frontiers in Genetics

METHODS
published: 17 March 2022
doi: 10.3389/fgene.2022.836108

An Integrated Algorithm for Designing Oligodeoxynucleotides for Gene Synthesis

Gang Fang and Hanjie Liang*

Institute of Computing Science and Technology, Guangzhou University, Guangzhou, China

The design and construction of large synthetic genes can be a slow, difficult, and confusing process, especially in the key step of oligodeoxynucleotide design. Herein we present an integrated algorithm to design oligonucleotide sets for gene synthesis by both ligase chain reaction and polymerase chain reaction. It offers much flexibility with no constraints on the gene to be synthesized. Firstly, it divides the long-input DNA sequence by a greedy algorithm based on the length of the oligodeoxynucleotide overlap region. Secondly, it tunes the length of the overlap region iteratively in an attempt to minimize the melting temperature variance of overlap. Thirdly, dynamic programming algorithm is used to achieve the uniform melting temperature of the oligodeoxynucleotide overlaps. Finally, the oligodeoxynucleotides with homologous melting temperature necessary for ligase chain reaction-based or two-step assembly PCR-based synthesis of the desired gene are outputted.

Keywords: gene synthesis, algorithm, oligodeoxynucleotide design, melting temperature, assembly

OPEN ACCESS

Edited by:
Youri I. Pavlov,
University of Nebraska Medical Center, United States

Reviewed by:
Hao Lin,
University of Electronic Science and Technology of China, China
Jianhua Xiao,
Nankai University, China

****Correspondence:***
Gang Fang
gangf@gzhu.edu.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *21 December 2021*
Accepted: *07 February 2022*
Published: *17 March 2022*

Citation:
Fang G and Liang H (2022) An Integrated Algorithm for Designing Oligodeoxynucleotides for Gene Synthesis. Front. Genet. 13:836108. doi: 10.3389/fgene.2022.836108

INTRODUCTION

Nowadays, gene synthesis is one of the most important biological technologies that can be used in the field of genome studies, gene expression studies, gene network studies, *etc*. This modern gene synthesis technique can synthesize a whole eukaryotic genome, and current gene synthesis methods rely on the use of overlapped oligonucleotides to construct large genes by ligase chain reaction (LCR) (Au et al., 1998) and polymerase chain reaction (PCR) (Stemmer et al., 1995). Algorithms and computer programs have been developed for gene synthesis to automatically design oligonucleotides to minimize the error of assembly and to optimize the LCR or PCR process. Programs such as TmPrime and DNAWorks, based on iteration algorithm, have been developed for gapless PCR assembly (David and Jacek, 2002; Marcus et al., 2009). Other programs, such as Gene2Oligo, Assembly PCR Oligo Maker, GeneDesign, and GeMS, also mainly based on iteration algorithm, have been developed for gap PCR assembly (Jean-Marie et al., 2004; Roman et al., 2005; Sebastian et al., 2005; Sarah et al., 2006). In the key step of oligodeoxynucleotide design, all the algorithms carried out in the programs will divide the input gene sequences into oligonucleotides with a homologous melting temperature, and the corresponding overlaps of these oligonucleotides possess uniform melting temperatures. The best result of these programs, which is attained by TmPrime, is less than 3°C in deviation of melting temperature (Marcus et al., 2009), but in optimization theory, it is not always the best solution to this sort of problem (Cormen et al., 2001). In order to prove this and minimize the error of assembly in gene synthesis, herein we present an integrated algorithm to solve this problem and attain a better result.

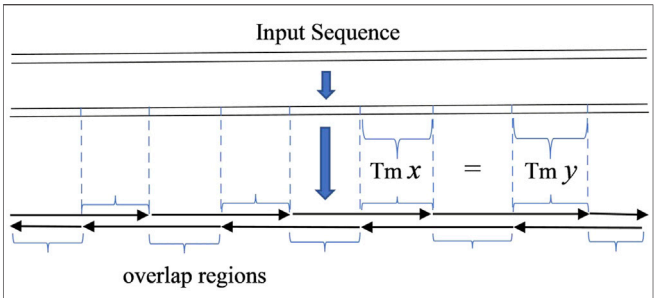

FIGURE 1 | Scheme of gapless PCR or ligase chain reaction assembly. The input sequence is regarded as the serial connection of the overlap regions of oligonucleotides.

In the key step of oligodeoxynucleotide design in gene synthesis for gapless PCR or LCR assembly, all oligonucleotides are designed to be exactly adjacent, with no gap between two consecutive oligonucleotides. The given sequence can be seen as the serial connection of all overlapping regions of oligonucleotides. With this simple observation, the problem of designing oligonucleotides with a uniform melting temperature in overlaps will be equivalent to dividing the given sequence into segments with a homologous melting temperature, with each segment representing an overlapping region (Marcus et al., 2009) (**Figure 1**).

In gapped PCR assembly, oligonucleotides are adjacent, with few base deletions (gaps) between two consecutive oligonucleotides. Compared to gapless assembly, gapped assembly can be more flexible and economical, with an insignificant rise in assembly errors (Xiong et al., 2004), though gapped assembly can only be carried out by PCR. In this paper, the presented algorithm can output oligonucleotides with a homologous melting temperature, not only for gapless assembly but also for gapped assembly.

MATERIALS AND METHODS

Based on above-mentioned simple observation, the presented integrated algorithm firstly starts from a greedy algorithm. The greedy algorithm always selects the apparent best result in every step, without considering other options. Thus, it cannot always obtain the best solution to the problem although it is a fast algorithm (Cormen et al., 2001). Secondly, the result obtained from the initial greedy algorithm is optimized by iteration algorithm in an attempt to minimize the melting temperature variance of overlap regions. In this step, the best solution is also not guaranteed. In order to decrease the melting temperature variance of the overlap regions further, a dynamic programming algorithm is carried out. This dynamic algorithm, which is adapted from Viterbi algorithm, has been successfully fulfilled to solve other optimization problems in synthetic biology (Fang et al., 2017).

The initial greedy algorithm processes the input DNA sequence by dividing the sequence into segments with a similar melting temperature. It cuts down the first segment whose length is from 20 to 30 bp and then cuts down the second consecutive segment whose length is also from 20 to 30 bp. The melting temperature of these segments was computed, and the segment combination with the least deviation in melting temperature was selected. In this way, the first two segments are determined. In **Table 1** the greedy algorithm is depicted in detail. Melting temperature is computed by using the nearest-neighbor model with SantaLucia's thermodynamic parameter (SantaLucia and Hicks, 2004), corrected with salt and oligonucleotide concentrations, and the total number of phosphates in the duplex (Owczarzy et al., 2008).

The greedy algorithm divides the input sequence into serial connections of segments with an approximately equal melting temperature. In theory, the algorithm cannot guarantee the best result. In an attempt to reduce the melting temperature variance

TABLE 1 | The Greedy algorithm.

Greedy algorithm for dividing the input sequence into segments with approximately equal melting temperature
Input: DNA sequence (e.g., 400—1000bp)
Output: Serial connections of segments with approximately equal melting temperature
1 Cut the first segment with length range from 20 to 30bp.
2 Cut the second consecutive segment with length range from 20 to 30bp. Compute melting temperature values of all these segments and select the combination with the least deviation in melting temperature.
3 Cut the next consecutive segment with length range from 20 to 30bp. Compute melting temperature values of the segments. Combine with the result obtained from last step and select the combination with the least deviation in melting temperature.
4 Repeat 3 until the sequence terminate.
Return the result.

TABLE 2 | The iteration algorithm.

Iteration algorithm to reduce the melting temperature variance of the serial connections of segments
Input: Serial connections of segments, the number of segments *n*. Output: Serial connections of segments with less deviation in melting temperature. *the least deviation = the deviation of all segments* **while Ture**:   **for** *i* **in range** (*n*):   Shift the *ith* boundary of the connections (1--4bp) and compute the deviation in melting temperature of all the segments   **if** *the newly computed deviation < the least deviation*:   *the least deviation = the newly computed deviation*   determine the new boundary of the connection with newly computed deviation   **if** *abs (the least deviation – the newly computed deviation)* <= 0.001:   **break** **Return** the result.

TABLE 3 | The dynamic programming algorithm.

Dynamic programming algorithm to minimize the melting temperature variance of the overlap regions
Input: Serial connections of segments obtained by iteration algorithm, the number of segments *n*. Output: Segments of overlap regions with less deviation in melting temperature. *the least deviation = 9999* **for** *i* **in range** (*1 to n*): /* To produce gap between segments, will produce n columns of segments */   Shrink *ith* segment 0--5bp in both ends respectively and deposit them in columns /* totally produce 6*6*n segments */ **for** *j* **in range** (*1 to n*):   **for** *k* **in range** (*1 to 36*):   Combine *kth* segment in *(j+2)th* column with every segment in *(j+1)th* and *jth* column and compute their deviation in melting temperature   **if** *the newly computed deviation < the least deviation*:   *the least deviation = the newly computed deviation*   select the combination with *the least deviation*   keep the address of the segments that consist of the combination recall from the segment which is in the least deviation combination in last column keep the trace of these segments **Return** the trace

of these serial connections of segments, iteration algorithm is carried out (**Table 2**).

After the iteration algorithm, a better result will be given. The result can be used to produce oligodeoxynucleotides for LCR or gapless PCR assembly. In previous study, the PKB2 gene is selected for synthesis based on the reported difficulty of assembly *via* PCR (Gao et al., 2003). Compared to the oligodeoxynucleotide set that TmPrime produces for *Escherichia coli* codon-optimized PKB2 gene, the integrated algorithm presented in this paper can produce

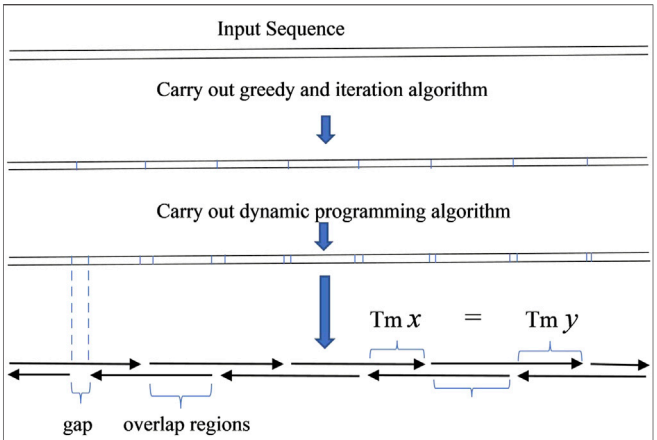

FIGURE 2 | Scheme of gapped PCR assembly.

TABLE 4 | Compared to the three genes designed by TmPrime, the oligonucleotide set designed by our algorithm possesses the least SD of overlap and oligonucleotides in Tm value (all the oligonucleotides were designed under the same conditions).

Gene	S100A4 (752 bp)		PKB2 (1,446 bp)		GFPuv (760 bp)	
	Overlap	Oligos	Overlap	Oligos	Overlap	Oligos
TmPrime	1.0980	1.5680	1.2004	1.7346	0.8660	0.9887
Ours (gapped)	0.2660	0.9690	0.4944	1.4888	0.1686	0.2796
Ours (gapless)	0.3890	1.0230	0.6860	1.5102	0.1890	0.3677

oligodeoxynucleotide sets for gapless PCR assembly with more uniform melting temperatures. The deviation in the melting temperature of oligodeoxynucleotide overlaps produced in this step by our algorithm is 0.6860 compared to 1.2004 in that produced by TmPrime (**Supplementary Data S1**).

In order to reduce the deviation in melting temperature of the overlap region further and produce more uniform oligodeoxynucleotide sets, a dynamic programming algorithm is employed. This type of algorithm is universally used in bioscience and other fields (Mount, 2001; Viterbi, 2006). In **Table 3**, the dynamic programming algorithm used to minimize the deviation in the melting temperature of oligodeoxynucleotide overlap regions is depicted in detail.

The result obtained from this algorithm can be used to produce oligodeoxynucleotides for gapped PCR assembly (**Figure 2**). Compared to the oligodeoxynucleotide sets that TmPrime produces for *E. coli* codon-optimized PKB2 gene, the result obtained by the dynamic programming algorithm can produce oligodeoxynucleotide sets for gapped PCR assembly with more uniform melting temperatures. The deviation in the melting temperature of oligodeoxynucleotide overlaps produced by the dynamic programming algorithm is 0.4944 compared to 1.2004 in that produced by TmPrime and 0.6860 by iteration algorithm (**Supplementary Data S1**). The presented integrated algorithm can produce oligodeoxynucleotide sets not only for gapless assembly but also for gapped assembly. If gapless assembly is needed, the result outputted by iteration algorithm can be used. If gapped assembly is needed, the step of iteration algorithm can be omitted, and the greedy algorithm and dynamic programming algorithm can just be integrated. In this way, the complexity of the whole algorithm can be decreased.

In order to produce an odd number of serial connections of segments, a small tail is added to the input sequence in some circumstances. This guarantees the imbricated structure of oligodeoxynucleotides as shown in diagrams in **Figures 1**, **2**. The added tail can be eliminated in a PCR reaction by using particular primers.

RESULTS

When designing oligodeoxynucleotides for gene synthesis, the uniformity of oligo melting temperature especially in an overlap region is the key factor that should be considered. The oligodeoxynucleotide sets designed by the integrated algorithm possess the least SD in overlap melting temperature and can be used for gapless and gapped assembly (**Table 4**). The SD in melting temperature of the designed oligodeoxynucleotides is also less than what another program produces (**Table 4**). The deviation in melting temperature of oligodeoxynucleotides for the gapless assembly of PKB2 produced by iteration algorithm is 1.5102 compared to 1.7346 in that produced by TmPrime, and the deviation in melting temperature of oligodeoxynucleotides for gapped assembly is 1.4888. Dynamic programming will shrink each end of every fragment to produce candidate fragment columns. In fact, the final oligodeoxynucleotides for gapped assembly are adjacent, with a few base deletions (gaps) or with no gap between two consecutive oligonucleotides. This result is produced by the inherent property of dynamic programming algorithm. One can adjust this result by changing the number of bases to be shrunk in the first *for* loop of dynamic programming algorithm (refer to **Supplementary Data S1** for more information). The algorithm is written in Python 3.7. The process of designing oligodeoxynucleotide sets for a multi-kilobase (<3 kb) gene takes less than 10 s when it is run on a Lenovo computer with dual 3.3-GHz Intel Xeons and 4 GB of RAM.

DISCUSSION

The integrated algorithm presented in this paper is fast and flexible, with no constraints on the input gene. It can design oligodeoxynucleotides for gene synthesis according to its inherent property. The annealing temperature of the assembly can be determined by the average melting temperature of the final oligodeoxynucleotide sets. The algorithm complexity of greedy algorithm is $O(n^2)$ if the first two segments contain n options, respectively. The complexity of iteration algorithm is difficult to determine, but one can change the threshold of the difference between *the least deviation* and *the newly computed deviation* for a rapid convergence. According to our experience, the threshold that is set to 0.001 will make a rapid convergence. If the length of a serial connection of segments is L and the number of potential segments in a column is N, the algorithm complexity of

dynamic programming algorithm will be $O(LN^3)$. On the contrary, the algorithm complexity of exhaustive algorithm is $O(N^L)$. The time complexity of dynamic programming algorithm can be more than $O(LN^3)$, but for the reason of feasibility, we just consider the combination of three columns in the algorithm, which caused the time complexity to be $O(LN^3)$ (**Table 3**). It is quite clear that the algorithm is technically feasible and can be run on an ordinary computing platform. Based on the simple observation, a greedy algorithm and iteration algorithm can be fulfilled to produce the oligodeoxynucleotides for gapless assembly. Despite the fact that a better result is obtained, in theory, the best solution to this sort of optimization problem cannot be guaranteed, and the best solution is extremely difficult to obtain. When dynamic programming algorithm is carried out after greedy and iteration algorithm, a better result than before is obtained. It can produce oligodeoxynucleotides with a more uniform melting temperature, and this will decrease the error of assembly, although it is only used for gapped PCR assembly. Compared to gapless assembly, gapped assembly may cause a rise in assembly error. The oligodeoxynucleotides designed by a dynamic programming algorithm possess a more uniform melting temperature, and this property can compensate the influence caused by the gaps between consecutive oligodeoxynucleotides. Furthermore, the number and the location of gaps can be adjusted by changing the number and the location of the bases to be shrunk in the first *for* loop of dynamic programming algorithm. The length of the oligodeoxynucleotides can be adjusted by changing the base number (length) of the segment to be cut in the beginning of the greedy algorithm (in this case, the base number to be cut is 20–30 bp, and it will produce an overlap length ranging from 20 to 30 bp, while the oligodeoxynucleotide length will approximately range from 40 to 60 bp. The dynamic programming algorithm is the core of this integrated algorithm. The first two algorithms just divide the input sequence approximately. The dynamic programming algorithm guarantees the uniformity of the melting temperature of the oligodeoxynucleotides that are produced for gene synthesis. Dynamic programming is universally used in optimization theory and solves a few optimization problems (Kececioglu and Myers, 1995). Its universality and feasibility lead to its application in the field of biology, especially in synthetic biology (Fang et al., 2017). This integrated algorithm will be simplified and written into a computer program, and a webserver will be built soon to facilitate the gene synthesis. Except the synchronization of the temperature of the overlap region, it is necessary to take into account other parameters, such as the presence of repeats, regions with high CG, *etc*. These factors will cause the formation of unwanted secondary structures. These problems have been successfully solved by FastPCR (Kalendar et al., 2011). In this paper, we emphasized the minimization of the melting temperature variance in the overlap region. We think that this can decrease the formation of unwanted secondary structures even if these cannot be eliminated. We do not consider other factors because we think that the minimization of the melting temperature variance in the overlap region can compensate for these to a certain extent.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

GF designed the study and wrote the paper. HL wrote the code.

FUNDING

This study was supported by the National Natural Science Foundation of China (grant number 61972107).

SUPPLEMENTARY MATERIAL

The supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.836108/full#supplementary-material

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 inEscherichia Coli. *Biochem. Biophysical Res. Commun.* 248, 200–203. doi:10.1006/bbrc.1998.8929

Cormen, T. H., Leieserson, C. L., Rivest, R. L., and Stein, C. (2001). *Introduction to Algorithms*. Cambridge MA: MIT Press.

David, M. H., and Jacek, L. (2002). DNAWorks: an Automated Method for Designing Oligonucleotides for PCR-Based Gene Synthesis. *Nucleic Acids Res.* 30, e43.

Fang, G., Zhang, S., and Dong, Y. (2017). Optimizing DNA Assembly Based on Statistical Language Modelling. *Nucleic Acids Res.* 45, e182. doi:10.1093/nar/gkx859

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. doi:10.1093/nar/gng143

Jean-Marie, R., Woonghee, L., Gilles, T., Xiaolian, G., Xiaochuan, Z., and Erdogan, G. (2004). Gene2Oligo: Oligonucleotide Design for *In Vitro* Gene Synthesis. *Nucleic Acids Res.* 32, W176–W180.

Kalendar, R., Lee, D., and Schulman, A. H. (2011). Java Web Tools for PCR, In Silico PCR, and Oligonucleotide Assembly and Analysis. *Genomics* 98 (2), 137–144. doi:10.1016/j.ygeno.2011.04.009

Kececioglu, J. D., and Myers, E. W. (1995). Combinatorial Algorithms for DNA Sequence Assembly. *Algorithmica* 13, 7–51. doi:10.1007/bf01188580

Marcus, B., Samuel, K., Hongye, Y., Mo-Huang, L., and Jackie, Y. Y. (2009). TmPrime: Fast, Flexible Oligonucleotide Design Software for Gene Synthesis. *Nucleic Acids Res.* 37, W214–W221.

Mount, D. (2001). *Bioinformatics: Sequence and Genome Analysis*. Cold Spring Harbor NY: Cold Spring Harbor Press.

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. doi:10.1021/bi702363u

Roman, R., Sharon, X. Z., and Philip, E. J. (2005). Assembly PCR Oligo Maker: a Tool for Designing Oligodeoxynucleotides for Constructing Long DNA Molecules for RNA Production. *Nucleic Acids Res.* 33, W521–W525.

SantaLucia, J., Jr., and Hicks, D. (2004). The Thermodynamics of DNA Structural Motifs. *Annu. Rev. Biophys. Biomol. Struct.* 33, 415–440. doi:10.1146/annurev.biophys.32.110601.141800

Sarah, M. R., Sarah, J. W., Robert, M. Y., and Boeke1, Jef. D. (2006). GeneDesign: Rapid, Automated Design of Multikilobase Synthetic Genes. *Genome Res.* 16, 550–556.

Sebastian, J., Ralph, R., and Daniel, V. S. (2005). GeMS: an Advanced Software Package for Designing Synthetic Genes. *Nucleic Acids Res.* 33, 3011–3016.

Stemmer, W. P. C., 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. doi:10.1016/0378-1119(95)00511-4

Viterbi, A. J. (2006). A Personal History of the Viterbi Algorithm. *IEEE Signal. Process. Mag.* 23, 120–142. doi:10.1109/msp.2006.1657823

Xiong, A.-S., Yao, Quan-Hong., Peng, Ri-He., Li, Xian., Fan, Hui-Qin., Cheng1, Zong-Ming., et al. (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. doi:10.1093/nar/gnh094

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.