

TmPrime: fast, flexible oligonucleotide design software for gene synthesis

Marcus Bode, Samuel Khor, Hongye Ye, Mo-Huang Li* and Jackie Y. Ying

Institute of Bioengineering and Nanotechnology, The Nanos, 138669, Singapore

Received January 27, 2009; Revised April 15, 2009; Accepted May 16, 2009

ABSTRACT

Herein we present TmPrime, a computer program to design oligonucleotide sets for gene assembly by both ligase chain reaction (LCR) and polymerase chain reaction (PCR). TmPrime offers much flexibility with no constraints on the gene and oligonucleotide lengths. The program divides the long input DNA sequence based on the input desired melting temperature, and dynamically optimizes the length of oligonucleotides to achieve homologous melting temperatures. The output reports the melting temperatures, oligonucleotide sequences and potential formation of secondary structures. Our program also provides functions on sequence pooling to separate long genes into smaller pieces for multi-pool assembly and codon optimization for expression. The software has been successfully used in the design and synthesis of green fluorescent protein fragment (GFPuv) (760 bp), human protein kinase B-2 (PKB2) (1446 bp) and the promoter of human calcium-binding protein A4 (S100A4) (752 bp) using real-time PCR assembly with LCGreen I, which offers a novel approach to compare the efficiency of gene synthesis. The purity of assembled products is successfully estimated with the use of melting curve analysis, which would potentially eliminate the necessity for agarose gel electrophoresis. This program is freely available at <http://prime.ibn.a-star.edu.sg>.

INTRODUCTION

Current gene synthesis methods include ligase chain reaction (LCR) (1) and polymerase chain reaction (PCR) assembly (2), relying on the use of overlapped oligonucleotides to construct genes. Various PCR-based methods have been proposed in attempt to optimize the PCR process for long DNA sequences and to enhance the accuracy of assembly. These methods are the thermodynamically balanced inside-out (TBIO) method (3), successive PCR

(4), dual asymmetrical PCR (DA-PCR) (5), overlap extension PCR (OE-PCR) (6,7) and PCR-based two-step DNA synthesis (PTDS) (4). In LCR assembly, adjacent oligonucleotides with no gap between consecutive oligonucleotides are ligated together, resulting in the extension of DNA length, whereas PCR assembly utilizes the DNA polymerase to extend the oligonucleotides. Regardless of LCR or PCR assembly, successful synthesis would require proper oligonucleotide design to ensure that the oligonucleotides are highly specific to their targets and have uniform hybridization temperature to maximize the assembly efficiency.

Programs have been developed for gene synthesis to automatically design oligonucleotides based on user-specific hybridization temperature and oligonucleotide length (8–13). Fairly good synthesis results have been achieved for DNA below 1 kb from oligonucleotides designed by DNAWorks (8), Gene2Oligo (9) and Assembly PCR Oligo Maker (10) programs. To reduce errors in the synthetic gene, genes with long sequences are usually split into fragments (~500 bp) and assembled in separated pools first, and then fragments are assembled into the full-length product in another PCR step. GeneDesign (11) and GeMS (12) have implemented this function for multikilobase gene synthesis. Except for the DNAWorks, the existing programs are developed for gapped PCR-based gene assembly, and are lacking in the function of predicting the potential mis-hybridization and secondary structures among oligonucleotides. Moreover, none of these oligonucleotide design programs is capable to design oligonucleotides or analyze potential mis-hybridization for aspiring applications in genome synthesis (13,14) with ultra-long DNA (>5 kb) or multiplex gene synthesis (15).

Herein we present TmPrime, a program that provides unique features that are lacking in the conventional gene synthesis software (Supplementary Table S1). It is able to design oligonucleotides with very long gene sequences (≤ 40 kb) for LCR and gapless PCR assembly. It uses a novel approach to divide the input gene sequences into oligonucleotides with homologous melting temperature. In addition, the TmPrime implements an automatic gene splitting for PCR-based long gene assembly (4,6,16).

*To whom correspondence should be addressed. Tel: +65 6824 7119; Fax: +65 6478 9085; Email: mhli@ibn.a-star.edu.sg

The potential mis-hybridization, hetero-dimer, homo-dimer and hairpin formations among oligonucleotides are screened by pair-wise sequence alignment. A mispriming reduction scheme is employed for constructing genes with highly repetitive DNA sequences using LCR assembly (17). We have employed TmPrime in the synthesis of three genes (752–1446 bp) using real-time PCR with LCGreen I detection. The melting curve analysis results of the assembled genes have been compared with agarose gel electrophoresis, as in quantitative competitive PCR (QC-PCR) (18).

MATERIALS AND METHODS

Fast and flexible oligonucleotide design

Our approach started from a simple observation of the gapless PCR or LCR assembly (Figure 1). For a given sequence, all oligonucleotides were exactly adjacent with no gap between two consecutive oligonucleotides. The given sequence could be seen as the serial connection of all overlapping regions of oligonucleotides. With this simple observation, the problem of designing oligonucleotides with uniform melting temperature would be equivalent to dividing the given sequence into fragments with homologous melting temperature, each fragment representing an overlapping region.

Central to TmPrime was the division of the given sequence into fragments, and equilibration of the melting temperature (T_m) across fragments. Two approaches, Equi-space and Equi- T_m , were developed. In the Equi-space method, the program first divided the given sequence into fragments with approximately equal lengths by markers (Equi-space in Figure 2), and computed the average and deviation in melting temperatures using the nearest-neighbor model with SantaLucia's thermodynamic parameter (19), corrected with salt and oligonucleotide concentrations, and the total number of phosphates in the duplex for the difference of LCR and PCR assemblies (20). One extra phosphate per oligonucleotide was contributed by phosphorylated oligonucleotides prerequisite for LCR. Next, the fragment sizes were adjusted by the TmPrime through shifting the marker positions to minimize the deviations in the overall melting temperature. In the Equi- T_m approach, the sequence was cut into fragments from the beginning until the end of the DNA sequence based on the user-specified melting temperature

(Equi- T_m in Figure 2). This process would usually leave a small DNA tail having T_m smaller than the user-specified T_m . The fragment boundaries were then shifted to accommodate this tail.

Once the melting temperatures of the fragments were equilibrated, the oligonucleotides for gapless PCR or LCR were formed by connecting two adjacent fragments along both the sense and antisense strands. Each oligonucleotide overlapped with its complementary neighbors by exactly one fragment. The two tail segments at the 3'-end of sense and antisense sequences were also included for LCR assembly.

Multiple-pool assembly

For a gene that could be too long to be efficiently constructed in a single synthesis, TmPrime could automatically split the gene into pools of shorter sequences based on the user-specified number of pools with the pool-pool overlap length automatically adjusted according to the annealing temperature of across-pool assembly out

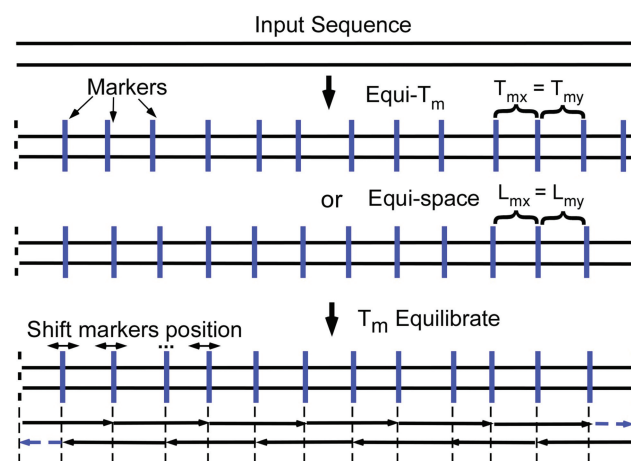


Figure 2. An overview of the oligonucleotide design scheme. The software first divides the input sequence into approximately equal-temperature (Equi- T_m) or equal-length fragments (Equi-space) using markers based on the user-specified melting temperature. The positions of the markers are iteratively shifted to globally minimize the deviation in melting temperature among the fragments (T_m Equilibrate). Two adjacent fragments are joined together to generate oligonucleotides for PCR gapless assembly. The two tail segments at 3'-end of sense and antisense sequences are also included for LCR assembly.

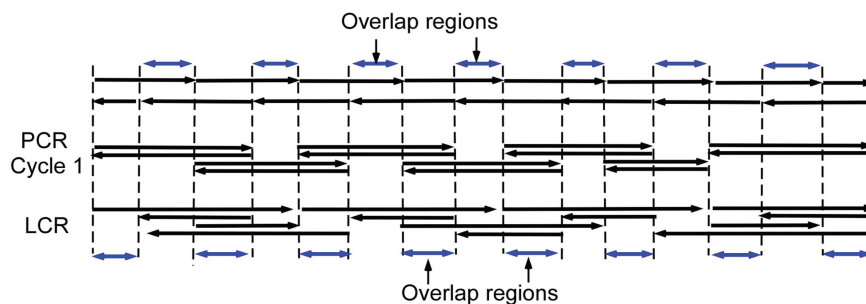


Figure 1. Scheme of LCR or gapless PCR assembly. The input sequence is the serial connection of overlap regions of oligonucleotides.

primers (4,6,16). Different annealing temperatures can be assigned for across-pool assembly outer primers, within-pool assembly outer primers and inner oligonucleotides, providing the flexibility for long gene construction. Oligonucleotides for each pool assembly could be optimized at the same melting temperature to allow the parallel synthesis of different segments or different genes simultaneously on a single thermal cyclers.

Mis-hybridization screening

TmPrime was implemented using functional modules. Beside the core oligonucleotide design module, various modules were constructed in a cohesive manner. Secondary structures, including hetero-dimers, homo-dimers and hairpin loops, were searched via local pair-wise sequence alignments with a score based on the number of matched bases, and G+C content. A window containing 18 bases was pair-wisely aligned with the entire forward and reverse DNA sequence, and the number of matched bases within this 18-base window was recorded. This pair-wise alignment was iteratively calculated with the window position shifted one base each time from the beginning until the end of the DNA sequence. The locations and types of secondary structures were reported when the score was higher than the user-defined threshold length.

TmPrime also implemented a mispriming reduction scheme for constructing genes with highly repetitive DNA sequences using LCR assembly, which represented one of the most complicated mis-hybridization scenarios where oligonucleotides can mispair at multiple sites. Our approach took the inherent advantages of LCR, which required stringent annealing conditions for effective assembly. Ligation was invoked after two adjacent oligonucleotides uniquely hybridized with an opposite pairing DNA. TmPrime utilized this property, and created two tables for each pair of mis-hybridization (consisting of two adjacent oligonucleotides and their opposite pairing template) based on the mis-hybridization screening results. One table recorded the coordinates and 18-base 5'- and 3'-ends sequences of participated oligonucleotides, whereas another table recorded the coordinate and sequence of each opposite pairing template (36 bp). This process recorded the 5'- and 3'-ends boundaries information of mispairing oligonucleotide fragments for LCR assembly. After that, each pairing template was pair-wisely aligned with a random pair of 18-base segments. A score was assigned if both the number of matched bases was higher than the user-defined threshold length, and the end coordinate of one segment was not the beginning of the other segment (i.e. two segments were from non-adjacent oligonucleotides). This step examined the possibility of non-adjacent oligonucleotides annealing with a pairing template, and the potential further extension of mis-hybridized fragments. This mis-hybridization analysis was iteratively calculated within the user-specified T_m range with an increment of 0.4°C per step. During each T_m optimization step, the boundaries of oligonucleotide fragments were also redefined, and the score was recalculated for each pairing template, resulting in a new overall

score for the entire sequence. The program output the oligonucleotide set with the lowest score.

TmPrime also included a codon optimized feature. It implemented codon optimization on the global optimization, which replaced each codon based on the organism-specific codon frequencies using the organism-specific codon data in Codon Usage Database (www.kazusa.or.jp/codon/).

Target proteins

To evaluate TmPrime, gene sequences for a segment of *Escherichia coli* codon-optimized GFPuv (760 bp) (21), *E. coli* codon-optimized PKB2 (1446 bp) (3) and S100A4 (752 bp; chr1:1503312036-1503311284) (22) were selected for synthesis via PCR assembly, using oligonucleotides derived by TmPrime. GFPuv was selected for its widespread use as a tagging protein, and due to the availability of a published oligonucleotide set suitable for use as an experimental control. PKB2 was selected based on the reported difficulty of assembly via PCR (3), while S100A4 has clusters of GC contents resulting in non-uniform regions of high melting temperatures. Published sequences of GFPuv (21) and PKB2 (3) were adapted for comparing assembly results. The oligonucleotide sets designed for the selected genes using the Equi-space method are shown in Supplementary Tables S1–S4.

Real-time gene assembly and amplification

All PCR reactions, whether for assembly or amplification, were run in Roche's LightCycler 1.5 real-time thermal cycling machine with a temperature transition of 20°C/s. The melting curves analyses of assembled genes were acquired using a ramp of 0.05°C/s for 72–99°C. The one-step process was performed with 10 µl of reaction mixture including 1× PCR buffer (Novagen), 2× LCGreen I (Idaho Technology Inc.), 4 mM of MgSO₄, 0.3 or 1 mM of each of deoxynucleotide precursor (dNTP) (Stratagene), 500 µg/ml of bovine serum albumin (BSA), 10 or 64 nM of oligonucleotides, 400 nM of forward and reverse primers and 1 U of KOD Hot Start (Novagen). The PCR was conducted under the following conditions: 2 min of initial denaturation at 95°C, 30 cycles of 95°C for 5 s, a variable annealing temperature (dependent on the target genes) for 10 s, 72°C for 10–30 s and a last extension of 72°C for 10 min. The PCR protocol for the two-step process was essentially the same as that for the one-step process, except for the concentration of oligonucleotides used. For PCR assembly, 64 nM of oligonucleotides were used. For gene amplification, 2 µl of the assembled product was diluted in an amplification reaction mixture of 25 µl with primers concentration of 4 µM each. Desalted oligonucleotides were obtained from Research Biolabs (Singapore) without additional purification.

RESULTS

The TmPrime was designed to be flexible and easy to use. Most of the parameters on the web interface were self-explanatory (Supplementary Figure S1). The user would be asked to provide the gene information, gene assembly

conditions (buffer condition and oligonucleotide concentrations), optional parameters for long DNA assembly and parameters for mispriming analysis. The T_m range was automatically set with $\pm 2^\circ\text{C}$, which would usually give us a rapid convergence. The software would report the melting temperatures, oligonucleotide sequences, potential formation of secondary structures and statistical information of the oligo sets of each pool in a PDF file. If the codon optimization function was selected, information regarding the codon optimization would also be reported in the same output file.

TmPrime was written in C++. The process of designing a multi-kilobase (<5 kb) sequence took <4 min when TmPrime was served over a Red Hot server from a Dell computer with dual 3.3-GHz Intel Xeons and 8 GB of RAM.

Designing oligonucleotides for target proteins

Three target proteins were designed for the gapless synthesis using TmPrime. Table 1 shows the melting temperature of the given genes, average melting temperatures and length of overlap regions, and oligonucleotides optimized using the Equi-space approach. A fixed-length control (40 nt) for GFPuv was also designed for comparison of assembly efficiency. The variations in overlap and oligonucleotide lengths were more than 14 bases for PKB2 and S100A4, with a melting temperature deviation smaller than 9°C . The Equi-space approach was sensitive to the distribution of sequence G+C content, which generated regions with elevated melting temperature (Supplementary Figure S2). Equi-space suffered from this drawback, whereby the T_m optimized tended to converge regionally. The Equi- T_m approach greatly improved the T_m uniformity to < 3°C deviation for all three genes (Supplementary Table S6), by dividing the sequence based on the melting temperature. Thus, the Equi- T_m approach was adopted for the latest TmPrime. No stable secondary structures were detected at the intended melting temperatures.

Oligonucleotide assembly and amplification

The gapless PCR-based gene syntheses have been performed using the real-time PCR method and the oligonucleotide sets derived from TmPrime. Performing gene synthesis using real-time PCR with LCGreen I detection enabled us to study the efficiency, quantity and quality of the gene synthesis process as in real-time PCR amplification (23,24). Full-length GFPuv fragments were achieved using one-step process with TmPrime optimized oligonucleotides (Lane 1) and fixed-length control

oligonucleotides (Lane 2) (control) as shown by gel electrophoresis (Figure 3). The optimized oligonucleotides derived from TmPrime with a uniform T_m improved the synthesis efficiency as indicated by the intensity difference of full-length gel bands. TmPrime-optimized oligonucleotides provided much better efficiency than the control. This trend was also confirmed by the ratio of area underneath the melting peak ($T_m = 86.5^\circ\text{C}$) of the real-time PCR results (Figure 4a). The integrated area of the melting peak in the negative derivative of fluorescence with respect to temperature ($-dF/dT$ versus T) gave the ratio of the assembled product quantity (25). It was observed that the DNA-melting curve over-presented the quantity of full-length DNA due to the nature of gene synthesis process. Besides the full-length DNA, the assembled mixture contained also long incomplete products, which have melting temperature indistinguishable from that of full-length DNA.

No full-length gene products were obtained with the one-step process on the relative long sequence (Lane 3; PKB2, 1446 bp) or cluster GC content (Lane 6; S100A4, 752 bp) (Figure 3), as consistent with the published results (3). However, successful gene syntheses were achieved with the two-step process for both PKB2 (Lane 5) and S100A4 (Lane 8), as demonstrated by the distinct bands on gel electrophoresis. The melting peak curves were acquired for both one-step and two-step processes (Figures 4b and c). The successful synthesis generated a single, sharp melting peak in the melting curve, which

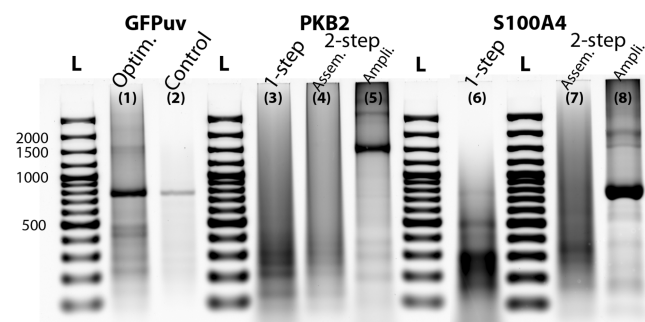


Figure 3. Agarose gel electrophoresis of assembled products. One-step synthesis of GFPuv (760 bp) from TmPrime: (Lane 1) optimized and (Lane 2) fixed-length control oligonucleotides. (Lane 3) One-step synthesis of PKB2 (1446 bp). Two-step synthesis of PKB2: (Lane 4) assembly and (Lane 5) amplification. (Lane 6) One-step synthesis of S100A4 (752 bp). Two-step synthesis of S100A4: (Lane 7) assembly and (Lane 8) amplification. The annealing temperatures for the PCR process are as follows: GFPuv, 50°C ; PKB2, 61°C ; S100A4, 58°C (assembly) and 49°C (amplification).

Table 1. Data on oligonucleotides

| Gene | Length (bp) | T_m of gene ($^\circ\text{C}$) | Average T_m (min, max) ($^\circ\text{C}$) | Number of oligonucleotides | Overlap length (nt) | Oligonucleotide length (nt) |
|---------------|-------------|------------------------------------|---|----------------------------|---------------------|-----------------------------|
| GFPuv TmPrime | 760 | 85.7 | 55.8 (52.2, 58.6) | 36 | 18–26 | 38–48 |
| GFPuv Control | 760 | 85.7 | 51.8 (40.6, 60.5) | 38 | 20 | 40 |
| PKB2 | 1446 | 93.5 | 66.6 (63.6, 68.8) | 60 | 17–31 | 20, 38–57 |
| S100A4 | 752 | 91.6 | 66.0 (61.1, 69.8) | 30 | 19–33 | 19, 41–66 |

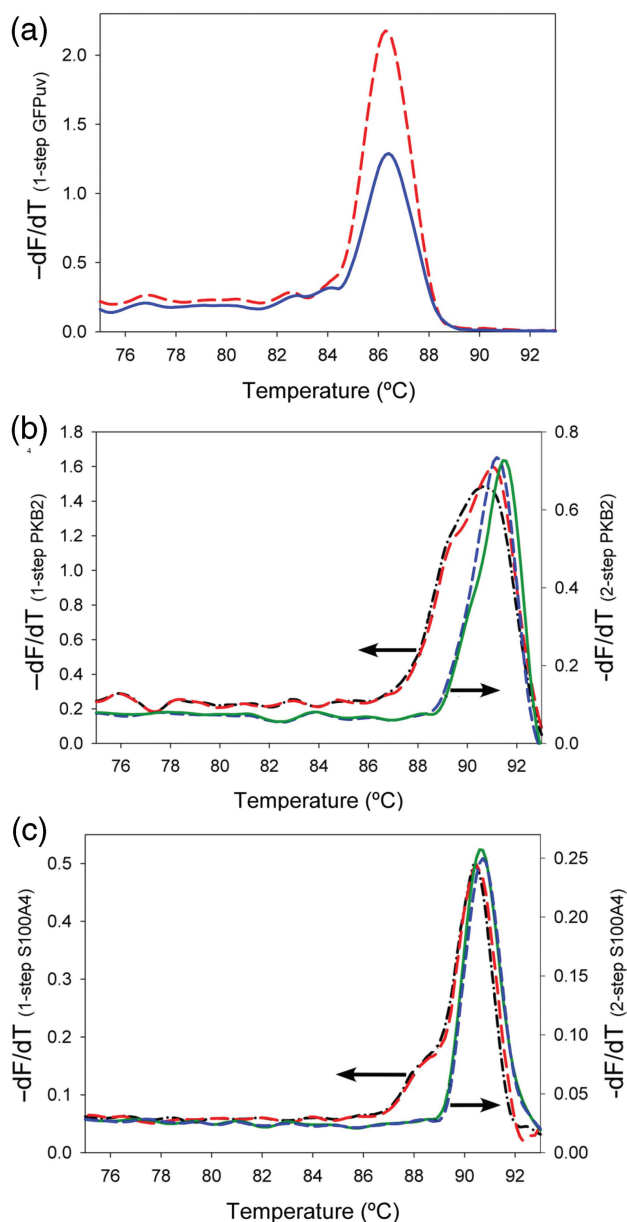


Figure 4. (a) Melting peak analyses of the assembled products for GFPuv from one-step synthesis: (dashed line) optimized and (solid line) fixed-length control oligonucleotides. Melting peak analyses of the assembled products for (b) PKB2 and (c) S100A4 from one-step and two-step syntheses; two replicas were performed for each set of oligonucleotides. The corresponding agarose gel electrophoresis results of the assembled products are shown in Figure 3. The measured T_m values are 86.5°C for GFPuv, 91.5°C for PKB2 and 90.5°C for S100A4.

corresponded to a distinct band in gel electrophoresis for the two-step process. In contrast, for the one-step process, the melting curve was broad, indicating that the product was a mixture of DNAs with intermediate lengths, as reflected in the smeared gel electrophoresis. The majority of one-step products were incomplete with DNA lengths of ~200–300 bp. The calculated T_m (Table 1) by TmPrime matched with the measured T_m for all three target genes within 2°C for the two-step

process (Figure 4). Successful LCR assembly was also demonstrated on GFPuv with comparable gel results to two-step PCR process (Supplementary Figure S3).

Comparison with existing oligonucleotide design programs

The performance of TmPrime was compared with conventional oligonucleotide design programs (8–11) for S100A4, PKB2, GFPuv and the whole genome of poliovirus (13) (Genbank FJ517648; 7418 bp) and ϕ X174 bacteriophage (14) (Genbank J02482; 5386 bp) (Supplementary Table S6). TmPrime offered the most homologous melting temperature for all three genes with $\Delta T_m < 3^\circ\text{C}$, about half of the values provided by DNAWorks. A wider range of annealing temperatures (50–70°C) was accepted for TmPrime than that of DNAWorks (58–70°C) and Assembly PCR Oligo Maker (50–60°C) for designing genes with various GC contents. GeneDesign only employs gapped design, and can not adjust the oligonucleotide concentration and PCR buffer conditions. The oligonucleotide design failed while the sequence of consecutive oligonucleotides collided (in PKB2 and GFPuv at T_m of 65°C). The Gene2Oligo has difficulty in designing S100A4, which failed to converge at specified annealing temperature. Moreover, only TmPrime can handle the poliovirus and ϕ X174 bacteriophage genomes.

DISCUSSION

TmPrime was fast and flexible with no inherent limitation on the gene length and GC content (the only constraint was the computer memory). It utilized an Equi- T_m oligonucleotide design approach to overcome the fundamental cluster melting temperature distribution problem, which was encountered by most programs (9–12) that divided DNA sequence based on the length of genes containing cluster GC content. Presumably, parsing the DNA sequence on length would be more vulnerable to the sequence context, and cause the T_m of oligonucleotides to be less uniform. Oligonucleotides with various lengths, but uniform T_m (ΔT_m usually $< 3^\circ\text{C}$), were generated for both gapless PCR and LCR assemblies. Long DNAs could be partitioned automatically into pools of smaller pieces based on the user-specified number of pools and across-pools assembly annealing temperature. This unique feature was essential for the whole genome synthesis (13,14,26) and artificial protein construction (27,28). All secondary structures including homo-dimers, hairpin loop and mis-hybridization were screened to reduce the possibility of incorrect synthesis. Unlike DNAWorks (8), TmPrime stitched the adjacent potential mis-hybridization regions together and reported the entire extended region. This feature allowed users to easily visualize and inspect the problematic DNA regions (such as hairpin and repetitive DNA sequence), which might cause problems during gene assembly. These key features were either lacking or inadequate in the conventional oligonucleotide design programs.

TmPrime is capable of handling ultra-long DNA up to 40 kb. This feature is particularly useful for genome synthesis (13,14) and multiplex gene synthesis with

oligonucleotides from DNA microarray chip (15). In the case of genome synthesis with ultra-long DNA, the mis-hybridization would become more prevalent. It would be beneficial to screen the whole genome for possible mis-hybridization, and divide the DNA into pools accordingly before synthesizing the oligonucleotides. The current microchips can provide hundreds of thousands of oligonucleotides per chip at a few hundred dollars, which would offer oligonucleotide at significantly reduced cost if one can utilize oligonucleotides directly from DNA microchip for *de novo* gene synthesis (29). To achieve this goal, two critical issues must be addressed, which are related to the inherent properties of DNA microchips. Current microchips have very low surface areas and hence only a small amount of oligonucleotides (0.1 pmol/mm^2) can be produced (30). Thus, the resulting concentration of eluted oligonucleotides ($<1 \text{ nM}$; $100 \mu\text{m} \times 100 \mu\text{m}$ spot size and $1 \mu\text{l}$ PCR volume) might be insufficient for effective hybridization. Moreover, the eluted solution contains hundreds of thousands of oligonucleotides, which increase the possibility of mis-hybridization. TmPrime, with its highly uniform T_m , would enhance the hybridization efficiency of oligonucleotide at ultra-low concentration ($<1 \text{ nM}$), while the mispriming analysis features would facilitate multiplex gene design.

The gapless assembly required shorter oligonucleotides than the gapped assembly at the same annealing temperature. Thus, the gapless oligonucleotide design would potentially provide an extra advantage in minimizing the error of gene synthesis, which was mainly introduced from the oligonucleotides (31) and increased with oligonucleotide length due to the inherently error-prone oligonucleotide chemical synthesis process pointed out by Xiong *et al.* (4). The error rate of PCR assembly using 90-nt oligonucleotides was three to four times higher than that of using 60-nt oligonucleotides. Oligonucleotides were chemically synthesized base-by-base with a step yield of $\sim 98.5\%$ (32). Beside the correct-length oligonucleotides, the population of other molecules including both truncated species capped at the growing end and uncapped molecules containing errors increased with the oligonucleotide length, which could have also participated in the PCR process and generated products of incorrect sequence. In addition, using short oligonucleotides with high melting temperature should also reduce the potential of forming secondary structures, and prevent the generation of faulty sequence. Moreover, the unique T_m -optimized LCR assembly option would enhance the designing of oligonucleotides for constructing difficult genes with highly repetitive DNA sequences, which could not be achieved via PCR method (17).

The quality and quantity of PCR-based gene synthesis are influenced by several factors, such as annealing temperature, concentration of oligonucleotides, DNA polymerase, concentration of monomers and number of PCR cycles. These factors have been systematically studied using real-time gene synthesis with TmPrime optimized oligonucleotides (S100A4), which has led to insights into the gene synthesis process (24). Depending on the complexity of target genes, the synthetic genes are often constructed with a one-step or two-step overlapping

process. The one-step process was preferred for short DNAs ($<500 \text{ bp}$), wherein the amplification primers were mixed with assembly oligonucleotides in a single PCR reaction. In the one-step process, the assembly and amplification were conducted simultaneously, which could compete for the fixed amount oligonucleotides and monomers (dNTPs), and render intermediate products with lower molecular weights. This competitive process was more critical for DNA with high GC content or length (3,4), and was minimized in the two-step PCR process. These GC content and length effects were demonstrated by syntheses of GFPuv (760 bp; GC: 30–50%, avg.: 42%), S100A4 (752 bp; GC: 40–68%, avg.: 56%) and PKB2 (1446 bp; GC: 50–68%, avg.: 58.4%). GFPuv synthesis was achieved with one-step PCR, while two-step process was more robust for S100A4 and PKB2. In addition, the reportedly difficult gene PKB2 with both cluster GC and high length (3) has been successfully synthesized from 60 oligonucleotides, demonstrating the capability of our program.

We have also reported a method on using real-time PCR to monitor the gene assembly results. The melting curve results were consistent with the agarose gel electrophoresis findings. Successful gene synthesis yielded product with a single, sharp melting peak, while incomplete synthesis resulted in a broad melting curve. It was noted that the melting curve analysis would over-estimate the quantity of full-length DNA when the assembled mixture contained intermediate DNAs. The accuracy and robustness of this method would be substantially improved by enzymatic impurities clean-up, whereby DNA is assembled from 5'-phosphorylated inner oligonucleotides and 5'-hydroxyl (5'-OH) outer primers, followed by λ exonuclease treatment to digest the phosphorylated truncated DNAs (17,33). Besides providing a tool for characterizing the synthesis products, the real-time method could be integrated with microfluidic gene synthesis (34) to develop automated gene synthesis.

The low fidelity of synthetic DNA is a substantial obstacle for *de novo* gene synthesis. Error rates of 1–11 errors per kb of DNA have been reported (4,8,15). Xiong *et al.* (4) pointed out that the error frequency was related to assembly methods. The two-step overlapping PCR has the lowest error rate (1.26 error/kb) compared to one-step overlapping (3.78 error/kb), two-step successive (4.62 error/kb) and one-step successive (7.98 error/kb) PCRs, studied using 60-nt PAGE purified oligonucleotides. This finding implied that the error rate increased as the DNA assembly efficiency decreased. It is conceivable that methods with lower assembly efficiency would produce more error-prone intermediate products, resulting in a higher error rate. In addition, we found that the assembly efficiency was largely influenced by oligonucleotide concentration (24). The assembly reached the plateau at around cycle 10 and cycle 30 with oligonucleotide concentrations of 80 nM and 10 nM, respectively. In short, with the help of our real-time gene synthesis and TmPrime, one could investigate the effect of oligonucleotide concentration on error rate, which is important for DNA-microarray gene synthesis.

We would like to point out that we have not incorporated the functions for automatic multiplex gene design. The user might have to stitch the multiple gene sequences together into a single DNA sequence by using a word processor, and conduct the oligonucleotide design and mis-hybridization analysis accordingly. An interface with multiple genes uploading will be added to support this feature. In addition, we hope to add in more features for codon optimization with global- and per-codon optimizations using the latest organism-specific codon data (35,36). The per-codon optimization would provide fine control, allowing the user to change specific codon by selecting one of its alternatives. These features are being introduced into our program.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors acknowledge Dr Shu Wang and Dr Jérôme Boulaire for their helpful suggestions on S100A4.

FUNDING

This work is funded by the Institute of Bioengineering and Nanotechnology [Biomedical Research Council, Agency for Science, Technology and Research, Singapore]. Funding for open access charge: Institute of Bioengineering and Nanotechnology.

Conflict of interest statement. None declared.

REFERENCES

- Au, L.C., Yang, F.Y., Yang, W.J., Lo, S.H. and Kao, C.F. (1998) Gene synthesis by a LCR-based approach: high-level production of leptin-L54 using synthetic gene in *Escherichia coli*. *Biochem. Biophys. Res. Commun.*, **248**, 200–203.
- Stemmer, W.P., Cramer, 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.
- Gao, X., Yo, P., Keith, A., Ragan, T.J. and Harris, T.K. (2003) Thermodynamically balanced inside-out (TBIO) PCR-based gene synthesis: a novel method of primer design for high-fidelity assembly of longer gene sequences. *Nucleic Acids Res.*, **31**, e143.
- Xiong, A.-S., Yao, Q.-H., Peng, R.-H., Li, X., Fan, H.-Q., Cheng, Z.-M. and Li, Y. (2004) A simple, rapid, high-fidelity and cost-effective PCR-based two-step DNA synthesis method for long gene sequences. *Nucleic Acids Res.*, **32**, e98.
- Sandhu, G.S., Aleff, R.A. and Kline, B.C. (1992) Dual asymmetric PCR: one-step construction of synthetic genes. *Biotechniques*, **12**, 14–16.
- Young, L. and Dong, Q. (2004) Two-step total gene synthesis method. *Nucleic Acids Res.*, **32**, e59.
- Prodromou, C. and Pearl, L. (1992) Recursive PCR: a novel technique for total gene synthesis. *Protein Eng.*, **5**, 827–829.
- Hoover, D.M. and Lubkowski, J. (2002) DNAWorks: an automated method for designing oligonucleotides for PCR-based gene synthesis. *Nucleic Acids Res.*, **30**, e43.
- 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.
- Rydzanicz, R., Zhao, X.S. and Johnson, P.E. (2005) Assembly PCR oligo maker: a tool for designing oligodeoxynucleotides for constructing long DNA molecules for RNA production. *Nucleic Acids Res.*, **33**, W521–W525.
- Richardson, S.M., Wheelan, S.J., Yarrington, R.M. and Boeke, J.D. (2006) GeneDesign: rapid, automated design of multikilobase synthetic genes. *Genome Res.*, **16**, 550–556.
- Jayaraj, S., Reid, R. and Santi, D.V. (2005) GeMS: an advanced software package for designing synthetic genes. *Nucleic Acids Res.*, **33**, 3011–3016.
- 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.
- Smith, H.O., Hutchison, C.A., Pfannkoch, C. III and Venter, J.C. (2003) Generating a synthetic genome by whole genome assembly: ϕ X174 bacteriophage from synthetic oligonucleotides. *Proc. Natl Acad. Sci. USA*, **100**, 15440–15445.
- Tian, J., Gong, H., Sheng, H., Zhou, X., Gulari, E., Gao, X. and Church, G. (2004) Accurate multiplex gene synthesis from programmable DNA microchips. *Nature*, **432**, 1050–1054.
- Shevchuk, N.A., Bryksin, A.V., Nusinovich, Y.A., Cabello, F.C., Sutherland, M. and Ladisch, S. (2004) Construction of long DNA molecules using long PCR-based fusion of several fragments simultaneously. *Nucleic Acids Res.*, **32**, e19.
- Bang, D. and Church, G.M. (2008) Gene synthesis by circular assembly amplification. *Nat. Methods*, **5**, 37–39.
- Wilhelm, J. and Pingoud, A. (2003) Real-time polymerase chain reaction. *ChemBioChem*, **4**, 1120–1128.
- SantaLucia, J. Jr. and Hicks, D. (2004) The thermodynamics of DNA structural motifs. *Annu. Rev. Biophys. Biomol. Struct.*, **33**, 415–440.
- Owczarzy, R., Moreira, B.G., You, Y., Behlke, M.A. and Walder, J.A. (2008) Predicting stability of DNA duplexes in solutions containing magnesium and monovalent cations. *Biochemistry*, **47**, 5336–5353.
- Binkowski, B.F., Richmond, K.E., Kaysen, J., Sussman, M.R. and Belshaw, P.J. (2005) Correcting errors in synthetic DNA through consensus shuffling. *Nucleic Acids Res.*, **33**, e55.
- Saleem, M., Kweon, M.-H., Johnson, J.J., Adhmi, V.M., Elcheva, I., Khan, N., Hafeez, B.B., Bhat, K.M.R., Sarfaraz, S. *et al.* (2007) S100A4 accelerates tumorigenesis and invasion of human prostate cancer through the transcriptional regulation of matrix metalloproteinase 9. *Proc. Natl Acad. Sci. USA*, **103**, 14825–14830.
- Wittwer, C.T., Herrmann, M.G., Moss, A.A. and Rasmussen, R.P. (1997) Continuous fluorescence monitoring of rapid cycle DNA amplification. *BioTechniques*, **22**, 130–138.
- Ye, H., Huang, M.C., Li, M.-H. and Ying, J.Y. (2009) Experimental analysis of gene assembly with TopDown one-step real-time gene synthesis. *Nucleic Acids Res.*, **37**, e51.
- Ririe, K.M., Rasmussen, R.P. and Wittwer, C.T. (1997) Product differentiation by analysis of DNA melting curves during the polymerase chain reaction. *Anal. Biochem.*, **245**, 154–160.
- Gibson, D.G., Benders, G.A., Andrews-Pfannkoch, C., Denisova, E.A., Baden-Tillson, H., Zaveri, J., Stockwell, T.B., Brownley, A., Thomas, D.W., Algire, M.A. *et al.* (2008) Complete chemical synthesis, assembly, and cloning of a *Mycoplasma genitalium* genome. *Science*, **319**, 1215–1220.
- Kodumal, S.J., Patel, K.G., Reid, R., Menzella, H.G., Welch, M. and Santi, D.V. (2004) Total synthesis of long DNA sequences: Synthesis of a contiguous 32-kb polyketide synthase gene cluster. *Proc. Natl Acad. Sci. USA*, **101**, 15573–15578.
- Cox, J.C., Lape, J., Sayed, M.A. and Hellinga, H.W. (2007) Protein fabrication automation. *Protein Sci.*, **16**, 379–390.
- Kong, D.S., Carr, P.A., Chen, L., Zhang, S. and Jacobson, J.M. (2007) Parallel gene synthesis in a microfluidic device. *Nucleic Acids Res.*, **35**, e61.
- Richmond, K.E., Li, M.-H., Rodesch, M.J., Patel, M., Lowe, A.M., Kim, C., Chu, L.L., Venkataramaiah, N., Flickinger, S.F., Kaysen, J. *et al.* (2004) Amplification and assembly of chip eluted DNA (AACED): a method for high throughput gene synthesis. *Nucleic Acids Res.*, **32**, 5011–5018.
- Carr, P.A., Park, J.S., Lee, Y.J., Yu, T., Zhang, S. and Jacobson, J.M. (2004) Protein-mediated error correction for *de novo* DNA synthesis. *Nucleic Acids Res.*, **32**, e162.

32. Hecker, K.H. and Rill, R.L. (1998) Error analysis of chemically synthesized polynucleotides. *BioTechniques*, **24**, 256–260.
33. Linshiz, G., Yehezkel, T.B., Kaplan, S., Gronau, I., Ravid, S., Adar, R. and Shapiro, E. (2008) Recursive construction of perfect DNA molecules from imperfect oligonucleotides. *Mol. Syst. Biol.*, **4**, 191.
34. Huang, M.C., Ye, H., Kuan, Y.K., Li, M.-H. and Ying, J.Y. (2009) Integrated two-step gene synthesis in a microfluidic device. *Lab. Chip*, **9**, 276–285.
35. Grote, A., Hiller, K., Scheer, M., Münch, R., Nörtemann, B., Hempel, D.C. and Jahn, D. (2005) JCat: a novel tool to adapt codon usage of a target gene to its potential expression host. *Nucleic Acids Res.*, **33**, W526–W531.
36. Puigbò, P., Guzmán, E., Romeu, A. and Garcia-Vallvé, S. (2007) OPTIMIZER: a web server for optimizing the codon usage of DNA sequences. *Nucleic Acids Res.*, **35**, W126–W131.