

Probe Design for Large-Scale Molecular Biology Applications

Vincent VanBuren, Toshiyuki Yoshikawa, Toshio Hamatani, and Minoru S.H. Ko
National Institutes of Health, National Institute on Aging, Laboratory of Genetics, Developmental Genomics and Aging Section, 333 Cassell Drive, Ste. 3000, Baltimore, MD 21224
{vanburenyi, yoshikawato, hamatanito, kom}@grc.nia.nih.gov

Abstract

Large-scale molecular biology technologies such as DNA microarrays and large-scale in situ hybridization (ISH) are used to gain an appreciation of global attributes in biological tissues and cells. Although many of these efforts use cDNA probes, an approach that makes use of designed oligo probes should offer improved consistency at uniform hybridization conditions and improved specificity, as demonstrated by various oligo microarray platforms. We describe a new Web-based application that takes FASTA-formatted sequences as input, and returns both a list of the best choices for probes and a full report containing possible alternatives. Probe design for microarrays may use a scoring routine that optimizes probe intensity based upon an artificial neural network (ANN) trained to predict the average probe intensity from the physical properties of the probe and a screen for possible cross-reactivity. This new tool should provide a reliable way to construct probes that maximize signal intensity while minimizing cross-reactivity.

1. Introduction

As the application of DNA microarrays becomes more widespread and interest in designing custom arrays increases, it is becoming increasingly clear that array sensitivity and our ability to analyze array data will have an enormous effect on the direction of future biomedical research. Several new and emerging large-scale technologies, including DNA microarrays, large-scale *in situ* hybridization, and RNAi, all make use of nucleotide sequences that may have enhanced performance when rational design is applied to the nucleotide sequence.

Previous work by others has approached probe design for microarrays with a screening process, where candidate probes are screened out based on known probe features. The disadvantage of this approach is that it assumes probe features are independent of one another, an assumption that fails for at least one example: GC content is known to positively affect the binding of probe to target, but

increasing GC content also increases the likelihood of strong hairpin formation, a feature that interferes with target binding. Screening these criteria independently will not allow a full appreciation of the interaction of these factors, and so the optimal probe may not be chosen by a screening routine.

Our approach to the problem of probe design uses screening criteria for technologies that are not presently “data-rich” (e.g. large scale ISH), but will use an artificial neural network (ANN) trained to predict average fluorescence intensity from probe qualities for an oligo DNA microarray. To satisfy our aim of automating sophisticated probe design, we created Probe Hunter, a new Web-based application for large-scale probe design with the following features:

- High-throughput capacity
- Parameters for physical constraints
- Hairpin-checking
- Cross-reactivity checking against the NIA Gene Index
- Detailed report generation for follow-up troubleshooting of hybridization problems
- Customizable weighting of parameter importance

2. Methods

The Probe Hunter application, part of the Probe and Primer Design Workshop (PPDW), was written in Perl with a front-end interface written in HTML. Probe Hunter was originally designed for ISH, but has broad applicability and so may be used to design probes for Northern blots or oligo DNA arrays, for example.

A specialized program for DNA microarray probe selection will be made available as part of the PPDW in the future. EasyNN was used to train the ANN that will be used to score probes for microarray probe selection.

3. Results

For large-scale ISH, the important screening criteria for probe selection were judged to be: minimal hairpin formation, minimal cross-reactivity with undesired target

sequences, preference for the relatively unique 3' untranslated region (3' UTR) of the target, and control over the range of physical parameters important for consistency when uniform hybridization conditions are used. Hairpin energy (ΔG) is calculated using the nearest-neighbor method. Cross-reactivity is determined by a parameter threshold for the maximum percent of complementary identity of the probe and the cross-reactive target across the length of the probe. These parameters, as well as parameters to prescribe the physical properties of the probe (GC content, length, maximum distance from the 3' end of the target sequence) may all be adjusted through the Probe Hunter interface (see Availability, below).

To maintain flexibility and broad applicability, Probe Hunter was created with customizable parameters that allow weighting of parameter importance. In the design of probes for Northern blots (cDNA sub clone), for example, one may wish to use a relatively large probe (300 nt), and for this length, hairpin formation may be of less concern to the user. In this instance, the user may assign a low weight to the screen of hairpin structures. This feature allows flexibility and incorporates the user's judgment. As this project continues to develop, default parameters are adjusted to optimize results according to our experience with the experimental results of probe design.

4. Discussion

Large-scale molecular biology technologies such as DNA microarrays and large-scale *in situ* hybridization require optimal probe quality to achieve the richest possible data set. This is crucial for the most comprehensive analysis and for the computational reconstruction of biochemical/genetic pathways. ANN training will allow probe selection by considering probe features as part of a complex probe description, rather than as independent features to be screened.

Microarray data provides a valuable source of feedback into probe design and will allow a more robust approach to the design process. Hughes, et al. [1] have shown that mismatches in up to about 18 nt at the 3' end of 60-mer oligo probes, which is the end fixed to the glass, has little effect on measured fluorescence intensity. This indicates that cross-reactivity checking at the 3' end may be relaxed in the design of oligo DNA array probes, thus allowing for better optimization of probe design.

An ANN (for scoring microarray probe selection only) was trained to predict average fluorescence intensity from probe qualities as input, and measured intensity as output. ANN prediction provides an estimate of mean intensity when the transcript is of average abundance. For a particular transcript, the predicted intensity for candidate probes may then be used to rank the candidates according to which have the best sensitivity for their target.

The ability to reconstruct biochemical/genetic pathways using microarray data will in part rely on the ability to glean the absolute abundance of transcripts from measured fluorescence on microarrays. By providing a mechanism of estimating average intensity from probe qualities, we have taken an important step towards understanding the relationship between measured fluorescence intensity and absolute transcript abundance.

Probe Hunter represents the first of a new suite of programs called the Probe and Primer Design Workshop. Many probe design efforts share the same basic goals: efficient target binding and low cross-reactivity. There are specific features of some applications, however, that benefit from the use of rules specific to probe choice for that application. As such, we have plans to use the basic framework of the Probe Hunter application to expand its applicability and offer more specialized versions for the following applications: oligo DNA microarrays, primer design for RT-PCR, PNA probe and primer design, and specialized ISH applications.

5. Availability

The Web-based Probe Hunter application is part of the Probe and Primer Design Workshop project and is available at: <http://probeworkshop.grc.nia.nih.gov>. Other probe design applications will go online at this location as they become available.

6. Acknowledgements

T.Y. provided the initial motivation for this project (large-scale ISH), helped develop criteria for ISH probe selection, and gave feedback from ISH experiments. T.H. provided the microarray data for ANN training. Thanks to Geppino Falco for feedback on probes designed for Northern blots. Thanks to Alexei A. Sharov, Yong Qian, and Dawood B. Dudekula for discussion about the NIA Mouse Gene Index. D.B.D. also helped with setting up the Web server. T.Y. is a JSPS Fellow. T.H. is supported by a fellowship from the Serono Foundation.

7. References

- [1] T. R. Hughes, M. Mao, A. R. Jones, J. Burchard, M. J. Marton, K. W. Shannon, S. M. Lefkowitz, M. Ziman, J. M. Schelter, M. R. Meyer, S. Kobayashi, C. Davis, H. Dai, Y. D. He, S. B. Stephanians, G. Cavet, W. L. Walker, A. West, E. Coffey, D. D. Shoemaker, R. Stoughton, A. P. Blanchard, S. H. Friend, and P. S. Linsley, "Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer," *Nat Biotechnol*, vol. 19, pp. 342-7, 2001.