Production and Characterization of MutS for use in Error Correction

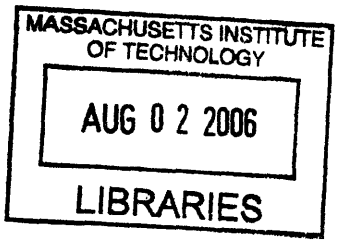
by

Samuel James Hwang

SUBMITTED TO THE DEPARTMENT OF MECHANICAL ENGINEERING IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

BACHELOR OF SCIENCE
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 2006



Signature of Author: _____
Department of Mechanical Engineering
May 11, 2006

Certified by: _____
Joseph Jacobson
Associate Professor of Media Arts and Sciences and Mechanical Engineering
Thesis Supervisor

Accepted by: _____
John H. Lienhard V
Professor of Mechanical Engineering
Chairman, Undergraduate Thesis Committee

Production and Characterization of MutS for use in Error Correction

by

Samuel J. Hwang

Submitted to the Department of Mechanical Engineering
on May 11, 2006 in Partial Fulfillment of the
Requirements for the Degree of Bachelor of Science in
Mechanical Engineering: Entrepreneurial Management and Bioengineering

ABSTRACT

The availability of inexpensive synthetic DNA (oligonucleotides) has allowed for the synthesis of longer, gene-length constructs of DNA. However, a critical barrier to making this technology a low-cost and high-throughput process has been due to the rate at which errors pervade the final product. The current state of the error reduction technology includes three different categories: error filtration, error correction, and error prevention. My research is a joint project as well as an addendum to the work done by Research Scientist Dr. Peter Carr and current MIT Department of Biological Engineering Masters Student Jason Park (MIT '05) who have been working on research in gene synthesis error correction over the past several years. I have been working very closely with both Dr. Carr and Jason Park on this research for the past two years. We have a publication we're about to submit in regards to optimizations of gene synthesis and a significant portion of my thesis deals with work done for the upcoming publication. My work includes optimizing the synthesis of large gene constructs, the synthesis of new hyper-thermophilic MutS proteins, characterizing these proteins using instruments such as the circular dichroism spectrophotometer and the Evotec MF20, as well as perfecting old error correction protocols while designing several new ones.

Thesis Supervisor: Joseph Jacobson
Title: Associate Professor of Media Arts and Sciences and Mechanical Engineering

I. Introduction

Basics of Gene Fabrication

The isolation and manipulation of genes and other large DNA molecules plays a critical role in many areas of molecular biology research. Unfortunately the current expense, time and labor of generating and modifying such constructs is substantial. An alternative is direct synthesis of a desired gene. But while the cost of synthetic oligonucleotides has approached the point (~$0.10 per base) where most laboratories simply purchase these, synthetic genes are still expensive, with typical costs of $1 to $2 per base (i.e. $1000 to $2000 per 1 kilobase gene) and a turn around time of 2-4 weeks. Regrettably such expenditures greatly limit the number and type of experiments which can be carried out in such areas as the study of gene pathways and de novo protein design, and all but preclude the construction of large gene libraries without extraordinary cost.

We set out to develop a technology platform that would make fast, economical high throughput gene synthesis a reality. Most needed at the inception of this project was a completely general means of DNA error reduction that could easily be implemented into a gene synthesis process. The error rates of current approaches add significantly to the effort expended towards quality control, and make direct synthesis of large DNA constructs especially troublesome. Thus to fabricate a desired piece of DNA larger than a few kilobases, one must typically first employ multiple costly and time-consuming cycles of assembly, cloning, and sequencing of smaller fragments.

The ultimate goal of gene fabrication research is to be able to have readily available and affordable methods and technologies that allow researchers and scientists to fabricate large DNA molecules. Current technologies are only effective for making single genes and can be slow and costly. As genetic engineering moves from a focus on single genes to designing complete biochemical pathways, genetic networks, and more complex systems, lower-cost, higher-throughput gene fabrication technology will become increasingly important. However, reaching this goal in genetic engineering is highly unlikely and even impossible without better error correction/reduction protocols.

Nature provides mechanisms and methods of synthesizing DNA with error rates ranging from 1 error in 10^8 base pairs to 1 error in 10^{10} base pairs. However there are no *in vitro* error correction methods for DNA synthesis with even remotely close fidelity. The lowest published error rates thus far have been 1 error in 10^5 base pairs. So there is still much improvement that can be made to current error correction protocols.

There are many applications of good gene fabrication technology ranging from synthesis of genes with novel functionalities that do not even exist in nature to synthesis of multiple genes at once. Also, the ability to synthesize a gene *de novo* eliminates the need to obtain an organism from its natural habitat in order to study one of its genes.

The Jacobson group of the MIT Media Lab has made some progress in the area of error correction by employing a DNA mismatch-binding protein, MutS (from *Thermus aquaticus*) to remove failure products from synthetic genes (Carr et al, 2004). I have been working with the Jacobson group for the past two years on optimizing the synthesis of large gene constructs, the synthesis of new hyper-thermophilic MutS proteins, characterizing these proteins using the circular dichroism spectrophotometer and the Evotec MF20, as well as designing more error correction protocols.

Current Methods of Gene Fabrication

Currently there are several protocols for *in vitro* synthesis of DNA in gene fabrication. The most prominent among these – and the protocol that we use – involves the use of the Polymerase Chain Reaction (PCR) (Park, 2005). The components of the reaction include: thermostable DNA polymerase, dNTPs, PCR buffer (including salts), a pool of oligonucleotides that together make up the entire sequence of the target DNA, and sometimes, oligonucleotide primers that define the ends of the target DNA (Park, 2005). The oligonucleotides in the starting pool are built up into longer constructs through successive rounds of PCR. The full-length constructs are amplified exponentially as in traditional PCR by the end oligonucleotide primers.

There are several methods to perform PCR for gene fabrication. For example, gene synthesis can be performed in a one-step process or two-step process (Park, 2005). In a one-step process, both the oligonucleotide pool and a high concentration of the end primers are included in the PCR. Thus, there is at once the linear build-up of the full-length product from the oligonucleotide pool and the end primers as well as exponential amplification of the full-length product once the first copy is synthesized. Though in general the one-step process requires much more fine-tuning of parameters than the two-step process and is difficult to do for longer gene products, it has the advantage of being quicker and reducing the amount of sample handling. For one thing, reducing the number of sample handling steps will be increasingly important as error correction protocols are ported from the lab bench to automatable microfluidic devices. In a two-step process, the first PCR assembles the oligonucleotide in the oligonucleotide pool together into the full-length product at some low frequency. The second PCR includes some of the first PCR product and adds primers that correspond to the ends of the full-length product sequence. This serves to exponentially amplify the full-length product sequence like a traditional PCR.

Optimization of Gene Fabrication

When a small piece of single-stranded DNA, (an oligonucleotide, "oligo") is desired, the process for obtaining it is as simple as filling out an online order form, waiting 1-2 days, and paying perhaps 10-20 U.S. dollars (depending on the oligo length). It is notable that roughly half of both the time and cost to receive a single oligo can be for shipping, not producing the molecules. This was not true as recently as 5-10 years ago, but costs have come down dramatically. A similar metric should be possible for obtaining pieces of DNA the lengths of single genes or longer: no more than a few days and a few tens of dollars per gene, producing virtually any sequence defined by the user. When this benchmark is achieved, many basic DNA manipulations now performed in the laboratory (for example mutagenesis, cloning, purifications) will often be replaced by simply ordering the exact DNA species desired.

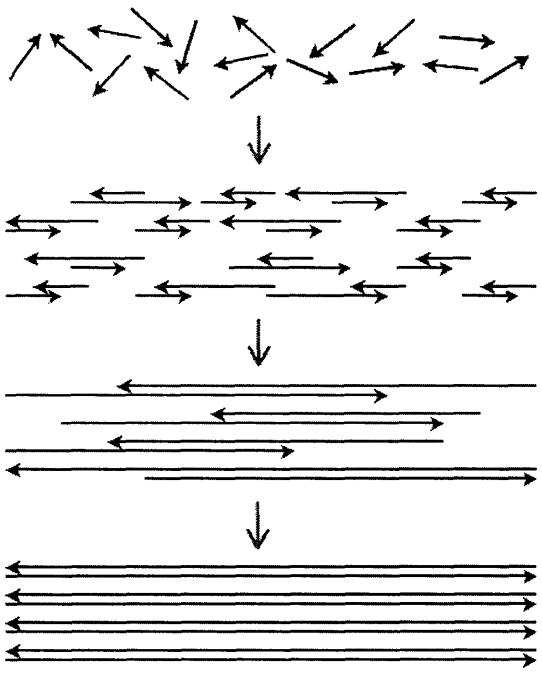


Figure 1. Schematic for gene synthesis by Polymerase Construction and Amplification. A pool of oligonucleotides is synthesized, representing the complete sequence of a desired gene. During ultiple rounds of oligo annealing and extension by DNA polymerase, each oligo acts as both primer and template, generating successively longer DNA assemblies, until the full length gene is produced. The pool of heterogeneous DNA products is enriched for the full length species by amplification via polymerase chain reaction (PCR, final step).

Gene synthesis companies have proliferated in the past five years, indicative of strong demand for synthetic genes. These companies have made substantial progress in reducing the costs and times involved in production, to where one may now expect to wait two to four weeks for a gene, and pay $1.00-$1.60 US dollars per base pair (bp), averaging roughly $1300 for 1000 bp. "Do it yourself" gene synthesis at the research bench is also feasible, and increasing in popularity (see Figure 1 for the most common technique). For a lab already equipped with conventional molecular biology equipment, the principal costs are the oligos—both strands of a 1000 bp gene cost $0.16-$0.20 per base of oligos (list price Invitrogen, IDT, if synthesized in high throughput format) i.e. $300 or more, a modest amount of labor, cloning reagents, and sequencing services. Figure 1 illustrates the most commonly employed general approach for gene synthesis, though many variants exist.

Strong demand for easy access to genes of interest is also evidenced by the emergence of large clone collections, both commercial (e.g. Invitrogen Ultimate ORF Clones, currently $765 each, or the OriGene TrueClone Collection, $195-$995 per clone) and nonprofit (e.g. the Mammalian Genome Collection, currently $89 per clone). These collections begin to address a tremendous need by making commercially available known genes found in living systems. However, synthetic genes have the capacity to reach far beyond these collections by providing the flexibility of user customization. Furthermore, low cost synthetic DNA on-demand would obviate the need for large centralized physical archives.

For a single gene of interest, the above costs fit well within many research budgets. But there are several research applications which would benefit from large sets of genes, such as studying interactions between every member of a particular functional class, many mutants of a single gene (for example, alanine scanning mutagenesis), a protein design project which requires many revisions of a gene, or labeling all the genes from a single genome (say, 4,000 to 30,000 genes) with an antigenic peptide tag or fluorescent protein for detection. If the synthetic capacity for generating such large sets of genes were available at low cost, it would be heavily utilized.

In addition to dramatically accelerating existing areas of experimentation, high-throughput DNA synthesis will enable new kinds of research and design projects. Already, those

working in the field dubbed Synthetic Biology are designing increasingly larger and more complex artificial genetic systems, built mostly from modified genes found in nature (Elowitz, 2000). Notable examples of completely non-biological DNA designs are beginning to emerge as well (Shih, 2004). The DNA constructs required for some proposed projects are so large that conventional gene synthesis becomes prohibitively expensive. These include complex *in vitro* genetics systems such as that proposed by Church and colleagues (Tian, 2004). Other groups are pursuing the direct synthesis of entire simple genomes (Smith, 2003). Such projects aim both to test hypotheses on the fundamental requirements for life, and to generate organisms that have been dramatically re-engineered for new purposes, such as waste processing, energy production, and complex syntheses of useful compounds.

Basic impediments to drastically reducing the time and costs of gene synthesis include 1) costs of raw materials, especially oligonucleotides or their precursors and enzymes; 2) the extensive sample handling required; 3) quality control (particularly DNA sequencing) mandated by the intrinsic error rates of the synthesis process (both of oligonucleotide synthesis and the assembly of oligonucleotides into genes).

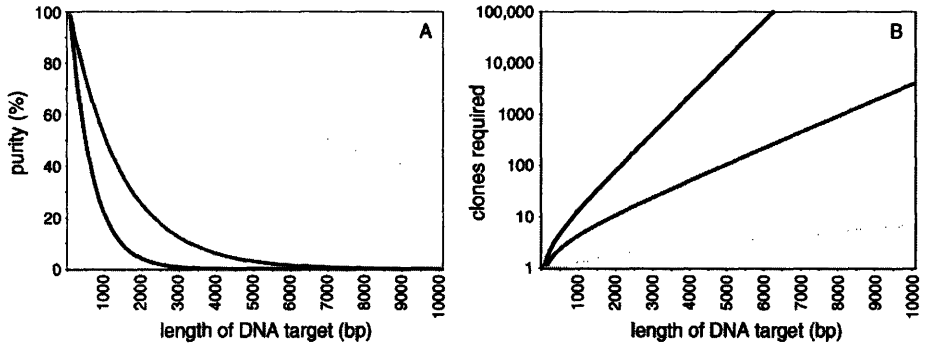


Figure 2. Influence of error rates on de novo DNA synthesis. A. For a given error rate, the purity of gene synthesis products (yield of error-free clones) decreases exponentially with the length of the product synthesized. Error rates shown are 1 in 600 bp (blue, lower curve) typical of conventional gene synthesis approaches, 1 in 1400 bp} (red, middle curve), and 1 in 10,000 bp (yellow, upper curve) achieved with our methods described below. B. The number of clones which must be sequenced to have a high (95%) probability of obtaining at least one which is error-free. The same three error rates as in (A) are indicated (lower, yellow curve is 1/10,000, etc)

To reduce materials costs, oligonucleotide microarrays show tremendous promise. In such arrays, large numbers of distinct oligos are synthesized in parallel, *in situ*—directly on a surface, typically a glass microscope slide. With spot sizes on the order of one hundred microns or less, these arrays can contain tens of thousands or hundreds of thousands of specific sequences. With the cost of an array ranging from a few hundred to a few thousand dollars, individual oligos on the array can effectively cost between $1x10^{-5}$ and $1x10^{-3}$ dollars per base pair—orders of magnitude less than conventionally synthesized oligos. Thus, the contribution of the cost of oligonucleotides to the overall cost of gene synthesis could be reduced to an almost insignificant amount. Progress in taking advantage of microarrays for gene synthesis has

recently been reported, cleaving oligos from the array surface, amplifying them via polymerase chain reaction (PCR) and assembling them into genes in test tubes.

The other key to the above challenges lies with controlling the errors implicit in gene synthesis. These errors begin with mistakes occurring during synthesis of the oligonucleotide precursors, some of which are propagated into the final product. Additional errors can occur during assembly and amplification (Figure 1), such as those introduced by DNA polymerase. Even after a synthetic gene is cloned, errors in biological replication can occur, though these rates are expected to be comparatively low relative to the other sources of error. The error rates of biological replication in fact set the standard to which gene synthesis technology should aspire to.

The consequences of errors in current gene synthesis approaches can be dramatic, especially when one desires genes greater than 1 kb in length. Figure 2A demonstrates the expected purity of a synthetic gene preparation with rates ranging from 1 error in 600 bp (fairly standard in our gene syntheses without any special error reduction, but equal or better than most reports in the literature) to 1 in 10,000 bp. For large constructs, at a conventional error rate of 1/600 very little of the mixture contains the correct product. A typical do-it-yourself user must choose a number of clones (colonies) to grow, isolate DNA from, and sequence—the number of isolates one must characterize rapidly increases beyond a practical range for greater than 1 kb (see Figure 2B).

Recent reports from four separate teams led by Venter, Cello, Church, and Santi have demonstrated excellent advances in the capacity to fabricate large molecules of DNA from oligonucleotides, with products 5.4, 7.5, 15 and 32 kb in length, respectively (Smith, 2003; Cello, 2002; Tian, 2004; Kodumal, 2004). In doing so they have showcased some of the ambitious research goals which can take advantage of this synthetic capacity. Yet despite their successes, they also underscore a severe need for DNA error reduction. Both the Venter and Church teams found it necessary to purify their oligonucleotides prior to assembly. In the former case this was a gel-based size separation needed prior to ligation (2 days added to the process). In the latter case a series of selective hybridization selections (1.25 days) was crucial to providing synthetic DNA of acceptable quality, as the oligonucleotides were synthesized in situ on microarrays, and were of lower purity than oligonucleotides from a conventional vendor. However, this procedure required the synthesis of an additional 200% more oligonucleotides to be used as the selective agents. Each of these teams was able to produce a DNA construct of fairly high quality, but in three cases only by first cloning and sequencing intermediate segments followed by later steps of assembly, cloning, and sequencing the full length product. In the remaining report, functional clones were achieved by natural selection in a manner reminiscent of Stemmer et al.: the full length product was a bacteriophage genome—most failure products could thus not replicate in a bacterial host. Even with this advantage (an estimated 20,000x enrichment over the 'fatal' failures), a substantial number of non-lethal mutations were present in the final sequenced products. This latter approach, while potent, will not generally be applicable to the majority of desired gene synthesis targets, which will not encode reproducing systems.

To optimize among the many choices required in gene synthesis, we first made a careful survey of the literature, chose initial protocols which showed the most promise for high throughput gene synthesis, and set about testing and optimizing these procedures.

The first substantial choice is whether the initial gene construction is performed with 1) DNA polymerase (i.e. PCA, polymerase construction and amplification, a general version is shown in Figure 1; also called polymerase cycling assembly) or 2) DNA ligase. In both cases, a

second step is required after the first assembly in order to amplify the desired full length gene, typically by standard PCR. However, recent work by Tian et al. demonstrated that with PCA, one may combine the two steps into a single reaction, where assembly predominates in the early cycles, and amplification dominates the late cycles (see Figure 3 for our simulations of this combined process). Consolidation of assembly and amplification is particularly attractive in the context of minimizing processing steps for high throughput production. This would not be possible with a ligase-based protocol, as polymerase extension products would frequently terminate at the wrong 3' position of the strand for effective ligation to other oligos.

Purification of oligos prior to assembly is certainly one route towards minimizing errors in gene synthesis. It can be argued that the amount of effort required for this type of error reduction is commensurate with those we discuss below. However there are important differences that require an understanding of the errors implicit in gene synthesis. The prominent error in conventional oligo synthesis is a single base deletion, caused by a failure to couple a particular base to the growing oligonucleotide chain. Most of these errors are "caught" by capping the end of the chain with an acetyl group, resulting in a termination of chain growth, i.e. a 5' truncation of the oligo. This is actually a very acceptable consequence for gene synthesis, as the truncation represents a small amount of missing information, which is supplied by other full length oligos in the assembly mixture. Assembly by polymerase or ligation approaches typically yields a mixture of assembly products anyway, with incomplete strands effectively diluted out by subsequent PCR amplification. Thus, while stepwise yields for oligo synthesis are typically reported around 99% (i.e. a 1% failure rate per coupling step, a highly optimized organic synthesis process) the majority of these errors are not propagated and the final synthetic genes can have per base error rates of 1 in 500 or better using *unpurified* oligos. We have observed this effect in our own gene syntheses, and similar results have been reported by others. Of greater concern for gene synthesis errors are internal deletions (most often a single base deletion) which can be the result of a failure to deprotect or combined failure to couple and failure to acetyl cap. These single base deletions (length n-1, where n is typically 40 to 50 for gene synthesis) are much harder to purify away from the full length product than the capped, truncated species described above. Also, bases damaged from exposure to the chemical conditions of oligo synthesis (e.g. depurination) can also give rise to errors in synthetic genes. Thus, optimization of oligo synthesis chemistry for the specific purpose of gene synthesis seems likely to yield more promising results than oligo purification.

Oligonucleotides can be purified collectively or individually, by high performance liquid chromatography (HPLC) or polyacrylamide gel electrophoresis (PAGE). Individual purification of each oligo (typically 20 or more) becomes impractical regardless of the method, and HPLC will generally give poor resolution if samples are pooled. Both HPLC and PAGE are more effort intensive than desired in an efficient, economical process. Thus, a purchased oligo that might normally cost $10 becomes a $50 expense if extra purification is desired (IDT, current pricing for 100 nmole scale synthesis).

Finally, all pre-assembly purification steps will not eliminate sufficient errors from the final product, since errors can be introduced during the assembly and amplification processes, such as from dNTP misincorporation by DNA polymerases. A pre-purification step may perform well in synergy with error reduction later in the process, but a sufficiently effective late stage treatment may make earlier purifications unnecessary.

The nature of the gene parse is also impacted by the choice of assembly protocol. The desired full length sequence is used to determine which oligos must be synthesized in order to

build the gene. 'Naive' parses have had some notable successes, simply chopping the sequence up into overlapping 40-mers or 50-mers (Stemmer, 1995). However more advanced software is now available to parse sequences with such considerations as normalizing melting temperatures, optimizing codon usage, avoiding undesirable oligo properties (such as hairpin and primer-dimer formation), and avoiding mispriming events that can lead to aberrations in the assembled gene.

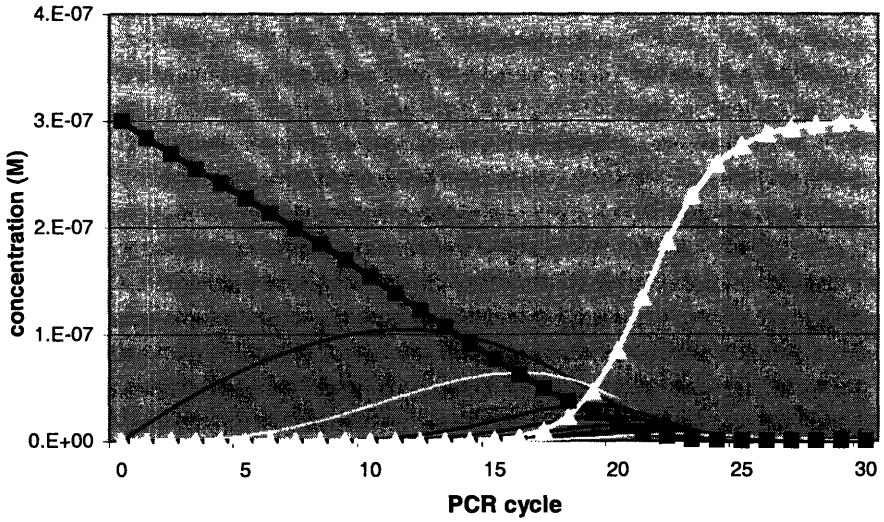


Figure 3. Simulation of a combined gene assembly and amplification reaction employing 20 oligonucleotides. Blue squares indicate the concentration of the amplifying oligonucleotide primer. Yellow triangles show the concentration of final full length product. Other curves represent intermediates in the assembly process. (Values are comparable to those observed in our gene assembly experiments.)

Thus far we have synthesized roughly two dozen genes of use to us in various research projects, and assisted colleagues in the synthesis of dozens more. These have ranged in size from 300 bp to 3.2 kb. More important for our purposes than the number or sizes of genes we make is the ability to easily evaluate success or failure of a given approach to gene synthesis or error reduction. Two particular targets we have synthesized repeatedly in different ways are gene constructs with straightforward readout: LacZ (blue/white screening, i.e. colony counts) and Green Fluorescent Protein (GFP, assessed by colony counts and flow cytometry, see Figures 7 and 13 respectively). These have been extremely useful to us in comparing a wide variety of experimental variables. In many cases these have been used to decide which datasets are worth examining further by DNA sequencing (>300,000 bp sequenced thus far).

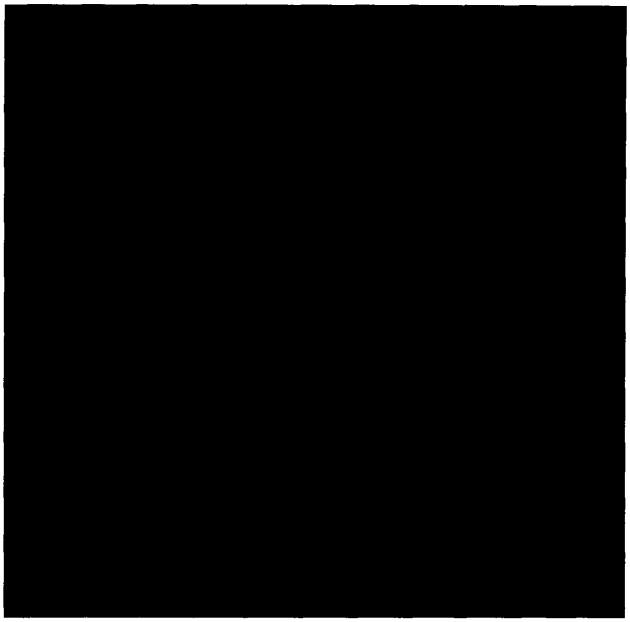


Figure 4. Error removal for green fluorescent protein (GFP) gene synthesis. Two different petrie dishes are shown (half of each). Left: E. coli cells expressing genes which have been treated by the method of Figure 12 to remove errors (>90% fluorescent, brighter colonies). Right: cells expressing the flawed genes extracted in the same process (<10% fluorescent). Not shown: untreated cells (~40% fluorescent). Images aqcuired in grayscale, and color-enhanced (all images enhanced with exactly the same parameters).

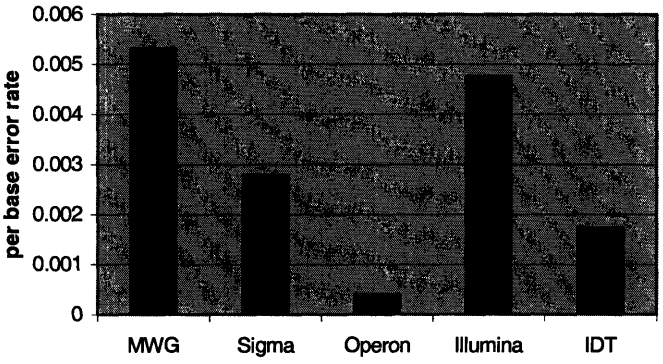


Figure 5. Choice of oligonucleotide vendor has a substantial effect on the error rate of gene synthesis. 50-mer oligonucleotides from each vendor were assembled into genes capable of expressing GFP. Clones were assayed by flow cytometry.

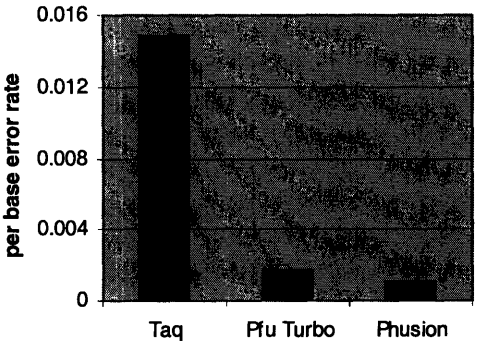


Figure 6. Choice of a high-fidelity proofreading polymerase (such as Pfu Turbo or Phusion) is important in for reducing gene synthesis error rates. Data acquired from synthesis of a GFP gene construct from a pool of 50-mer oligonucleotides, employing various PCR polymerases, and assayed by flow cytometry.

Our most common basis set for these experiments has been a pool of forty 50-mer oligonucleotides used to assemble the GFP construct, using a two PCR assembly reaction, with the high-fidelity polymerase Pfu Turbo Hotstart (Stratagene). The parse for this set is 'naive' in that the sequences have not been optimized in any way, despite known possible sites for misassembly in our construct. Thus errors that may be due to the parse can be revealed, and we can observe if our error reduction protocols are able to handle these errors. (We have detected a 50 bp deletion caused by a mispriming event, recurring at low frequency.) The construct is cloned using the Gateway cloning system (Invitrogen), with no cloning errors observed in over 200 clones sequenced. Error rates for this basis set were measured by sequencing to be 1.8×10^{-3} per base pair synthesized. Flow cytometry analysis of this set leaves us with a dynamic range of measurement for error rates which are both higher and lower than this figure. Even our evaluation of simple choices for gene synthesis has been informative. Among other parameters, we have examined the influence of oligo vendor (Figure 5), DNA polymerase (Figure 6), and oligo length (Figure 7).

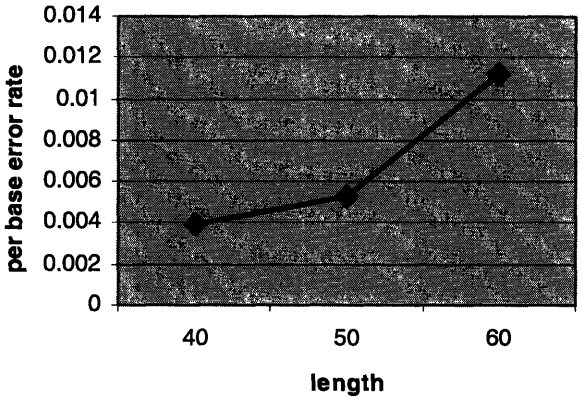


Figure 7. Effect of oligonucleotide length on gene synthesis error rate. Oligonucleotides of various lengths were purchased from MWG-Biotech and assembled into gene constructs capable of expressing GFP. Clones were assayed by flow cytometry.

We were surprised by the large impact of oligo vendor. Each of these reputable companies produces oligos very satisfactory for most applications, and can provide standard measures of quality control (e.g. mass spectroscopy, analytical HPLC). However, since the errors that survive PCA gene synthesis are only a minority of the total, these lower frequency errors are not readily apparent, nor are they a concern to most users. This observation further underscores the potential for optimization of oligo synthesis chemistry to improve the quality of synthetic genes. Oligo length was expected to have some impact, as longer oligos have been subjected to more cycles of harsh chemical exposure. Thus in addition to the stepwise yield reported per base coupling, an additional source of error seems a function of overall exposure. This influence is important to grasp: many users otherwise prefer to use longer oligos when possible, leaving longer single stranded gaps in the overall gene parse as a way of saving on costs of oligo synthesis. Third, the consequence of PCR polymerase is dramatic—a high-fidelity proofreading polymerase is essential. Taq polymerase, with a known high error rate for DNA synthesis, was expected to perform poorly. Still many users have attempted to build genes with this enzyme, or with blends of Taq and a proofreading polymerase, with only a moderate improvement in fidelity over Taq polymerase by itself. We expected a larger contribution from oligo synthesis errors would mostly mask observable differences between a high-fidelity polymerase (Pfu Turbo, Stratagene) and one of the new "ultra high-fidelity" enzymes (such as Phusion from Finnzymes, Pfu UltraII from Stratagene, or Pfx50 from Invitrogen). Still, we found a roughly two-fold difference in synthetic gene errors between these two categories, indicating that the polymerases may be contributing more errors than initially expected. However, in this case our interpretation of the flow cytometry data will need to be followed up with further sequencing, as the proportion of silent errors could be different between these polymerases. Despite their impact, we have not seen any of the above factors reported substantially in the literature.

Error Correction for Gene Fabrication

The optimizations from the above gene fabrication may be able to reduce the error rates in gene synthesis, but we do not expect them to yield error rates lower than perhaps 0.0005 per base. Especially in the cases of assembling long DNA constructs, or of reliably assembling large numbers of high-fidelity genes in parallel, further error reduction will be required (See Figure 2). What are the properties of our ideal error reduction system? It should be:

1. Capable of handling all major types of errors
2. Rapid to perform
3. Need only be performed once (or be easy to iterate)
4. Simple to implement into an automated system

The following error correction involving the use of the mismatch recognition protein MutS was chosen as our most likely candidate for attaining a broadly applicable error reduction system by Dr. Peter Carr and Jason Park several years ago. Initial studies focused on criteria 1, an essential step before optimizing criteria 2-4. Commercial sources were available for MutS from *E. coli* ("Eco MutS," USB) and *T. aquaticus* ("Taq MutS," Epicentre). Principal concerns were the specificity of MutS proteins, specifically a relatively high degree of nonspecific DNA binding (i.e. in the absence of mismatches) and poor binding to CC mismatches in the case of the *E. coli* protein (Brown 354). The Taq MutS protein had been previously observed to not have CC binding deficiency. Since then, I have helped work on improving gel filtration error correction through the use of higher fidelity MutS proteins as well as optimizing incubation times of the MutS protein with the error-containing DNA. I will first describe the gel filtration error correction and then move onto a discussion of the new proteins and new error correction methods we have developed since the original gel filtration paper published by Carr et al. in 2004.

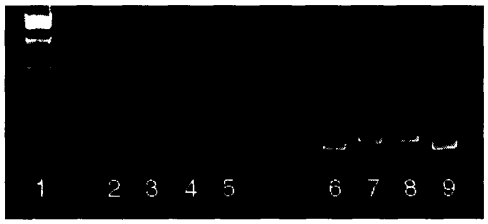


Figure 8. Polyacrylamide gel electrophoresis (PAGE) of DNA segments used to assemble the GFP gene construct. Lane 1: size standard (kb DNA Ladder, Stratagene; from bottom, sizes are 250, 500, 750, and 1000 bp). Lanes 2-5: the four segments, each complexed with MutS. Lower bands are the error-depleted fractions; upper bands are the error-enriched (MutS-bound) fractions. Lanes 6-9: the same four segments, with no MutS present. Some smearing of the DNA is consistently observed between the two bands in all lanes containing MutS, likely representing protein-DNA complexes which have dissociated.

Both proteins ("Taq MutS" and "Eco MutS") were tested in simple gel mobility shift assays as employed in the literature, assaying binding to different types of single base mismatches constructed from oligonucleotide duplexes. Taq MutS proved more reliable in these studies, though initial binding tests with some types of mismatches appeared discouraging. We demonstrated that Taq MutS was capable of binding well to a single base deletion mismatch, and separating these complexes from those without mismatches. In a small sample, cloning and sequencing of these constructs confirmed that the mismatches had been removed (10 clones analyzed, no mismatches present from an initial 50/50 mixture).

Given that single base deletions are the dominant error we have observed in gene synthesis, we used this procedure to separate mismatches from a pool of synthetic fragments used to build our ~1kb GFP gene construct (see Figure 9). Early attempts to perform this procedure with the full length construct gave results which proved difficult to reproduce, thus smaller fragments were used. (Note: this was indeed a length-dependent effect. At the same per base error rate, longer DNA is more likely to contain a mismatch and thus bind MutS, leaving behind a smaller unbound fraction.

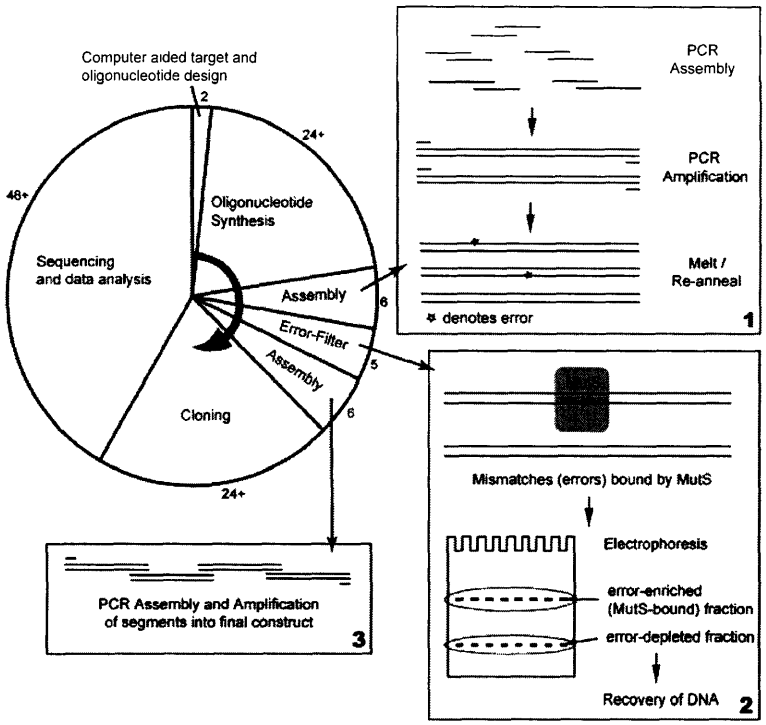


Figure 9. Construction of synthetic genes employing MutS protein for error-reduction. The pie chart indicates the approximate amount of time consumed by each step (in hours), with a red arrow indicating the order of operations. The most time consuming steps in this process are often oligonucleotide synthesis and DNA sequencing (including plasmid production). The 24+ and 48+ hours indicated for each of these represent lower bounds on these processes, possible if performed with immediate access to the appropriate equipment. If these steps are performed by outside providers, 3-5 days are typical of each step. Box 1: gene segments are synthesized and amplified using conventional PCR protocols. The resulting products are dissociated and re-annealed so that errors are present as DNA heteroduplexes (mismatches). Box 2: MutS protein is mixed with this pool of molecules and binds to mismatches. The error-enriched (MutS-bound) fraction is resolved from the error-depleted fraction by electrophoresis. Box 3: The error-depleted segments are assembled into the desired gene and amplified by PCR prior to cloning.

We have since optimized this procedure so that it works quite well with 1 kb genes). The overall process used for error removal and assembly is diagrammed in Figure 9. Competent cells were transformed with vector (pDONR221) carrying the synthetic GFP genes, and both plated (for colony counts, Figure 4), and analyzed by flow cytometry. Various sets of data were analyzed from this procedure, including DNA that was put through the error removal process a second time (this time as the assembled 1 kb construct) and products built from the fragments that were specifically bound by MutS (i.e. an error-enriched pool).

Approximately 40 clones were sequenced from each of four sets: error-enriched, untreated, error-depleted, and error-depleted twice. A strong trend was evident for error removal by this MutS procedure, including a highly error-enriched pool from the MutS-bound fraction. All categories of errors first observed in the untreated pool were substantially diminished with each cycle of error reduction.

The first cycle of error removal reduced the overall error rate 6.9-fold, to 0.00026 per base. The following cycle reduced this figure a further 2.6-fold, to 0.00010 per base. No other approach has achieved such an effective result applicable to general gene synthesis without first cloning and sequencing fragments, followed by subsequent assembly steps and later rounds of cloning and sequencing. To demonstrate the

usefulness of this method, we applied a single cycle of error depletion to the construction of our own Taq MutS gene (2.5kb), sent a single clone for sequencing, and found it to contain the correct product (an estimated 50% likelihood). In the absence of error removal, we would expect to sequence ~100 clones to reliably achieve this result (Figure 2). We now express and purify our own supplies of Taq MutS protein, which we have used in our technique to produce other genes of comparable size. For these constructs, we typically sequence 3 clones and find one or more to be correct.

II. Current and Future Work

New Proteins Synthesized for potential use in Error Correction

In the interests of devising improved methods for DNA error reduction, we have designed and synthesized the genes for several new proteins, some of which have been only minimally studied in the literature. These include new species of MutS, intended for optimization of our current methods. Others possess new functions. Proteins that have a potential role in mismatch biding or mismatch cleavage are of particular interest. Often initial reports exist where such proteins have been characterized for a related purpose: mismatch detection for analysis of genetic variation. In principle, functions which detect mismatches also have potential for error elimination as well. An important difference, however, is that for detection often only a small fraction of a sample need be bound or cleaved, whereas for error elimination the entire sample must be processed thoroughly. Thus if a putative mismatch nuclease cleaves very inefficiently, it may still serve well for mutation detection but not overall error correction.

The likelihood of any new protein being superior to MutS for error reduction is difficult to assess. Certainly, the effort to obtain these proteins carries with it some risk. However, it is precisely our capacity to synthesize genes quickly and easily that makes the cost of taking such risks relatively small. Some of them have truly been "genes of convenience" i.e. when we have required a new or different gene to test an assembly method, we have selected one from our wish list. Protein production is typically with established T7 expression systems in *E. coli*. Purification is generally aided by protease-cleavable polyhistidine tags and by the thermostable properties many of these proteins possess. Table I presents many of the proteins we have produced thus far, several of which are in the early testing phase or are about to be tested.

Table I. Proteins produced for error reduction. Those protein features which have gone through at least an initial screen for function have that feature or use shown in bold face.

Protein	Source Organism	Modifications	Expressed?	Purified?	Thermostability		Putative Uses
					predicted	measured	
MutS-FokIN	*E. coli F.okeanokoites*	his-tag, protein fusion	Y	insoluble	no		mismatch endonuclease
MutS	*T. aquaticus*	his-tag	Y	Y	>60 C	75 C	MutS gel, columns
MutS	*T. maritima*	his-tag, linker	Y	Y	>75 C	82 C	MutS gel, columns
MutS(K620M)	*T. maritima*	his-tag, linker, -ATPase	Y	Y	>75 C		error-correcting PCR

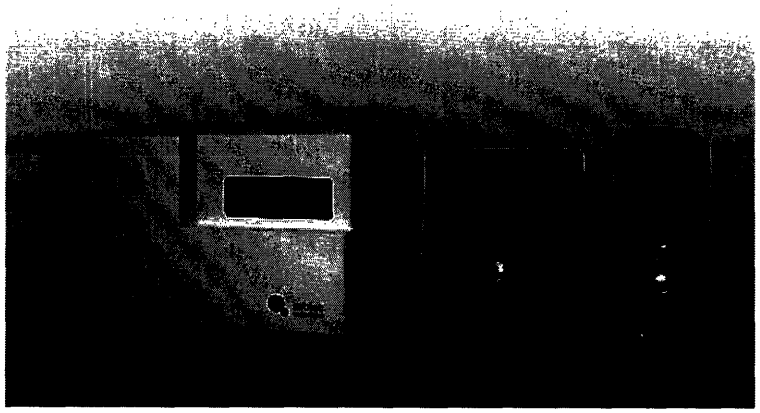
MutS	*A. aeolicus*	his-tag linker	Y	Y	>90 C	>95 C	MutS gel, columns
MutS(K620M)	*A. aeolicus*	his-tag, linker, -ATPase	Y	Y	>90 C		error-correcting PCR
Hjc	*SIRV-1*	his-tag	Y	in progress	>70 C		mismatch endonuclease
EndoV	*T. Thermophilus*	his-tag	Y	Y	>60 C	>60 C	mismatch endonuclease
SSB	*S. solfataricus*	his-tag	Y	Y	>70 C		mismatch binding

Tools to Characterize our Proteins

In attempting to characterize all of our proteins used in our research, we have identified two instruments for use. We are currently finishing up the characterization of all of our proteins through the employment of both of these measurement tools and are in the final stages of our data analysis. We plan to publish our results soon of our analysis of the various proteins we have built and used in our error correction protocols over the years.

Circular dichroism (CD) spectroscopy measures differences in the absorption of left-handed polarized light versus right-handed polarized light which arise due to structural asymmetry. The absence of regular structure results in zero CD intensity, while an ordered structure results in a spectrum which can contain both positive and negative signals (APL). We are using this tool to test the DNA and MutS interaction through various temperatures (ranging from room temperature to 95 degrees Celsius) and for various wavelengths (from 220nm to 320nm).

Figure 10. The Olympus Evotech MF20. The only system in the United States is in the Zhang Lab at 500 Technology Square. Professor Jacobson's Biology Lab work is all done in the Zhang Lab space.



The Olympus Evotec MF20 system is a useful tool that employs a technique known as single molecular fluorescence spectroscopy that does not require the sample to be in a solid

phase. Instead, interactions between biological molecules can be studied directly in buffer solution. High-speed analysis yields data on functional biomolecular interactions and binding within a short period of time ranging from a few seconds to less than a minute. A combination of confocal laser optical system with a high-sensitivity fluorescence detection device permits measurements in an extremely small volume of around 1 femtoliter (1 fL = 1 x 10^{-15} L), thereby capturing any interactions at the level of single molecules. This set-up operates at low noise and detects fluorescent signals with extremely high sensitivity. In addition, since measurements are performed in solution, damage to the sample is minimal and artifacts from unspecific surface interaction are avoided. This allows the same sample to be rapidly put through other tests subsequently. We have designed an oligonucleotide sequence with a fluorescent tag on one end which can be picked up by the MF20.

Through our characterization and tests, we have found one of our new strains of MutS (Aae) to give us higher fidelity in error correction and to work better over a wider range of temperatures.

Methods in Error Correction

MutS has been used as our primary tool for error correction. We have built and tested several different strains of the MutS protein. Below, I will elaborate on some protocols proposed by Jason Park (MIT '05) in his thesis since which we have done a lot more work on.

MutS pull-down – As described in Jason Park's 2A thesis, *MutS pull-down* is a protocol in which newly synthesized DNA product (or fragments thereof, in the case of gene targets too large for this procedure) are melted and re-annealed to re-assort errors and create heterodimers of errors as described earlier. In this protocol, the DNA is then exposed to MutS (*T. aquaticus*) under a certain set of reaction conditions (temperature, time, etc.) at which point DNA with errors is selectively bound by the MutS. Figure 11 shows a TBE non-denaturing polyacrylamide gel showing the difference of running DNA alone and DNA with MutS. A shifted band of DNA bound to MutS can be seen, due to the reduced mobility of the complex relative to unbound DNA. After performing a *MutS pull-down* one can excise the error-depleted DNA by using the crush and soak method and then use PCR to amplify this "good DNA." Since the Carr et al. paper was published, I have been helping test the efficacy of our other MutS proteins (the *A. aeolicus* and the *T. thermophilus*). Initial tests have shown the *A. aeolicus* protein to give us better efficacy and than the *T. aquaticus*. The *T. thermophilus* MutS protein had only a slightly better efficacy in error binding than the *T. aquaticus*. In our protein characterization that we plan to submit for review in a month or so we will quantify these differences in protein efficacy.

Figure 11 MutS pull-down filter. Lane 1: kb ladder. Lanes 2,3,4,5: ~300mer pieces of GFP (993bp), treated with MutS. Lanes 6,7,8,9: Same as lanes 2,3,4,5, except without MutS treatment. (From Carr et al., 2004)

MutS-FokI fusion – The major design component of my 2A thesis involves the design of the *MutS-FokI fusion* design. I have worked with Dr. Carr and Jason Park to design and build this fusion protein. MutS (Figure 12) as previously mentioned is a protein that binds to heteroduplexes in DNA. FokI (Figure 13) is a member of the type IIS "outside cutter" class of endonucleases, which bind to a specific DNA sequence but cleave at a remote site (Modrich, 1991). The original goal of the *MutS-FokI fusion* was to have the protein bind to the error and then cleave around the error after which we could separate the two pieces. The original design that Jason Park proposed in his 2A thesis had some major problems. I have been helping to design a new and improved version of *MutS-FokI fusion* composed of fusing the FokI nuclease domain to the N-terminus of the *Aquifex aeolicus* MutS via a flexible and soluble $(Gly\text{-}Ser\text{-}Gly)_n$ linker of variable-length. We have decided to use the ATPase-deficient mutant version of the *Aquifex aeolicus* MutS as it is the most thermostable of the MutS protein variants we have synthesized in our lab.

We have decided to express the fusion protein in a plasmid vector system such as pET-32, pET-41, pET-42, pET-43.1, or pET-44. These vectors fuse a protein such as thioredoxin, NusA, or GST to the protein of interest for enhanced production and solubility. A $(His)_6$ tag, which will be used in protein purification, is also included. These proteins can later be cleaved off at a known site after expression and purification. Improved solubility is an important consideration in building the MutS-FokIN fusion construct, especially since the earlier version of the protein was shown to be insoluble.

We have decided to take a modular approach to building the MutS-FokIN fusion construct in order to facilitate the making of versions of the construct with different length linkers. By minimizing the amount of necessary PCR and sequencing, we can cut down on time and cost expended.

We have also decided to attach flanking sequences to the 5' and 3' ends of the existing MutS and FokIN DNA constructs to insert appropriate restriction sites. These constructs will be ligated sequentially into a plasmid vector after restriction endonuclease digestion. Clones will be screened and sequenced at this point. After verifying that the desired construct has been made successfully, the $(Gly\text{-}Ser\text{-}Gly)_n$ linkers of various lengths – which will have been made with appropriate sticky ends by annealing synthesized oligonucleotides – will be ligated into the new plasmid construct. After one more round of screening and sequencing clones, we will proceed with protein purification and expression.

Now that the design is complete we hope to build and test the *MutS-FokI fusion* protein in the next several weeks.

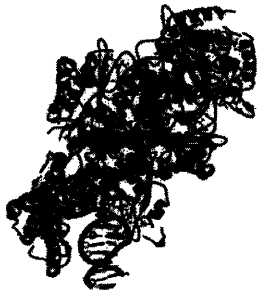


Figure 12 MutS (Sixma, 2001)

Figure 13 FokI nuclease (Wah et al., 1997)

MSPCR – MutS-in-PCR (MSPCR) is a novel idea for error prevention that involves the use of the MutS protein in the gene fabrication PCR reaction that was originally proposed by Jason Park in his 2A thesis. The main idea behind MSPCR is to address errors before they become integrated into synthesized DNA. Originally in his thesis, Jason Park outlines the idea of using thermostable *T. aquaticus* MutS for use in this protocol (Park, 2005). However, initial tests of the *Taq* MutS in PCR did not prove to be very effective. After some tests we found that *Taq* becomes denatured in the low 80 degrees Celsius. This is where our new hyper-thermophilic MutS proteins have been extremely effective and useful. Especially our *A. aeolicus* MutS protein has been effective up to mid 90 degrees Celsius without denaturation. *Aae* also seems to bind better to DNA mismatches than our other MutS proteins.

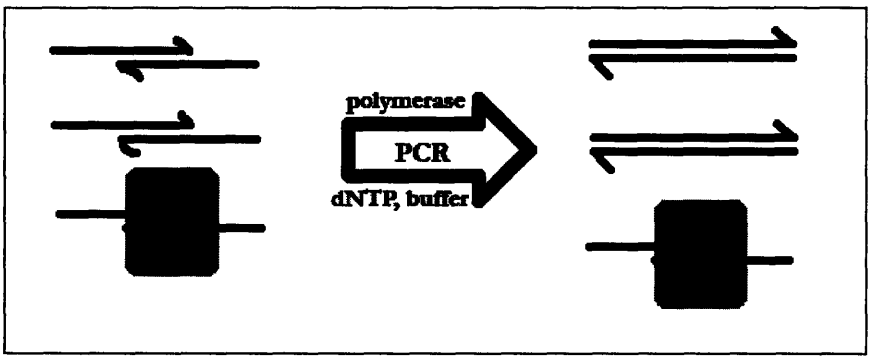


Figure 14 One proposed mechanism of action of MutS in MSPCR - Steric blocking of polymerase for short pieces of DNA and for mismatches near 3' ends (* denotes error)

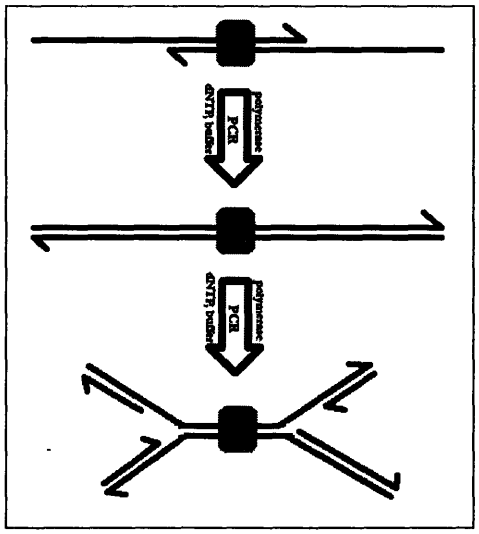


Figure 15 Another proposed mechanism of action of MutS in MSPCR - MutS binds error; polymerase copies everywhere but mismatch (falls off once it reaches MutS)

All of the abovementioned protocols are error correction/reduction/prevention protocols developed as a result of the work of Dr. Peter Carr, Jason Park, myself, and Professor Joseph Jacobson. We will be continuing to work to perfect our current methods as well as to design new protocols in the future.

III. Closing Statement

Continued improvement in error-correction methods is pivotal in furthering the field of biological engineering and more specifically in the field of gene fabrication. A good error correction protocol should be a quick, easy, robust, economical, and effective method by which errors in a gene product synthesized in gene fabrication is reduced (Park, 2005). Some of the protocols and methods mentioned above are effective tools to accomplish error correction/reduction while others are still very much a "work in progress."

In the future, I plan on finishing up the projects that I have mentioned in the "current and future work" section that are yet perfected over the summer and as a Masters Degree student next year. The development of error correction and prevention protocols and methods involves engineering optimization and design, as well as a solid background and understanding of biological processes and systems and my mechanical engineering background has been invaluable in our research.

IV. References

Alliance Protein Laboratories. Circular Dichroism. http://www.ap-lab.com/circular_dichroism.htm

Brown, J., Brown, T., and Fox, K.R., Affinity of mismatch-binding protein MutS for heteroduplexes containing different mismatches. Biochem J, 2001. **354**(Pt 3): p. 627-33.

Carr, P.A., Park J.S., Lee Y.J., Yu T., Zhang, S., Jacobson J.M. (2004) Protein-mediated error correction for *de novo* DNA synthesis. Nucleic Acids Research. Vol. 32 No. 20 e162

Cello, J., Paul, A.V., and Wimmer, E., Chemical synthesis of poliovirus cDNA: generation of infectious virus in the absence of natural template. Science, 2002. **297**(5583): p. 1016-8.

Elowitz, M.B. and Leibler, S., *A synthetic oscillatory network of transcriptional regulators.* Nature, 2000. **403**(6767): p. 335-8.

Kodumal, S.J., Patel, K.G., Reid, R., Menzella, H.G., Welch, M., and Santi, D.V., *Total synthesis of long DNA sequences: synthesis of a contiguous 32-kb polyketide synthase gene cluster.* Proc Natl Acad Sci U S A, 2004. **101**(44): p. 15573-8.

Modrich, P. Mechanisms and biological effects of mismatch repair. (1991) *Annu. Rev. Genet.* 25:229-53.

Park, J.S. 2A Thesis: Error Correction Methods in Gene Fabrication. May 2005. https://web.mit.edu/2.tha/paperwork/fall2004/Jason_Park_Proposal.pdf.

Shih, W.M., Quispe, J.D., and Joyce, G.F., *A 1.7-kilobase single-stranded DNA that folds into a nanoscale octahedron.* Nature, 2004. **427**(6975): p. 618-21.

Sixma, T.K (2001) DNA mismatch repair: MutS structures bound to mismatches. Curr. Op. Struct. Biol. 11:47–52.

Smith, H.O., Hutchison, C.A., 3rd, Pfannkoch, C., and Venter, J.C., *Generating a synthetic genome by whole genome assembly: phiX174 bacteriophage from synthetic oligonucleotides.* Proc Natl Acad Sci U S A, 2003. **100**(26): p. 15440-5.

Tian, J., Gong, H., Sheng, N., Zhou, X., Gulari, E., Gao, X., and Church, G., *Accurate multiplex gene synthesis from programmable DNA microchips.* Nature, 2004. **432**(7020): p. 1050-4.

Wah, D.A. et al. (1997) Structure of themultimodular endonuclease FokI bound to DNA Nature 388:97.