Katie M. Shelef
Biochemistry 218, Fall 2007

A Critical Analysis of Degenerate Primer Design Programs

Introduction

The ability to target and amplify DNA sequences of interest using the polymerase chain reaction (PCR) has revolutionized the biological sciences. PCR can be used as a diagnostic tool in pathogen detection, genotyping, screening genes for mutations and identification of single nucleotide polymorphisms (SNPs), determining molecularly-defined evolutionary relationships among organisms, and analysis of prokaryotic and microbial eukaryotic diversity, among many, many other applications.

Traditional PCR allows millions of copies of the target sequence. Sequences flanking the target sequence can be used to design PCR primers, which hybridize to the flanking sequences. These primers are added to a mixture of template DNA, all four deoxynucelotide triphosphates (dNTPs), and a heat-stable DNA polymerase (other reagents may also be included). The mixture then undergoes several cycles comprised of three basic steps: strand separation (denaturing), primer hybridization (annealing), and DNA synthesis (extension). After n cycles of PCR, the target sequence is amplified 2^n-fold, essentially “drowning out” the non-target DNA that was in the original template. While PCR is a relatively straightforward procedure, the key to successful amplification of the target DNA is the knowledge of the flanking sequences on which primer are designed. Therefore, much effort has gone into creating programs for primer design.

PCR is used for a variety of applications depending upon the question of biological interest. For instance, variations on the PCR protocol include multiplex PCR for multiple genes, bisulfite PCR for methylated DNA, and genome-scale PCR [18]. Therefore, a wide variety of programs have been developed to design primers or probes for specific PCR protocol needs. These include, but are not limited to, Primer3 for single gene amplification [14], Primegens for design of gene-specific probes and primers for microarray analysis [17], PRIDE and PRIMO for large-scale DNA sequencing [2, 8], MultiPLX and MIPS for Multiplex PCR [7, 15], MethPrimer for bisulfite PCR [9].

One particular PCR application of interest is the use of degenerate primers to amplify related genomic sequences or gene families for the determination of evolutionary relationships among homologous genes. One can examine a single gene family among different organisms (orthologous genes) or examine several genes within the same organism (paralogous genes). In addition, the use of degenerate primers on complex samples may allow the determination of the diversity of sequences of the gene family within the sample or the identification and characterization of unknown, related members of the gene family [13]. Homologous genes typically contain a mix of highly conserved regions and divergent regions. Evolutionary distances are determined from the nucleotide or amino acid differences within the divergent regions, while primers are created from the highly conserved regions flanking a divergent region. The ideal primer for a particular protein or gene family would amplify every member of the group within a sample and nothing more. However, one unique nucleotide sequences is rarely sufficient for amplifying homologous genes from different organisms because small amounts of mutations in the conserved regions may exist [16]. Degeneracy in the primers is required in order for all possible variants of a gene family to be amplified and compared.

The Degeneracy Problem

The degeneracy of a primer is the number of different sequences that it represents [16]. Primers with a degeneracy greater than one have one or more positions with several possible nucleotide bases. A more formal definition of degeneracy is as follows: A degenerate primer is a string P with several possible characters at each position. $P = p_1,p_2, \ldots p_l$ where $p_i \subseteq \Sigma$, $p_i \neq 0$ and $\Sigma = \{A, C, G, T\}$. The degeneracy of primer P, $d(P) = \prod_{i=1}^{l} |p_i|$, where l = length of primer P [15].

The need for degeneracy in primer design is partially due to the degeneracy of the genetic code. Because several nucleotide codons can code for the same amino acid, target DNA sequences that have differing codon usages may require degeneracy in the primers even if the sequence is conserved on the amino acid level. In addition, some amino acid residues may not be conserved in the regions of general conservation. This may be especially true in trying to design broad-range primers that cover all possible sequence diversity in a gene family derived from a wide taxonomic range of organisms.

Linhart and Shamir define the unique challenges facing degenerate primer design as the Degenerate Primer Design (DPD) problem [10]. At the crux of the problem is the tradeoff between degeneracy and coverage (the number of input sequences that can be amplified by the primer). From one perspective, the goal of degenerate primer design is to match and amplify as many of the input sequences as possible. Primers of high degeneracy are favored because this allows a greater number of input sequences to be matched, and the probability of detecting new, related sequences from the same gene family in your sample would increase.

However, too much degeneracy presents another problem. The greater the degeneracy of primers, the more likely it is that primer pairing to non-target sequences will occur. The low stringency of annealing conditions needed to amplify the diversity of the target sequences further adds to the problem. Thus, high degeneracy causes a loss in PCR specificity due to the increase in probability that unrelated sequences will be amplified in addition to sequences of interest [10, 15]. An additional factor in the relationship between degeneracy and PCR sensitivity is the loss of ability to amplify rare target sequences in a sample. As degeneracy increases, the concentration of any one particular primer drops relative to the total amount of primer. Each primer is therefore at a suboptimal concentration, and the amplification of template DNA that is in low copy number suffers as a result [11, 13, 15]. The degenerate primer design problem is one of optimization: primers must amplify as many of the genes in a gene family as possible while having a degeneracy under a predefined limit. Linhart and Shamir define this as the "Maximum Coverage DPD: *Given a set of strings of length k and an integer d, find a primer of length k and degeneracy at most d that matches a maximum number of input strings*" [10]. They expand this idea to include the need for primer pairs that are spaced such that a target DNA sequence for amplification lies between them. This is called the "Maximum Coverage degenerate primer pair design: *Given a set of n strings and integers k, d, find two primers, P_1, P_2, each one of length k and degeneracy at most d, so that a maximum number of input strings match both primers, and the match site of P_1 occurs in all covered strings to the left of the match site of P_2 without overlap between them*" [10].

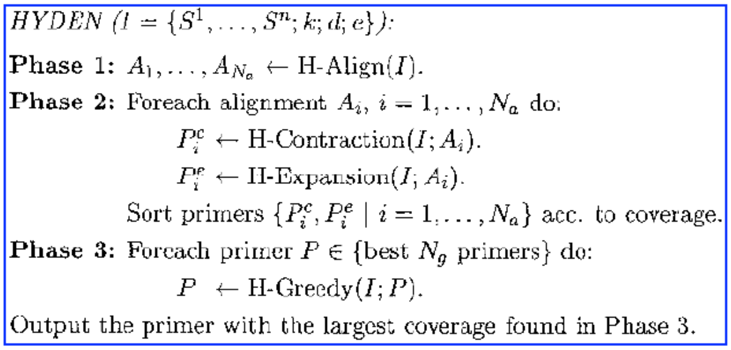
In addition to the degeneracy problem, other aspects of general PCR primer design must be taken into account. These include proximity between forward and reverse primer melting temperatures, minimization of hybridization effects between forward and reverse primers, and minimization of hybridization effects between a primer and itself [6]. The distance between the two primer sites must also be large enough so that the amplified gene product is sufficiently long for biologically meaningful comparisons to be made [10].

Programs for Degenerate Primer Design

Degenerate primer design was originally based on a manual inspection of a multiple alignment of a small number of sequences. Regions of conservation could be determined by eye, and primers were designed based on these regions of conservation. For instance, when Hales et al. published primers for the gene family of methyl-coM reductase, there were only five sequences currently in GenBank that were available for alignment [3]. Hales et al. were able to choose two conserved regions from this alignment via visual inspection and design primers based upon the conserved regions. However, many more input sequences for a gene family of interest are available today. The increase in sequence data and development of computational bioinformatics tools allows for a more formal, systematic approach to degenerate primer design. Some of these bioinformatics tools are reviewed below.

Hyden (Linhart and Shamir, 2005)

HYDEN was developed by Chaim Linhart and Ron Shamir, who formally defined the degenerate primer design problem (DPD) [10]. HYDEN tackles the DPD problem with a three-phase algorithm using a DNA nucleotide-based multiple sequence alignment and a set of integers that specify the primer length, maximum degeneracy, and the number of mismatches allowed as input. The first phase of the algorithm is to locate conserved regions in the DNA sequences by finding ungapped local alignments with a low entropy score, H_A, which is used as a metric to determine the level of conservation in the local alignment. The lower the entropy, the greater the chance of finding a primer that covers most of the input sequences. The second phase of the algorithm is to design primers from these local alignments using the H-Contraction and H-Expansion algorithms, discussed below. Finally, the primers are improved using a greedy hill-climbing procedure and the primer with the largest coverage is selected as the output (Figure 1).



```
HYDEN ($I = \{S^1, \dots, S^n; k; d; e\}$):
Phase 1: $A_1, \dots, A_{N_a} \leftarrow$ H-Align($I$).
Phase 2: Foreach alignment $A_i$, $i = 1, \dots, N_a$ do:
            $P_i^c \leftarrow$ H-Contraction($I; A_i$).
            $P_i^e \leftarrow$ H-Expansion($I; A_i$).
         Sort primers $\{P_i^c, P_i^e \mid i = 1, \dots, N_a\}$ acc. to coverage.
Phase 3: Foreach primer $P \in$ {best $N_g$ primers} do:
            $P \leftarrow$ H-Greedy($I; P$).
Output the primer with the largest coverage found in Phase 3.
```

Figure 1: Summary of the HYDEN three-phase algorithm [10]

HYDEN takes a minimization approach to degenerate primer design (i.e. minimize the number of sequences that the primer does not match). HYDEN is unique in that it takes both a destructive and constructive approach, using H-Contraction and H-Expansion, respectively. Both algorithms employed by HYDEN use a column distribution matrix, D(b, i) that contains the count of each character (A, T, C, or G) at each position. The two algorithms then take opposite approaches to find potential degenerate primers. The H-Contraction algorithm starts with a fully degenerate primer. The algorithm then examines all degenerate positions and chooses a position i, with character b, whose count D(b, i) is the smallest and discards b from position i in the primer. It then repeats this process iteratively until the primer reaches the maximum amount of allowed degeneracy, d, which is set by the researcher.

In contrast, the H-Expansion algorithm starts with an individual input sequence (or "string"). For each iteration, it degenerates a different input sequence, for a total of n iterations (n = number of input strings). It does so by using a substring T of each input string, A, as an initial nondegenerate primer and repeatedly adds to it a character with the largest count as long as its degeneracy does not exceed the maximum degeneracy threshold, d. It calculates a matrix, $D'_j(b, i)$, which is the number of strings that will be mismatched due to setting the ith position in the primer to one particular character, denoted "s_i," while the actual character in the ith position is b. If $s_i = b$ for a particular input sequence, a score of 0 is given in $D'_j(b, i)$, otherwise a positive score of D(b, i) is given. The program tries to minimize the number of strings that will be mismatched, and the output primer is the best primer the algorithm found in the n iterations.

The final step in the program uses a hill-climbing procedure, H-Greedy. H-Greedy seeks to improve the primers derived from the above algorithms by checking whether it can remove a character in a degenerate position of a given primer, P, and add a different character in any position instead so that the coverage of the primer increases. The process is repeated until coverage stops improving.

One major advantage of HYDEN is its ability to handle large numbers of sequences with long lengths (~1 Kbp). To show this, Linhart and Shamir tested Hyden on 127 human olfactory receptor genes (OR) that were 1 Kbp in length. The program output 13 degenerate primers for their sequences, and they assessed the value of each primer pair in two ways: The primers' sequencing efficacy, or percentage of distinct genes that were obtained (i.e. sequence richness) out of the total number of clones sequenced for that pair, and the percentage of the training-set genes, or input gene sequences, each primer pair covered [10] Several of the primer pairs were successful in both respects. In addition, HYDEN allows the creation of primers with very high degeneracy, with one primer for the OR genes reaching a degeneracy of $4x10^{12}$. Although primers with very high degeneracy may result in problems with PCR sensitivity, they may be the only solution when dealing with a large number of relatively divergent input sequences.

CODEHOP (Rose et al 1998, 2003; Rose 2005)

CODEHOP stands for "Consensus-degenerate hybrid oligonucleotide primer." The CODEHOP program brings together two previous solutions for primer design for "distantly related" input sequences and combines them into one, "hybrid" primer [11]. One of these solutions is the degenerate approach, which uses a pool of degenerate primers containing all of the possible nucleotide sequences that comprise a conserved amino acid motif from a multiple alignment. The other approach uses a single consensus sequence primer across a highly conserved region of an amino acid alignment by choosing the most common nucleotide at each position (which results in lots of potential primer-template mismatches.). The CODEHOP designed primers contain elements of both approaches. Specifically, each primer consists of a 3' degenerate core region and a 5' consensus clamp region.

The input for CODEHOP is either a multiple sequence alignment or a set of conserved BLOCKS from BLOCKmaker. The BLOCKS created by BLOCKmaker are aligned array of amino acid sequences without gaps that represents highly conserved regions of homologous proteins [4].

CODEHOP converts the multiple sequence alignment or BLOCKS into an amino acid position-specific scoring matrix (PSSM) using an odds-ratio method that considers sequence redundancy and amino acid conservation. An amino acid consensus sequence is formed, where the consensus amino acid at each position is the highest-scoring residue in the matrix. A DNA PSSM is calculated from the amino acid PSSM using a user-selected codon usage tables and a DNA consensus sequence is calculated.

The output of CODEHOP is a set of potential primers with a degenerate region on the 3' end of each primer and a consensus region on the 5' end of each primer. The 3' degenerate core region derives from 3-4 highly conserved amino acid residues resulting in a pool of all possible 11 or 12-mers. The primer's degeneracy is determined by the number of possible nucleotides at each position in the 3' core. The 5' consensus clamp region is an 18-25 base pair stretch (from 5 or more conserved amino acids) containing the most probable nucleotide predicted for each position, based on a combination of the most probable amino acid in each position and the most common codon corresponding to each amino acid chosen based on a user-selected codon usage table. The clamp regions are scored by the quality of the match between the 5' clamp and the sequence block given in a codon usage table. CODEHOP doesn't explicitly suggest forward and reverse primers, but primers designed from BLOCKS conserved in different parts of the amino acid alignment can be used for this purpose.

One advantage of CODEHOP primers is the ability of the different regions of each primer to stabilize the PCR during different points of the reaction (Figure 2). At the beginning of the PCR, the precise matches at the 3' end between one of the degenerate primers and the template DNA stabilizes any mismatches on the 5' consensus clamp end of the sequence. The consensus technique also ensures that the 5' end of the PCR product will contain a precise match to the 5' clamp region of the primer. Therefore, the non-degenerate, 5' portion of the primer stabilizes the hybridization of the degenerate, 3' portion of the primer, which allows for higher annealing temperatures. In other words, the hybridization is more stringent than a highly degenerate primer because there are no mismatches between the PCR product and 5' clamp region. CODEHOP also gives the researcher some flexibility in primer design by allowing the researcher to change the weights for each sequence segment to favor the contribution of selected sequences in primer design. In addition, a weight threshold can be specified so nucleotide bases that contribute less than a minimum weight are ignored to create less degeneracy.

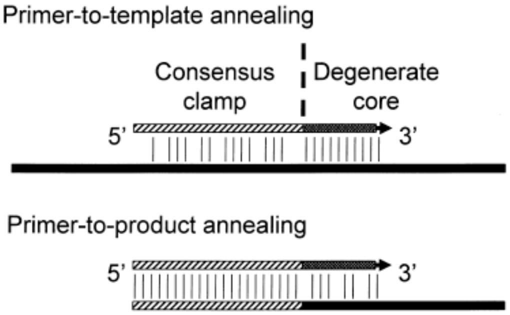


Figure 2: Stabilization of CODEHOP primer hybridization during different phases of the PCR process [12]

CODEHOP works well for designing primers from small sets of proteins. For instance, Rose used CODEHOP designed primers to identify and characterize 14 previously unknown DNA polymerase sequences from members of the alpha, beta and gamma herpesvirus subfamilies [13]. He was able to analyze amplified sequences via phylogenetic comparison. However, CODEHOP is not ideal for designing primers requiring high degeneracy and handling large sets of long input sequences [10]. Another limitation of CODEHOP is the requirement of codon usage knowledge. If the codon usage of the target genome is unknown or the desired primers need to span a gene family found in multiple organisms with different codon usages, the CODEHOP method might not produce the most optimal primers.

DePiCt (Wei et al., 2003)

A third option for degenerate primer design is DePiCt. The program uses a multiple sequence alignment of amino acid sequences or DNA sequences as input then implements clustering techniques to perform an automatic grouping of the input sequences based on the presence of conserved regions for the members of the cluster. DePiCt detects conserved regions for each individual group using a similarity metric called a BlockSimilarity score, a novel scoring scheme implemented as part of the DePiCt program. The BlockSimilarity score is a measure of the length of a conserved sequence to determine blocks that are sufficiently long, or greater or equal to a prescribed minimum primer length, k. Sufficiently long blocks are then scored by the program's algorithm, which determines the conservation of amino acids occupying the same position in the alignment.

DePiCt takes a unique approach to defining amino acid conservation. A position in the block is said to be "conserved" if the amino acids occupying that position in all the sequences are either identical or "similar." Amino acid "similarity" is based upon the closeness of the codon usage of the amino acids, not the similarity of the physiochemical properties of the compounds themselves. For instance, Cysteine (C) and Tyrosine (Y) are considered similar because they are represented by nucleotide sequences that differ by one base, TGY and TAY, respectively.

DePiCt uses a hierarchical clustering technique to group together sequences that have high BlockSimilarity scores so a primer pair can be designed for each block (Figure 3). Each sequence starts as its own singleton group, and in each subsequent iteration the two groups with the highest similarity are merged to form one grouping, if the grouping is considered "feasible." A "feasible" group is one that has at least one block of length greater than or equal to the parameter of Minimum Primer Product Length, or at least tow blocks separated by a length in the range between Minimum Primer Product Length and Maximum Primer Product Length. This process continues until groups can no longer be merged. The groups are output as clusters by the algorithm. Finally, the conserved amino acid blocks are then reverse translated into degenerate nucleotide primers. The program takes into consideration the maximum degeneracy of the primer, which is specified by the researcher. Degeneracy can be decreased by splitting a single degenerate primer into multiple degenerate primers with reduced degeneracy at some positions or accounting for codon biases, if they are known.

The main advantage of DePiCt is the flexibility that is build into the program. By taking a bottom-up hierarchical approach, the program merges sequences into as many groups as possible. Ideally all sequences in the alignment would form one large group with conserved blocks that are sufficiently long, but if this is not possible it outputs the minimum number of subgroups across all sequences in the alignment with their respective conserved blocks. In other programs, if conserved blocks are not found across all sequences, the primer design will fail and the researcher must manually go back, choose subgroups in the alignment, and run the program again. Wei et al give provide an example of this concept by developing six groups of primers necessary to amplify all genes from a group of resistance gene homologues derived from several plant families [16]. Other programs, including CODEHOP, were unable to find primers on their data set.

Primaclade (Gadberry et al., 2005)

Primaclade designs minimally degenerate primers using the authors' BioPerl based executable file and the Primer3 software [14]. The input for Primaclade is a DNA nucleotide-based multiple sequence alignment file. First a consensus sequence is computed from all sequences in the multiple sequence alignment, although how the consensus sequence is computed is not specified. The alignment is then split into each individual sequence. Primer3 is run 11 times for each sequence in the alignment, starting with an 18-mer primer search and ending with a 28-mer primer search. A maximum of 20 primers are returned per run of Primer3 for a maximum total of 11x20x(# of sequences) primers. For each primer, the starting location and length are calculated. The primer is then compared to the corresponding nucleotides in the consensus sequence. If the consensus sequence contains the same number or fewer degenerate nucleotides compared to the primer, the primer is saved. Duplicate primers are removed, and the results are checked to see if they match the input criteria.

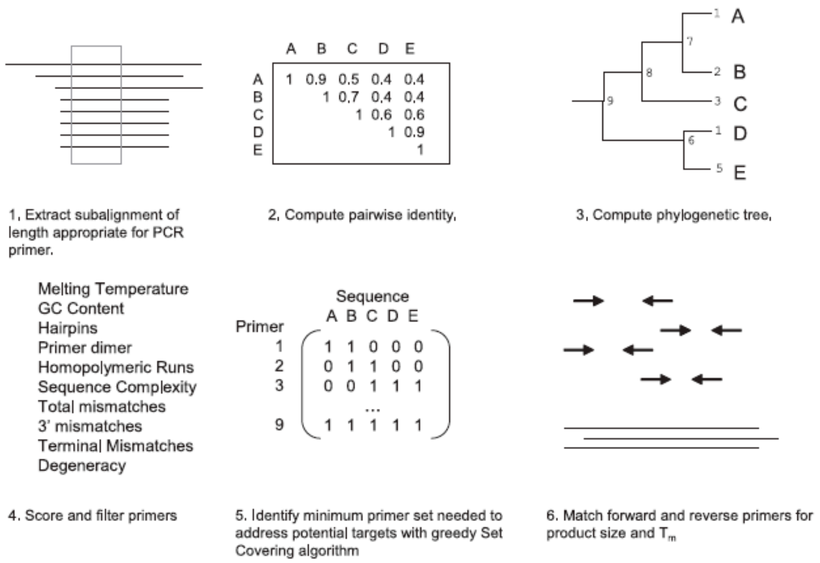
Unlike other programs, which design primers based on ungapped, conserved locally aligned regions, Primaclade uses Primer3 to design a set of unique primer sequences for each individual input string in the alignment. The resulting primers are then compared to the corresponding hybridization sites in the sequences of the multiple alignment. However, if some of the sequences do contain gaps at this site, the primer might not be able to bind to these sequences and PCR product would not be obtained.

Primaclade's strength is its ability to give the researcher flexibility in designating potential biologically relevant information. The researcher can specify the maximum number of degenerate base pairs (up to 5) per primer, the number of gapped sequence lines to ignore, and a single region in the alignment to exclude. For instance, the exclusion feature allows the exclusion of areas so highly conserved that they may be shared with homologous genes in organisms not of interest (e.g. human genes in an examination of human-associated microbial communities) or shared with paralogous genes of the same organisms of interest.

On the other hand, the "brute force" approach of Primaclade limits the amount of input data that can successfully be utilized. The authors report that Primaclade worked best when the number of input sequences was small (up to eight) and closely related (up to 29% sequence divergence). Like CODEHOP, Primaclade would not be a good choice for highly divergent gene families or a gene of interest that exists in a wide diversity of taxonomic groups. Part of this limitation is the time required for the program to run; a larger number of sequences causes the program to run more slowly [1]. The quality of the alignment is vital as well. Primers will not be found in regions of ambiguous alignment or poor consensus. If primers can't be found, the authors suggest running Primaclade again with less stringent parameters. Alternatively, the multiple sequence alignment can be divided into sub-files and re-run, but this increases the number of primer pairs needed to capture all the sequence diversity within your gene family of interest. Finding forward and reverse primers is also not clearly integrated into the program. To find forward and reverse primers, the authors suggest running Primaclade twice, first excluding all the right-handed portion of the sequences to find a forward primer and then excluding the left-handed portion of the sequences to find the reverse primer.

Greene SCPrimer (Jabado et al., 2006)

Greene SCPrimer approaches the degenerate primer design problem as an optimization problem which asks: "given a set of *n* strings (the input sequences), is there a primer *P* of length *k* and degeneracy of utmost *d* that matches *m* of the strings?" To answer this question, Greene SCPrimer uses a phylogenetic tree-building approach followed by a set-covering algorithm to find optimal primers. The program first takes sub-alignments of a DNA nucleotide-based multiple sequence alignment of the specified primer length, k. The sub-alignments are extracted from the entire alignment and filtered for uniqueness. An all-against-all pairwise comparison of the potential primers is used to generate a similarity matrix for each subalignment. The matrix is used to generate a phylogenetic tree using a hierarchical clustering algorithm based on Euclidean distance. A consensus sequence containing all nucleotides in the subalignment is computed for each node in the tree for a total of 2n-1 primers, where n = the number of sequences in the subalignment. Potential primers are checked for physical parameters such as melting temperature, GC content, and degeneracy. Primers that fit within the specified parameters are then compared to all sequences in the subalignment to determine if hybridization is likely, and the number of mismatches (both total and on the 3' end) are determined. If the primer can be extended it will assign it a score of 1, if not it will be assigned a score of 0, and these numbers will be organized into a primer extension matrix with sequences as columns and primers as rows. This matrix is the input for the greedy SCP algorithm. This algorithm aims to minimize the primer count using a basic cost function for minimization while simultaneously accounting for primer quality (achieved by "augmenting [the cost function] to reflect primer quality by adding a negative weight for deviation from ideal parameters" [5]). The output is the minimal set of primers required to amplify all sequences in the subalignment. Finally, the program selects the minimum number of forward and reverse primer sets that are optimized for the specified physical parameters (Figure 3).



1. Extract subalignment of length appropriate for PCR primer.

2. Compute pairwise identity.

3. Compute phylogenetic tree.

4. Score and filter primers

5. Identify minimum primer set needed to address potential targets with greedy Set Covering algorithm

6. Match forward and reverse primers for product size and T_m

Figure 3: Schematic diagram of the SCPrimer design method [5]

The use of the Set Covering Problem (SCP) for degenerate primer design is a unique application of a computer science method. An important strength of the set covering algorithm is its ability to factor in negative weights for deviations from ideal parameters. Another major advantage of the program is that it presents the output as primer pairs and selects the minimum number of primer pairs needed for optimal amplification. However, if the Greene SCPrimer program results in a pool of degenerate primers, problems may arise. For instance, adding all suggested primers in one PCR reaction (multiplex PCR) may not pick up on rare sequences of interest or each individual primer might get used up early in the reaction. Alternatively, each primer pair could be run on a sub-aliquot of the sample, but this has the disadvantage of being costly, time-consuming, and using up sample.

Summary and comparison

Each of the programs reviewed above take a unique approach to degenerate primer design, but comparisons can be made. For example, all of the programs use a multiple sequence alignment of the researcher's sequences of interest as input. However, some of the programs (CODEHOP) use an amino acid-based multiple sequence alignment, while others (HYDEN, Primaclade, Greene SCPrimer) use a DNA nucleotide-based multiple sequence alignment. DePiCt is unique in that it will accept either type. Using each type of alignment has advantages. Multiple alignments of amino acid sequences may permit the inclusion of more diverse organisms because divergence (and therefore primer degeneracy) is more likely to be more pronounced on the DNA level compared to the amino acid level. However, organisms with similar codon biases may be more conserved at the DNA nucleotide level than would be expected using amino acid degeneracy alone. Therefore, using DNA sequences may allow primers of a lower degeneracy to be found.

The programs also differ in the number and length of sequences that can be used as input. HYDEN allows creation of highly degenerate primers from large numbers of sequences with long lengths. Out of all the programs I reviewed, HYDEN may be the only choice for successful degenerate primer design for large quantities of input sequences deriving from a wide array of taxa. If a smaller number of input sequences are used, a program such as CODEHOP may be more desirable because it allows for degeneracy via the 3' degenerate core region while controlling for PCR sensitivity via the 5' consensus clamp. However, CODEHOP has a difficult time designing primers for a large number of input sequences [10], so this option is not always feasible. Primaclade is also limited by the number of input sequences; it has optimum of eight [1].

Locations of the mismatches within the primer must be taken into account because mismatches occurring near the 3' extension site are more disruptive than mismatches at the 5' end [10]. One limitation of HYDEN and DePiCt is that they do not consider where mismatches occur within the primer. In contrast, CODEHOP accounts for this by combining a stabilizing consensus clamp on the 5' end with the shorter degenerate portion on the 3' end. Greene SCPrimer takes into account mismatches overall and at the 3' end of the primer, so mismatches at the 3' end can be given a more negative weight than mismatches at the 5' end [5].

Finally, the ability to choose primer sets that have enough target sequence length between them so that biologically meaningful comparisons among sequences can be made is an important requirement. DePiCt solves this issue by requiring the minimum and maximum primer product lengths as researcher-specified parameters [16]. CODEHOP allows the researcher to do this manually by choosing primers designed from different BLOCKS at different locations in the input sequences [11, 12,13]. Both Primaclade and HYDEN must be run twice to find forward and reverse primers a desired length apart; for example, the first and last x-number of base pairs of an alignment of interest [1, 10].

Testing degenerate primer design programs on bacterial lactate dehydrogenase genes

One important application of degenerate primer design is the exploration of evolutionary relationships and diversity within a functionally important gene family. To compare the degenerate primer designing software and offer an applied critique of each program, I chose twelve bacterial lactate dehydrogenase genes from the following organisms: *Lactobacillus reuteri, Bacillus cereus, Streptococcus mutans, Streptococcus agalactiae, Streptococcus pneumoniae, Enterococcus faecium, Streptococcus sanguinis, Geobacillus stearothermophilus, Listeria monocytogenes, Lactococcus lactis, Staphylococcus aureus,* and *Clostridium novyi.* Lactate dehydrogenase is the enzyme that catalyzes the oxidation of lactate to pyruvate. This enzyme is found in a variety of organisms including plants and animals, but in bacteria it acts as an important catabolic enzyme in carbohydrate fermentation (Note: These were the same protein sequences I used in Homework 6).

Since some of the programs require an amino acid multiple sequence alignment and others a nucleotide multiple sequence alignment, I performed both multiple sequence alignments using ClustalW. I retrieved DNA sequences for each gene by entering the protein accession number in the NCBI database and clicking on the "CDS" link, which took me to the nucleotide sequence for that gene. I then chose to view the sequence in FastA format and copied and pasted each gene into a new file. I then used this FastA list of sequences to perform the nucleotide-based multiple sequence alignment. While all of the amino acid sequences were able to be aligned, four of the twelve DNA sequences were not (they showed up in the ClustalW output as all gaps). Therefore, for the DNA sequence-dependent programs, I used only the eight lactate dehydrogenase gene sequences that aligned: *L. monocytogenes, E. faecium, S. agalactiae, C. novyi, L. lactis, S. mutans, S. pneumoniae, and L. reuteri*.

I first used the CODEHOP program to design degenerate primers for my twelve sequences [11, 12]. CODEHOP first required the input of an amino acid multiple sequence alignment in order to create conserved amino acid blocks via Blockmaker's Multiple Alignment Processor. The program found four conserved blocks consisting of 9 to 12 amino acids each. The resulting blocks were automatically pasted into the CODEHOP program. I used all default settings for primer design, and since I was aligning sequences from a wide variety of bacterial genomes, I chose the "equal" codon usage table. Out of four conserved blocks, three sets of primers were found corresponding to three of the conserved blocks. Each primer had a degenerate 3' core of 11 or 12 nucelotides and a 5' consensus clamp of 16 to 22 nucleotides. Furthermore, each block had four to six suggested forward primers and three to four suggested reverse, or "complement" primers, of degeneracies ranging from 32 to 128. In sum, 25 potential primers were output by the program.

I then tried the DePiCt program for degenerate primer design [16]. DePiCt will take either amino acid or nucleotide multiple sequence alignments. I first tried the program with my amino acid alignment and the program's default parameters, including a minimum primer length of 18, a maximum degeneracy of 1024, and the incorporation of sticky ends into the primers. The program automatically threw out one sequence, *Streptococcus sanguinis*, because of its short length and used the remaining eleven sequences. Primer design failed with these default parameters. However, I was successful when I did not require the incorporation of sticky ends and I changed the minimum primer length to 15. The resulting forward and reverse primers had the maximum degeneracy of 1024. When the minimum primer length was decreased to 12, however, degeneracy of the forward and reverse primers decreased to 64 and 32, respectively. I also ran DePiCt on the DNA nucleotide alignment, which excluded four of the eight sequences that were unable to be aligned. In contrast to the amino acid alignment-based primers, the DNA alignment-based primers split the eight sequences into two groups and suggested forward and reverse primers for each group. Although split into two groups, the primers were much less degenerate. The first group comprised three sequences and had primer degeneracies of 48 (forward) and 1 (reverse). The second group comprised five sequences, and had primer degeneracies of 6 (forward) and 8 (reverse). Another advantage of the DNA alignment-based primers is the length of the PCR product covered – 335 bases and 444 bases for the two DNA alignment-based groups, respectively, versus only 18 bases for the amino acid-based primers.

The use of Primaclade to design degenerate primers was also attempted. The input for Primaclade was a nucleotide-based multiple sequence alignment. However, no primers were found for any combination of input parameters. T_m range, %GC content range, degeneracy, etc. were all relaxed compared to the default parameters, but the program could not create primers based on the eight sequence in my alignment. One great weakness of Primaclade for the design of degenerate primers is the maximum amount of degeneracy allowed. A degeneracy of up to five is permitted, but as evidenced by CODEHOP and DePiCt, a much higher degeneracy is needed to create a primer pair that would capture all the sequence diversity in my input sequences. Primaclade may be a useful option for degenerate primer design only when the sequences are very closely related (i.e. have a high sequence similarity). However, the program is not a viable option for more broad-range primer design.

Unfortunately, both the HYDEN and Greene SCPrimer websites were unavailable. I e-mailed Dr. Shamir and Dr. Linhart to try to gain access to HYDEN, but I was told I needed to fill out a license agreement, which needed to be signed by a legal representative, to gain access to the software. I would predict that HYDEN would have done an equally good job compared CODEHOP at designing primers for this example because HYDEN effectively designed highly degenerate, highly specific primers for 127 human olfactory receptor genes with known regions of high variability [10].The Greene SCPrimer website has been “experiencing electrical problems” since August and the program was not working. The authors of SCPrimer claim that CODEHOP is the program “most similar to SCPrimer,” so results may have been comparable [5].

Of the three programs I was able to access, CODEHOP was the most successful program for degenerate primer design for my twelve lactate dehydrogenase sequences. Because the program uses amino acid alignments rather than nucleotide alignments, I was able to find primers that covered all twelve of my genes of interest. Furthermore, CODEHOP gave me the most options in terms of the amount of sequence I could cover. Three sets of primers deriving from three conserved amino acid blocks were returned, and for each set four forward primers and four reverse (or “compliment”) primers were suggested. From these, I was able to choose two primers, each with a low degeneracy of 32 that covered a 128 base pair stretch of the lactate dehydrogenase gene and had complimentary melting temperatures (the forward primer had a t_m = 60.9, the reverse primer had a t_m = 60.3). CODEHOP could have potentially returned less degenerate primer results had I known the codon usage patterns among my sequences. Although DePiCt gave several options, the researcher would have to choose between having highly degenerate primers based on an alignment of eleven of the twelve amino acid sequences, or two sets of less degenerate primers covering only eight of the DNA nucleotide sequences. Neither option would be as ideal as the CODEHOP primers. Primaclade failed to suggest any primers, and was therefore the least-suited program to this example.

Future directions

While current degenerate primer designing programs encompass a wide range of options, future programs could include novel options such as the explicit incorporation of protein structure data to the primer design process. This would allow the inclusion of functionally-important amino acid residues in the PCR product, such as the residues comprising the active site of an enzyme, and data could be queried from protein databases like Swissprot or PDB. The ability to incorporate functionally important residues would allow new, biologically meaningful questions to be asked. For example, the determination of variants of an enzyme-encoding gene that contain different residues in or around the active site might behave very differently in terms of the kinetics of the reaction they catalyze. The difference in kinetics of the same enzyme deriving from two different organisms could potentially have ecological implications by influencing how these two organisms compete with each other or coexist in the same habitat. The ability to handle even larger data sets from more even more divergent groups of organisms would also be a welcome addition. All in all, it is important to recognize that any single degenerate primer design program cannot optimize every parameter – each has its own unique set of strengths and weaknesses. Therefore, choosing the "best" bioinformatics-based tool for primer design depends on the researcher knowing the nature of the experiment and the biological question that he or she is asking.

Works Cited

1. Gadberry, M.D., S.T. Malcomber, A.N. Doust and E.A. Kellogg. Primaclade – a flexible tool to find conserved PCR primers across multiple species. Bioinformatics 21 (7), 1263-1264 (2005).
2. Haas, S., M. Vingron, A. Poustka, and S. Weimann. Primer design for large scale sequencing. Nucleic Acids Research 26 (12), 3006-3012 (1998).
3. Hales, B.A., C. Edwards, D.A. Ritchie, G. Hall, R.W. Pickup, and J.R. Saunders. Isolation and identification of methanogen-specific DNA from blanket bog peat by PCR amplification and sequence analysis. Applied and Environmental Biology 62 (2), 668-675 (1996).
4. Henikoff, S., J.G. Henikoff, W.J. Alford, and S. Pietrokovski. Automated construction and graphical presentation of protein blocks from unaligned sequences. Gene 163, GC17-GC26 (1995).
5. Jabado, O.J., G. Palacios, V. Kapoor, J. Hui, N. Renwick, J. Zhai, T. Briese, and W.I. Lipkin. Greene SCPrimer: a rapid comprehensive tool for designing degenerate primers from multiple sequence alignments. Nucleic Acids Research 34 (22), 6605-6611 (2006).
6. Kampke, T. The reference point method in primer design. In *PCR Primer Design: Methods in Molecular Biology* (ed. Yuryev, A.) 75-91 (Humana Press, Totowa, NJ, 2007).
7. Kaplinsky L. and M. Remm. MultiPLX: Automatic grouping and evaluation of PCR Primers. In *PCR Primer Design: Methods in Molecular Biology* (ed. Yuryev, A.) 287-303 (Humana Press, Totowa, NJ, 2007).
8. Li, P., K.C. Kupfer, C.J. Davies, D. Burbee, G.A. Evans, and H.R. Garner. PRIMO: A primer design program that applies base quality statistics for automated large-scale DNA sequencing. Genomics 40, 476-485 (1997).
9. Li, L. Designing PCR primer for DNA methylation mapping. In *PCR Primer Design: Methods in Molecular Biology* (ed. Yuryev, A.) 371-383 (Humana Press, Totowa, NJ, 2007).
10. Linhart, C. and R. Shamir. The degenerate primer design problem: Theory and applications. Journal of Computational Biology 12 (4), 431-456 (2005).
11. Rose, T.M., E.R. Schultz, J.G. Henikoff, S. Pietrokovski, C.M. McCallum, and S. Henikoff. Consensus-degenerate hybrid oligonucelotide primers for amplification of distantly related sequences. Nucleic Acids Research 26 (7), 1628-1635 (1998).
12. Rose, T.M., J.G. Henikoff, and S. Henikoff. CODEHOP (COnsensus-DEgenerate Hybrid Oligonucleotide Primer) PCR primer design. Nucleic Acids Research 31 (13), 3763-3766 (2003).
13. Rose, T.M. CODEHOP-mediated PCR – A powerful technique for the identification and characterization of viral genomes. Virology Journal, 2-20 (2005).
14. Rozen, S. and H.J. Skaletsky. Primer3 on the WWW for general users and for biological programmers. In *Bioinformatics Methods and Protocols: Methods in Molecular Biology* (eds. Krawetz, S. and S. Misener) 365-386 (Humana Press, Totowa, NJ, 2000).
15. Souvenir, R., J. Buhler, G. Stormo, and W. Zhang. Selecting degenerate multiplex PCR primers. In Proc. 3rd Workshop on Algorithms in Bioinformatics (WABI 2003), 512-526 (2003).
16. Wei, X., D. Huhn, and G. Narasimhan. Degenerate primer design via clustering. In *Proc. 2nd IEEE Computer Society Bioinformatics Conference (CSB 2003)*, 78-83 (2003).

17. Xu, D., G. Li, L. Wu, J. Zhou, and Y. Xu. PRIMEGENS: Robust and efficient design of gene-specific probes for microarray analysis. Bioinformatics 18 (11), 1432-1437 (2002).
18. Yuryev, A. *PCR Primer Design: Methods in Molecular Biology* (Humana Press, Totowa, NJ, 2007).