W176–W180 *Nucleic Acids Research, 2004, Vol. 32, Web Server issue*
DOI: 10.1093/nar/gkh401

Gene2Oligo: oligonucleotide design for *in vitro* gene synthesis

Jean-Marie Rouillard*, Woonghee Lee, Gilles Truan[1], Xiaolian Gao[2], Xiaochuan Zhou[3] and Erdogan Gulari

Department of Chemical Engineering, University of Michigan, Ann Arbor, MI, USA, [1]CGM-CNRS, Gif-sur-Yvette, France, [2]Chemistry Department, University of Houston, Houston, TX, USA and [3]Atactic Technologies Inc., Houston, TX, USA

Received February 20, 2004; Revised and Accepted March 25, 2004

ABSTRACT

There is substantial interest in implementing a bioinformatics tool that allows the design of oligonucleotides to support the development of *in vitro* gene synthesis. Current protocols to make long synthetic DNA molecules rely on the *in vitro* assembly of a set of short oligonucleotides, either by ligase chain reaction (LCR) or by assembly PCR. Ideally, such oligonucleotides should represent both strands of the final DNA molecule. They should be adjacent on the same strand and overlap the complementary oligonucleotides from the second strand to ensure good hybridization during assembly. This implies that the thermodynamic properties of each oligonucleotide have to be consistent across the set. Furthermore, any given oligonucleotide has to be totally specific to its target to avoid the creation of incorrectly assembled sequences. We have developed Gene2Oligo (http://berry.engin.umich.edu/gene2oligo/), a web-based tool that divides a long input DNA sequence into a set of adjacent oligonucleotides representing both DNA strands. The length of the oligonucleotides is dynamically optimized to ensure both the specificity and the uniform melting temperatures necessary for *in vitro* gene synthesis. We have successfully designed and used a set of oligonucleotides to synthesize the *Saccharomyces cerevisiae* cytochrome b5 by using both LCR and assembly PCR.

INTRODUCTION

There are many applications for the *in vitro* synthesis of long DNA molecules ranging from the introduction of point mutations or restriction sites into existing sequences to the *de novo* creation of engineered genes. Advances in the technology of oligonucleotide parallel synthesis (1) and purification make possible large-scale gene synthesis at low cost. One can now envisage synthesizing not only a new gene, but also a whole library of genes or mutants of a single gene for bioengineering applications, structural studies, drug development, combinatorial biology and so on.

Various methods have been proposed over the last three decades for *in vitro* gene synthesis, including oligonucleotide ligation (2–5), the '*Fok* I method' (6), DNA shuffling (7–9) and PCR (10–12). Among them, assembly PCR (11,12) and ligase chain reaction (LCR) coupled to PCR (5) are most suitable for large-scale gene synthesis starting from short synthetic oligonucleotides.

Most of these publications do not emphasize the oligonucleotide design and rely on oligonucleotides of the same length without any attempt to optimize them to enhance the accuracy of the assembly. While this is valid when a single sequence is considered for synthesis and one can afford the time to optimize the assembly protocol, it is highly probable that a poor oligonucleotide design will lead to many failures during parallel gene synthesis. Hoover and Lubkowski have developed DNAWorks, an automated method for designing optimized oligonucleotides for PCR-based gene synthesis (11). This software accepts either DNA or protein sequences as input. In the latter case, it designs optimized oligonucleotides matching the codon bias of the chosen host for expression.

There are two major parameters to consider when designing oligonucleotides for gene synthesis by LCR or assembly PCR: first, all the oligonucleotides should share similar thermodynamic properties (i.e. melting temperature) to ensure uniform hybridization during assembly; second, a given oligonucleotide should be highly specific to its target to avoid incorrect assembly. Here, we present Gene2Oligo, a web-based tool to design optimized oligonucleotides for *in vitro* gene assembly by LCR or PCR. By dynamically choosing the length of the

*To whom correspondence should be addressed. Tel: +1 734 764 0111; Fax: +1 734 763 0459; Email: jmrouill@umich.edu

The online version of this article has been published under an open access model. Users are entitled to use, reproduce, disseminate, or display the open access version of this article provided that: the original authorship is properly and fully attributed; the Journal and Oxford University Press are attributed as the original place of publication with the correct citation details given; if an article is subsequently reproduced or disseminated not in its entirety but only in part or as a derivative work this must be clearly indicated.

© 2004, the authors

oligonucleotide, this software ensures both the specificity and the melting temperature uniformity necessary for *in vitro* gene synthesis. For maximum versatility, this tool accepts DNA input sequences.

MATERIALS AND METHODS

All of the oligonucleotides were obtained from IDT. A list of the sequences can be obtained by repeating the design described in the Application section.

Assembly PCR

The assembly PCR was carried out in a volume of 50 µl containing 200 µM of each dNTP, 0.1 µM of each oligonucleotide, 0.5 U of *Taq* DNA polymerase (Promega) or 1 U of *Vent* DNA polymerase (NEB) in a buffer containing 2.5 mM $MgCl_2$. Assembly PCR was conducted as follows: denaturation at 95°C for 2 min followed by 30 cycles at 95°C for 30 s, 52°C for 30 s and 72°C for 1 min, and terminated by an incubation at 72°C for 5 min.

LCR

The LCR was carried out in a final volume of 25 µl containing 0.1 µM of each oligonucleotide phosphorylated using the T4 Polynucleotide Kinase (NEB), and 10 U of *Taq* DNA Ligase (NEB). LCR was conducted as follows: 94°C for 2 min; 40 cycles at 94°C for 30 s, 51°C for 4 min.

Second PCR

The full-length product from LCR or assembly PCR reactions was PCR amplified using primers yb5-EcoRI-5′, 5′-GCGCGCGAATTCCATGCCTAAAGTTTACAGTTACCA-AGAAGTTGC-3′ and yb5-SacI-3′, 5′-GCGCGCGAGCTCT-TATTCGTTCAAC-AAATAATAAGCAACACCTAG-3′ in a 50 µl reaction containing 3 µl of assembly PCR mix or 5 µl of LCR mix, 200 µM of each dNTP, 1 µM of each primer, 0.5 U of *Taq* DNA polymerase (Promega) or 1 U of *Vent* DNA polymerase (NEB) and a buffer containing 2.5 mM of $MgCl_2$. PCR was conducted as follows: denaturation at 95°C for 2 min followed by 30 cycles at 95°C for 30 s, 55°C for 30 s and 72°C for 1 min, and terminated by an incubation at 72°C for 5 min.

ALGORITHM

We offer three different methods to partition the input sequence into oligonucleotides. The first one emphasizes the length of the oligonucleotides, the second one gives priority to the melting temperature of the oligonucleotides and the third one only allows a simple regular cut of the input sequence. All of these methods are described in detail below. They all share some definitions and algorithms.

The input sequence will be considered as the forward (or positive) strand (see Figure 1). Its complementary strand will be called the reverse (or negative) strand. On a given strand, all oligonucleotides should be exactly adjacent (no gap between two consecutive oligonucleotides). While this is not necessary for assembly PCR, it is indispensable for LCR. A given oligonucleotide will overlap two oligonucleotides from the complementary strand; thus a single oligonucleotide could be seen as the combination of two independent sequences that we call hybridization units. The melting temperatures (T_m) of the hybridization units are computed using the nearest-neighbor model with corrections based on salt and DNA concentration (13).

Sequence processing

Since the oligonucleotide size is dynamically modified during the design, the first step is to add short header and trail sequences to the input sequence. This gives a higher degree of freedom when selecting the oligonucleotides. The first and last oligonucleotides do not need to exactly start and finish at both ends of the input sequence. These extra sequences can be removed during the final PCR amplification by the careful selection of PCR primers.

Gene2Oligo can design oligonucleotides within a range of ±4 nt around a median size. We call s this median size and l the total length of the input sequence including header and tail. The program first computes a matrix of T_m for each position of the input sequence ranging from 0 to $l - (s - 4)$ and for hybridization unit lengths ranging from $s - 4$ to $s + 4$. Then it computes the mean of all of these T_m to get $T_{m\text{-}avg}$. This will be the default value in the absence of user-defined T_m.

For each position ranging from 0 to $l - (s + 4)$ and for both the forward and reverse strands of the input sequence, Gene2Oligo generates a collection of substrings of length $s + 4$ that

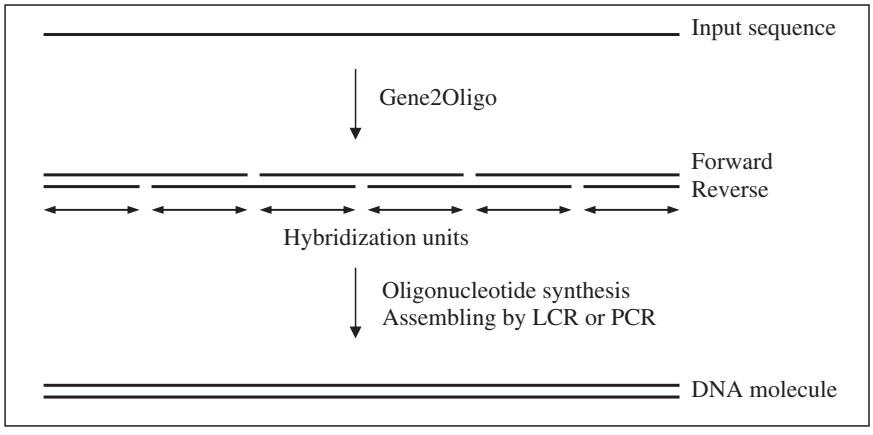

Figure 1. An overview of the experimental process from the input sequence to the final *in vitro* synthesized gene.

will be used to create a blast database. The entire input sequence is then blasted against this database using WU-Blast (14,15) to search for sequence similarity. WU-Blast is tuned as follows: E5000 V1 W5 B100000 W5 gapS2 = 1 S2 = 1 hspmax = 0 warnings matrix = GT. The first eight options are set for reporting any alignments spanning at least five consecutive nucleotides while not reporting single line descriptions that are not relevant to our case. The warnings option is to suppress warning messages. The GT matrix allows substitution of G by A and T by C with no extra weight given to stable GT pairing in DNA (16). The blast output is parsed to detect oligonucleotides matching more than one region of the input sequence. This is done to ensure the specificity of a candidate oligonucleotide to its target. T_m is computed for each putative cross-hybridization for any oligonucleotide size ranging from $s-4$ to $s+4$. If this T_m is above a threshold set at $T_{m\text{-}avg}-20$, then this candidate is considered to be non-specific and reported as such in the specificity matrix.

Then, Gene2Oligo builds a length matrix by parsing both the T_m and the specificity matrices. For each position of the input sequence ranging from 0 to $l-(s-4)$ the program selects two oligonucleotides satisfying all the following conditions:

- minimum$|T_m - T_{m\text{-}avg}|$,
- $T_m > T_{m\text{-}avg} - dT_m$,
- $T_m < T_{m\text{-}avg} + dT_m$,

where dT_m is the T_m range defined by the user (or set to four by default). If these candidates are not reported as excluded in the specificity matrix, then their sizes are stored in the length matrix. If no oligonucleotide is found at a given position, then this position is flagged as non-valid in the length matrix.

The final selection of the oligonucleotides consists of the analysis of the length matrix, which could be seen as a tree. The reading starts at the position corresponding to the first nucleotide of the sequence entered in the interface, excluding the header sequence. The first oligonucleotide length is read and used to point to the starting position of the next candidate. The length of the second candidate determines the starting position of the third candidate, and so forth until the end of the input sequence is reached or a non-valid position is encountered. In the latter case, the reader goes one step back and picks the second-best oligonucleotide size and tests whether it leads to a valid solution. If not, the reader will take another step back to explore a new branch of the tree. The reader can go back all the way to its starting position and so explore all possible branches of the tree. In the worst case, where no valid set of oligonucleotides is found, the process restarts from the first base of the header sequence (position −1 relative to the beginning of the user's input sequence). This iteration is repeated until a complete set of oligonucleotides is found or the end of the header sequence is reached. The latter represents the case where no solution exists to satisfy the design criteria.

Length priority mode

In the length priority mode, one can choose the median size for the hybridization unit. By default, the software will use the $T_{m\text{-}avg}$ calculated for the given median size and a dT_m of ±4°C to search for a set of oligonucleotides satisfying all the conditions described above. For higher flexibility, it is also possible to force the program to use user-defined values for $T_{m\text{-}avg}$ and dT_m.

Melting temperature priority mode

In this mode, one can set a value for $T_{m\text{-}avg}$ and dT_m. The program will automatically compute which median hybridization unit size will lead to the closest T_m to the input T_m. Then all computation will be the same as described above. This could be seen as the length priority mode using the optimal median size for this T_m.

Basic cutting mode

We also offer a tool to chop down the sequence into oligonucleotides of equal length. There is no optimization in this mode to avoid non-specific hybridization between different oligonucleotides or to ensure a good uniformity of T_m. No header sequence will be added, and only a short tail sequence will be used to get the correct size for the last oligonucleotide.

Implementation

Gene2Oligo was developed using the Java programming language. It is controlled via an HTML form used to choose between the different design modes and options, and to enter the input sequence. The address of the web server is http://berry.engin.umich.edu/gene2oligo.

USAGE

Most of the parameters on the web interface are self-explanatory. The user is asked to provide a name for the sequence and the sequence itself, oriented from 5′ to 3′. Then, there is a choice between the three design modes. For inexperienced users, we strongly recommend using the length priority mode with the software-optimized T_m or the T_m priority mode if there is a need to focus on a special temperature. Setting both hybridization unit size and T_m in the length priority mode requires special attention to avoid selecting mutually exclusive values. For example, there is no possibility of getting any oligonucleotides if the temperature selected is too high or too low for a given oligonucleotide size (e.g. a length of 15 nt and a T_m of 80°C are not compatible).

It is also important to allow a certain degree of freedom to the system to succeed in the design. If the T_m range required is too narrow, say ±1°C, there is almost no chance of obtaining a set of oligonucleotides. From our experience, 3–4°C seems to be a good compromise.

The interface also allows one to set DNA and sodium concentrations according to the experimental conditions used during assembly. Changing these parameters will affect the T_m computation. There is no guarantee of obtaining the same set of oligonucleotides when using different DNA or salt concentrations.

APPLICATION

Design of an oligonucleotide set for the yeast cytochrome b5

In order to test and evaluate Gene2Oligo, we have designed a set of oligonucleotides representing the coding region of the *Saccharomyces cerevisiae* cytochrome b5 gene (*CYB5*, GenBank accession no. L22494) (17) with a GCTCATC sequence added to its 3′ end. This gene encodes a 14 kDa membrane-bound hemoprotein involved in the electron transfer to various

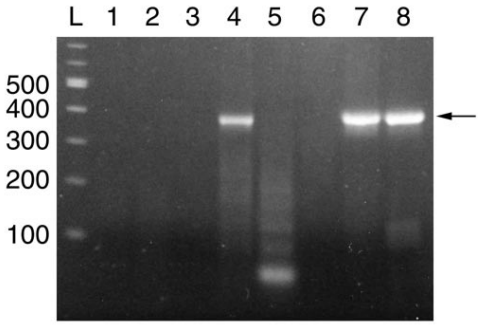

Figure 2. Analysis of the synthetic *CYB5* gene. About 10 µl of each sample are analyzed on a 1% agarose gel. Lanes: (L) 100 bp DNA ladder (NEB); (1) negative control, second PCR without template DNA; (2) LCR product before second PCR; (3) second PCR without primer after LCR; (4) second PCR after LCR; (5) assembly PCR product before second PCR; (6) second PCR without primer after assembly PCR; (7) second PCR after assembly PCR; and (8) positive control, second PCR using yeast genomic DNA as template. The arrow indicates the 388 bp *CYB5* PCR product.

acceptors including cytochrome P450 and enzymes involved in lipid biosynthesis. The design was done using the oligonucleotide length priority with a median size of 20 nt and a software-optimized T_m. The sodium and DNA concentrations were left unchanged. The program produced a total of 21 oligonucleotides with hybridization unit sizes ranging from 16 to 22 nt and T_m ranging from 62.5 to 69.6°C.

In vitro synthesis of the yeast cytochrome b5 gene

We have used this set of oligonucleotides in two independent gene synthesis experiments using either LCR or assembly PCR. Both approaches were followed by a PCR, referred to as the second PCR, to amplify the full-length product. Results are shown in Figure 2. As a positive control, Lane 8 shows the expected PCR product when the reaction is carried out with yeast genomic DNA as template. After LCR only, there is no detectable product in a volume equal to the one used as template for the second PCR (Lane 2). The second PCR amplifies a fragment of the expected size from the LCR product (Lane 4), while a negative control second PCR carried out without primer (Lane 3) shows no amplification, demonstrating that the amplification was due to the PCR primers and not to the oligonucleotides present in the LCR mix. We obtained comparable results with assembly PCR. Lane 5 shows the short products generated by assembly PCR. When used in the second PCR, these oligonucleotides alone are not sufficient to obtain the full-length product (Lane 6). Only the second PCR carried out in the presence of primers leads to a correct amplification (Lane 7). We have also cloned and sequenced the second PCR products amplified from both LCR and assembly PCR. The sequences confirm that all oligonucleotides of the set have been assembled in the correct order. These results are consistent with previously published data (5,11,12).

DISCUSSION

The strategy we have developed to design optimized oligonucleotides for *in vitro* gene synthesis offers several advantages compared with a simpler model where the sequence is just cut into oligonucleotides of the same size. The use of a dynamic hybridization unit size allows Gene2Oligo to select oligonucleotides having a much greater uniformity of thermodynamic properties. This will result in a better hybridization during assembling.

Oligonucleotide synthesis is prone to error. While it is relatively easy to purify an oligonucleotide of the correct length from shorter abortive molecules, it becomes much more difficult to exclude molecules having nucleotide substitutions. These mutations will lead to a mismatch during hybridization, which will destabilize the hybrid DNA. By having a uniform T_m across the oligonucleotide set, one can perform LCR or assembly PCR at a temperature closer to the mean T_m without affecting too much the hybrids having the lower T_m. At a higher temperature, the mismatches have a stronger destabilizing effect which will diminish the probability of incorporation of the mutated oligonucleotides into a final product.

Gene2Oligo designs oligonucleotides specific to a single target. This is of particular interest when the sequence to be synthesized contains repeated regions. The algorithm pays particular attention to the avoidance of hybridization of two oligonucleotides from different regions. This is true to a certain limit. Indeed, in the worst possible case of exact repeats spanning more than several times the length of the hybridization unit, it will be difficult to find specific oligonucleotides. One solution to circumvent this problem is to divide the sequence into short segments, design a set of oligonucleotides for each segment, assemble them *in vitro* in parallel and combine all these segments in the second PCR for final amplification. In this case, the second PCR acts also as an assembly PCR.

This strategy can be applied to the synthesis of long genes. While Stemmer *et al*. reported the synthesis of a 2.7 kb long sequence in a single step (12), it is likely that one will want to verify the synthesis products' sequences at intermediary stages. With current sequencing technologies, it is easier to assemble and sequence fragments of DNA <1 kb. Only the correct fragments will be combined during the final assembly step. For a given error rate during synthesis, the probability of having the right sequence after assembly decreases when the length increases. Thus, subdividing a long input sequence into fragments <1 kb will have several advantages. First, it will increase the probability of obtaining the correct sequence; second, it will increase the chance of designing a more uniform T_m for the oligonucleotide set; and third, it will make possible the parallel synthesis of different genes simultaneously on a single device. If a 4 and 2 kb sequence are reduced to six 1 kb sequences, then it will be possible to process them in parallel until the sequence verification step. Only the final assembly will need to be carried out separately. The possibility of spliting a long sequence into shorter segments will be implemented in a further version of Gene2Oligo.

ACKNOWLEDGEMENTS

This project was supported in part by grants from DARPA, NIH and Michigan Life Sciences Corridor.

REFERENCES

1. Gao,X., LeProust,E., Zhang,H., Srivannavit,O., Gulari,E., Yu,P., Nishiguchi,C., Xiang,Q. and Zhou,X. (2001) A flexible light-directed

DNA chip synthesis gated by deprotection using solution photogenerated acids. *Nucleic Acids Res*., **29**, 4744–4750.
2. Goeddel,D.V., Kleid,D.G., Bolivar,F., Heyneker,H.L., Yansura,D.G., Crea,R., Hirose,T., Kraszewski,A., Itakura,K. and Riggs,A.D. (1979) Expression in *Escherichia coli* of chemically synthesized genes for human insulin. *Proc. Natl Acad. Sci., USA*, **76**, 106–110.
3. Heyneker,H.L., Shine,J., Goodman,H.M., Boyer,H.W., Rosenberg,J., Dickerson,R.E., Narang,S.A., Itakura,K., Lin,S. and Riggs,A.D. (1976) Synthetic lac operator DNA is functional *in vivo*. *Nature*, **263**, 748–752.
4. 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.
5. 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.
6. Mandecki,W. and Bolling,T.J. (1988) FokI method of gene synthesis. *Gene*, **68**, 101–107.
7. Crameri,A. and Stemmer,W.P. (1995) Combinatorial multiple cassette mutagenesis creates all the permutations of mutant and wild-type sequences. *Biotechniques*, **18**, 194–196.
8. Stemmer,W.P. (1994) DNA shuffling by random fragmentation and reassembly: in vitro recombination for molecular evolution. *Proc. Natl Acad. Sci., USA*, **91**, 10747–10751.
9. Stemmer,W.P. (1994) Rapid evolution of a protein *in vitro* by DNA shuffling. *Nature*, **370**, 389–391.
10. Dillon,P.J. and Rosen,C.A. (1990) A rapid method for the construction of synthetic genes using the polymerase chain reaction. *Biotechniques*, **9**, 298–300.
11. Hoover,D.M. and Lubkowski,J. (2002) DNAWorks: an automated method for designing oligonucleotides for PCR-based gene synthesis. *Nucleic Acids Res*., **30**, e43.
12. 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.
13. SantaLucia,J.,Jr. (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl Acad. Sci., USA*, **95**, 1460–1465.
14. Gish,W. and States,D.J. (1993) Identification of protein coding regions by database similarity search. *Nature Genet*., **3**, 266–272.
15. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.
16. Allawi,H.T. and SantaLucia,J.,Jr (1998) NMR solution structure of a DNA dodecamer containing single G*T mismatches. *Nucleic Acids Res*., **26**, 4925–4934.
17. Truan,G., Epinat,J.C., Rougeulle,C., Cullin,C. and Pompon,D. (1994) Cloning and characterization of a yeast cytochrome b5-encoding gene which suppresses ketoconazole hypersensitivity in a NADPH-P-450 reductase-deficient strain. *Gene*, **142**, 123–127.