MPrime-DEG: Multiple Degenerate Primer Design for Amplifying Homologous Sequences

Eric C Rouchka, Yamini Latha Rudraraju

Address: Department of Computer Engineering and Computer Science, Speed School of Engineering, University of Louisville, Louisville, Kentucky, USA and Bioinformatics Research Group, University of Louisville, Louisville, Kentucky, USA

Email: Eric C Rouchka – eric.rouchka@louisville.edu; Yamini L Rudraraju – yamini_latha@yahoo.com

ABSTRACT

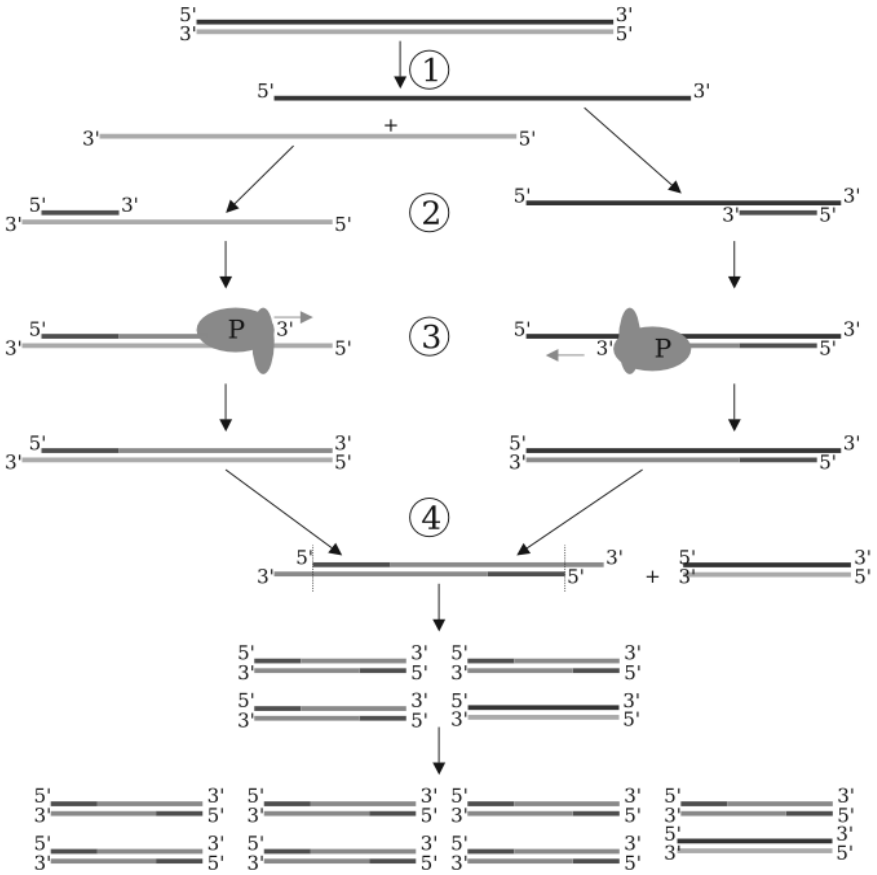
Primers are the most important factors of the Polymerase Chain Reaction (PCR), which is the most widely, used laboratory technique for amplifying DNA sequences. Degenerate primers are mixed PCR primers with different options at several positions in the sequence, which can anneal to and amplify a variety of related sequences. Degenerate primers are widely used when the related genomic sequences are unknown or known only in related organisms. A degenerate primer program was created using the previously developed MPrime software for traditional primer construction. The degenerate primer design program produces all possible primer pairs that together amplify a targeted multiple sequence alignment. The degenerate primer design approach begins with a multiple nucleotide sequence alignment. All the conserved regions are identified and the user selects the regions to amplify based on conserved blocks. The list of all possible degenerate primers that can amplify the region are sent out to the user. The program was written in PERL and is the expansion of the MPrime program which designs large scale multiple primers and oligonucleotides for customized gene microarrays.

The degenerate primer program will be freely available along with MPrime program at http://kbrin.a-bldg.louisville.edu/Tools/MPrime/ for academic and commercial users.

INTRODUCTION

Polymerase chain reaction (PCR) is the widely used technique for creating copies of specific fragments of a DNA sequence. PCR is used in many biological and medical research labs for performing wide variety of tasks such as the detection of hereditary diseases, the identification of genetic fingerprints, the diagnosis of infectious diseases, the cloning of genes, paternity testing, and DNA computing (1).

Primers are the most important factors of a PCR process. Primers are short nucleotide sequences designed to bind to a desired segment of DNA and amplify the target region (1). They are synthetically synthesized oligonucleotides usually shorter than 50 nucleotides (often 18-25 nucleotides) (2). PCR primers are important in polymerase chain reaction, as they are complementary to the beginning and end of the DNA fragment to be amplified. During the PCR annealing cycle, PCR primers anneal to the complementary region of the DNA. DNA polymerase binding and the 3' OH of the oligo allows the synthesis of DNA to occur (1).



Source: http://en.wikipedia.org/wiki/Polymerase_chain_reaction

Fig: DNA amplification in PCR cycle. (1) Denaturing at 94-96°C. (2) Annealing at (eg) 68°C. (3) Elongation at 72°C (P=Polymerase). (4) The first cycle is complete. The two resulting DNA strands make up the template DNA for the next cycle, thus doubling the amount of DNA duplicated for each new cycle.

A PCR process follows in three steps, denaturation, annealing and polymerization (3). In the first step, double helix of the DNA is separated into single strands by heating it at 94°C. In the second step, the temperature is lowered to 54°C to allow primers to anneal with the complementary regions of the single stranded DNA molecules. In the third step, DNA polymerase attaches to the primers at 72°C and starts copying the strands. Good primer design is very essential for successful PCR reactions. There are several properties that are to be carefully considered while designing primers (4). Most important of them are:

- Primer length
- GC clamp at 3' end
- G/C content
- Melting/Annealing temperatures
- Self-complementary of primers

Primer length: As other factors of the primers like specificity and annealing temperature are dependent on primer length, this is the most important factor that is to be considered while designing primers. The optimal primer length of 18-22 bp is long enough for adequate specificity, and short enough for annealing temperature (3).

GC Clamp: It is well known that 3' end of PCR primers is essential for the control of mis-priming (4). Presence of any GC nucleotide pair, CC, GG, GC, or CG at the 3' end of primers creates stable hybridization due to the stronger bonding of G and C bases. The GC Clamp at 3' end also helps the efficiency of the reaction by minimizing any breathing that might occur.

G/C content: The percentage of number of G's and C's of the total bases in the primer should be 40-60%. An optimal GC content enables specific binding and efficient melting temperature of the primer.

Melting/Annealing temperatures: The melting or annealing temperature of a primer is the temperature at which the primer anneals to a DNA template. The melting temperature required by the primer increases with the length of the primer. Primers with low melting temperatures and too high temperatures are not desirable, because DNA-Polymerase will be less active at such temperatures (5). There is more than one way for calculating melting temperature of the primer.

1. $Tm = 2 * (A + T) + 4 * (G + C)$, is the simplest formula of all for calculating melting temperature which is valid for oligos up to 18-24 bases long (6). This formula is calculated based upon the base compositions in the primer sequence.
2. $Tm = 64.9°C + 41°C \times ((G + C) - 16.4)/N$, is another basic method for calculating melting temperature which is based on the GC concentration of the primer and also length of the primer (7).
3. Another accurate method which uses nearest neighbor thermodynamic calculations is give by the formula (6)

$$Tm = \frac{-1000 * \Delta H}{-10.8 - \Delta S + R\ln(c/4)} - 273.15 - 16.6 * \log_{10} M$$

Where ΔH is the enthalpy of helix formation, ΔS is entropy for helix formation, c is the molar concentration of the primer (set ar 250 pM), M is the molar concentration of Na+ (set at 50mM) and R is the gas constant (1.987 cal/degree * mol) (6).

Self-Complementarity of primer sequences: Self-complementarity of primers enables the formation of secondary structures and primer dimmers (8). The presence of large number of complementary bases within the primer sequences, allows them to self anneal to form stem loops and hairpin loops (6). The interaction of primers with each other at 3' end produces primer dimers. The runs of three or more Cs or Gs at the 3'-ends of primers may promote mispriming at G or C-rich sequences (because of stability of annealing), and should be avoided.

Degenerate primers are mixed PCR primers with different options at several positions in the sequence, which can anneal to and amplify a variety of related sequences (9, 10). Degenerate primers are mixtures of similar, but not identical, primers. These mixed primers are used to find new genes or gene families, amplify same gene from different organisms, as the genes themselves are similar but not identical (1). These primers can also be used to amplify conserved sequences of a gene or genes from the genome of an organism or to get the nucleotide sequence after having sequenced some amino acids from a protein of interest. The most common use of degenerate primers is when the amino acid sequence of protein is known. One can reverse translate this sequence to determine all of the possible nucleotide sequences that could encode that amino acid sequence (1).

Degenerate PCR is a very powerful tool for amplification of unknown sequences and identification of novel members of gene families. Designing primers for degenerate PCR requires following steps: In the first step, the homologous nucleotide sequences for which degenerate primers are to be designed are obtained form a nucleotide sequence database. These sequences are then aligned using a multiple sequence alignment program such as ClustalW. Two blocks of conserved regions are identified to design the forward and reverse primers in order to amplify the region between these blocks. Primers of 18-22bp length are obtained from these blocks are tested for several properties. These properties include, GC clamp at 3'end, G/C content, melting temperature and self complementarity. Finally the primer pairs that satisfy all the primer properties are chosen to amplify the homologous sequences.

Degenerate Primer programs:

Other primer programs that are available for designing degenerate primers include CODEHOP (11), GeneFisher (12), and HYDEN (13). CODEHOP designs primers for distantly related gene sequences based on consensus-degenerate hybrid oligonucleotide primers from conserved blocks of amino acid sequences. All the degenerate primers designed by this program contain a short 3' degenerate core region and a longer 5' consensus clamp. CODEHOP takes a set of conserved blocks as an input. These blocks are derived from either the BlockMaker program or multiple alignment processor available at the Blocks Database website. The program computes a position-specific scoring matrix for each block and designs primers with low degeneracy based on the PSSM of aligned nucleotide sequences.

The HYDEN algorithm is a heuristic designed for finding approximate solutions to the degenerate primer design problem. HYDEN first locates conserved regions in the DNA sequences by finding ungapped local alignments with a low entropy score. Then designs

primers by using simple approximation methods and uses a greedy hill-climbing procedure to improve the primers and selects the best one as the output.

GeneFisher is the simplest amongst all the degenerate primer design programs. It takes amino acid sequence alignment as input and back translates the conserved portions of the alignment to obtain degenerate primers.

MATERIALS AND METHODS

The degenerate primer design for amplifying homologous sequences requires multiple sequence alignment of the sequences of interest. The input to the MPrime-DEG is a multiple sequence alignment computed using ClustalW (14). MPrime-DEG allows the user to enter the name of the text file with multiple sequence alignment or name of the file containing multiple sequences stored in FASTA (15) format. When the file with multiple sequences is entered, the program creates the alignment by directly calling the ClustalW into the program. Along with the nucleotide sequence alignment obtained form the ClustalW, this program also requires user to input other parameters like minimum, maximum and optimal values for primer length, G/C percentage and melting temperatures (Tm). In the first step the program parses the text file with nucleotide sequence alignment, eliminating all the unwanted details. The conserved regions of length 20bp are more are sent as output to the user along with the size, starting and ending position of each conserved region. The user selects the regions to be amplified based on the conserved blocks in the alignment. The conserved blocks of interest are then translated into their exact reading frames and then translated into a protein alignment. The forward and reverse primers from these blocks are obtained and the possible degenerate primer pairs that satisfy all the primer properties are given out to the user.

Implementation Details:

1. Sequence Alignment:

   In the first step, homologous nucleotide sequences of interest are aligned using multiple sequence alignment tools such as ClustalW.

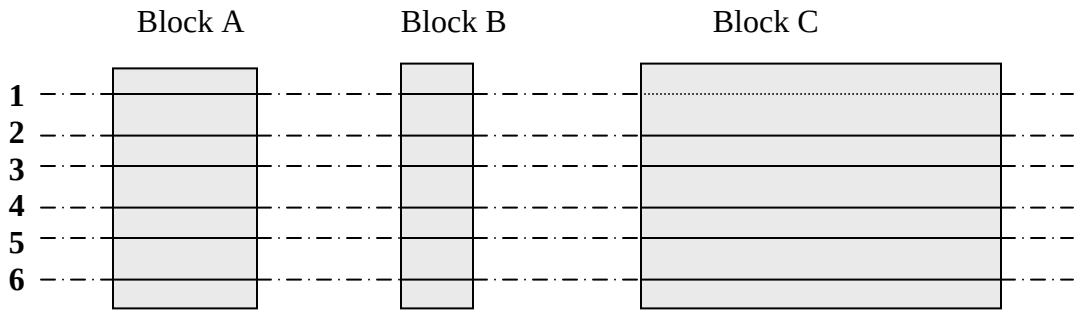


Figure 1: Conserved regions more than 20 bases long in a multiple sequence alignment.

2. Conserved Blocks:

   From the multiple sequence alignment, all the conserved regions with minimum of 20 bases are identified and are sent to the user along with the size, starting and ending positions of the conserved blocks. These blocks are ungapped alignments for designing forward and reverse primers to amplify the region between them.

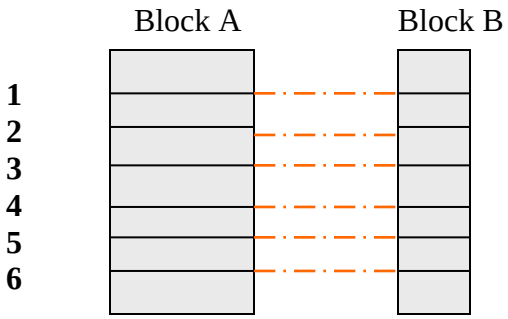


Figure 2: Region to be amplified between the conserved blocks.

3. Choosing correct reading frame of each block and translating into protein alignment :

To get the correct reading frame of the block, the block is first translated into amino acid alignment. Starting from the first column of the alignment, the change in the amino acid residues from one sequence to the other sequence is calculated using a PAM 250 matrix. The PAM250 matrix (16) is used to score the amino acid alignment to determine the similarity of sequences.

4. Calculating degeneracy of amino acid alignment:

The degeneracy of the amino acid alignment is then calculated by multiplying the degeneracies at all positions in the alignment. The degeneracy at each position is calculated by adding up the degenerate values of different amino acids using the values given in the table. The number of unique sequences that can be obtained from the amino acid alignment is given by the final degeneracy value

Amino Acids	Degeneracy Value
M W	1
F Y H Q N K D E C	2
I	3
V P T A G	4
L S R	6

Alignment: A M KC L A
A M K C L A
A M R C L T

Degeneracy: A M K C L A
R T
4*1*8*2*6*8 = 3072

Figure 3: Calculating degeneracy of an amino acid alignment.

5. Obtaining unique nucleotide sequences of 20bp from each block:

The different amino acids at each position of the alignment are recorded and their corresponding codons are obtained to get all the unique sequence combinations that the protein sequence alignment contains. From each unique sequence all the 20 bp sequences are obtained moving 1 base at a time till the end of the sequence.

```
Sequence:            GCTATGAAATGTCTTCTTCTTGCTCTT
20 base sequences:   GCTATGAAATGTCTTCTTCT
                      CTATGAAATGTCTTCTTCTT
                       TATGAAATGTCTTCTTCTTG
                        ATGAAATGTCTTCTTCTTGC
                         TGAAATGTCTTCTTCTTGCT
                          GAAATGTCTTCTTCTTGCTC
                           AAATGTCTTCTTCTTGCTCT
                            AATGTCTTCTTCTTGCTCTT
```

Figure 4: List of all 20 base sequences moving one base at a time from original sequence

6. Testing for primer properties:

All these 20bp sequences are then tested for the primer properties like GC Clamp at 3'end, GC content and Melting temperature. Inputs given and calculations performed for all these properties are similar to MPrime.

GC clamp at 3' end: All the 20bp sequences are tested for a GC clamp at 3'end by checking the last two bases of the sequence. Any sequences that do not posses any of the GC, CG, GG or CC pairs are filtered out.

G/C content: All the sequences that have passed the GC clamp test are tested for this property. G+C content of primers is the percentage of number of G's and C's present in the primer sequence. The user inputs the minimum, maximum and optimal G/C content values that are required. All the primer sequences that do not fall in this range of values are filtered out.

Melting temperature: For testing melting temperatures of the primer, user needs to input the minimum, maximum and optimal values. Melting temperatures are calculated by using the formula $Tm = 2 * (A + T) + 4 * (G + C)$. Any of the primers that do not fall into the range of melting temperatures are filtered out.

Since there is a possibility of having millions of 20bp sequences, many of these are eliminated to minimize the primer pairs list using all the above properties.

7. Obtaining best primer pairs:

Primer pairs for amplifying target regions are then designed by testing each primer sequence of the forward block with each and every primer sequence from the reverse block for formation of secondary or tertiary structures. MPrime-DEG uses the same approaches that are used in MPrime to overcome the problems of unwanted annealing between primer pairs. Like MPrime, MPrime-DEG incorporates an overall scoring scheme for paired-end and self-end annealing to reduce primer dimers, paired annealing to reduce partial double stranded structures, and self

annealing to reduce secondary structure formation of single stranded sequences. The reason behind this scoring scheme is to find the best scoring primer pairs with the smallest deviation from the overall optimal values (4). The parameters GC content, melting temperature, end annealing, self end annealing are weighted evenly to get the overall score for a primer pair. The smaller the score of a primer pair, the more that primer will function as desired.

RESULTS

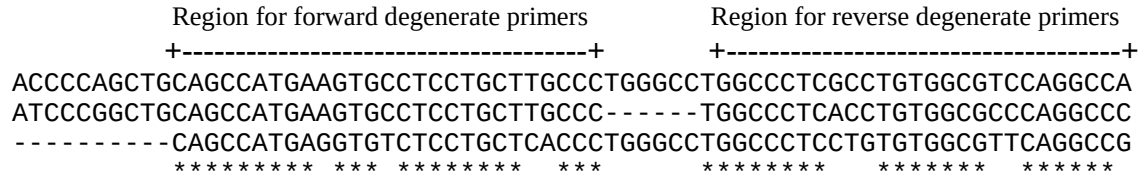


Figure 2: Sample alignment with two conserved regions for designing forward and reverse primers

CONCLUSION

Careful design of PCR primers is always an essential factor for successful PCR. The strategy used in this project carefully selects PCR primer pairs from a multiple nucleotide sequence alignment by identifying multiple conserved regions on the alignment. All the individual primers are carefully tested for their properties and primer pairs for formation of secondary and tertiary structures. The degenerate primer pairs designed by this technique can be used to amplify conserved sequences of genes from same organism or different organisms and to identify novel members of gene families.

Parallelizing the source code and developing a web interface are the two important future goals of this project. The source code of this project is implemented in a sequential manner. Each module of the code is carried out in a sequence, executing one instruction at a time using single processor. Parallelizing the source code and making it computationally effective is the most important aim of this project. This can be achieved by understanding the internal communication between each module of the program and decomposing the program into several single pieces which can be executed by multiple processors simultaneously.

Developing a web interface makes the project user friendly. It also makes the software platform independent and eliminates the need for local installations. In the web interface of this project, the user can enter the FASTA format of the sequences of interest into a text box to generate multiple alignment. All the conserved blocks are highlighted in the alignment and sent back to the user along with starting position and ending position of each block. The output is given in another text box where user can scan through the whole alignment and can choose two blocks of interest. The starting and ending positions of the conserved blocks along with other inputs such as Primer GC%, Primer Tm and GC clamp are entered as input to get the multiple degenerate primer pairs for the region of interest. All the possible degenerate primes are sent back to the user in the results box, where user can download them by using a save option.

ACKNOWLEDGEMENTS

We would like to thank members of the University of Louisville Bioinformatics Research Group (BRG) for their support. This project was supported by grant number P20 RR-16481 from the National Center for Research resources (NCRR), a component of National Institutes of Health. Its contents are solely the responsibility of the authors and are do not necessarily represent the official views of NCRR or NIH.

REFERENCES

1. Polymerase Chain Reaction. Retrieved June 27, 2006, from Wikipedia, The Free Encyclopedia Web site: http://en.wikipedia.org/wiki/Polymerase_chain_reaction

2. Definition of PCR primers. Retrieved June 27, 2006, from PCR Primers Web site: http://www.pcrstation.com/pcr-primer/

3. PCR primer design guidelines. Retrieved June 27, 2006, from PCR Primer Design Web site: http://www.premierbiosoft.com/tech_notes/PCR_Primer_Design.html

4. General notes on primer desing in PCR. Retrieved July 1, 2006, from Eppendorf North America:::General Notes on Primer Design in PCR Web site: http://www.eppendorfna.com/applications/PCR_appl_primer.asp

5. Primer design. Retrieved June 6, 2006, from Primer Design Web site: http://bioweb.uwlax.edu/GenWeb/Molecular/Seq_Anal/Primer_Design/primer_design.htm

6. Rouchka, E.C., Khalyfa, A., & Cooper, N.GF. MPrime: efficient large scale multiple primer and oligonucleotide design for customized gene microarrys. *BMC Bioinformatics*, *6:175*, Retrieved June 6, 2006, from http://www.biomedcentral.com/1471-2105/6/175.

7. Tm Calculations for Oligos. Retrieved July 10, 2006, from BioMath-Tm Calculations for Oligos Web site: http://www.promega.com/biomath/calc11.htm

8. Burpo, F.J. (August 11, 2001). A critical review of PCR primer design algorithms and cross hybridization case study. *Biochemistry 218*, Retrieved June 6, 2006, from http://biochem218.stanford.edu/Projects%202001/Burpo.pdf

9. Molecular biology techniques manual. Retrieved June 22, 2006, from PCR Primer Design Web site: http://www.mcb.uct.ac.za/pcroptim.htm

10. Degenerate PCR. Retrieved October 26, 2006, from MBCC Protease Research Group Web site: http://cgat.ukm.my/protease/degpcr.html

11. Rose, T.M., Henikoff, J.G., & Henikoff, S. (2003). CODEHOP(COnsensus-DEgenerate Hybrid Oligonucleotide Prime PCR primer design. *Nucleic Acids Research*. *31*, 13.

12. Giegerich, R., Meyer, F., & Schleiermacher, C. (1996). GeneFIsher-oftware support for the detection of postulated genes. *Mol Biol*. *4*, 68-77.

13. Linhart, AuthorC., & Shamir, R. (2002). The degenerate primer design problem. *Bioinformatics*. *18 Suppl. 1*, S172-S180.

14. FASTA format. Web site: http://www.cmbi.kun.nl/bioinf/tools/crab_fasta.html

15. ClustalW 2006. Web site: http://www.ebi.ac.uk/clustalw/

16. Amino acid distance measures. Web site: http://helix.biology.mcmaster.ca/721/outline2/node45.html