Degenerate Primer Design using Computational Tools
Computational Molecular Biology
Veronica Brand
11 December 2011

The polymerase chain reaction (PCR) is widely used to uncover new information about genes and genomes, but is limited in that it is highly dependent on the use of specific and sensitive primers to yield good results. Primers can target a specific, known sequence, but what is often more interesting is the use of degenerate primers to target unknown sequences. Degenerate primers are designed based on multiple known sequences for a gene family, and can be used to discover new homologs in other species. Here, I review three programs that focus on the design of degenerate primers to search for unknown genes: HYDEN, SCPrimer, and iCODEHOP. The algorithms, advantages, and limitations of each are discussed. Although HYDEN and SCPrimer are promising programs in terms of optimizing the degeneracy of each primer, CODEHOP is the most widespread in use (Chakravorty and Vigoreaux 2010; McMahon *et al.* 2002; Pereyra *et al.* 2010). This is likely due to its focus on biologically relevant constraints and ease of use in designing primers (Staheli *et al.* 2011).

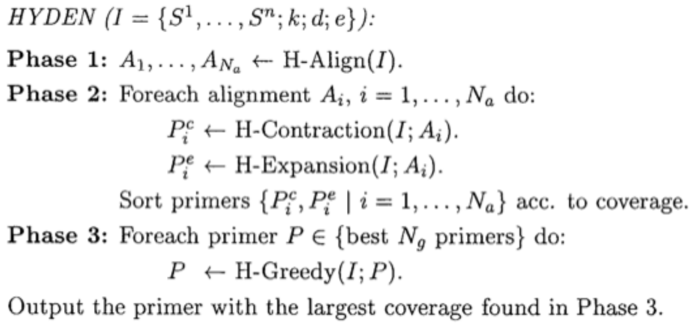
INTRODUCTION

The goal of primer design is to create a short oligonucleotide that can be used to amplify a specific sequence of DNA. The primer sequence must meet a set of given constraints, including correct annealing temperature to its target, GC content that matches the target organism, no secondary structure in each primer, no annealing to itself or creation of primer dimer structures (Giegerich *et al.* 1996). A quick search in Pubmed for "DNA Primers" and "Software" reveals that several software programs are available to complete this task. However, in the case of designing primers to match a variety of sequences so that the primers can amplify unknown targets as well as known targets, the problem becomes trickier. In a thorough description of the problem and several variations, Linhart and Shamir describe the problem as such: “Given a training set of known genes, design a pair of primers, one for the 5’ side and another for the 3’ side, so that the primers would amplify many of the genes and would have degeneracy that does not exceed a predefined limit” (2005). A degenerate primer in this case is a primer where multiple nucleotide bases are possible for a given position in the primer. A primer’s total degeneracy, then, is the total number of specific primers that a degenerate sequence can have, or, mathematically, is the product of the number of possible bases at each position in the primer (Linhart and Shamir 2005). It is possible to design non-degenerate primers for multiple sequences based on a consensus sequence alone, but these will have too many mismatches to distantly related sequences to amplify unknown targets efficiently (Rose *et al.* 1998). In contrast to consensus primers, those that are too highly degenerate face several problems. For one, as a primer sequence increases in degeneracy, the concentration of each specific primer will drop, so that they are quickly used up in early rounds of a PCR reaction. In later rounds, then, no additional amplification will occur (Rose *et al*. 1998).

Thus, it is important to be able to design specific and sensitive degenerate primers for amplifying distantly related sequences. Although one can often do this manually for well-conserved sequences (Lang and Orgogozo 2011), computational methods are available to systematically look for conserved sections in the sequence and to design primers. Although several methods have been suggested in the literature, here, I will focus on three methods that have been experimentally validated: HYDEN, SCPrimer, and iCODEHOP.

HYDEN

HYDEN is a primer design algorithm created as part of a larger scheme, DEFOG, to uncover novel members of a gene family (Fuchs *et al.* 2002). The goal of this program is to create a highly degenerate primer that will amplify a maximum number of input sequences (Linhart and Shamir 2005). Degeneracy must be bound to reduce the probability of amplifying unrelated sequences. The general scheme of this program is depicted in Figure 1.



```
HYDEN ($I = \{S^1, \ldots, S^n; k; d; e\}$):
Phase 1: $A_1, \ldots, A_{N_a} \leftarrow$ H-Align($I$).
Phase 2: Foreach alignment $A_i$, $i = 1, \ldots, N_a$ do:
            $P_i^c \leftarrow$ H-Contraction($I; A_i$).
            $P_i^e \leftarrow$ H-Expansion($I; A_i$).
         Sort primers $\{P_i^c, P_i^e \mid i = 1, \ldots, N_a\}$ 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. General overview of HYDEN algorithm. From Linhart and Shamir 2005.

Methods. HYDEN proceeds in three stages: a local alignment stage to search for conserved subsequences of length *k*, a creation of a primer with degeneracy *d*, and an optimization of that primer (Linhart and Shamir 2005). In the first stage, HYDEN accepts as input, *I*, *n* nucleotide sequences (S^1-S^n) that do not need to be previously aligned, a specified length of the primer, *k*, a limit on the degeneracy of the primer, *d*, and a limit on the number of mismatches that are permitted between primer and template, *e*. The program then finds ungapped local alignments of length *k* and for each alignment, calculates a distribution matrix of the nucleotide bases at each position, as well as an entropy score for each alignment. In this case the entropy score is a measure of how variable each position is in the alignment (Linhart and Shamir 2005). The local alignments with low entropy scores are passed on to the second stage, where sample primers are actually designed. Construction of the primers proceeds using two methods. In H-Contraction, the program starts with a fully degenerate primer (all possibilities are considered in each position), and discards characters at a degenerate position with the lowest count in the distribution matrix, until the required degeneracy is reached. In H-Expansion, the program starts with a non-degenerate primer, and adds degenerate positions based on the distribution matrix until the degeneracy is attained (Linhart and Shamir 2005). After all possible primers have the required degeneracy, the program computes the coverage of each primer relative to the input set (ie: how many input sequences it matches with fewer than e mismatches), and passes only the primers with the highest coverage to the third stage. The third stage is a refinement stage that uses a hill-climbing procedure to incrementally change two positions: one character is removed from a degenerate position and added at a different position as long as coverage of the primer increases (Linhart and Shamir 2005). This optimizes the location of the degeneracy, and creates primers that match a maximum of input sequences, while still being constrained by the maximum degeneracy.

Advantages. By using nucleotide sequences as a basis for primer construction, HYDEN uses the information stored in the DNA sequence to decide which positions should be degenerate and does not include extra degeneracy that is not found in the input sequence. For conserved motifs found in the local alignment (stage 1), the program considers a large subset of possible primer sequences, and iteratively refines them to optimize the positions of degeneracy. As such, it effectively utilizes computational power to search through a large number of possibilities, and can handle a large set of input sequences. The process is methodical, and has been verified experimentally to achieve good amplification of both known and unknown targets (Fuchs *et al.* 2002). HYDEN was used to design several primers for human olfactory receptor (OR) genes based on a set of 127 genes from the draft human genome (Fuchs *et al.* 2002). Each primer matched 76-90% of sequences with up to 2 mismatches, and use of 20 combinations of 13 primers resulted in both high sensitivity of amplification—300 unique genes were amplified—and high specificity—only 0.4% of sequenced clones were non-OR products (Fuchs *et al.* 2002).

It is also available on the web as a downloadable program at http://acgt.cs.tau.ac.il/hyden/, but is only available for Windows XP. Overall, this program is able to design highly degenerate primers to match the greatest number of input sequences.

Limitations. However, a major limitation of this program is that while it seems to be computationally sound, it does not consider the biological parameters associated with primer design. Although it takes parameters of the length of the primer, the desired degeneracy, and the number of acceptable mismatches as part of its input, these parameters may have to be iteratively changed to achieve, for example, the desired melting temperature of the primer. Additionally, this program does not take into account GC content, secondary structure, self-annealing, or the location or type of mismatches that may affect the efficiency of the PCR reaction. While future implementations of this program may consider these biological parameters, the original implementation that is currently available does not (Linhart and Shamir 2005). Thus, it may not be quite as effective in designing optimal PCR primers than other programs that take these biological considerations into account.

SCPrimer

SCPrimer is another computational program that determines optimal primer pairs from multiple nucleic acid sequence alignments. It differs from HYDEN in that its goal is to create a set of multiple primers for use in multiplex PCR. It should be accessible on a web server, available at http://scprimer.cpmc.columbia.edu/SCPrimerApp.cgi. However, I encountered an internal server error in trying to access this page, so this program may no longer be available to the general public.

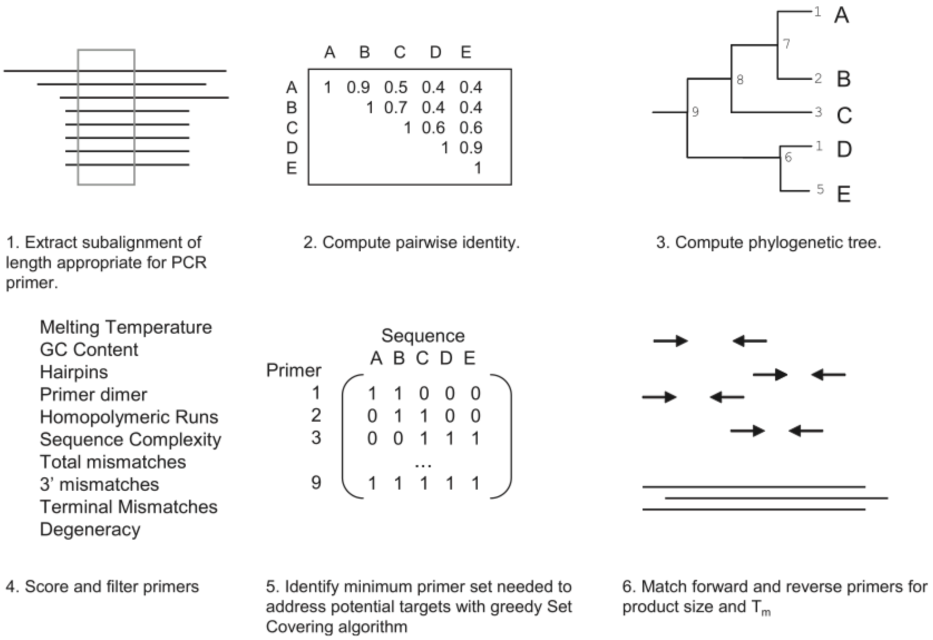

A B C D E
A 1 0.9 0.5 0.4 0.4
B 1 0.7 0.4 0.4
C 1 0.6 0.6
D 1 0.9
E 1
1 A
7
8
2 B
9
3 C
1 D
6
5 E
1. Extract subalignment of length appropriate for PCR primer.
2. Compute pairwise identity.
3. Compute phylogenetic tree.
Melting Temperature
GC Content
Hairpins
Primer dimer
Homopolymeric Runs
Sequence Complexity
Total mismatches
3' mismatches
Terminal Mismatches
Degeneracy
Sequence
A B C D E
Primer
1 1 1 0 0 0
2 0 1 1 0 0
3 0 0 1 1 1
...
9 1 1 1 1 1
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 2. General scheme of SCPrimer. From Jabado *et al.* 2006.

Methods. SCPrimer consists of four steps, as seen in Figure 2. Like HYDEN, it also accepts DNA sequences as input, and creates a local multiple sequence alignment. SCPrimer then uses the alignment to create a phylogenetic tree of the sequences. A degenerate consensus sequence is then made for each node of the tree. It then scores the consensus sequence based on biological parameters, such as melting temperature, GC content, homopolymeric runs, hairpin/primer-dimer formation and degeneracy (Jabado *et al.* 2006). Those that fit these physical constraints are then checked against the sequences to determine if they are likely to amplify the template. Using those that will amplify a large set, a set covering algorithm solves for the minimum number of primers needed to cover the complete set of sequences, and outputs of pairs of primers that are optimized for product length, cross reactivity and melting temperature are provided (Jabado *et al.* 2006).

Advantages. SCPrimer seems to be a straightforward program for finding multiple primers to cover a whole set of sequences. Unlike HYDEN, each primer is checked so that it fits biological parameters for efficient PCR analysis. It has a different purpose than HYDEN; its goal is not so much to optimize one degenerate primer pair for a set of sequences, but to look at a set of primers as a whole. As such, each primer can be a little more specific for its target sequence. This method can also accept a large number of sequences as input; it was validated by creating primers to target all influenza HA5 sequences in the database using 449 sequences, and found that a set of 4 primer pairs was all that was needed to cover all sequences (Jabado *et al.* 2006). By using information at the DNA level, it tries to find a primer that best matches those sequences with a degeneracy limit set by the user, and keeps the number of mismatches to the template sequence at a minimum, thus ensuring that PCR can be used as a sensitive, diagnostic tool (Jabado *et al.* 2006).

Limitations. SCPrimer is a software program designed to generate multiple PCR primers to cover a set of given sequences. As such, it may not necessarily optimize one specific primer set to the extent that HYDEN does, so multiple PCR reactions would be needed to obtain full coverage of a gene. Also, since it uses only the nucleic acid sequence and does not consider the amino acid sequence, it may not be able to efficiently design primers for more distantly-related sequences that may be conserved at the protein level, but not at the DNA level.

iCODEHOP

iCODEHOP is a web-based program (available at http://dbmi-icode-01.dbmi.pitt.edu/i-codehop-context/iCODEHOP/Welcome) that implements an updated version of an older degenerate primer design program known as CODEHOP. Using this program, Consensus-Degenerate Hybrid Oligonucleotide Primers (CODEHOP) are generated based on conserved regions on the amino acid level. Primers consist of a relatively short 3’ degenerate core that is based on a conserved motif of 3-4 amino acids, and a 5’ non-degenerate consensus clamp that stabilizes hybridization to the target template (see Figure 3; Boyce *et al.* 2009). Though it does not computationally optimize solutions as intensively as HYDEN or SCPrimer, it provides a perhaps more elegant solution to the problem of degenerate primer design by constraining parts of the method to biologically relevant approaches.

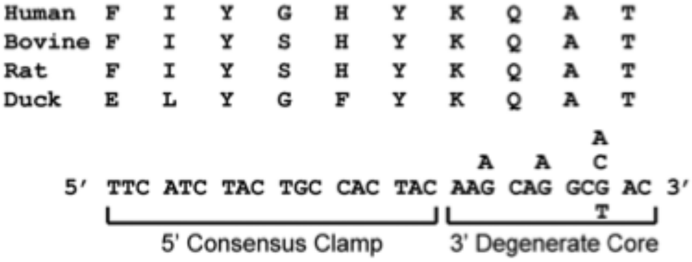

Human F I Y G H Y K Q A T
Bovine F I Y S H Y K Q A T
Rat F I Y S H Y K Q A T
Duck E L Y G F Y K Q A T
A
A A C
5′ TTC ATC TAC TGC CAC TAC AAG CAG GCG AC 3′
T
5' Consensus Clamp
3' Degenerate Core

Figure 3. CODEHOP reverse translates conserved amino acid sequences to a 5' Consensus Clamp and a 3' Degenerate Core. From Boyce *et al*. 2009.

Methods. iCODEHOP is an interactive web-based program that incorporates multiple web-based programs and exhibits transparency in the information it uses to design the primers. Thus, the user has access to a large number of inputs and parameters. Although many more options are available, I will only outline a general workflow using the program that begins with inputting individual amino acid sequences. In this case, the program uses the ClustalW algorithm to align the sequences, with alignment parameters set by the user (Boyce *et al.* 2009). After an alignment is generated, the next step is to find blocks in the multiple sequence alignment representing conserved motifs, using the Block Maker program available at http://blocks.fhcrc.org/. After this step, primers are designed from the most highly conserved amino acid sequences within the block (Rose *et al.* 1998). The user must specify a codon usage table for a particular organism, in addition to specifying parameters such as melting temperature. Based on this information, iCODEHOP will reverse translate the amino acid sequence to a DNA coding sequence. It will design a 3' degenerate core by utilizing all possible codons to form the nucleotide sequence of 3-

4 highly conserved amino acids. A 5' consensus clamp is then generated by selecting the most likely codon for the consensus amino acid sequence immediately upstream of the core region (See Figure 3; Boyce *et al.* 2009). The length of the primers are selected based on desired melting temperature. Several primers for the sequence are shown to the user in both the forward and reverse direction, and more detailed information about melting temperature and sequence information is available to the user (Boyce *et al*. 2009). As the name implies, these primers are a hybrid of a consensus region and a degenerate 3' end, and can be widely used to detect functional genes from unknown genomes (Chakravorty and Vigoreaux 2010).

Advantages. There are several advantages to using iCODEHOP. On a practical level, the interface has recently been updated to enhance usability so that the user has several options in changing parameters and deciding how sequences should be weighted in designing primers. At each step, the user can modify the input to the next step.

From a biological perspective, iCODEHOP may also present primers that are more likely to amplify a diverse array of targets in a PCR reaction. Unlike HYDEN and SCPrimer, it takes the amino acid sequence as input rather than the DNA sequence. Since sequences are often much more highly conserved at the amino acid rather than at the nucleotide sequence level, this program parses the sequences into conserved blocks that may be more biologically meaningful. For example, these may be conserved motifs that may include an active site or another structurally important site. In terms of looking at the primer itself, hybrid consensus-degenerate primers may be more efficient in achieving high levels of amplification in a PCR reaction. In initial rounds of PCR, the degenerate 3' end will be effective in capturing a large proportion of the diversity of the sample. In later rounds, the consensus clamp will ensure that primers can bind to the initial product sequences, since it is uniform among all primers (see Figure 4; Rose *et* *al.* 1998). Even though mismatches of the consensus sequence to the target template will occur, these mismatches are not as egregious as mismatches at the 3' end of a primer, and so amplification may still occur (Rose *et al.* 1998).

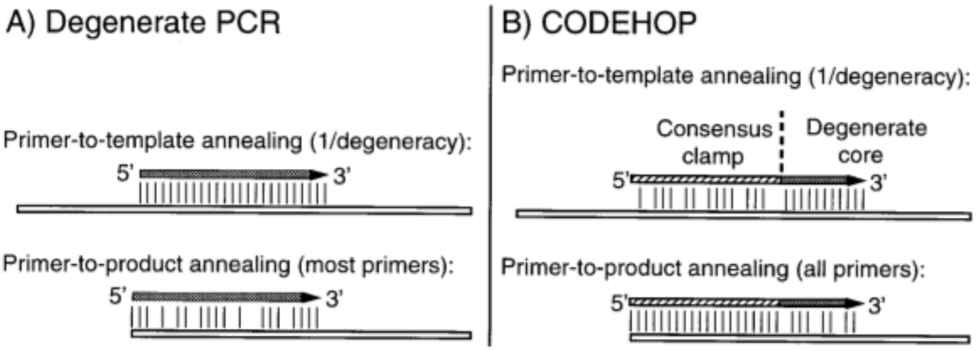

A) Degenerate PCR
Primer-to-template annealing (1/degeneracy):
5'
3'
Primer-to-product annealing (most primers):
5'
3'
B) CODEHOP
Primer-to-template annealing (1/degeneracy):
Consensus clamp
Degenerate core
5'
3'
Primer-to-product annealing (all primers):
5'
3'

Figure 4. Comparison of A) Degenerate PCR using highly degenerate primers, and B) CODEHOP PCR. From Rose *et al*. 1998.

This program has been used widely, and has been validated not only by the authors of the software, but also by various researchers who have used the program to look for functional genes in samples as diverse as activated sludge and engineered anaerobic bioreactors (McMahon *et al.* 2002; Pereyra *et al.* 2010).

Limitations. iCODEHOP has been widely used to design degenerate primers for various functional genes. However, since it relies on amino acid sequence, it cannot be used to design degenerate primers targeting 16S rRNA, as one might do for studying microbial diversity. This may not be a major limitation since 16S rRNA genes are generally more highly conserved than proteins; sequences may be sufficiently related to design primers manually.

More significant limitations include that the nucleotide sequences are not used at any point by the program to optimize the reverse-translation of amino acid to nucleotide base. Although an organism's specific codon usage table can be selected, if no table is available for a particular organism, the primer sequence, especially at the 5' consensus end, may have several mismatches to even the input sequences. Future versions of iCODEHOP may implement use of DNA sequence information (Staheli *et al.* 2011). Additionally, CODEHOP does not make an attempt to optimize the degeneracy found in the primers; it is simply in the location of the wobble base of a codon. This may be an advantage if this extra degeneracy allows amplification of novel sequences, but may just be unnecessary degeneracy (Jabado *et al.* 2006).

DISCUSSION

Although I have outlined the methods of three programs that design degenerate primers, only iCODEHOP seems to have been widely adopted. And yet, even in most cases, primers are still designed manually, though this practice does rely on computational methods for creating multiple sequence alignments at the protein or nucleic acid level (Chakravory and Vigoreaux 2010; Lang and Orgogozo 2011). In the case of less conserved regions, CODEHOP or its newer web-implementation iCODEHOP seems to be the program of choice (Chakravory and Vigoreaux 2010; McMahon *et al*. 2001; Pereyra *et al.* 2011).

What can we learn from other software programs that have been less successful in being adopted by a wide variety of researchers? HYDEN, though not optimal, gives us an idea of what computation can do for us: we can optimize the total degeneracy of a primer by carefully weighing the advantage of including or removing extra degeneracy at each position. SCPrimer also includes an optimization algorithm, in this case, to find a set of degenerate primers that covers a set of input sequences most effectively. An improvement over the HYDEN software is that SCPrimer also incorporates a scoring function to evaluate the primer based on biological constraints, so that primers are screened for annealing temperature, secondary structure, and other parameters important for PCR. CODEHOP stands out from these programs, however, because it uses biological parameters not just as a scoring function, but as a basis for the primer design itself. By aligning amino acids, not just the nucleotide bases, CODEHOP is able to find conserved motifs that may not be as immediately apparent in the nucleotide sequence, but that would be important in grouping homologous genes together. Additionally, by using a 5' consensus clamp and a 3' degenerate core, the CODEHOP method effectively weights the 3' end of the primer to have fewer mismatches with the template. Thus, we are more likely to achieve more reliable PCR amplification from these primers.

To design a better degenerate primer design algorithm, I would attempt to utilize the strengths of the above algorithms: the methodical approach for optimizing primers found in HYDEN; the scoring function used by SCPrimer; and the biological constraints imposed by CODEHOP. For functional gene analysis, amino acid sequences should be used to look for conserved motifs, but it also makes sense to incorporate the sequence data available in the DNA sequences when reverse-translating amino acids to codons. These primers could include a few degenerate positions in the 5' region, which would be optimized using a hill-climbing procedure similar to that found in HYDEN. Finally, an additional step that a program could take would be to check primer sequences for specificity. Currently, users must check specificity of the primers separately from designing them, often by doing a BLAST search using the primers as a query. However, I would incorporate this as an additional step to a software program. The input would include not only a set of target sequences, but a set of similar, but non-target sequence. Primers should match to the target sequence, but should not match the non-target sequences. In this way degenerate primers could be designed that were sensitive and specific to for a particular subset of sequences.

Ultimately, software programs have been effective in constructing primers to discover novel members of gene families across species. However, each has limitations that must continue to be addressed in future implementations of the software.

REFERENCES

Boyce,R., Chilana,P., and Rose,T.M. (2009) iCODEHOP: a new interactive program for designing Consensus-Degenerate Hybrid Oligonucleotide Primers from multiply aligned protein sequences. Nucleic Acids Res., 37, W222-W228.

Chakravorty,S., and Vigoreaux,J.O. (2010) Amplification of Orthologous Genes Using Degenerate Primers. Methods Mol. Bio., 634, 175-185.

Fuchs,T., Malecova,B., Linhart,C., Sharan,R., Khen,M., Herwig,R., Shmulevich,D., Elkon,R., Steinfath,M., O'Brien,J.K., Radelof,U., Lehrach,H., Lancet,D. and Shamir,R. (2002) DEFOG: A Practical Scheme for Deciphering Families of Genes. Genomics, 80, 295-302.

Giegerich,R., Meyer,F. and Schleiermacher,C. (1996) GeneFisher- Software Support for the Detection of postulated genes. Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology. AAAI Press, Menlo Park, CA 94025, pp. 68–77.

Jabado,O.J., Palacios,G., Kapoor,V., Hui,J., Renwick,N., Zhai,J., Briese,T. and Lipkin,W.I. (2006) Greene SCPrimer: a rapid comprehensive tool for designing degenerate primers from multiple sequence alignments. Nucleic Acids Res., 34, 6605–6611.

Lang,M., and Orgogozo,V. (2011) Identification of Homologoous Gene Sequences by PCR with Degenerate Primers. Methods Mol. Bio., 772, 245-256.

Linhart,C. and Shamir,R. (2005) The degenerate primer design problem: theory and applications. J. Comput. Biol., 12, 431–456.

McMahon KD, Dojka MA, Pace NR, Jenkins D, Keasling JD (2002) Polyphosphate kinase from activated sludge performing enhanced biological phosphorus removal. Appl. Environ. Microbiol. 68:4971–4978

Pereyra,L.P., Hilbel,S.R., Prieto Riquelme,M.V., Reardon,K.F., and Pruden,A. (2010) Detection and Quantification of Functional Genes of Cellulose-Degrading, Fermentative, and Sulfate-Reducing Bacteria and Methanogenic Archaea. Appl. Environ. Microbiol., 76, 2192-2202.

Rose,T.M., Schultz,E.R., Henikoff,J.G., Pietrokovski,S., McCallum,C.M. and Henikoff,S. (1998) Consensus-degenerate hybrid oligonucleotide primers for amplification of distantly related sequences. Nucleic Acids Res., 26, 1628–1635.

Staheli,J.P., Boyce,R., Kovarik,D., and Rose,T.M. (2011) CODEHOP PCR and CODEHOP PCR Primer Design. Methods Mol. Bio., 687, 57-73.