Margaret Gentile

Computational Methods for the Design of PCR Primers for the Amplification of Functional Markers from Environmental Samples

Introduction

Molecular techniques are becoming increasingly popular for exploring the diversity, function, and structure of microbial communities. Looking at DNA sequences from environmental samples with molecular techniques allows researchers to understand the physiology of organisms that cannot be cultured in the lab. Functional markers are genes specific to a particular metabolic function of interest. For example, ammonia monoxygenase (amo) is a functional marker for nitrification and nitrite reductase (nirS) serves as a marker for denitrification. To assess the diversity of species with a particular metabolic function in a community, functional markers are amplified by PCR, cloned and sequenced (Braker *et al.,* 2000). Functional gene microarrays can then be constructed and used to study community composition (DNA) and functioning (cDNA) (Wu *et al.,* 2001).

The PCR amplification of a functional marker requires primers. The design of primers for the amplification of a specific gene from many different species is not a trivial task. The functional markers in the sample can be highly divergent from known sequences, but primers must be very similar to target sequences for efficient amplification. The methods for the design of primers for the amplification of a functional marker from many bacteria in an environmental sample have been ad hoc to date (Braker *et al.,* 1998, Hallin and Lindgren, 1999). This paper reviews the current state of computational methods for PCR primer design and analyzes how these methods with improvements can be incorporated into the design of primers for the amplification of divergent functional markers.

Lack of computational methods in current designs

Studies amplifying sequences from known environmental samples have not been computational to date. As a result, the results may have underestimated diversity. In general, known sequences have been globally aligned , and primers designed for regions which appear to be conserved. The following two studies designed primers for the gene nirS for use in assessing diversity of denitrifiers. They illustrate the weakness of current primer design methods.

In a study by Braker *et al.*, primers were designed from conserved sequence segments identified by inspection of six EMBL nirS sequences aligned with MULTIALIGN. (Braker *et al.,* 1998). The specificity of the primers was checked by doing a BLASTN search which revealed significant similarity only to nirS sequences. When these primers were used to assess the diversity of denitrifiers in a marine sediment community, the resulting clone library contained 228 putative clones, few of which were redundant or matched previously seen nirS sequences. (Braker *et al,* 2000) A similar strategy was employed in a study by Hallin and Lindgren with the addition of adding some degenerate primers to account for some of the wobble positions. (Hallin and

Lindgren, 1999) These primers were found to amplify nirS from known denitrifying isolates and did not produce products for non-denitrifying isolates. (Hallin and Lindgren, 1999).

From these studies, it is apparent that primers can be designed which are gene specific and yet are able to amplify a diverse set of sequences for a particular gene. What is not clear is whether these primers are able to capture all of the diversity that exists or are merely sampling a subset of the actual diversity present. Primer designs relying heavily on consensus nucleotide sequences determined by non-computational methods may fail to amplify all of the probably degenerate sequences of a given gene. Computational methods for designing PCR primers for a variety of applications have been developed. Many of the ideas from these methods could be incorporated into the design of PCR primers for the amplification of degenerate functional markers. These methods include; calculation of parameters important for primer efficiency such as melting temperature and GC content, determination of consensus sequence information more rigorously from local alignments on the protein level and from biological information, determination of degenerate nucleotide sequences from probabilistic methods, and the use of novel primers composed of consensus and degenerate segments.

Basics of computational primer design

Design Parameters

Regardless of the application for which a primer is designed, several parameters are used in the design process to quantify its annealing properties and efficiency. These parameters include melting temperature, GC content, and the primer-primer interactions. The melting temperature is that at which a primer will anneal or break away from the template DNA. It depends upon the amino acid sequence and length. This temperature is often used as an input for a primer design program, because the researcher requires a primer that will work under specified reaction conditions. The melting temperature is also important for applications with greater than one primer, because primers with different melting temperatures will have different efficiencies. One method for calculating melting temperature is the nearest neighbor method. Melting temperature is calculated as a function of the sums of the entropy and enthalpy of the consecutive pairs of amino acids (Kampke *et al.*, 2001). The stability of the primer DNA duplex is important for primer design, because it will affect the efficiency of priming. The GC content describes the stability of the primer template duplex, because different energies are required to break apart GC pairs which have three hydrogen bonds and AT pairs which have only two (Kampke *et al.*, 2001). Interactions between the forward and reverse primer or a primer with itself are evaluated, because these interactions reduce amplification efficiency.

Algorithms for amplification of known gene

The complexity of designing an appropriate primer varies across applications. In many applications, the DNA sequence is known, and the design of primers is simply the identification of an appropriate segment of the known sequence. Such applications include sequencing, specific gene detection, and whole genome microarray construction. In sequencing, an unknown segment of DNA is amplified for subsequent sequencing by designing primers in known segments that bracket the unknown segment. Detecting a gene in a sample is often done by PCR amplification of that gene using primers designed from the known sequence for that gene. In whole genome microarray construction, dots containing amplified fragments of the genome of a sequenced organism are spotted onto a microarray. Primers must be designed to amplify the various regions and give full coverage of the genome.

The algorithm for these applications is fairly similar. The program PRIMEARRAY is an example of such an algorithm. This program is specifically for whole genome microarray construction. In short, the program shifts along the sequence evaluating chunks of the specified primer length according to the criteria: melting temperature, GC-content, and interactions with self and other primers. When primers that meet all of the criteria are found, they are recorded into the output file (Raddatz *et al.*, 2001). Other available methods have improved upon this brute force method in an effort to speed up the evaluation of criteria. DOPRIMER is a faster dynamic programming algorithm. It approximates numeric values for the criteria and selects a list of best candidates. More rigorous, time consuming calculations for the criteria are then only done for these best candidates (Kampke *et al.*, 2001).

Challenges of primer design for unknown, diverse sequences

The design of a primer to amplify a gene of interest from all species present differs from the applications described above, because the sequence to be amplified is not actually known and can be quite different from known sequences of the gene. This challenge also arises when designing primers to amplify unknown members of a gene family. The process for designing this type of primer is much more complex than for the amplification of known sequences. The primers must operate on generally conserved regions and yet amplify very divergent sequences. The obvious strategy for designing the primers is to look for conserved regions within known sequences for the gene of interest and then design the primers within those regions.

Previously, we discussed two attempts at designing primers in conserved regions (Braker *et al*, 1998 and Hallin and Lindgren, 1999). However, both of these attempts neglected to deal with the challenges which complicate the search for so called "conserved regions". On the protein level, many different amino acid sequences could yield the same functional protein, because amino acids in some regions of the protein can replace each other without affecting activity. Further variation occurs due to the degeneracy of the genetic code. Many different nucleotide sequences can be translated into the same amino acid sequence. Primers work with varying efficiencies based upon how similar they are to the target sequence. If the primer matches the target sequence perfectly, it will anneal more strongly and amplify more efficiently than if there are base pair mismatches. This presents quite a dilemma for designing primers for diverse sequences. If certain sequences are favored, they will be preferentially amplified, and then other equally important sequences will be missed.

Methods of primer design for unknown, diverse sequences

Primers must be designed in regions of DNA that are highly conserved. In order to find a conserved region, several sequences of the gene of interest must be studied. The sequences are aligned, and conserved regions identified. But which sequences should be aligned, nucleotide or amino acid? How should the alignment be done? How are conserved regions identified? What is the biological basis for these procedures? Upon the identification of a conserved region, how will the primer be designed? This section will seek to answer these questions.

Nucleotide versus Amino Acid

Computational methods have been reported for designing PCR primers for divergent sequences from multiple sequence alignments from both amino acid sequences (Kariko, 1995) and from protein sequences (Rose *et al.,*1998). In the study using amino acid sequences, primers were designed that could amplify the same gene from many species which, "are known to preserve the primer-targeted sequences," (Kariko, 1995), but not for species in which the primer targeted sequences are different from known sequences. When amino acid sequences are used, the biological significance of sequence conservation is more apparent. Candidate conserved regions can be further evaluated based upon the role they play in the function of the enzyme. Specific residues may be more or less likely to be conserved based upon whether they are stabilizing a secondary structure, binding a cofactor, or participating in a reaction at the active site. Depending upon the application, the choice of which conserved region to use can be aided with this additional information. Also, the possible codons can be predicted from the amino acid sequence allowing for the design of degenerate primers. If the correct reading frame is known for a DNA sequence, the same analyses can be made after translation of a nucleotide sequence as for an amino acid sequence. However, it is less useful to base the design on solely the nucleotide sequences, because much of the possible diversity will be missed.

Alignment Method

In previous methods for primer design for amplification of functional markers, global alignment methods were used for sequence alignment (Braker *et al.*, 1998 and Hallin and Lindgren, 1999). For example, the MULTIALIGN algorithm used by Braker *et al.* is a global algorithm which seeks to make the best possible alignment over the entire length of the gene (Mount, 2001). However, a local alignment is actually more appropriate for primer design, because it is only important to identify short continuous segments of conserved sequence. The program COnsensus-DEgenerate Hybrid Oligonucleotide Primers (CODEHOP) uses sequences aligned with BLOCKMAKER which yields short highly conserved segments, more suitable for primers design (Rose *et al.*, 1998). Using blocks as an alignment method is also advantageous, because the blocks are also fairly conserved regions. If a global alignment is done, further computation must be done to identify conserved regions.

Primer Design from Conserved Segments

Once a conserved amino acid sequence has been established, the issue of degeneracy in the genetic code still plagues the design of flexible primers. One strategy to get around the degeneracy issue is to synthesize many different primers for all of the possible amino acid sequences and codon usages. Using too many primers within one PCR reaction creates problems, because the concentration of each individual primer and efficiency decreases (Rose *et al.*, 1998). The opposite strategy is to design one primer that has the most common amino acids or nucleotides at each position (Rose *et al.*, 1998). This strategy is not appropriate for amplifying unknown sequences, because distantly related sequences will not be amplified (Rose *et al.*, 1998). The CODEHOP method has been developed to deal with the degeneracy issues in primer design for amplification of divergent sequences. It is discussed in the next section.

CODEHOP

CODEHOP is a program specifically developed to create primers for the amplification of distantly related sequences (Rose *et al.*, 1998). The method was developed to amplify unknown genes within gene families. These sequences may be highly divergent from the known sequences of the family. Although the program was developed with a different goal in mind, the same issues face the design of primers for the same gene from different organisms. The program is novel and powerful both because innovative primers are designed and because creative bioinformatics tools are used.

Innovative primers tackle degeneracy problems

This method is basically a hybridization of making many degenerate primers and designing a single consensus primer. Primers are designed to have both a degenerate segment at the 3' end and a consensus clamp at the 5' end. The degenerate segment is designed in a highly conserved segment of amino acids as short as 3 residues. Then primers will be designed with all of the possible combinations of codons which code for these amino acids. Creating the possible primers for only 3 residues yields a much smaller pool than if creating degenerate primers for a typical primer length of 7 amino acids. The consensus segment of the clamp ensures strong annealing over rounds of replication and lengthens the primer making it more specific. When annealing to an original strand of sample DNA, there may be mismatches in this consensus region and reliance is placed upon the degenerate segment and the non wobble amino acids of the clamp. In subsequent rounds of PCR, the primer will anneal very well to the end of the segment that was a primer in the previous round. These features of the CODEHOP primers are discussed in Rose *et al.*, 1998.

Computational methods of CODEHOP

The computational methods used in CODEHOP, along with the novel consensus-degenerate regions, are very rigorous and allow for amplification of much more diversity. The method was outlined in *Nucleic Acids Research* (Rose *et al.*, 1998). CODEHOP works with amino acid sequences that have first been made into BLOCKS with BLOCKMAKER. This is a local alignment method that focuses on finding really short and highly conserved amino acid sequences. This is an improvement over other mentioned primer design algorithms that preformed global alignments and then looked for conserved regions. The program then allows for sequences to be weighted for importance in the design. Using a position specific scoring matrix (PSSM), a consensus amino acid is calculated as the highest scoring amino acid at each position in the aligned sequence. A DNA PSSM is calculated by using the amino acid specific scoring matrix and an appropriately chosen codon usage table. The codon usage table tabulates how many times each codon is used for a given amino acid in the organism (Nakamura *et al.*, 1997). Using this table, probabilities of nucleotides in all positions including the wobble position are calculated. The sequence is then scanned for possible degenerate regions and a value calculated for degeneracy from the nucleotide position specific scoring matrix. Degenerate segments beginning with a highly conserved 3' end, 11-12 nucleotides long, and with a degeneracy score below some maximum (default=128) are selected as candidates. The consensus clamp is then designed as an extension of the degenerate region based upon the most common codon for each consensus amino acid. The length of the consensus clamp is determined by melting temperature calculations.

Analysis of existing methods and suggestions for improvement

Out of these computation methods that have been developed for designing primers from multiple sequence alignments, none specifically amplify one gene of interest from different organisms. The CODEHOP method is the most appropriate, because it develops primers that can amplify unknown sequences that may be quite different from those used to develop the primers. However, there are several differences between amplifying an unknown that belongs to a gene family and amplifying a particular gene from unknown organisms.

Codon Usage Table

When looking for new genes that belong to a family, the search is often within the same organism. In the CODEHOP program, the codon usage table used for creating the DNA PSSM is organism specific, because the primers are supposed to look for members in the gene family within a given organism. The codon usage tables have been tabulated for specific proteins, for organisms, and for larger divisions such as primates, rodents, and bacteria. When designing primers for amplification from many different organisms, it is more appropriate to use tables that have been tabulated for bacteria rather than for any specific organism.

Searching for a specific gene vs a family of genes

The incorporation of biological information along the way in CODEHOP could make it suitable for designing primers to amplify one gene rather than a gene family. If the gene of interest belongs to a family of genes which share a binding site, or some such feature it is possible that the primer designed will be too general and could amplify sequences of other members of the family besides the gene of interest. Earlier cited studies were amplifying the gene nirS. This gene belongs to the cytochrome c and heme d1 binding families. Primers designed in the regions which bind the heme group may result in amplification of other heme binding proteins. A check should be added to the program to check for possible amplification of related genes. This could be done by comparing the blocks to sequences in families and looking for sequences of overlap with distant members. Alternatively, proposed degenerate primers could be compared to the sequences in the family to check for possibilities of nonspecific priming.

Conclusion

There is considerable room for improvement in the methods used to design primers for amplification of functional markers from environmental samples. Primers that include a degenerate core and consensus clamp will be capable of amplifying functional markers in the environment which may be highly divergent from known sequences. To design such primers, degenerate regions and consensus sequences should based upon probabilistic computations of locally aligned sequences. Codon usage tables for bacteria in general should be used for construction of DNA position specific scoring matrices. Biological information should be incorporated where appropriate to prevent indiscriminate priming. These improvements will yield primers more capable of amplifying divergent functional markers and unfold new possibilities for the study of diversity, function, and structure of microbial communities.

References

Braker, G., Fesefeldt, A., and Witzel, K. 1998. “Development of PCR Primer Systems for Amplification of Nitrite reductase Genes (nirK and nirS) to Detect Denitrifying Bacteria in Environmental samples.” *Applied and Environmental Microbiology.* 64(10):3769-3775.

Braker, G., Zhou, J., Wu, L., Devol, A., and Tiedje, J. 1000. “Nitrite Reductase Genes nirK and nirS) as Funcitonal Markers to Investigate Diversity of Denitrifying Bacteria in Pacific Northwest Marine Sediment Communities”. *Applied and Environmental Microbiology.* 66(5):2096-2104.

Hallin, S., and Lindgren, P. 1999. “PCR Detection of Genes Encoding Nitrite Reductase in Denitrifying Bacteria.” *Applied and Environmental Microbiology.* 65(4):1642-1657.

Kampke, T., Kieninger, M., and Mecklenburg, M. 2001. “Efficient Primer Design Algorithm.” *Bioinformatics.* 17(3):214-225.

Kariko, Katalin. 1995. “Identification of Conserved Sequences for PCR Primer Design by Multiple Alignments of Dot Matrix Plots.” *Biotechniques*. 18(6):1048-1049.

Mount, D. W. (2000). Bioinformatics: Sequence and Genome Analysis. Cold Spring Harbor Hew York, Cold Spring Harbor Press. 564 Pages

Nakamura, Y., Gojobori, T., and Ikemura, T. “Codon Usage Tabulated from the International DNA Sequence Databases.” *Nucleic Acids Research.* 25(1):244-245.

Raddatz, G., Dehio, M., Meyer, T., and Dehio, C. 2001. “PrimeArray: Genome-Scale Primer Design for DNA-Microarray Construction.” *Bioinformatics* 17(1):98-99.

Rose, T., Schultz, E., Henikoff, J., Pietrokovski, S., McCallum,C., and Henikoff, S. 1998. "Consensus-Degenerate Hybrid Oligonucleotide Primers for Amplification of Distantly Related Sequences." *Nucleic Acids Research.* 26(7):1628-1635.

Wu, L., Thompson, D., Li, G., Hurt, R., Tiedje, J., and Zhou, J. 2001. "Development nd Evaluation of Functional Gene Arrays for Detection of Selected Genes in the Environment." *Applied and Environmental Microbiology*. 67(12):5780-5790.