GeneDesign: Rapid, automated design of multikilobase synthetic genes

Sarah M. Richardson, Sarah J. Wheelan, Robert M. Yarrington, et al.

Genome Res. 2006 16: 550-556
Access the most recent version at doi:10.1101/gr.4431306

Resource

GeneDesign: Rapid, automated design of multikilobase synthetic genes

Sarah M. Richardson, Sarah J. Wheelan, Robert M. Yarrington, and Jef D. Boeke[1]

High Throughput Biology Center, The Johns Hopkins University School of Medicine, Baltimore Maryland 21205, USA

Modern molecular biology has brought many new tools to the geneticist as well as an exponentially expanding database of genomes and new genes for study. Of particular use in the analysis of these genes is the synthetic gene, a nucleotide sequence designed to the specifications of the investigator. Typically, synthetic genes encode the same product as the gene of interest, but the synthetic nucleotide sequence for that protein may contain modifications affecting expression or base composition. Other desirable changes typically involve the revision of restriction sites. Designing synthetic genes by hand is a time-consuming and error-prone process that may involve several computer programs. We have developed a tools environment that combines many modules to provide a platform for rapid synthetic gene design for multikilobase sequences. We have used GeneDesign to successfully design a synthetic Tyl element and a large variety of other synthetic sequences. GeneDesign has been implemented as a publicly accessible Web-based resource and can be found at http://slam.bs.jhmi.edu/gd.

The power and flexibility of gene synthesis is increasingly being recognized (Han and Boeke 2004; Neves et al. 2004; Tian et al. 2004; Patterson et al. 2005). Traditional gene synthesis applications include expression of exotic genes in a model organism (Itakura et al. 1977), facilitation of site-directed mutagenesis (Nambiar et al. 1984), structural analysis (Jay et al. 1984), investigation of transcriptional regulation (Krieg et al. 1991), and the generation of de novo proteins (Quinn et al. 1994).

The theory of gene design when high expression levels are desired is relatively uncomplicated. First, the desired protein sequence should be reverse translated into a nucleotide sequence. This step allows codon usage to be optimized for the host organism to be used for expression, or changed completely to accommodate a variety of constraints (Fig. 1). While there are enormous numbers of possible synthetic sequences that can be made, and could in principle lead to increased expression, we have used the highly simplifying method of choosing the single most abundant codon specifying each amino acid in highly expressed genes for the host organism of choice. Codon optimization can be an important factor in establishing gene expression, although generally it is less significant than are promoter strength, position in the genome, etc. Second, the new nucleotide sequence may be analyzed for the strategic introduction and removal of restriction sites (Fig. 2). A useful strategy is to space sites evenly throughout the gene. Both introduction and removal of sites are done without altering the amino acid sequence. Finally, the sequence to be made should be minced into small oligonucleotides for assembly by PCR as described by Stemmer and others (Fig. 3; Stemmer et al. 1995).

While the above theory is indeed relatively uncomplicated, manual design is a complex, tedious, and error-prone process. In the past, researchers used many different programs to address the requirements of the separate steps of synthetic gene design. Alternatively, they sent off their requirements to a black box provided by a gene synthesis company and let it use its proprietary programs to design genes. Today there are two publicly accessible computer programs that perform synthesis-related oligonucleotide design: Gene2Oligo (Rouillard et al. 2004) designs short oligonucleotides for gene synthesis, and DNAWorks (Hoover and Lubkowski 2002) performs reverse translation as well as oligonucleotide design. However there is only one publicly available program, GeMS, that performs all of the major tasks of gene synthesis (Jayaraj et al. 2005). To facilitate the use of synthetic genes in both traditional and high-throughput applications, new and more flexible solutions are required. GeneDesign is a useful tool for investigators who wish to optimize protein expression and/or redesign their gene of interest for detailed structure/function studies (e.g., mutagenesis).

In this article we describe a suite of Web-based programs that is able to perform all of the functions outlined above for gene design in a directed, step-wise manner. It accepts as input either amino acid or nucleotide sequences and allows users to move through the process of design in a series of modules that address practical issues surrounding cloning vector sequences, restriction site placement, and oligonucleotide design. Users can follow the main "design a gene" path or use the modules individually as needed. We have tested this program with the 5.2-kb gag-pol gene on the yeast retrotransposon Ty1 and a 0.6-kb nucleotide fragment of the human retrotransposon L1 that was difficult to break into oligos manually (Han and Boeke 2004). Members of our laboratory and other colleagues have also extensively tested the program, and our group alone has successfully designed ~30 kb of synthetic DNA by using GeneDesign.

Results

Workflow with GeneDesign

GeneDesign consists of six modules that may be used individually or in series to automate the tasks associated with the design and manipulation of synthetic sequences (Fig. 4). The modules are, in typical order of use: reverse translation, codon juggling, silent restriction site insertion, silent restriction site removal, oligo design, and sequence analysis. Although the modular design allows any number of permutations, we anticipate that most users will be interested in the design a gene pathway. For in-

[1]Corresponding author.
E-mail jboeke@jhmi.edu; fax (410) 502-1872.
Article published online ahead of print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/ gr.4431306. Freely available online through the *Genome Research* Open Access option.

16:550–556 ©2006 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/06; www.genome.org

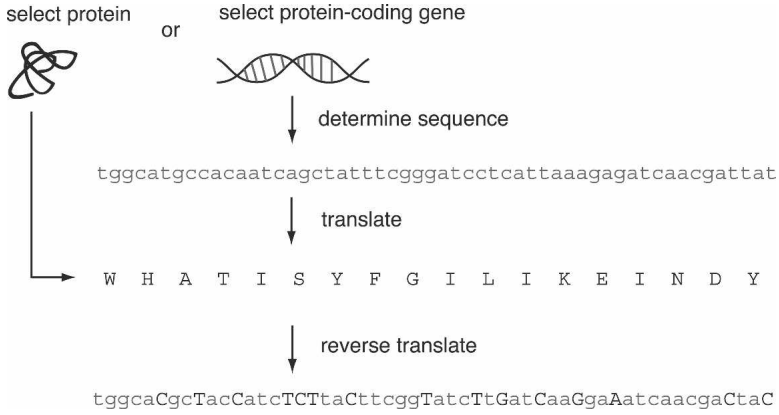

Figure 1. Reverse translation and codon optimization. The flowchart indicates the steps involved in reverse translation of a sample sequence.

stance, an investigator with a 500-amino-acid human gene to be expressed in yeast for modular mutagenesis would use the design a gene path. She would begin with the reverse translation module, yielding a 1500-bp nucleotide sequence that is optimized for expression in yeast. She then takes the sequence to the silent site insertion module, where she is able to define the qualities of the landmark sites to be used for modular mutagenesis and to select their locations. Finally, she takes the sequence to the oligonucleotide design module, which breaks the synthetic sequence into three 500-bp "chunks" (separated by unique restriction sites) and each of those three chunks into ~12 overlapping 60mers for PCR assembly and amplification. Another researcher with a 600-bp nucleotide sequence from yeast that is to be cloned into bacteria would begin with the codon juggling module to optimize the nucleotide sequence for expression in *Escherichia coli*, and then take the sequence to silent site removal to knock out any instances of internal restriction sites that conflict with his choice of cloning vector. Finally, he would use the oligonucleotide design module, which would leave him two 300-bp chunks containing a total of 16 60mers encoding his new synthetic gene.

The sequence analysis module is accessible from all of the other modules and is designed to provide useful information about the nascent synthetic sequence during the design process. A manual describing each module in detail is available online and as a PDF, and the user interface includes guidelines for use as well.

Reverse translation module

The reverse translation module allows a user to convert a protein sequence to nucleotides. Given a target organism, GeneDesign's default is to use the most optimal codon for each amino acid as defined by the highest relative synonymous codon usage (RSCU) value in highly expressed genes in that organism (Sharp et al. 1988). The user has the opportunity to view each codon in an organism's optimal set and make changes. Alternatively, the user may supply any codon table of interest.

Codon juggling module

The codon juggling module offers a choice of several algorithms to alter a protein-coding nucleotide sequence. The optimization algorithm replaces each codon with the optimal codon for expression; the "next most optimal" algorithm replaces each codon with the most optimal codon that is *not* the original codon; and the "most different" algorithm replaces each codon with the most optimal codon that is most dissimilar to the original codon. In this way it is possible to design the synthetic nucleotide sequence most different from the native sequence that encodes the identical protein. We are using this to eliminate the hypothetical possibility of nucleic acid signals imbedded within a protein coding sequence. These three algorithms use the RSCU data from highly expressed genes for determining which codons are more optimal (Sharp et al. 1988). The random algorithm replaces each codon with a randomly chosen codon from the same family that is not the original codon. The most different algorithm uses transversions as often as possible at the wobble position of two-, three-, or four-codon families and at all three positions of six-codon families.

Silent restriction site insertion module

To facilitate modular mutagenesis, the silent site insertion module allows the researcher to populate a synthetic coding sequence

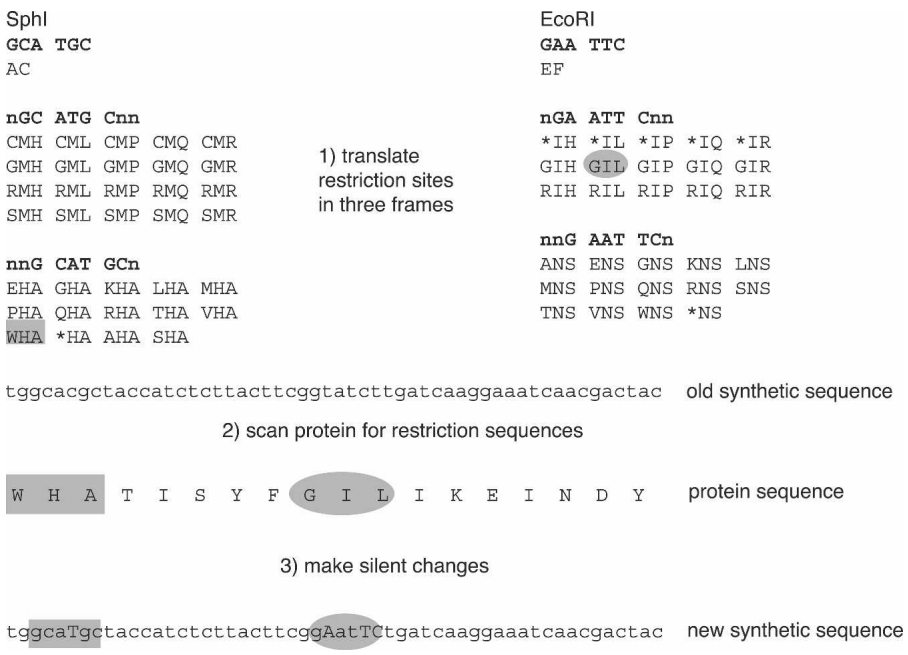

Figure 2. Silent restriction site insertion. The chart indicates the steps involved in inserting new restriction sites into the sample sequence from Figure 1.

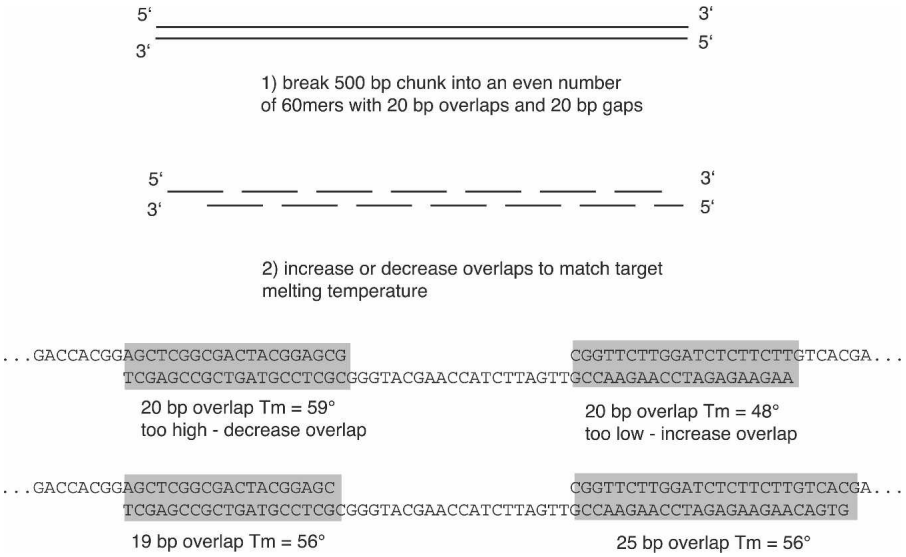

Figure 3. Oligonucleotide design and melting temperature optimization. The general schema of oligonucleotide selection used by GeneDesign.

with unique restriction sites, also called "landmark" restriction sites. In our laboratory, we have found it convenient to separate chunks of 500-bp sequence by these landmark sites, which greatly facilitates assembly of large sequences. We often find that it is interesting to assemble chimeras between native and synthetic sequences if there are biological differences between them to crudely map the source of variation. To do this, we swap segments of synthetic sequence with the corresponding native sequence, exploiting existing unique sites. In order for such "segment swapping" between a native sequence and a synthetic congener to be convenient, the landmark sites must be absent from the vector hosting the synthetic gene; therefore, sites already present in the vector should not be considered for use as landmarks. To avoid site duplication, the user can select specific enzymes for exclusion and consideration, select a vector sequence from a list of commonly used vectors, or provide a vector sequence. If a vector is used, GeneDesign will automatically consider for landmark use only sites absent from that vector. In addition, sites that should not be considered at all may be specified.

Because no changes are ever considered that alter the first-frame amino acid sequence in any way, the encoding of second- and third-frame ORFs is not preserved. GeneDesign will check that the sequence submitted for silent site insertion is a simple coding sequence in the first frame of translation; it is recommended (but not required) that landmark sites be inserted only into ORFs because the effects of inserting restriction sites into noncoding sequences is difficult to predict.

GeneDesign consults a list containing every possible amino acid permutation that could be encoded by each frame of each restriction enzyme recognition sequence, searches the translated nucleotide sequence for these short amino acid sequences, and presents the results as a display of all possible silent site introductions. Sites that are defined as interesting by the user or are absent from the user's vector are presented in red, and all other sites are presented in black. At this point the user may go through the display and prepare a solution manually or give the program an amino acid interval at which sites are desired and have the program select the enzyme sites automatically.

Before entering automatic design mode, the user is given the opportunity to rank enzymes by overhang, recognition site length, recognition sequence, and price. Only enzymes that fit the provided criteria will be considered. By default the program will not consider enzymes that leave blunt ends or single base pair overhangs, as these are more difficult to ligate in the assembly of the synthetic gene.

In automatic design mode, GeneDesign breaks the nucleotide sequence into pieces according the user-defined interval and then ranks each piece by the number of possible restriction site introductions. The chunks with the fewest possible introductions are processed first, and the highest-ranking enzyme present is chosen for a landmark. This enzyme is added to the list of used cutters so that any enzyme with an ambiguous site that could resolve to the same sequence will not be considered for the rest of the sequence. The program processes each piece this way, attempting to space consecutive landmark choices by at least half the interval length to avoid an unnecessary clustering of sites that would only remove more sites from consideration. The solution is presented to the user with the same graphic that was used to list all possible introductions, with the program's landmark selections presented in blue. The user can make changes to the program's choices or have the program re-evaluate the sequence completely.

After a solution is reached that is satisfactory to the user, the sites are processed for introduction. The nucleotide sequence is compared to the sequence needed to introduce each enzyme and changed accordingly, with care taken to preserve the amino acid sequence. Important practical considerations at this step are the length of the segment to be synthesized and the type of vector to be used. The larger the insert and vector sizes, the less sites will be available. Small vectors with very few sites have been described (Mandecki et al. 1990) and should be very useful to minimize this problem.

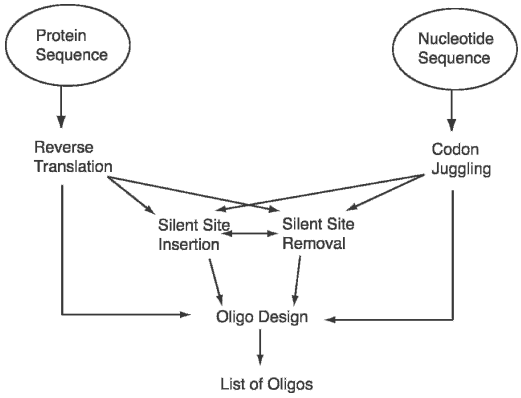

Figure 4. Workflow of GeneDesign. The modules of GeneDesign are designed to smoothly interact with each other. It is also possible to enter any of these modules separately for individual use.

The solution is summarized for the user in the final screen, where he or she has the opportunity to check the properties of the introduced enzymes. If undesirable enzymes have been included, the user is able to select those enzymes and begin the process over again with them automatically added to the list of banned sites.

Silent restriction site removal module

Just as restriction sites can be silently added to a sequence, they can be silently removed. GeneDesign's silent site removal module accepts a nucleotide sequence and parses it for all present restriction sites. It then displays the site names and the number of times they occur, allowing the user to select as many as needed for removal. Because there are many ways to silently destroy a restriction site, the program also allows the host organism to be defined for optimization purposes. When changing codons to break a restriction site, more optimal codons will be used first. If no optimization is selected, all codon changes will be random.

Oligonucleotide design module

The synthetic gene cannot usually be constructed in a single synthesis, because of inherent limitations in oligonucleotide synthesis, but instead is assembled from a collection of smaller oligos. GeneDesign's oligo design module splits sequences >800 bp into "chunks" of ~500 bp that overlap by a user-defined length and are flanked by unique restriction sites. Sequences between 600 and 800 bp long are split into two smaller chunks at a unique site, and sequences <600 bp are left as is. By default, each unique restriction enzyme is a nonblunt, ligation-friendly (i.e., not a one base-pair overhang) enzyme with a single cleavage site internal to its recognition site.

Within each of these chunks, the oligos are then designed. The user defines an oligo length and a target annealing temperature for the oligo overlaps. The defaults are 60-bp oligos with 56°C overlaps, which works well for us with yeast and mammalian sequences that are ~40% G+C. It is important to note that with GeneDesign we use a partially overlapping oligonucleotide strategy (Fig. 4), which allows gaps of 0–20 nucleotides at the end of each oligonucleotide. These gaps allow the program a certain flexibility in optimizing the melting temperatures (Tms) of all of the double-stranded regions. Empirically, this overlapping oligo strategy has worked well for us in terms of accurate gene synthesis, and we show in a later section of this article that the presence of the single-strand gaps only modestly elevates error rates.

GeneDesign uses the formula ((chunk length − overlap length)/(oligo length − overlap length)) to determine the number of oligos of the requested size that will actually fit in the chunk. Chunk length is always 500, and overlap length is always 20. Only a few oligo lengths are suggested to the user because only a few lengths will, in this formula, result in an even number of oligos (Table 1). Oligo lengths of ≥60 bp leave >20-bp gaps, which are ideal for temperature optimization. Oligo lengths of ≤50 bp leave <10-bp gaps, which are difficult to standardize; chunks designed with 40- or 50-bp oligos are treated as gapless and are not optimized for melting temperature.

If oligos ≥60 bp are requested, GeneDesign first breaks the chunk into an even number of oligos of the requested length with 20-bp overlaps. After adjusting every oligo in length to evenly make up the difference between 500-bp and the actual chunk length (thus ensuring that no oligo is a grossly different length), it analyzes the average melting temperature of the overlaps and adjusts the target melting temperature for that chunk. This on-the-fly adjustment allows every chunk to have an internally consistent Tm for assembly and prevents the program from stalling because of an impossible design requirement. Once the target Tm for the chunk has been determined, the oligo lengths are adjusted so that the Tms of each overlap are consistent with the target.

Table 1. Possible oligo lengths given a chunk size of 500 bp and an oligo overlap length of 20 bp

Chunk length (bp)	Overlap length (bp)	Gap length (bp)	Oligo length (bp)	No. of oligos
500	20	60	100	6
		40	80	8
		20	60	12
		10	50	16
		0	40	24

The formula is [(chunk length − overlap length)/(oligo length − overlap length)]. The formula for length is oligo length = 2 × overlap length + gap length. In order for oligo assembly to work, there must be an even number of oligos in the chunk, and the same number on each strand. GeneDesign only offers oligo lengths that lead this formula to an even number of oligos.

The Tms of the oligo overlaps are calculated as an average of three formulas: two salt-adjusted equations (Baldino Jr. et al. 1989) and a formula (Rychlik et al. 1990) based on the nearest neighbor thermodynamic model (Borer et al. 1974), using adjusted thermodynamic values (Sugimoto et al. 1996).

Every oligo is displayed for the user's approval, and when ready, the user can export them as a tab-delimited text file for ordering.

Sequence analysis module

The sequence analysis module provides brief information about the nucleotide sequences it is given. It counts the number of bases of each type, displaying length and composition. It applies a number of Tm formulas to sequences that are <100 bp in length. It will also create a map of the unique restriction sites present in a sequence and a list of the restriction sites that are absent. For every sequence it is given, it creates a vertical chart of the ORFs present in the three forward translation frames.

Program output

The design of a 5000-bp synthetic gene using the design a gene path (from reverse translation through automatic site insertion and gapped oligo design) takes <5 min when GeneDesign is served over a LAN from a 500-MHz PowerPC G4 computer with 1 GB of RAM. The customized use of modules causes processing time to vary slightly. In short, the execution time of the program is typically far less than is the time required to make decisions regarding the design of the gene.

Oligo design and A+T content

The GeneDesign oligo design module was tested with random 500-bp sequences of varying A+T content by using an average oligo length setting of 60 nucleotides. The resulting gapped oligos were analyzed for annealing temperature, length, and coverage of the original sequence. We defined a "collision" as the overlap of two nonconsecutive oligos and a "touch" as the lack of a gap between two nonconsecutive oligos. Although collisions and touches will not hinder the in vitro assembly of a chunk,

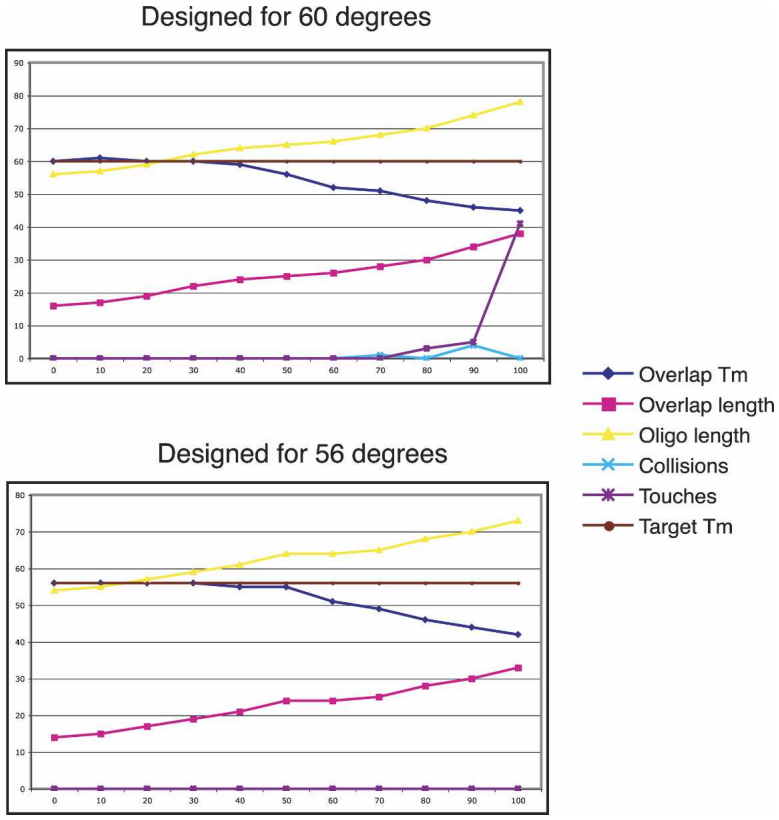

Figure 5. Oligo design and A+T content. To test the robustness of GeneDesign's oligo design algorithm, 100 random 500-bp sequences with A+T content varying from 0% to 100% were generated by a Perl script and run through the oligo design module. The resulting chunks were analyzed for the length and melting temperature of their constituent oligos. The number of times consecutive oligos on the same strand overlapped (collision) or met without a gap (touch) was counted by a Perl script.

their presence indicates inefficient design and may represent a slightly higher cost. GeneDesign successfully designed the oligos of both A+T-rich and A+T-poor sequences with consistent annealing temperatures and a minimum of collisions between nonconsecutive oligos. When designing oligos for a 60°C melting temperature, there were no collisions at 100% A+T and oligos collided 5% of the time at 90% A+T. When the target melting temperature was 56°C, there were no collisions at any base composition (Fig. 5).

Oligo assembly and amplification

We addressed the question of to what degree fluctuations in annealing temperature would affect the efficiency and the accuracy of gene synthesis by using the PCR assembly technique as a way to evaluate its robustness. To do this, we chose three 600-bp chunks of synthetic human L1 retrotransposon (58% GC) and eight 500-bp chunks of synthetic yeast Ty1 retrotransposon (44% GC). The L1 chunks were designed to have mean annealing temperatures of 50°C, 53°C, or 56°C, and the Ty1 chunks were designed to have an annealing temperature of 56°C. Each chunk was assembled and amplified across a 20°C gradient of annealing temperatures centered on the mean annealing temperature, and we were able to obtain a band of the proper size in every case (Fig. 6). Thus we conclude that the process is remarkably robust and that small variations in PCR machines or oligonucleotide melting temperatures are unlikely to create problems. We were able to transform plasmids containing the amplified DNA into competent cells and obtain DNA sequences from both high and low temperature endpoints.

Rate of mutation

In GeneDesign's default oligonucleotide design strategy, oligos are overlapped, leaving 1–25 base gaps throughout each ~500-bp chunk. Because annealing in double-stranded regions is expected to reduce mutation frequencies by selecting against incorrectly base-paired molecules (i.e., molecules containing incorrect bases), we performed an experiment to evaluate the mutation frequencies in the single- and double-stranded segments of a chunk. We determined the mutation rate of synthetic sequence from double-stranded and single-stranded oligo coverage by aligning the sequenced clones with the set of oligos from which it was assembled and locating the mutations. We found that the ratio of mutations per kilobase in single-stranded to double-stranded regions was, as expected, elevated. However, the elevation in the single-strand regions was only ~44% higher than that of double-stranded regions on average. The total number of mutations per kilobase was <5 using our conditions and this particular preparation of oligonucleotides (Table 2). In practice, we sequenced 24 clones per 500-bp chunk, and on average, four of these are perfect (no substitutions) and the nearly all instances have at least one perfect clone.

Discussion

Not all combinations of oligo length, annealing temperature, and base composition are possible in gapped oligo design. In gapless design, conflicts of annealing temperature and oligo

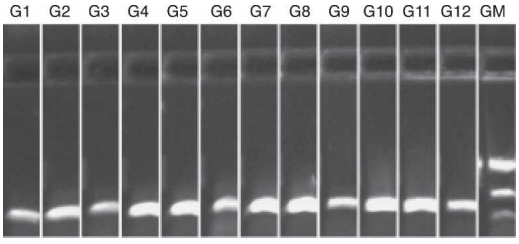

Figure 6. Amplification PCRs of a 500-bp synthetic chunk of yeast retrotransposon Ty1. The oligos used to assemble this single 500-bp chunk were optimized for a melting temperature of 56°C and then assembled with gradient PCR with annealing temperatures ranging from 45°C–65°C. Each assembly was then amplified at the same annealing temperatures. This gel shows the amplification results on the 20°C gradient with the lowest temperature on the *right*.

Table 2. Mutations in single-stranded and double-stranded regions of a synthetic gene assembled by gapped oligo assembly

	Double-stranded (43,918)		Single-stranded (41,371)		Total (85,289)	
	No.	Per kilobase	No.	Per kilobase	No.	Per kilobase
Deletions	91	2.07	83	2.01	174	2.04
Insertions	10	0.23	17	0.41	27	0.32
Transitions	50	1.14	89	2.15	139	1.63
Transversions	15	0.34	36	0.87	51	0.60
	166	3.78	225	5.44	391	4.58
Sequencing reads				177		
Ratio of single-stranded to double-stranded mutant				1.4389		

Number of bases are in parentheses. PCR products were transformed into competent cells in Topo vector and sequenced in the forward direction; 177 sequencing reads were analyzed, totaling 85,289 bases.

length do not usually arise because temperature optimization is not carried out by oligo length adjustment. In order to ensure that GeneDesign could still perform well on sequences with unusually high (or low) A+T content, the oligonucleotide design algorithm is designed to sample A+T content and readjust the design parameters accordingly. Especially when designing oligos for larger genes, this allows the program to come as close as possible to the annealing temperature the user requested and still find an oligo design solution, no matter the base composition of the sequence. Individual 500-bp pieces of the gene can have their constituent oligo annealing temperature adjusted specifically to their A+T content.

In terms of fidelity, we have noted only a marginally significant benefit to gapless oligo design. As has been noted before (Hoover and Lubkowski 2002), the rate of error is owed in large part to the fidelity of the oligonucleotides used because the rate of PCR polymerase error is relatively very low. By sequencing a few extra clones, we find we can overcome the slightly higher mutation rate of gapped oligos in the single-stranded region and still benefit from the lower cost of ordering fewer oligonucleotides.

Comparison with existing oligo design programs

GeneDesign's oligonucleotide design module handles much longer sequences than do the other oligonucleotide programs DNAWorks (Hoover and Lubkowski 2002) and Gene2Oligo (Rouillard et al. 2004). These older programs do not make any provision for splitting long sequences into smaller pieces for individual assembly. GeMS (Jayaraj et al. 2005) only employs gapless design, does not generate oligos longer than 40 bp, and cannot adjust the annealing temperature of oligo overlaps.

Applications of GeneDesign

As we use GeneDesign, we find ourselves continually finding new applications for synthetic DNA. Sometimes it is as simple as solving nagging cloning problems in the molecular biology laboratory. Programs such as GeneDesign facilitate the assembly of individual yeast genes or mammalian cDNAs. We have used the program to synthesize retrotransposons of ≥5 kb (Han and Boeke 2004), and others have synthesized a bacterial plasmid lacking restriction sites (Mandecki et al. 1990) and small viral genomes (Cello et al. 2002; Smith et al. 2003). We are also experimenting with the use of synthetic DNA to remove segments of yeast chromosomes at will to facilitate their re-engineering. Clearly, one of the next frontiers will be synthesizing bacterial genomes and smaller eukaryotic chromosomes. GeneDesign and programs like it will evolve to meet these new needs.

Next steps

We hope to develop a command line interface for high-throughput applications. Also, we would like to create a module for inserting or removing sites other than restriction sites, for example, known transcription factor binding sites, as supplied by the user. Another frontier will be to develop a software package that designs all genes within a single pathway or that encode a large multicomponent macromolecular complex, entire chromosomes, or genomes.

Methods

GeneDesign is written in Perl and C. The source code is available as a link from the GeneDesign home page. All output is displayed in HTML friendly to JavaScript activated browsers. Safari 1.3 and Firefox are the recommended browsers for Macintosh and Windows platforms, respectively. All oligos in this study were synthesized by Integrated DNA Technologies on a 100-nm scale with standard desalting purification.

A+T content and oligo design

To test the flexibility of the oligo design algorithm when stressed with sequences with varying base compositions, we generated 100 random nucleotide sequences. All sequences were 500 bp long and ranged from 0% to 100% AT content. GeneDesign's oligonucleotide design module was used with the default settings (60mers with 56°C overlaps) to generate a list of oligos for analysis.

Assembly and amplification PCR

We performed "assembly PCR" (Stemmer et al. 1995) in a 25 µL volume consisting of 1 U of buffered ExTaq polymerase (Takara) and 2.5 µL of a 300 nM oligo mix. The reaction consisted of 4 min of denaturation at 94°C followed by 20 cycles of 30 sec at 94°C for denaturation, 30 sec at a variable annealing temperature, and 30 sec at 72°C for elongation. Gradient PCR covered a 20°C range with a median temperature of 56°C.

We performed "amplification PCR" (Stemmer et al. 1995) in a 25 µL volume consisting of 1 U of buffered ExTaq polymerase, 1.3 µL of the assembly PCR product, and 0.7 µL of a 10 µM mix of the outer primers. The reaction consisted of 4 min of denaturation at 94°C followed by 25 cycles of 30 sec at 94°C for denaturation, 30 sec at a variable annealing temperature, and 30 sec at 72°C for elongation, terminated by a 7-min incubation at 72°C. The gradient PCR covered a range of 20°C with a median temperature of 56°C.

Cloning and sequencing

We transformed PCR products into competent DH5α cells with Topo TA vector (Invitrogen) by heat shock for 45 sec at 42°C.

Colonies that contained inserts were frozen in glycerol stock and sequenced.

Acknowledgments

We thank Brian Greenlee for helpful graphic advice, Brian Olson and Mark Forrer for programing advice, and Daniel Yuan for help with Web serving. Supported in part by NIH grants CA16519, GM36481, and Roadmap grant RR020839 to J.D.B.

References

Baldino Jr., F., Chesselet, M.F., and Lewis, M.E. 1989. High-resolution in situ hybridization histochemistry. *Methods Enzymol.* **168:** 761–777.

Borer, P.N., Dengler, B., Tinoco Jr., I., and Uhlenbeck, O.C. 1974. Stability of ribonucleic acid double-stranded helices. *J. Mol. Biol.* **86:** 843–853.

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

Han, J.S. and Boeke, J.D. 2004. A highly active synthetic mammalian retrotransposon. *Nature* **429:** 314–318.

Hoover, D.M. and Lubkowski, J. 2002. DNAWorks: An automated method for designing oligonucleotides for PCR-based gene synthesis. *Nucleic Acids Res.* **30:** e43.

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

Jay, E., MacKnight, D., Lutze-Wallace, C., Harrison, D., Wishart, P., Liu, W.Y., Asundi, V., Pomeroy-Cloney, L., Rommens, J., Eglington, L., et al. 1984. Chemical synthesis of a biologically active gene for human immune interferon-γ: Prospect for site-specific mutagenesis and structure-function studies. *J. Biol. Chem.* **259:** 6311–6317.

Jayaraj, S., Reid, R., and Santi, D.V. 2005. GeMS: An advanced software package for designing synthetic genes. *Nucleic Acids Res.* **33:** 3011–3016.

Krieg, R., Stucka, R., Clark, S., and Feldmann, H. 1991. The use of a synthetic tRNA gene as a novel approach to study in vivo transcription and chromatin structure in yeast. *Nucleic Acids Res.* **19:** 3849–3855.

Mandecki, W., Hayden, M.A., Shallcross, M.A., and Stotland, E. 1990. A totally synthetic plasmid for general cloning, gene expression and mutagenesis in *Escherichia coli*. *Gene* **94:** 103–107.

Nambiar, K.P., Stackhouse, J., Stauffer, D.M., Kennedy, W.P., Eldredge, J.K., and Benner, S.A. 1984. Total synthesis and cloning of a gene coding for the ribonuclease S protein. *Science* **223:** 1299–1301.

Neves, F.O., Ho, P.L., Raw, I., Pereira, C.A., Moreira, C., and Nascimento, A.L. 2004. Overexpression of a synthetic gene encoding human α interferon in *Escherichia coli*. *Protein Expr. Purif.* **35:** 353–359.

Patterson, S.S., Dionisi, H.M., Gupta, R.K., and Sayler, G.S. 2005. Codon optimization of bacterial luciferase (lux) for expression in mammalian cells. *J. Ind. Microbiol. Biotechnol.* **32:** 115–123.

Quinn, T.P., Tweedy, N.B., Williams, R.W., Richardson, J.S., and Richardson, D.C. 1994. Betadoublet: De novo design, synthesis, and characterization of a β-sandwich protein. *Proc. Natl. Acad. Sci.* **91:** 8747–8751.

Rouillard, J.M., Lee, W., Truan, G., Gao, X., Zhou, X., and Gulari, E. 2004. Gene2Oligo: Oligonucleotide design for in vitro gene synthesis. *Nucleic Acids Res.* **32:** W176–W180.

Rychlik, W., Spencer, W.J., and Rhoads, R.E. 1990. Optimization of the annealing temperature for DNA amplification in vitro. *Nucleic Acids Res.* **18:** 6409–6412.

Sharp, P.M., Cowe, E., Higgins, D.G., Shields, D.C., Wolfe, K.H., and Wright, F. 1988. Codon usage patterns in *Escherichia coli*, *Bacillus subtilis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Drosophila melanogaster* and *Homo sapiens*: A review of the considerable within-species diversity. *Nucleic Acids Res.* **16:** 8207–8211.

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

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

Sugimoto, N., Nakano, S., Yoneyama, M., and Honda, K. 1996. Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes. *Nucleic Acids Res.* **24:** 4501–4505.

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

Received July 14, 2005; accepted in revised form November 9, 2005.