Bioinformatics Advance Access published June 1, 2006

BIOINFORMATICS

PROBER : Oligonucleotide FISH Probe Design Software

Nicholas Navin [1,2], Vladimir Grubor [2], Jim Hicks [2], Evan Leibu [2], Elizabeth Thomas [1], Jennifer Troge [2], Michael Riggs [2], Pär Lundin [3], Susanne Månér [3], Jonathan Sebat [2], Anders Zetterberg [3] and Michael Wigler [2]

[1] Watson School of Biological Sciences, Cold Spring Harbor, NY 11724, USA

[2] Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA

[3] Karolinska Institutet, Cancer Center Karolinska, Stockholm, Sweden

Associate Editor: Martin Bishop

ABSTRACT

PROBER is an oligonucleotide primer design software application that designs multiple primer pairs for generating PCR probes useful for fluorescence in-situ hybridization (FISH). PROBER generates Tiling Oligonucleotide Probes (TOPs) by masking repetitive genomic sequences and delineating essentially unique regions that can be amplified to yield small (100-2000bp) DNA probes that in aggregate will generate a single, strong fluorescent signal for regions as small as a single gene. TOPs are an alternative to bacterial artificial chromosomes (BACs) that are commonly used for FISH but may be unstable, unavailable, chimeric, or non-specific to small (10-100kb) genomic regions. PROBER can be applied to any genomic locus, with the limitation that the locus must contain at least 10kb of essentially unique blocks. To test the software, we designed a number of probes for genomic amplifications and hemizygous deletions that were initially detected by Representational Oligonucleotide Microarray Analysis (ROMA) of breast cancer tumors.

Availability: http://prober.cshl.edu

Contact: navin@cshl.edu

1 INTRODUCTION

Identification of submicroscopic chromosome abnormalities is useful in the clinical diagnosis of diseases, including mental retardation, autism and cancer. The detection of heritable copy number polymorphisms (CNPs) in the normal population (Sebat et al., 2004) and in cancer amplifications and deletions (Lucito et al., 2003) may be important for studying human disease and genome evolution. Whole genome microarray analysis using Comparative Genomic Hybridization (CGH) or ROMA provides a method for initial discovery of these variations, and create a corresponding need for validation and more accurate quantification by interphase or metaphase FISH. In order to target very specific locations of the genome that are separated by as little as 50kb, we have developed a method for designing Tiling Oligonucleotide Probes for any specified genomic region. Coverage of as little as 20% of a 100kb region with essentially unique short sequences provides hybridization probes sufficient for robust FISH analysis.

DESIGN OVERVIEW

Genomic DNA sequences are retrieved from a server, masked for repetitive exact string matches in the human genome, and analyzed for contiguously amplifiable, nearly repeat free regions of sufficient aggregate length. These regions are then searched for optimized PCR forward and reverse primers, resulting in a collection of oligonucleotide probes that can be used to PCR amplify and purify a collection of longer probes that we combine into a cocktail for FISH analysis.

Contact : Navin@cshl.edu

MERMATCH

PROBER initiates probe designs by requesting a target genomic sequence 10-100kb in length from "DAS.DNA", a Distributed Annotation Sever specific to a human genome freeze from UCSC **(**Dowell et al., 2001**)**. Short sequence substrings of a specified ("*mer.match.length*") length in the target DNA sequence having multiple exact matches elsewhere in the genome are masked using the "MerMatch" algorithm. This algorithm is based on the "MerEngine" (Healy et al. 2003). The MerEngine marks every substring of *mer.match.length* in the target sequence with the number of its exact matches in the human genome. To operate this algorithm, and other algorithms that we use routinely for probe design, a database of the human genome is compressed using a Wheeler-Burrows transformation into a suffix array that is stored in an external file. The database is loaded into 1gigabyte of RAM minimizing execution time. MerMatch masks the "frequent mers" in the human genome, where "frequent" is defined as the number of exact matches greater than a user-specified parameter ("*mer.count.cutoff*").

TOLERANCE

"Tolerance" is a program that finds regions "suitable" to be hybridization probes. We first convert the masked sequence output of MerMatch into a binary string, with 0's indicating the frequent mers. Positions within the string with "*consec.freq*" consecutive frequent mers are then marked as "condemned zones" by setting them to a large negative number, and no region overlapping a condemned zone is ever considered suitable to be a hybridization probe. Using successive cumulative sums, we mark a region suitable to be a hybridization probe if it has a specified ("*min.length*") minimal length, but less than a specified ("*repeat.tolerance*") proportion of frequent mers. Our default values are 0.2 for *repeat.tolerance*, 100 for *min.length*, 18 for *mer.match.length*, 1 for *mer.count.cutoff*, 5 for *consec.freq*. By setting *repeat.tolerance* higher, *consec.freq* higher, *mer.match.length* longer, or *mer.count.cutoff* higher, we increase the tolerance for repeats in the regions considered suitable as a hybridization probe.

PROBE DESIGN

The desired probe size range (100-2000bp) for Tier 1 and for Tier 2 probe selection are specified along with the primer Tm range (55-80 ° C), *mer.match.length* (15,18, 21 mer), maximum number of nucleotide repeats (n < 4) and base pair spacer (if a distance between probes is desired). Every possible primer sequence is extracted from the masked DNA sequence within a size range of 15-30bp and placed in a 3D matrix. Primer melting temperature (Tm) is calculated using the Rychlik method (Rychlik et al.)

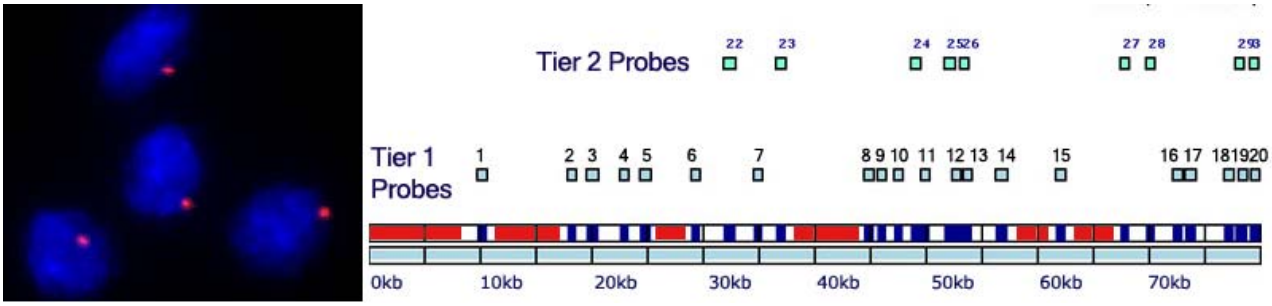

Fig. 1. A cocktail of probes 1-29 from two tiers of probe selection within an 80kb region generates a highly specific single fluorescent signal by FISH. Highly repetitive areas (red) are avoided. Blue areas are covered by Tier1 or Tier2 probes. White areas did not have suitable probe primers. FISH analysis shows a hemizygous loss of an 80kb region on chromosome 16q1 in a homogenous population of breast tumor cells.

which is based on the nearest neighbor Borer method (Borer et al.) (Tm = 81.5 + 16.6(log[Na+]) + 0.41(%GC) – 675/probe length). Primer pairs are matched according to minimal Tm deviation and primers outside of a specified Tm range or GC percentage are eliminated. The remaining primers are subjected to the G/C clamp rule (must end in G/C at the 3' end to control mispriming) and must contain no polypyrimidines or polypurines that could promote non-specific annealing (max repeat nucleotides < 4 by default) **(**Dieffenbach et al., 1995**)**. Additionally, the three nucleotides at the 3' end of each primer are scored according to the presence of a GC clamp, but absence of any GC dinucleotides that may facilitate primer dimerization.

Probe selection proceeds by selecting the forward set of primers for a single base pair position and then jumping ahead by the probe length distance (100-2000bp) in the matrix until the highest scoring set of the reverse primers are located. If the primers at either base pair position do not meet the primer rules, then the next forward or reverse primer set is considered (n+1). The two columns of forward and reverse primers are compared and the primers with the closest Tm match are selected, resulting in a final probe sequence. Probe sequences that have been utilized are marked in the DNA sequence, so that they will not be reused during 'Tier 2' probe selection, where more relaxed parameters are used to identify additional probes.

Finally the Percent Genome Coverage (PCG) [(Bp Sequence covered with probes / Total Bp) * 100] is calculated and the probe distribution is visualized in a graphical plot. We have determined that a PGC > 20.00 % of a 100kb sequence will not compromise the fluorescent probe signal in FISH. The output can be saved as a full report or short report (forward/reverse primer sequences) formatted text file.

2 RESULTS

SIMULATIONS

http://prober.cshl.edu/simulations.html

APPLICATION

http://prober.cshl.edu/applications.html

3 IMPLEMENTATION

PROBER was written in C# 2.0 for Microsoft Windows and requires installation of the dot net framework 2.0 for runtime.

ACKNOWLEDGEMENTS

John Healy, Lakshmi Muthuswamy, Joan Alexander, Eirny Tholsdorf and Kristin Anna. This work was supported by the Simon's Foundation, The Lindsey-Goldberg Foundation, grants to AZ from the Swedish Cancer Society (grant number 0046-B04-38XAC) and from the Stockholm Cancer Society (grant number 03:171 and 02:144)

REFERENCES

Dieffenbach, C.W. et al. (1995) General Concepts for PCR Primer Design, in PCR Primer, A Laboratory Manual. Cold Spring Harbor Laboratory Press, New York, pp.133-155.

Dowell, R.D. et al. (2001) The Distributed Annotation System. BMC Bioinformatics. 2:7. Epub 2001 Oct 10.

Lucito R, et al. (2003) Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. Genome Research. 10, 2291-305.

Healy J, et al. (2003) Annotating large genomes with exact word matches. Genome Research, 10, 2306-15.

Rychlik W, et al. Optimization of the annealing temperature for DNA amplification in vitro. Nucleic Acids Res 18, 6409-12 (1990).

Sebat J, et al. (2004) Large-scale copy number polymorphism in the human genome. Science, 23;305 (5683):525-8

Wallace, R.B. et al (1979) Hybridization of synthetic oligodeoxyribonu cleotide to phi chi 174 DNA: the effect of single base pair mismatch. Nucleic Acids Research, 6 (11), 3543-57