificity of bacteriophage T4 endonuclease VII and bacteriophage T7 endonuclease I on synthetic branch migratable Holliday junctions. J. Mol. Biol. *212*, 723–735.

Singh-Gasson, S., Green, R.D., Yue, Y., Nelson, C., Blattner, F., Sussman, M.R., and Cerrina, F. (1999). Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. Nat. Biotechnol. *17*, 974–978.

Smith, H.O., Hutchison, C.A., 3rd, Pfannkoch, C., and Venter, J.C. (2003). Generating a synthetic genome by whole genome assembly: phiX174 bacteriophage from synthetic oligonucleotides. Proc. Natl. Acad. Sci. USA *100*, 15440–15445.

Soll, D., Ohtsuka, E., Jones, D.S., Lohrmann, R., Hayatsu, H., Nishimura, S., and Khorana, H.G. (1965). Studies on polynucleotides, XLIX. Stimulation of the binding of aminoacyl-sRNA's to ribosomes by ribotrinucleotides and a survey of codon assignments for 20 amino acids. Proc. Natl. Acad. Sci. USA *54*, 1378–1385.

Stemmer, W.P., Crameri, A., Ha, K.D., Brennan, T.M., and Heyneker, H.L. (1995). Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. Gene *164*, 49–53.

Takehisa, J., Kraus, M.H., Decker, J.M., Li, Y., Keele, B.F., Bibollet-Ruche, F., Zammit, K.P., Weng, Z., Santiago, M.L., Kamenya, S., et al. (2007). Generation of infectious molecular clones of simian immunodeficiency virus from fecal consensus sequences of wild chimpanzees. J. Virol. *81*, 7463–7475.

Taubenberger, J.K., Reid, A.H., Krafft, A.E., Bijwaard, K.E., and Fanning, T.G. (1997). Initial genetic characterization of the 1918 "Spanish" influenza virus. Science *275*, 1793–1796.

Tian, J., Gong, H., Sheng, N., Zhou, X., Gulari, E., Gao, X., and Church, G. (2004). Accurate multiplex gene synthesis from programmable DNA microchips. Nature *432*, 1050–1054.

Tumpey, T.M., Basler, C.F., Aguilar, P.V., Zeng, H., Solorzano, A., Swayne, D.E., Cox, N.J., Katz, J.M., Taubenberger, J.K., Palese, P., and Garcia-Sastre, A. (2005). Characterization of the reconstructed 1918 Spanish influenza pandemic virus. Science *310*, 77–80.

WHO. (2003). Fact Sheet No. 211—Influenza (Geneva: World Health Organization).

Young, L., and Dong, Q. (2004). Two-step total gene synthesis method. Nucleic Acids Res. *32*, e59.

Zolotukhin, S., Potter, M., Hauswirth, W.W., Guy, J., and Muzyczka, N. (1996). A "humanized" green fluorescent protein cDNA adapted for high-level expression in mammalian cells. J. Virol. *70*, 4646–4654.