

Toward Automated Design of Cell State Detectors

Jacob Beal
Raytheon BBN Technologies
10 Moulton Street
Cambridge, MA, USA 02138
jakebeal@bbn.com

Fusun Yaman
Raytheon BBN Technologies
10 Moulton Street
Cambridge, MA, USA 02138
fusun@bbn.com

1. MOTIVATION

There are a wide range of applications in which it would be useful to have a small synthetic biology circuit that could reliably classify cell state. For example, in [5], the authors propose cancer therapy based on a circuit that uses miRNA markers to test whether a cell belongs to a particular type of cancer and then kills only those cells. The authors then demonstrate an miRNA classifier that can distinguish between HeLa cells and several other cell lines. This same approach might be applied to therapeutics for many other diseases, as well as for high-precision assays that can monitor the cell-by-cell progress of a disease being studied, and for many other possible applications.

There are a very large number of possible markers and an even greater number of combinatorial circuits that can be used to test for particular cell states of interest. Effective design and optimization of such circuits therefore demands the application of design automation techniques. We are developing an information-based technique for designing cell class detectors, which is further compatible with the automated design approaches presented in [1] and [2], thereby offering the potential for rapid design and prototyping of cell-state classifier applications.

2. FINDING CANDIDATE MARKERS

We began by evaluating the information content of individual miRNA markers from the data set in [4] of cancerous and healthy cell types in humans. This data set contains 172 cell types and 708 miRNA markers. We evaluated information content with respect to the problem of fully differentiating all cell lines, both cancerous and healthy. For this evaluation, we used a standard measure of information gain and a multiplicative Gaussian noise model for threshold testing. With each marker, we considered all thresholds at the geometric mean of two data points and selected the best threshold for each of a logarithmically scaled range of noise levels from $\sigma = 0$ to $\sigma = 10$.

The available information is shown in Figure 1, which

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

IWBDA '12 San Francisco, California, USA

Work partially sponsored by DARPA DSO; the views and conclusions contained in this document are those of the authors and not DARPA or the U.S. Government..

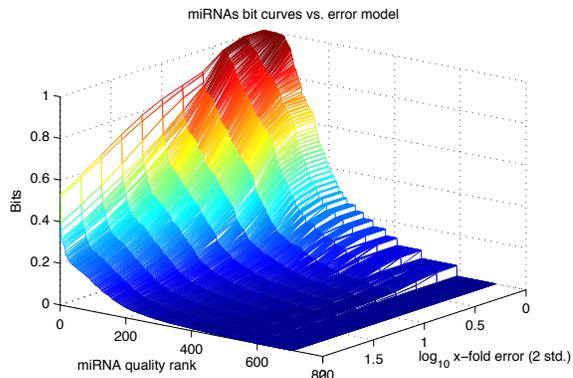


Figure 1: Bits for miRNAs in data set from [4] vs noise model, sorted by miRNA rank.

plots rank-sorted information level of miRNAs vs. noise. Importantly, we find that there are a large number of high-information candidate miRNA markers, and the information content of most of these markers does not degrade rapidly with the injection of noise. This means that the cell-state detection approach of [5] is likely to be applicable to a broad spectrum of cell classes, and that automated techniques are likely to be able to design such detectors well.

3. DESIGN OF CELL CLASS DETECTORS

Having determined that there was a large supply of miRNAs likely to be good detectors of cell state, we next constructed a method for automatic construction of a detector for a class of cells. Our preliminary method for constructing detectors uses greedy elimination, as follows:

1. Partition cells into k classes to be distinguished.
2. Compute candidate marker information gain and optimal thresholds with respect to these classes.
3. Discard all except for the C candidates with the highest information.
4. Determine which candidate can be discarded with minimal rise in the predicted misclassification rate, and discard that candidate.
5. Repeat until there are only n candidate markers remaining or until misclassification rate would rise beyond acceptable limits.

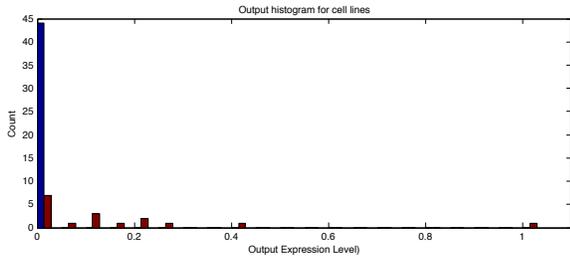


Figure 2: Automated design of a 5-marker detector to differentiate “B-ALL” cancer (red) from normal cells (blue), using our preliminary method for threshold selection, produces a circuit expected to produce no false positives (blue > 0.2) and a moderate number of false negatives (red < 0.2).

We tested this method by constructing a detector for differentiating the “B-ALL” class of cancers from all healthy cells in the data set from [4], using $C = 50$ and $n = 5$. Figure 2 shows the result of applying the model used in [5] to the thresholds computed using a digital test model. Our detector is predicted to produce no false positives and a moderate number of false negatives. Moreover, adjusting the method to match the threshold selection model with the computational model is expected to greatly reduce the number of false negatives.

This design approach could easily be combined with the automatic compilation techniques presented in [1]. The template for thresholded miRNA sensor tests might be expressed in Proto as:

```
(def miRNA< (mir|symbol threshold|scalar) boolean
:grn-motif ((RXN mir|scalar represses value)
(P|threshold value T)))

(def miRNA> (mir|symbol threshold|scalar) boolean
:grn-motif ((RXN mir|scalar represses value)
(RXN mir|scalar represses ?X)
(P|threshold ?X T)
(P R- ?X value T)))
```

and the full detector circuit as:

```
(and (miRNA< 'hsa-let-7c 5.0)
(miRNA> 'hsa-miR-130a 0.3)
(miRNA< 'hsa-miR-29b 27.0)
(miRNA< 'hsa-miR-154 0.3)
(miRNA< 'hsa-miR-197 0.3))
```

Linking marker selection to program synthesis and GRN design in this way would then allow a tool-chain approach (e.g., [2]) to rapidly construct plasmids for candidate detector circuits with a high probability of correct operation.

4. APPLICATION TO OTHER DISEASES

Although the data set discussed so far deals only with human cancers, the approach is expected to generalize. Since cell state is generally tightly linked with gene expression, it is likely that the same approach should be applicable to detect a broad spectrum of other cell states of interest, including many types of disease in many types of organisms.

To test this hypothesis, we applied our algorithm for detecting candidate markers to a transcriptome data set for a different disease. In particular, we consider Series GSE12254 RNA microarray data on the progress of LCMV arenavirus

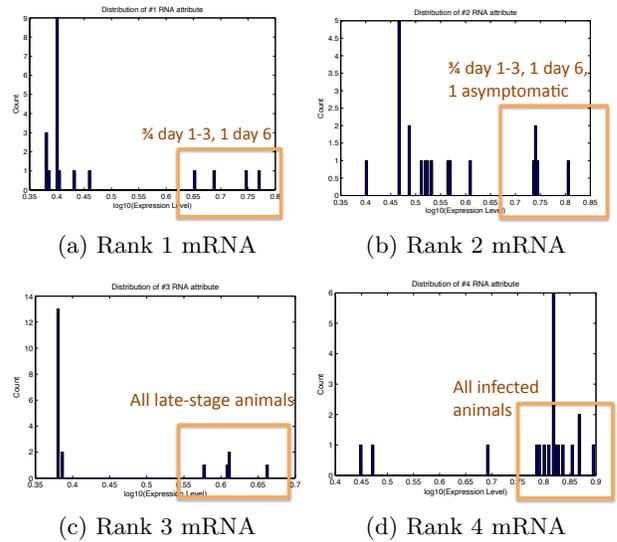


Figure 3: The highest information mRNAs in microarray data for macaques infected with arenavirus are all strongly correlated with disease states (label describes samples in above-threshold box).

in the liver of macaques, from [3]. Note that this data set contains mRNA rather than miRNA markers, and hence would require different sensor designs, though the approach is otherwise the same.

As with the human cancer data set, our algorithm finds a number of high-information candidate markers. Particularly interestingly, even when classification targets are not provided, the disease shows up quite clearly in this data set. Figure 3 shows the top four candidates, all of which are strongly correlated with particular disease states.

5. CONTRIBUTIONS AND FUTURE WORK

We have demonstrated that it is possible to use informational measures to design candidate cell-state detection circuits from miRNA array data. We are already working towards the clear next steps in developing this approach: 1) tuning the circuit thresholds for better compatibility with available sensors and computational parts and 2) realization of cell-state detector designs into plasmids that can be used for testing of the circuits under transfection into living cells. We expect that this capability, when fully developed, may enable a wide array of high-fidelity disease assays and possible new therapeutic approaches.

6. ACKNOWLEDGMENTS

We would like to thank Ron Weiss and Ryan Gill for providing us with pointers to interesting data sets for investigation, and also thank Ron Weiss for help interpreting the annotations of the Landgraf miRNA supplementary material.

7. REFERENCES

- [1] J. Beal, T. Lu, and R. Weiss. Automatic compilation from high-level biologically-oriented programming language to genetic regulatory networks. *PLoS ONE*, 6(8):e22490, August 2011.
- [2] J. Beal, R. Weiss, D. Densmore, A. Adler, J. Babb, S. Bhatia, N. Davidsohn, T. Haddock, F. Yaman, R. Schantz, and J. Loyall. TASBE: A tool-chain to accelerate synthetic biological engineering. In *3rd International Workshop on Bio-Design Automation*, June 2011.
- [3] M. Djavani, O. R. Crasta, Y. Zhang, J. C. Zapata, B. Sobral, M. G. Lechner, J. Bryant, H. Davis, and M. S. Salvato. Gene expression in primate liver during viral hemorrhagic fever. *Virology Journal*, February 2009.
- [4] P. Landgraf, M. Rusu, R. Sheridan, A. Sewer, N. Iovino, A. Aravin, S. Pfeffer, A. Rice, A. Kamphorst, M. Landthaler, C. Lin, N. Socci, L. Hermida, V. Fulci, S. Chiaretti, R. Foa, J. Schliwka, U. Fuchs, A. Novosel, R. Muller, B. Schermer, U. Bissels, J. Inman, Q. Phan, M. Chien, D. Weir, R. Choksi, G. D. Vita, D. Frezzetti, H. Trompeter, V. Hornung, G. Teng, G. Hartmann, M. Palkovits, R. D. Lauro, P. Wernet, G. Macino, C. Rogler, J. Nagle, J. Ju, F. Papavasiliou, T. Benzinger, P. Lichter, W. Tam, M. Brownstein, A. Bosio, A. Borkhardt, J. Russo, C. Sander, M. Zavolan, and T. Tuschl. A mammalian microRNA expression atlas based on small rna library sequencing. *Cell*, 129(7):1401–14, June 2007.
- [5] Z. Xie, L. Wroblewska, L. Prochazka, R. Weiss, and Y. Benenson. Multi-input rna-based logic circuit for identification of specific cancer cells. *Science*, 333(6047):1307–1311, 2011.