BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 19 2006, pages 2437–2438
doi:10.1093/bioinformatics/btl273

Genome analysis

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

Received on February 22, 2006; revised on May 6, 2006; accepted on May 24, 2006

Advance Access publication June 1, 2006
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 (100–2000 bp) 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 (10–100 kb) genomic regions. PROBER can be applied to any genomic locus, with the limitation that the locus must contain at least 10 kb 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 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 Representational Oligonucleotide Microarray Analysis (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 50 kb, we have developed a method for designing Tiling Oligonucleotide Probes for any specified genomic region. Coverage of as little as 20% of a 100 kb 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 searched for optimized PCR forward and reverse primers, resulting in a collection of oligonucleotide probes. Individual tiling probes are then PCR amplified and combined into a cocktail for FISH analysis.

MerMatch. PROBER initiates probe designs by requesting a target genomic sequence 10–100 kb 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 1 Gb 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 0s 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 0.8 for repeat.tolerance, 100 for min.length, 18 for mer.match.length, 1 for mer.count.cutoff, 10 for min.length. By setting repeat.tolerance lower, min.length higher, mer.match.length longer or mer.count.cutoff higher, we increase the

[*]To whom correspondence should be addressed.

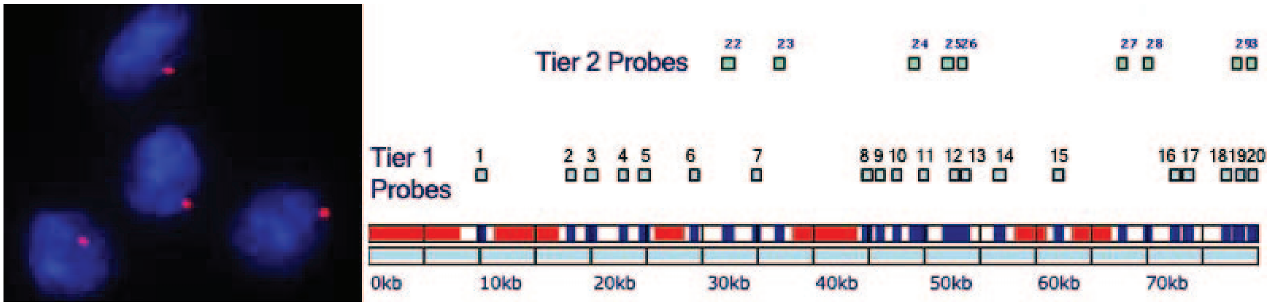

Fig. 1. A cocktail of probes 1–29 from two tiers of probe selection within an 80 kb region generates a highly specific single fluorescent signal by FISH. Highly repetitive areas (red) are avoided. Blue areas are covered by Tier1 or Tier2 probes. White areas did not have suitable probe primers. FISH analysis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

tolerance for repeats in the regions considered suitable as a hybridization probe.

Probe design. The desired probe size range (100–2000 bp) for Tier 1 and for Tier 2 probe selection are specified along with the primer T_m range (55–80°C), mer.match.length (15,18, 21mer), 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 15–30 bp and placed in a 3D matrix. Primer melting temperature (T_m) is calculated using the Rychlik method (Rychlik *et al*., 1990) which is based on the nearest neighbor Borer method (Borer *et al.*) ($T_m = 81.5 + 16.6(\log[\mathrm{Na}^+]) + 0.41(\%\mathrm{GC}) - 675/$ probe length). Primer pairs are matched according to minimal T_m deviation and primers outside of a specified T_m 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 (maximum repeat nucleotides <4 by default) (Dieffenbach *et al*., 1995). In addition, 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 (100–2000 bp) 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 T_m 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 100 kb 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

The authors thank John Healy, Lakshmi Muthuswamy, Joan Alexander, Eirny Tholsdorf and Kristin Anna for their help. This work was supported by a fellowship from the Lindsey-Goldberg Foundation to N.N. and grants to A.Z. from the Swedish Cancer Society (0046-B04-38XAC) and from the Stockholm Cancer Society (03:171 and 02:144). This work was also supported by grants to M.W. from the National Institutes of Health (5R01-CA078544-07), Dept. of the Army (W81XWH04-1-0477), (W81XWH-05-1-0068), (W81XWH-04-0905), The Simons Foundation, Miracle Foundation, Breast Cancer Research Foundation, Long Islanders Against Breast Cancer, West Islip Breast Cancer Foundation, Long Island Breast Cancer (1 in 9), Elizabeth McFarland Breast Cancer Research Grant, Breast Cancer Help Inc. M.W. is an American Cancer Society Research Professor. Funding to pay the Open Access publication charges was provided by The Simons Foundation.

Conflict of Interest: none declared.

REFERENCES

Borer,P.N. *et al*. (1974) Stability of ribonucleic acid double-stranded helices. *J. Mol. Biol*., **86**, 843–853.

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.

Lucito,R. *et al*. (2003) Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. *Genome Res*., **10**, 2291–2305.

Healy,J. *et al*. (2003) Annotating large genomes with exact word matches. *Genome Res*., **10**, 2306–2315.

Rychlik,W. *et al*. (1990) Optimization of the annealing temperature for DNA amplification *in vitro*. *Nucleic Acids Res*., **18**, 6409–6412.

Sebat,J. *et al*. (2004) Large-scale copy number polymorphism in the human genome. *Science*, **305**, 525–528.