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