



Lawrence Livermore National Laboratory

UCRL-TR-201189

Oligonucleotide and Long Polymeric DNA Encoding

R.P. Mariella Jr., A.T. Christian, J.A. Young,
S.N. Gardner, D.S. Clague,
J.M. Williams, and E.L. Miller

December 15, 2003

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

This work was performed under the auspices of the U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

Oligonucleotide and Long Polymeric DNA Encoding

Mariella Jr., Dr. Raymond
Biology & Biotechnology Research Program
Lawrence Livermore National Laboratory

LLNL Team:

Dr. Allen Christian

Dr. Jennifer Young

Dr. Shea Gardner

Dr. David Clague

Ms. Jennifer Williams

Mr. Edward Miller

Abstract

This report summarizes the work done at Lawrence Livermore National Laboratory for the Oligonucleotide and Long Polymeric DNA Encoding project, part of the Microelectronic Bioprocesses Program at DARPA. The goal of the project was to develop a process by which long (circa 10,000 base-pair) synthetic DNA molecules could be synthesized in a timely and economic manner. During construction of the long molecule, errors in DNA sequence occur during hybridization and/or the subsequent enzymatic process. The work done on this project has resulted in a novel synthesis scheme that we call the parallel pyramid synthesis protocol, the

development of a suit of computational tools to minimize and quantify errors in the synthesized DNA sequence, and experimental proof of this technique. The modeling consists of three interrelated modules: the bioinformatics code which determines the specifics of parallel pyramid synthesis for a given chain of long DNA, the thermodynamics code which tracks the products of DNA hybridization and polymerase extension during the later steps in the process, and the kinetics model which examines the temporal and spatial processes during one thermocycle. Most importantly, we conducted the first successful syntheses of a gene using small starting oligomers (tetramers). The synthesized sequence, 813 base pairs long, contained a 725 base pair gene, modified green fluorescent protein (mGFP), which has been shown to be a functional gene by cloning into cells and observing its green fluorescent product.

Introduction

The DARPA Microelectronic Bioprocesses Program directive was meant to create a rapid, efficient process for creating very long, user-defined DNA sequences. While an *in vivo* mechanism for doing so would be ideal in terms of production rate and accuracy (bacteria are able to replicate 4 million bases in 20 minutes), there is no known mechanism by which an organism can produce a “user-defined” nucleic-acid sequence in the absence of a template. Chains of DNA can be produced by a variety of methods, including *de novo* phosphoramidite synthesis, or enzymatic processing of short DNA molecules into longer ones using polymerization or ligation. While the nucleotide-by-nucleotide extension of DNA using traditional phosphoramidite synthesis may appear to be the simplest procedure, it is not viable for lengths greater than 100 base pairs due to the high failure rate of approximately 2% per base (Hecker and Rill, 1998). Therefore, current *de novo* synthesis of DNA typically involves starting with relatively long oligomers (roughly 40 bases) which have been synthesized via phosphoramidite chemistry (Caruthers *et al.*, 1983) and joining them together using DNA ligase and/or polymerase. This research focuses on the development of a novel synthesis technique which will advance the state of the art in DNA synthesis by assembling long DNA of a user-defined sequence (i.e. any possible sequence) from short starting oligonucleotides (i.e. down to 4 bases long) produced via phosphoramidite synthesis.

Current state-of-the-art commercial enterprises use ligation, polymerase, or a combination of the two to create DNA from starting oligos, 40-100 base pairs long, to act as either the starting material or as templates. Shorter starting oligos are not only less expensive to acquire, but they also enable more DNA sequence variations to be synthesized, in house. For example, if one had all 256 (4^4) possible 4-mers on hand, in principle one could make any DNA sequence. In contrast, 1×10^{24} more expensive 40-mers would be necessary to create any DNA sequence. While the desire to make long DNA of user-defined sequence may seem insurmountable and only of academic interest, this is what is needed in the areas of biology (gene therapy, mutagenesis, protein crystallography, protein function), detection of biological agents, and data encoding. The key factors in the synthesis of DNA are the length of the starting oligos and the length, purity, and yield of the final product. The following sections outline the limits of the current, widely-used methods, and describe the novel method that we have successfully developed using polymerase based parallel pyramid synthesis.

Work of Others

Once the starting oligos have been obtained via phosphoramidite synthesis, genes are commonly assembled by combining overlapping complementary oligos via hybridization and linking or extending them by ligation or polymerization respectively. While ligation is a popular method it has several drawbacks in regards to this application. In most experiments involving DNA ligase two or more small oligomers are joined together by hybridizing them to a much longer piece of DNA (i.e. a template) (Leitzel and Lynn, 2001). While this method allows the use of short starting DNA, a lower length limit is set by the size of the footprint of the ligase (typically 6-10 bases), and it still requires a hybridization template that is much longer (e.g. 50 bases) than the starting oligomers. Therefore, although it is possible to use 6-mers in ligation, one must first have the complete set of much longer hybridization templates on hand. This greatly limits the ability of ligation-based technology to synthesize user defined DNA sequences on demand.

Polymerase based extension, also known as DNA shuffling, is regularly used to reassemble complete genes from a pool of enzymatically-cut DNA or synthetic oligomers (Stemmer, 1994, Stemmer, Cramer *et al.*, 1995 and Zhou *et al.*, 2003). This method is similar to PCR in that the polymerase attaches to the 3' end of the DNA and fills in the complementary

bases using the other strand as a template. Yet unlike PCR it is an autoprimering reaction in which the oligos act as their own templates. Unlike the ligase enzyme, the polymerase enzyme is not known to have a minimum footprint size and DNA is created during the process. These factors lead us to believe that polymerase based extension will allow us to minimize the length of our starting oligos as well as maximize the yield of our final product.

While DNA synthesis via ligation (i.e. the creation of a physical bond between two separate pieces of DNA) and polymerization (the creation of DNA by extension as dictated by the complementary strand) are different enzymatic processes, there is a fundamental similarity. This similarity is the association, or hybridization, of single strands of DNA by Watson-Crick base-pairing. This phenomenon has been widely studied in the context of DNA melting, and the same associations are assumed to occur in the presence of an enzyme. Acting as a catalyst, however, the enzyme may enhance the rates of association above those predicted based on the thermodynamics of hybridization alone. Consistent with the aims of this effort, we have developed models and computational capabilities to examine polymerase-based extension. These new capabilities, however, have been developed in a framework that can be easily modified and applied to ligase-based synthesis also.

DNA synthesis via ligation and polymerization appears to be straightforward, but in practice, errors in DNA synthesis commonly occur. These errors are a result of impure starting oligos, errant enzymatic processes and incorrect or undesired hybridizations. Fortunately, the incorrect incorporation of bases during polymerase extension is extremely low, 10^{-4} - 10^{-5} errors/base pair for many polymerases (Matilla *et al.*, 1991 and Lundberg *et al.*, 1991). These errors were not included in the current generation of software but could easily be added in future generations. The polymerase also acts as a key error checking device in that it halts or severely retards chain extension if there is a mismatch at the 3' end (Petruska *et al.*, 1988). Errors also arise during synthesis if there are mismatched base pairs in the hybridized region or if undesired yet mis-match free hybridizations occur. Each of these potential sources of error have been considered in both our computational and experimental efforts by choosing the starting oligos and optimally distributing them in reaction wells as well as establishing reaction conditions which minimize the possibility of undesired hybridizations. These considerations are key in developing a successful experimental synthesis scheme, as discussed below, and require predictive modeling approaches to minimize the incorporation of errors.

The length of the starting oligos and the length of the desired product play a central role in the development of a synthesis method which will produce DNA with minimal errors at the maximum yield. To date, others have taken the 135 40-mers which comprise a 2700 base plasmid, placed them in a single reaction tube, and assembled the desired product using polymerase extension (Stemmer, Cramer *et al.*, 1995). This approach works because each of the 40-mers overlaps by 20bp with its neighbors, thus affording 4^{20} unique overlaps of which only 135 are used in this assembly. So, barring any repeated sequence segments, one may assume that the overlap regions are sufficiently unique that incorrect and undesired hybridizations are extremely unlikely. In contrast, if one were to attempt to make the same 2700 bp plasmid by starting with shorter oligos such as 6-mers, requiring 900 starting 6-mers, one would be faced with the fact that only $4^3 = 64$ unique overlaps exist. Clearly, the attempt to synthesize the plasmid using 6-mers in a single reaction would fail. Therefore, we have adapted the approach for shorter starting oligos of first subdividing the process into a number of reaction wells. The longer products of each well are then mixed with the products of other wells until the final gene is assembled. This process is called the parallel pyramid synthesis process and is shown schematically in Figure 1.

In this approach, a bioinformatics-based code is used to choose the placement of oligos of a specified length and sequence into multiple wells which contain the reaction buffer and polymerase enzyme. These wells are then thermocycled, as is done in PCR, to anneal and extend the DNA. This process occurs simultaneously for all of the wells, thus significantly reducing the overall processing time. The products of these wells are then successively mixed with the products of other wells as guided by the bioinformatics analyses. The process is then repeated until one well remains, which generates the desired long-chain product, after its thermocycling. The algorithms developed for the bioinformatics code were designed to minimize the occurrence of errors during the parallel pyramid synthesis process by examining the melting temperatures of the interacting strands. Not only are the sequence and distribution of oligos critical, but also the processing conditions (temperature, number of thermocycles, and concentration). Thermodynamics and kinetics based codes are used to examine how these variables affect the specific distribution of products that result. The thermodynamics code tracks both the exact sequences of the products and the concentrations of each DNA strand through the after the first few steps in the synthesis process. This is accomplished by modeling the equilibrium

hybridization of single strands based on nearest neighbor thermodynamics (Bommarito *et al.*, 2000, Peyret *et al.*, 1999, SantaLucia, 1998 and SantaLucia *et al.*, 1996) and performing the appropriate polymerase extension. This allows us to track the specific sequences of errant products, how they originate, and how they propagate through the synthesis process. The kinetics code is then used to examine important reaction wells in more detail by accounting for the effects of both time and the spatial distribution of DNA fragments. Like the thermodynamics code, the kinetics code also provides a prediction of annealed products. Because both polymerization and ligation are enzymatic processes, understanding the kinetics of both hybridization and the catalyzed reaction are important in fine tuning the synthesis process. Figure 2 shows how these aspects of the model are interrelated. The following sections discuss in detail how they minimize and quantify the errors in the final product.

While the considerations which comprise the parallel pyramid synthesis approach to DNA synthesis are important, ultimately, it is the ability of this procedure to produce the desired long chain DNA in the lab that is of utmost importance. To this end, we have run experiments to both examine our ability to produce the desired DNA and to determine key issues encountered in the lab which will affect the fidelity of our final product. Experimental data, obtained through the literature, was used to establish values for parameters used in the models. Not only does the production of correct sequences depend on the hybridization of DNA strands, it also hinges on the activity of the enzyme. Experimental data shows that polymerase enzymes aid in the reduction of errors by not extending hybridized strands that have mismatches at key positions or by doing so at a greatly reduced rate (Petruska *et al.*, 1988). Yet ligase has been shown to tolerate mismatches to a sufficient degree that serious problems in fidelity result (Housby and Southern, 1998 and James *et al.*, 1998). While the exact mechanism of this selective enzymatic action is not clear, these observations have been included in our models. To determine whether we were able to synthesize the desired product, we have turned to sequencing techniques and when the product was a gene, it was cloned into a cell and the gene expression monitored. Due to the ease of cloning and observation of expression, we focused on the synthesis of mGFP. This gene is 725 base pairs long; the synthesized sequence was 813 base pairs long. Flanking the gene sequence were sites for PCR primers, to allow post-synthesis amplification, and restriction endonuclease cut sites, for cloning purposes. Previous to our success, no one had achieved *de novo* gene synthesis (without a template) starting from small oligos only four base pairs long.

We have decreased the size of the starting oligos by an order of magnitude compared to what had been accomplished before, four versus 40 bps, and through reduction of starting oligo size, we anticipate concomitant improvements in cost and speed in gene synthesis. Traditionally, the processes of hybridization and polymerization are usually improved upon largely through experimental iteration with minimal guidance from computational efforts. Recent efforts such as the development of DNA melting programs have demonstrated the potential importance of calculations yet detailed modeling of more complicated processes such as parallel pyramid synthesis has not been attempted. This is because current models 1) need to address more of the underlying physics and biology of the problem; 2) only examine isolated aspects of the systems; 3) do not follow all significant reaction pathways which arise during the process. To this end, we developed an integrated computational suite that captures the problem from three integrated perspectives: Bioinformatics, Thermodynamics, and Kinetics. This three-pronged approach not only models relevant parts of the process under investigation and the resulting reaction pathways, but also *integrates* the results, feeding the output of one calculation into the subsequent calculation.

Methods, Assumptions, Procedures

Bioinformatics

The goal of the bioinformatics code for DNA polymerase-based synthesis is to minimize synthesis errors by determining the optimal subdivision of the full-length sequence into starting oligos of length n (n -mers) and the optimal distribution of oligos into wells in the first and subsequent rounds of self-priming PCR reactions. This code has two main capabilities:

Selection of initial oligos and their allocation within wells:

Input to the first part of the bioinformatics module is the full length DNA sequence that is the target for synthesis and the size ($=n$) of oligos to start the process. The program outputs the total number of wells (each containing a spatially separated PCR reaction) into which the starting oligos should be divided. Also output are the sequences of every oligo that should be mixed in each well, and predicted melting temperatures, T_m 's (Deaton *et al.*, 1998), of the hybridizations that are desired to occur in that well.

The process that is used here is to cut (*in silico*) the full length DNA sequence into pieces of size n , for every reading frame from 1 to n . Starting at the beginning (5' end of the positive sense strand) of the sequence, the program cycles through reading frames to find which will enable one to combine the maximum number of starting oligos of size n in a given well without mixing in error-prone sequences, as defined by the melting temperatures of the oligos. This reading frame is chosen as the optimal reading frame for that well, as it will maximize the length of the synthesized fragment in that well. All n -mers in the same well are in the same reading frame, in order to maintain constant overlap distances between each sequential pair of oligos, although each well may be in a different reading frame. In the first well, the reading frame was automatically set to start at the beginning of the chain, i.e. reading frame one. The reading frames of the subsequent wells varied to optimize the process.

The program carries out the same process for each well, advancing to a new well whenever further addition to that well would result in errors for the given reading frame. Oligos in the new well are selected so that there is sequence overlap between wells ranging from length n to $3n+1$.

There are a number of conditions that could result in sequence errors, that can be minimized by appropriate subdivision into wells. These include the following:

- 1) Perfect but incorrect hybridizations that are greater than or equal to the length that is the maximum of 2 or the integer part of $n/4$ that will allow extension from a 3' end. The quantity $n/4$ is half the length of the desired overlap, $n/2$. Setting a lower bound on this number of 2 is chosen so that with tetramers as starting oligos, one does not check for matches of only 1 base. Perfect but undesired hybridizations can occur for a number of reasons, for example, as a result of self-complementary sequence, repetitive sequence, or complementary 3' ends of oligos (that make up either the positive or negative sense strands) that are not adjacent in the target sequence. Self-complementary sequences not only reduce the desired yield by consuming reactants but also produce incorrect sequences when extended.
- 2) T_m 's of desired hybridizations that are too low compared to other desired hybridizations in the same well. This can occur if the GC ratios differ substantially between desired hybridizations. This is important because it ensures that all of the desired products will have similar yields for the given processing conditions.

- 3) T_m 's of unwanted hybridizations (those that can result in sequence errors) are too close to those of the desired hybridizations. T_m 's of unwanted hybridizations must be sufficiently far below (in these results we used 5 °C) those of the desired reactions to reduce the likelihood of them occurring.
- 4) Oligos that contain sequence patterns that are difficult to create (and thus error prone) in the chemical synthesis of the starting oligos, such as the sequence GGG, are avoided.

In some cases, it is not possible to subdivide a sequence to avoid errors. For example, with $n=6$, the following sequence cannot be subdivided into wells without breaking one of the above rules:

forward: TCGGGAGATCTA

reverse: AGCCCTCTAGAT

has the following problems in each reading frame:

frame 1: TCGGGA problem with GGG

frame 2: CGGGAG problem with GGG

frame 3: GGGAGA problem with GGG

frame 4: GGAGAT minus strand is 3'CTAGAT'5 which is self complementary with an overlap of 4:

5'TAGATC'3

3' CTAGAT'5

frame 5: 5'GAGATC'3 is self complementary with overlap of 4

3'CTAGAG'5

frame 6: 5'AGATCT'3 will bind to itself with an overlap of 6 but will not polymerize (6 base overlap). Also it will bind to itself with an overlap of 3:

5'AGATCT'3

5'AGATCT'3

In such cases where it is not possible to subdivide sequence into wells while avoiding all potential errors, we allowed starting oligos to contain GGG's and self-complementary sequence. If the sequence could not be split so as to simultaneously avoid mixing oligos in error-prone combinations while also generating sequence overlap between wells, then the requirement for sequence overlap between wells prevailed. Such wells containing error-prone oligos were noted.

Although incorrect hybridizations will compete with the desired reactions, it is hoped that at least some of the desired hybridizations will occur, albeit at a lower efficiency.

Combining wells

The second part of the bioinformatics code specifies how wells containing the products of the first tier of reactions should be mixed in the second and subsequent rounds. The series of products from one tier become the starting products for the next round. This is repeated in multiple rounds, until finally the full target DNA sequence is produced. We call this the “parallel pyramid plan”. Each step entails determining how many wells are required in that tier, and how the products of the previous round should be allocated among those wells.

This program takes as input the desired products from the previous tier or the range of products predicted from thermodynamic calculations. As output, the program produces well allocations for the next level of parallel pyramid synthesis and their desired products, as well as the predicted T_m 's of the desired hybridizations occurring in each well.

The wells are combined sequentially along the strand until adding the products of another well would result in sequence errors due to one of the conditions (1-3) described above in the first part of the bioinformatics module, at which point a new well is started. The code ensures that there is increasing overlap between the sequences within and between each well as the tier level increases, thus the desired T_m 's increase for higher specificity binding with each round of synthesis. This is accomplished by, for example, mixing the products of wells 1-4 and 4-6 from the previous cycle into wells 1 and 2, respectively, of the next cycle. Thus, that the products of wells 1 and 2 in the next cycle will overlap by the length of the desired product of well 4 in the previous cycle.

Thermodynamics

The goal of the thermodynamics module is to predict the equilibrium products of each PCR synthesis step as a function of variables such as temperature, concentration and PCR cycle. The sequence identities of the distribution of predicted products are examined to see how they compared with the desired product, i.e. exactly what errors in the synthesis are likely to occur and how they can be minimized. The input to these calculations includes the sequences and well allocations determined from the bioinformatics code as well as the experimentally-determined

annealing temperatures from the near neighbor model (SantaLucia *et al.*, 1996) and the concentration of each DNA fragment.

Equilibrium Product Prediction

The products of later PCR thermocycles are determined by considering all of the possible hybridizations between the oligos in solution which could result in DNA extension, i.e. extension from one or both ends of the duplex as a result of having a perfect base pair match at the 3' end as well as the accompanying strand to use as a template. Using parameters for the enthalpy and entropy of hybridization (Bommarito *et al.*, 2000, Peyret *et al.*, 1999, SantaLucia, 1998 and SantaLucia *et al.*, 1996), the free energies of hybridizations between each DNA fragment and all the others in the solution are calculated for all overlaps of v_{\min} or greater. The number of possible reaction pathways depends of the number of oligos present (i) and length (l) of the oligos. Assuming that all oligos are of equal length, which is only true at the beginning of the first thermocycle, the following equation estimates the number of reaction pathways.

$$pathways = (2 * (l - v_{\min}) + 1) \left(\frac{i!}{2!(i-2)!} + i \right).$$

As products are produced, the number of pathways quickly increases as a function of thermocycle. While the minimum allowed overlap is an adjustable parameter in the code, the thermodynamic parameters that are used have only been validated for overlaps of six or more. With this in mind, current calculations have focused on predicting the products of tiers two and higher which have overlaps of this size. In an effort to focus on the relevant reactions, reaction pathways are followed only if their free energy is below a predefined limit. While free energy limits of 0 kcal/mol are commonly used, positive free energies are considered when the focus is on short oligos with weaker annealing energies. Due to the catalytic nature of the enzyme, allowing positive free energies is a reasonable assumption. Those reaction pathways which have favorable free energies are used to generate the simultaneous equations which relate the concentrations of single and double stranded DNA to their free energies of hybridization. Each reaction is described by the following equation:

$$\frac{x_{i,j,v}}{x_i x_j} = \exp\left(\frac{-\Delta G(T)}{RT}\right)$$

where x_i and x_j are the concentrations of the two single strands, $x_{i,j,v}$ is the concentration of the hybridized duplex having an overlap of v and ΔG is the free energy of hybridization which is a

function of temperature and ion concentration. Using the initial concentrations of the species or their concentrations from the previous thermocycle, the set of simultaneous equations is numerically solved using a modified Newton method to give the new equilibrium concentrations of the species. Once the concentrations are calculated, only species with concentrations above a specified cut off were retained for the next thermocycle. To account for the error checking of the polymerase, extension of the strand occurs if there is not a 3' mismatch (Petruska *et al.*, 1988). The products and concentrations of the products which result from this thermocycle are then used as input for the following thermocycle. The code is written in Mathematica 5.0 and runs on a PC, but is limited in its ability to solve large numbers of highly non-linear simultaneous equations. It has proven adequate, however, for examining sections of the synthesis process in detail. To allow continuous simulation of the entire parallel pyramid synthesis process we are in the process of porting the code to a more robust system, the details of which will be described later.

Kinetics

The goal of the kinetics module is to track both the temporal and spatial production of annealed products and DNA products produced during PCR based synthesis as function of operating conditions. Specifically, the kinetics and thermodynamics of DNA annealing, along with the kinetics of polymerase association and extension, have been modeled for *all* significant reaction pathways that arise during parallel pyramid synthesis. The probability of a particular reaction is a function of the base-pair-specific thermodynamic free energies and oligo proximities, relative orientations, and mobilities. To determine rates of reaction as a function of the free energy of each reaction pathway and oligo-oligo proximity, we adapted Gillespie's (Gillespie, 1977) stochastic kinetics formulation and developed a novel Monte Carlo simulation capability to predict the probability of pairs of oligos being in the proper position to anneal. This simulation capability takes into account all possible reaction pathways, the concentration of starting oligos, and the reaction volume.

The DNA fragments from the bioinformatics code, the free energies from the thermodynamics code, and experimental oligo concentrations are accounted for in the stochastic kinetics formulation in the following manner.

$$P_v \alpha (P_{\Delta G} * P_{prox})_v$$

where P_v is the over all probability of reaction pathway v occurring, which is a function of the product of the $P_{\Delta G}$ and P_{prox} , the reaction probabilities based on the free energy and physical proximity of oligos respectively for reaction pathway v . The probability based on physical proximity is determined through a psuedo Monte Carlo technique. The overall probabilities for each reaction pathway is used in Gillespie's Stochastic Kinitics formulation to update product concentrations and distributions as a function of concentration, time, and processing conditions.

Experimental

We have designed and performed experiments to test and validate the approach described above. The first step in this effort has been to show that the proposed parallel pyramid synthesis protocol for producing long DNA of a specified sequence from short staring oligos is viable. To do this we have chosen to focus our efforts on the production of the modified green fluorescent protein (mGFP) gene. This choice in genes allows us two methods of checking the sequence of the products. First, we sequenced the results using standard capillary electrophoretic sequencing methods at the Joint Genome Institute to obtain detailed sequence information. Second, we determined that the desired gene was functional by cloning the product into a shuttle plasmid and transforming it into bacterial cells, which were expanded to a plateau-phase culture. The mGFP gene was isolated from the bacteria, and transfected into human cells using a cytomegaloviral constitutive promoter. Gene expression was then visualized using a fluorescent microscope.

Results and Discussion

Need for parallel synthesis

The length of the starting oligos, the length of the desired long DNA chain, and the sequence of the chain are important in determining the specifics of the synthesis protocol. When long starting oligos such as 40-mers are used, synthesis of the mGFP gene can be done in a single reaction well. Experiments done in our laboratories have successfully produced the gene from both 100-mers as well as 40-mers in a single reaction chamber. These results have been verified by sequencing the products at the DOE Production Sequencing Facility in Walnut Creek, CA, using Sanger sequencing chemistry. In addition, subsequent cloning followed by analysis by fluorescence microscopy demonstrated that a functional mGFP gene had been successfully

assembled by this method. While these results validated published studies on the use of polymerase based synthesis using 40-mers as the starting oligos, it does not demonstrate the use of shorter starting oligos. The current literature does not establish a lower limit for the size of the starting oligos used in self-priming PCR. To establish such a limit, our laboratory investigated starting oligos of various lengths. Experiments were able to produce a 96-mer from 6-mers in two hours with no intermediate purification steps. The synthesis of a 40-mer from 4-mers was undertaken and the sequence of the product was verified using HPLC methods. Having shown that we were able to use 4-mers for polymerase based auto-priming extension, we began developing a parallel pyramid protocol which would exploit the use of these short starting oligos.

As previously stated, the synthesis of long DNA from short starting oligos requires the use of the parallel pyramid protocol. This is demonstrated computationally by considering the reactions which occur when using 6-mers as starting oligos. The selectivity of the desired reaction, i.e. hybridization, is defined as

$$S_{i,j,v}(T) = \frac{x_{i,j,v}}{\sum_{j=1}^{\nu_{\max}} \sum_{\nu_{\min}} x_{i,j,v}}$$

and serves as a useful characterization of the potential for the incorporation of errors as a result of undesired hybridizations. Selectivities of one (zero) indicate that the correct hybridization will occur (will not occur). Figure 3 shows that if one were to attempt to make mGFP from 6-mers in a single reaction, numerous errors would be produced in the first thermocycle alone. When the sequence is optimally divided into 6-mers that are allocated to multiple wells as dictated by the first module of the bioinformatics code, the selectivity is greatly enhanced.

The synthesis protocol and analysis procedure presented in this report can be applied to oligos and DNA chains of any length, considering current computer power. Therefore, we examined how the length of the starting oligos affects both the number of wells in the first tier as well as the total number of tiers for mGFP, figures 4a and 4b. These figures show that by using 20-mers or longer as starting material, one can synthesize mGFP in two reaction wells. The reason that there are two reaction wells is that the T_m of desired hybridizations is lower in one well than the other (rule (2) in the Bioinformatics section above). One well is likely to be

adequate, and would be predicted by the bioinformatics code, if the T_m range parameter for desired hybridizations within a well is increased. Yet one must recall that mGFP is less than 1 kb long. Analysis shows that to synthesize longer genes of 10 Kb using 40-mers, approximately 5 wells would be necessary, or starting with 20-mers might require 20 wells. To obtain information about the protocol necessary to make genes other than GFP, we examined the effects of gene length on the number of required wells and tiers, diagrammed in Figure 5. It can be seen that there is a strong linear relationship between the number of wells in the first tier and the size of the gene. Because the number of tiers necessary for synthesis does not rapidly increase, the total synthesis time for longer genes will not increase rapidly.

Synthesis plan

Starting tetramer sequences and well allocations were selected using the bioinformatics module. The details of the sequences and well allocations are included in the appendices. Figure 1 summarizes the protocol and highlights the growth of the DNA chains during the process. On the right of this figure we see the average length of the desired reactants of each tier. In reality, all tiers after the first will receive as reactants a distribution of product lengths from the previous tier.

Experimentally it has been found that selection of the correct enzyme is crucial to the success of the process. We tried a variety of enzymes, including those with 5' – 3' exonuclease capability (Vent, Deep Vent) and those deficient in 5' – 3' repair capability (pfu-, Vent-, Deep Vent-). We found that when the reaction was initiated with tetramers, only those *without* the repair function were capable of producing any product at all. This may indicate that there are many errors in the extension of very short sequences. The fact that the 'ground state' for tetramer annealing is likely to be the correct sequence is what allows for the prevailing product to be the correct one. However, it is very difficult to analyze such short sequences, and in all cases, a large distribution of sizes was seen following gel analysis of the products of each Tier. This large size distribution presents a certain amount of difficulty in the eventual automation of the process, but it should be noted that there were no purification steps between the multi-Tiered reactions; products from one Tier were added directly to the wells making up the next Tiers.

Synthesis of mGFP using 4-mers as the starting oligos was completed in the following manner. The 4-mers determined from the bioinformatics code were resuspended and combined

in the specified well which contained a reaction mixture. Each well was then thermocycled to produce longer fragments via polymerase extension. The products of these first tier reactions were then used as the template in a second set of PCR reactions. Upon completion of the parallel pyramid synthesis protocol the product was digested with restriction enzymes and the fragment DNA purified from a 1.2% agarose gel. After ethanol precipitation, the DNA was dissolved and restriction enzymes were also used to digest the pcDNA3.1(+) vector (Invitrogen). Phenol-chloroform extraction and ethanol precipitation followed. The PCR product fragment and vector fragment were ligated with T4 DNA ligase. This assembled product was used to transform *E.Coli* Max Efficiency® DH5 α F'IQ competent cells (Invitrogen) and the bacterial colonies grown. The cells were then transfected, subcultured and grown. GFP expression was visualized after three days using a fluorescent microscope with a filter for FITC. Figure 6 shows the glowing mGFP cells which indicated the success of this protocol.

This novel result of making a \approx 1-Kb gene from tetramers using the parallel pyramid synthesis protocol is important for the several reasons. It demonstrates the ability to use tetramers as starting material in polymerase-based reactions. This work shows that the enzyme facilitates the interaction between short oligos and extends them. As discussed in the Introduction, the ability to use short instead of longer oligos as starting materials, in principle adds greatly to the viability of delivering on-demand gene synthesis to the mass market. Second, this demonstrates the ability of the bioinformatics code to organize the synthesis protocol in a manner which reduces the number of errors in the produced sequence enough to allow PCR amplification of the desired product. This first stage of error reduction is critical.

Having successfully tested the initial oligo selection and allocation code we began to examine in more detail what DNA sequences are produced during synthesis. The current model for the mixing of wells assumes that only the desired product is made in the preceding tier. To examine how this assumption affects the calculation of errors, the thermodynamics code is used to track the products of the first well in Tier 3 as a function of thermocycle. Figure 7 shows two cases. First, when only the desired products from the wells in Tier 2 are considered at the initial concentrations of 0.4mM, we see that no errant sequences are produced and that there is a high concentration of desired product. However, when all of the products produced in Tier 2 wells, both desired and undesired, as predicted by the thermodynamics code, are mixed into the Tier 3 well, the concentrations of the desired products decrease and the concentration of errant

sequences increases. This is a result of continuing to amplify errant sequences from previous tiers as well as the introduction of more reaction pathways as a result of mixing the fragments in each well. While this may appear alarming, it is important to recall that after the parallel synthesis protocol has been completed, PCR is used to amplify the desired product. Although not shown, it is also worth noting that after multiple thermocycles there is a broad distribution of product lengths in each well. These sequences are a combination of unreacted oligos, partially reacted fragments, fully reacted fragments, errant sequences and fragments which, although not the desired product of the given well, do appear elsewhere in the gene. Because errors are propagated and possibly even amplified as one moves up the pyramid, it is important to minimize the errors at every step of the process.

While each well is unique in what specific reactions occur and how incorrect sequences are produced, we will highlight some situations which are commonly observed. If there are more than two starting oligos in a first tier well (or more than four in higher level wells) multiple hybridizations/extensions are necessary to produce the final product. Because all of the necessary reaction pathways are unlikely to occur with the same probability due to differences in AT/GC content, it is common to produce the desired top and bottom strands in different concentrations. Figure 8 shows how the concentrations of the desired top and bottom strands evolves as a function of thermocycle. The number of thermocycles necessary to reach a steady state concentration of products is a function of the extent of overlap between neighboring strands, the AT/GC content of the overlaps, and the number of reacting fragments in the well. The lower the free energy of the necessary hybridizations (as a result of longer overlaps and higher GC content) and the fewer the number of reactions necessary to reach the desired product, the more quickly a steady state is reached.

Figure 9 shows what can occur when additional reaction pathways are created after multiple thermocycles have been completed. Here the top and bottom strand are initially produced at the same rate but then the concentration of the bottom strand decreases while the concentration of strands containing sequence errors begins to rise. Detailed examination of the reaction pathways reveals that in the first cycle an undesired hybridization and extension occurs along a low probability pathway ($\Delta G = -0.84$ kcal/mol). After denaturing, one of these strands hybridizes to the bottom strand with a large overlap of 14 bp. Due to the low free energy ($\Delta G = -$

10 kcal/mol) of this reaction the bottom desired strand is quickly consumed in a reaction which creates errant DNA.

Temperature is an effective tool in controlling the production of errant DNA sequences. Figure 10a shows that as one increases the annealing temperature, the concentration of error containing DNA strands decreases. Yet the cost of increasing the temperature is generally a decrease in the overall yield of desired products. Due to the non-linear relationship between the free energies and the resultant concentrations, and the strong sequence-dependence of the reactions, no sweeping generalization can be made. This ability to control the incorporation of errors via changing the temperature is likely to be most valuable in the upper tiers of the parallel pyramid protocol. The bioinformatics code has been designed such that the average overlap between neighboring wells increases as one moves up the parallel pyramid protocol. The final levels of the pyramid have overlaps of 20-30 bp, thus giving melting temperatures around 70 °C. Therefore, annealing the fragments at 40 °C will activate unwanted reaction pathways. Examining the errors that arise in just the first round of thermocycling of tier 5 wells, one sees that the concentration of errors drops from an average of 1.85 mM, to 0.15 mM to 0 mM in going from 40 °C to 50 °C and 60 °C. Simulations such as these offer a powerful tool to investigate and optimize the multiple interrelated non-linear variables which control the synthesis of DNA.

It has been shown experimentally that the purity of the starting oligos is an important factor in determining the success or failure of the synthesis. As an example, in some cases when starting oligos were obtained from a single source no product of the expected length was produced. Yet the exact same protocol but with oligos obtained from multiple vendors yielded the desired product. Having tested for the activity of the polymerase it was determined that the key difference in the runs was the source of the starting materials. While it is not possible to first sequence every starting oligo that will be used, errors in phosphoramidite synthesis tend to be consistent, i.e. human error may lead to AATT being made instead of AATA or errors in the chemistry may always occur during GGG synthesis. The thermodynamic code is capable of examining how such errors affect the fidelity of the final product. For example, if six perfectly synthesized starting oligos are placed in a well, the final concentration of the desired product will be 0.57 mM and the concentration of the errant fragments will be 0mM. Now if one of those oligos were mis-synthesized such that half were incorrect by one base, the concentration of the desired product is reduced to 0.36 mM and the concentration of the errant fragments increases to

0.2 mM. While this simple example highlights the problems associated with impure starting oligos, the magnitude of the effects will be highly dependant upon the exact nature of the reactants.

While errors can arise as a result of hybridizations that contain one or more mismatched base pairs, the majority of errors arise from undesired hybridizations and extensions. While the bioinformatics code greatly reduces the occurrence of these unwanted reactions, some still occur as new reaction pathways are activated during both thermocycling and the mixing of products from wells in lower tiers. Determination of the optimal processing conditions for the synthesis of a specific gene, as well as obtaining data about the type and quantity of errant sequences, requires a full simulation of each well, in addition to carrying all of the products from previous tiers. At the current time we are capable of doing only parts of this process. The main obstacle is solving all of the simultaneous equations which arise in an efficient and robust manner. To address this issue we have begun migrating the Mathematica program to another language that affords us more control over the numerical methods employed, improved code maintainability, and facilitates the development of parallel versions of the thermodynamics analyses. Perl was selected as the primary language for its rapid development potential, powerful string manipulation, ease of extensibility, and its wide familiarity within the programming community. An external numerical library written in C, KINSOLV, can be externally called from Perl to solve the large number of simultaneous equations required by the problem, utilizing adaptive Newton-Krylov methods. This library has been developed with parallelism in mind, and ultimately will allow us to run our program distributed over multiple processors or machines via MPI. Ultimately this shift in development strategy should yield a faster, more robust code base which can be easily extended by many members of the bioinformatics community.

The thermodynamics simulations examine the production of products as a function of thermocycle and the kinetics code allows for a more thorough examination of the evolution of products in a single well. Specifically, the kinetics allows for the temporal and spatial tracking of annealed products and DNA products produced during PCR based synthesis as function of operating conditions. Figure 11 shows the time dependence of the reactant and product concentrations when 20-mers are run through one PCR thermocycle. These results highlight the rapid uptake of reactants and may be used to optimize the necessary annealing times.

Summary

It is surprising to discover that 4-mers are capable of being used to produce a 1-kb base pair gene. This is possible because of the novel, computationally-based synthesis protocol that we have developed called parallel pyramid synthesis. During this project we have not only shown the viability of this approach but we have also developed the tools necessary to examine the application of this synthesis protocol to other long chain DNA sequences. As the simulations presented in this report demonstrate, there is much room for further optimization of the procedure. Due to the shortened timeline of this project we were unable to do a full experimental verification of the computationally predicted products as a function of the key variables (i.e. temperature, PCR cycle, etc.). Additional areas of study which would greatly improve this technique include the examination and further quantification of the role of the enzyme in the polymerization process, the development of a set of thermodynamic hybridization parameters geared specifically towards short oligos (i.e. less than 10 base pairs) in the presence of the polymerase as well as probing the use of novel filtering techniques for purification between mixing steps. While further optimization is necessary, the ability to produce synthetic long DNA and artificial genes by this method will accelerate the pace of biological research by making gene synthesis more affordable and less time consuming.

References

- Bommarito, S., Peyret, N., and SantaLucia, J. (2000) "Thermodynamic parameters for DNA sequences with dangling ends." Nucleic Acids Research **28**(9): 1929-1934.
- Caruthers, M.H., Beaucage, S.L., Becker, C., Efcavitch, J.W., Fisher, E.F., Galluppi, G, Goldman, R, deHaseth, P., Matteucci, M., McBride, L., *et al* (1983) "Deoxyoligonucleotide synthesis via the phosphoramidite method." Gene Amplif. Anal. **3**: 1-26.
- Deaton, R., Garzon, M., Murphy, R. C., Rose, J. A., Franceschetti, D. R., and Stevens Jr., S. E. (1998) "Reliability and efficiency of a DNA-based computation." Physical Review Letters **80**: 417-420.
- Gillespie, D. (1977) "Exact stochastic simulation of coupled chemical reactions." The Journal of Physical Chemistry **81**(25): 2340-2361.
- Hecker, K. and Rill, R, (1998) "Error analysis of chemically synthesized polynucleotides." Biotechniques **24** (2): 256-260.
- Housby, J.N., and Southern, E.M. (1998) "Fidelity of DNA ligation: A novel experimental approach based on the polymerization of libraries of oligonucleotides." Nucleic Acids Research **26**: 4259-4266.
- James, K.D., Boles, A.R., Henckel, D., and Ellington, A.D. (1998) "The fidelity of template-directed oligonucleotide ligation and its relevance to DNA computation." Nucleic Acids Research **26**: 5203-5211.
- Leitzel, J.C. and Lynn D.G. (2001) "Template directed ligations: from DNA towards different versatile templates." Chem. Rec. **1**(1): 53-62.
- Lundberg, K.S., Shoemaker, D.D., Adams, M.W., Short, J.M., Sorge, J.A., and Mathur, E.J. (1991) "High-fidelity amplification using a thermostable DNA polymerase isolated from *Pyrococcus furiosus*." Gene **108**(1): 1-6.
- Matilla, P., Korpela, J., Tenkanen, T., and Pitkanen, K. (1991) "Fidelity of DNA synthesis by the *Thermococcus litoralis* DNA polymerase--an extremely heat stable enzyme with proofreading activity." Nucleic Acids Research **19**(18): 4967-4973.
- Petruska, J., Goodman, M.F., Boosalis, M.S., Sowers, L.C., Cheong, C., and Tinoco, I. (1988) "Comparison between DNA melting thermodynamics and DNA polymerase fidelity." Proc. Natl. Acad. Sci. USA **85**: 6252-6256.

Peyret, N., Seneviratne, A., Allawi, H., and SantaLucia, J. (1999) "Nearest-neighbor thermodynamics and NMR of DNA sequences with internal AA, CC,GG, and TT mismatches." Biochemistry **38**: 3468-3477.

SantaLucia, J., Allawi, H.T., and Seneviratne, P.A. (1996) "Improved nearest-neighbor parameters for predicting DNA duplex stability." Biochemistry **35**: 3555-3562.

SantaLucia, J. (1998) "A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics." Proc. Natl. Acad. Sci. USA **95**: 1460-1465.

Stemmer, W. (1994). "DNA shuffling by random fragmentation and reassembly: *in vitro* recombination for molecular evolution." Proc. Natl Acad. Sci. U.S.A. **91**: 10747-10751.

Stemmer, W. and Cramer, A. (1995) "Single-step assembly of a gene and entire plasmid from large number of oligodeoxyribonucleotides." Gene **164** (1): 49-53.

Zhou, Z., Zhang, A.H., Wang, J.R., Chen, M.L., Li, R.B., Yang, S., Yuan, Z.Y. (2003) "Improving the specific synthetic activity of a penicillin G acylase using DNA family shuffling." Acta Biochimica Et Biophysica Sinica **35** (6): 573-579.