

Classification Method for Microarray Probe Selection using Sequence, Thermodynamics and Secondary Structure Parameters

Lalit Gupta¹, Sunil Kumar¹, Randeep Singh¹, Rafi Shaik¹, Nevenka Dimitrova¹, Aparna Gorthi¹
B. Lakshmi², Deepa Pai², Sitharthan Kamalakaran¹, Xiaoyue Zhao² and Michael Wigler²
*Philips Research Asia - Bangalore, Philips Innovation Campus, Nagavara, Bangalore, India*¹
*Cold Spring Harbor Laboratory: 1, Bungtown Road Cold Spring Harbor, NY 11724 USA*²
{lalit.gupta, sunil.jaglan, randeep.singh, rafi.shaik, nevenka.dimitrova}@philips.com¹
{muthusla, wigler}@cshl.edu², {aparna.gorthi, sitharthan.kamalakaran}@philips.com¹

Abstract

Probe design is the most important step for any microarray based assay. Accurate and efficient probe design and selection for the target sequence is critical in generating reliable and useful results. Several different approaches for probe design are reported in literature and an increasing number of bioinformatics tools are available for the same. However, based on the reported low accuracy, determining the hybridization efficiency of the probes is still a big computational challenge. Present study deals with the extraction of various novel features related to sequence composition, thermodynamics and secondary structure that may be essential for designing good probes. A feature selection method has been used to assess the relative importance of all these features. In this paper, we validate the importance of various features currently used for designing an oligonucleotide probe. Finally, a classification methodology is presented that can be used to predict the hybridization quality of a probe.

1. Introduction

Microarrays are the gold standard technology for high-throughput genomic studies. For these arrays accurate probe design and selection which is critical in generating reliable results, is influenced by various sequence and structural properties of the probes and also thermodynamics and kinetics of hybridization between the probe and target sequences. A study has been presented in this paper to analyze the relative importance of different features for probe selection.

Various machine learning approaches are proposed in literature for probe design and automated analysis of

microarray data [7]. Molla et al. [12] describe the challenges in machine learning that arise with microarray technology and also review recent prominent applications of machine learning to microarray data. However, machine learning has not been fully explored for microarray probe selection. Liu et al. [10] present a new algorithm that performs Integration of Artificial neural networks (ANN) and BLAST (IAB). In this study, input vectors generated from a unique marker database for human and rat genes are used to train and test the IAB. The IAB performance was 7.1 times faster than BLAST search without ANN and the predicted oligos were found to specifically amplify the corresponding gene. Tobler et al. found naive bayes and artificial neural networks to be effective in probe selection based on a set of 67 features derived from probe sequences pertaining to eight bacterial genes [18]. They also found Decision-tree induction and the simple approach of using predicted melting temperature to rank probes perform significantly worse than these two algorithms. However, this problem is still open and provides a big challenge to the researchers.

Present study deals with the identification of several novel parameters that may affect the hybridization efficiency and then determining their relative importance in the hybridization process. We have used probe data generated using the method proposed in [16]. In brief, the probes are designed by Array Designer 4 (Premier Biosoft Intl) and their ratings are obtained through modified NetPrimer [1]. Rest of this paper is organized as follows; section 2 discusses the complete feature selection and classification methodology for “good” and “bad” probe classification. Section 3 discusses the experimental results and section 4 concludes the paper with a discussion on the future work.

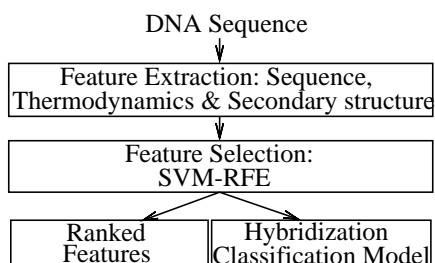


Figure 1. Typical classification Methodology

2. Classification Methodology

We model the hybridization problem as a two class pattern classification problem. Our goal is twofold: determine sequence composition characteristics for probe sequences that perform well during the hybridization process and generate a computational classification model that will be able to predict if a new sequence has a high probability of hybridizing well with the target sequence. The classification methodology for this problem is shown in Fig. 1. Initially, various features are extracted from the probes. The most relevant features are then selected from all the features using feature selection technique. Finally, classification is done to discriminate between a good probe (probes with higher hybridization efficiency) from a bad probe (probes with low hybridization efficiency).

2.1. Feature Extraction

There are several important parameters/ features that need to be considered when designing oligonucleotide probes. We have divided various features into three categories (1) sequence and composition parameters, (2) thermodynamics parameters and (3) secondary structure parameters. The features extracted under each of these categories are as follows:

2.1.1 Sequence and composition parameters:

1. Fraction of A, C, G, or T in the probe and at different positions (5' end (25% of the probe length), 3'end (25% of the probe length) and in the middle of the sequence (50% of the probe length)).
2. Fraction of dimers in the probe and at different positions (5' end, 3'end and in the middle of the sequence).
3. Fraction of trimers in the probe and at 5' end, 3'end and in the middle of the sequence.
4. T_m uniformity: It is calculated from the base composition. There are several methods to calculate T_m . We have used the Nearest Neighbor method [14]:

$$T_m = \frac{\Delta H}{\Delta S + R(\ln C_1 - \frac{C_2}{2})} - 273.15 \quad (1)$$
 where, ΔH and ΔS are the standard enthalpy and entropy, C_1 and C_2 are the initial concentration of single and complementary strand, and R is the universal gas constant.

Overall 337 features are obtained using sequence and composition parameters. The distribution of features is 16 from monomers, 64 from dimers, 256 from trimers and T_m .

2.1.2 Thermodynamics parameters

There are many thermodynamics parameters, that play an important role in determining the stability of the oligonucleotide probes. Some of them are:

1. **Stacking Energy:** Dinucleotide base stacking energy represents how easily parts of the DNA de-stack. High value represents an unstable region [13].
2. **Propeller twist:** The dinucleotide propeller twist is the value for the flexibility of the helix. Low values indicate more flexibility [8].
3. **Bendability:** Sections with high values are more bendable than regions with a low value. Trinucleotide bendability model models the bendability of the DNA towards the major groove [6].
4. **Duplex Stability Disrupt Energy:** Regions with a high disrupt energy value will be more stable than regions with a lower energy value [5]
5. **Duplex Stability Free Energy:** Regions with low free energy content will be more stable than regions with high thermodynamic energy content [17].
6. **DNA denaturation:** DNA regions with a low value are more likely to denature than regions with a higher value [4] [3].
7. **DNA Bending stiffness:** High values correspond to DNA regions that are more rigid, while low values correspond to regions that will bend more easily [2].

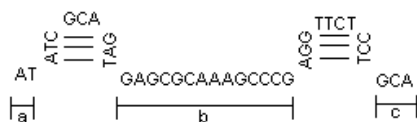


Figure 2. Secondary structure

2.1.3 Secondary structure parameters:

Secondary structure formation is considered as one of the biggest problems associated with any kind of hybridization

- **Hairpin loop formation:** A 3' end hairpin with a ΔG of -2 kcal/mol and an internal hairpin with a ΔG of -3 kcal/mol is generally tolerated. The features are computed as:
 - Number of nucleotides which do not form a loop formed by trimers (Further in the paper we will call it SS feature 1) or quadrimers (SS feature 2). Fig. 2 shows an example of a probe with two loops formed by trimers. The feature value for this probe is computed as $a + b + c$.
 - The length of the longest sequence with the loops formed by trimers (SS feature 3) or quadrimers (SS feature 4). In an example shown in Fig 2 the feature will be computed as $\max(a, b, c)$.

We have generated an overall 390 features for various regions of probes including the 3' end, 5' end and in the middle of the probe. These features are normalized between 0 and 1 and concatenated to form a long feature vector, which is then used for classification.

2.2. Feature Ranking and Classification

A desired algorithm should extract the most relevant features (give higher rank) and eliminate the irrelevant (lower rank) and redundant ones. It is important because throwing away irrelevant features reduces the risk of overfitting, decreases computational complexity and increases overall accuracy. The selection of an optimal subset of features can be carried out by using an appropriately designed performance measure to evaluate their ability to classify the samples. It could not be done using a brute force method, if number of features are huge.

In this paper, we have used SVM-RFE for feature selection as proposed in [11]. We have used Support Vector Machine (SVM) [9] for classification.

Table 1. Classification results (%) with different SVM Parameters

SVM Parameters	Accuracy (%) Synthetic Data
Polynomial with Degree 1	80.1
Polynomial with Degree 2	59.6
RBF with spread 1	97.9
RBF with spread 0.1	96.8

3. Results and Discussions

We have used synthetic as well as experimental data for experimentation. The synthetic data used for experimentation was generated using the method proposed in [16] and experimental data was obtained from Cold Spring Harbor Laboratory (CSHL). Initial analysis has been done on synthetic data and the results of analysis have directly been used to compute accuracy on experimental data. Each of the probes is of 25 and 50 base pairs length in synthetic database and experimental database, respectively. Each probe is differentiated in two class namely good and bad probe. Present study is an attempt to analyze the relative importance of all these parameters for probe design and to come up with a classification methodology to classify a probe as being good or bad. We have used 8000 probes (in each synthetic database and experimental database) out of which 4000 are good (net primer ranking = 100 in case of synthetic database) and 4000 bad (net primer ranking < 100 in case of synthetic database) probes. Results obtained using all features with different SVM parameters on synthetic data are shown in Table 1.

Table 2. Description of results

Type of feature	Higher Ranked Features (In descending order)
Sequence	CG, AG, ACT, AAG, AA, TC, CT, T_m
Thermodynamics	Stacking, Entropy, B-DNA twist, Propeller twist, Bendability
Secondary structure	SS feature 1
All	Secondary Structure, Trimers, Dimers, Thermodynamics

3.1. Performance of Various Features

Several features (390) were extracted from the probes. These features were broadly grouped into se-

Table 3. Classification results (%) using different features

Type of feature	Size	Accuracy (%) Synthetic Data	Accuracy (%) experimental data
Sequence and T_m	337	80.3	58.4
Thermodynamics	49	73.7	56.8
Secondary structure	4	65.2	54.0
All	390	97.9	61.0

quence, composition, thermodynamics and secondary structure. Relative importance of each of these features was then determined on synthetic data using feature selection methods and is presented in Table 2. Some of the important observations included:

- In sequence and composition features, it has been observed that the T_m and GC% plays significant role in determining the overall hybridization efficiency of the probe. Most important sequence feature was the dimer and trimer fraction of G and C in the middle and at 3' of the probe. The importance of these parameters has already been well studied in literature. Our results also validate these studies by determining the importance of these parameters in the hybridization. However unlike Shin et al. [15], where T_m was considered as the most important parameter for probe design, our model demonstrates that the secondary structure plays a bigger role in determining greater hybridization efficiency.
- Most of the current literature highlights only the significance of free energy on hybridization. However in our study it has also been observed that the less studied *thermodynamic parameters including the DNA twist, stacking energy, propeller twist, bendability, etc. shows relatively high importance in determining the goodness of a probe.*
- In case of secondary structure formation, current literature analyzes the free energy of the nucleotide base pairs involved in loop formation. However, in our model we have also taken into consideration the sequence composition of the bases which are not involved in loop formation. Interestingly, the analysis revealed that *these base pairs are significant in determining the stability of hybridization.*

When considering the relative importance of various groups in determining the hybridization efficiency of the probe, it was observed that the most important parameter, contributing significantly to the accuracy is secondary structure formation. These structures are the energy minimized state of oligonucleotide folds, thereby, making thermodynamics another impor-

tant feature for hybridization.

3.2. Performance of Hybridization Classification Model

Table 3 shows results obtained using different types of features. Column 1 shows the different categories of features. Column 2 shows the number of features under the category listed in Column 1. Column 3 and 4 show the classification results obtained using proposed methodology on synthetic and experimental data, respectively.

The main observations from Table 3 are as

- Classification accuracy on synthetic data, achieved using SVM-RFE for feature selection and SVM for classification is 97.9% which is significantly higher than any other methods used. This involved combining all the features and selecting the best set of features. In this case top 370 features were selected, in order to achieve this accuracy.
- Classification accuracy of 61.0% is achieved on experimental data.
- The secondary structure parameters were found to be relatively more important than other feature group.

To examine the robustness of the algorithm, Receiver Operating Characteristic (ROC) curves has been used for study. Fig. 3 shows the ROC curves generated using all the features on synthetic data. It can be observed that ROC curve generated using the proposed method is more closer to its ideal shape. It is also evident that after a particular threshold true positive rate (TPR) of the system does not change significantly and false positive rate (FPR) starts increasing. Higher TPR can also be obtained at the cost of increase in FPR by varying the threshold in the binary classifier as shown in Fig. 3.

The present study validates the importance of various parameters currently being used for designing an oligonucleotide probe. In addition, it also provides additional new features that play a significantly higher role in determining the overall hybridization efficiency of the probe. Further work is being carried out on the complementary sequence of the oligonucleotide studied in

order to analyze the feature of the target DNA which will bind to these probes.

4. Conclusion

Present study evaluates the relative importance of various sequences, thermodynamics and secondary structure features that need to be considered when designing an oligonucleotide probe for microarray applications. Already existing features that are in current practice were validated and several new features were identified and analyzed. An interesting observation from our work is that the secondary structure features play significantly greater role than the currently used sequence parameters. We have also presented a classification methodology that can be used to discriminate a good probe from bad probe. Our future work scope includes exploring ways to improve the accuracy of hybrid classification system.

5. Acknowledgements

We are grateful to Jen Troge and Lisa Hufnagel for technical assistance.

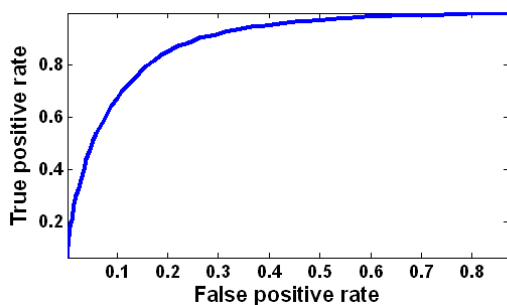


Figure 3. ROC curve the variation between TPR and FPR.

References

[1] Netprimer. <http://www.premierbiosoft.com/netprimer>.
 [2] S. N. K. A. V. Sivolob. Translational positioning of nucleosomes on DNA: the role of sequence-dependent isotropic DNA bending stiffness. *J Mol Biol.*, 247(5):918–31, April 1995.
 [3] R. D. Blake, J. W. Bizzaro, J. D. Blake, G. R. Day, S. G. Delcourt, J. Knowles, K. A. Marx, and J. J. SantaLucia. Statistical mechanical simulation of polymeric DNA melting with meltsim. *Bioinformatics*, 15(5):370–5, May 1999.
 [4] R. D. Blake and S. G. Delcourt. Thermal stability of DNA. *Nucleic Acids Res.*, 26(14):3323–32, July 1998.

[5] K. J. Breslauer, R. Frank, H. Blocker, and L. A. Marky. Predicting DNA duplex stability from the base sequence. *Proc Natl Acad Sci. USA*, 83(11):3746–50, June 1986.
 [6] I. Brukner, R. Sanchez, D. Suck, and S. Pongor. Sequence-dependent bending propensity of DNA as revealed by dnase i: parameters for trinucleotides. *EMBO J.*, 18(8):1812–8, April 1995.
 [7] S. Graf, F. G. Nielsen, S. Kurtz, M. A. Huynen, E. Birney, H. Stunnenberg, and P. Flicek. Optimized design and assessment of whole genome tiling arrays. *Bioinformatics*, 23:195–204, 2007.
 [8] M. A. Hassan and C. R. Calladine. Propeller-twisting of base-pairs and the conformational mobility of dinucleotide steps in DNA. *J Mol Biol.*, 259(1):95–103, May 1996.
 [9] S. Kumar. *Neural Networks - A Classroom Approach*. Tata McGraw Hill, New Delhi, 2004.
 [10] C. C. Liu, C. C. Lin, K. C. Li, W. S. Chen, J. C. Chen, M. T. Yang, P. C. Yang, P. C. Chang, and J. J. Chen. Genome-wide identification of specific oligonucleotides using artificial neural network and computational genomic analysis. In *BMC Bioinformatics*, volume 22, page 164, 2007.
 [11] S. K. Majumdar, N. Ghosh, and P. K. Gupta. Support vector machine for optical diagnosis of cancer. *J. of Biomedical Optics*, 10(2):024034–1–14, 2005.
 [12] M. Molla, M. Waddell, D. Page, and J. Shavlik. Using machine learning to design and interpret gene-expression microarrays. *AI Magazine*, 25(1):23 – 44, 2004.
 [13] R. L. Ornstein, R. Rein, D. L. Breen, and R. D. Macelroy. Optimized potential function for calculation of nucleic-acid interaction energies .I. base stacking. *Biopolymers*, 17(10):2341–2360, 1978.
 [14] J. J. SantaLucia, H. A. T. Allawi, and P. A. Seneviratne. Improved nearest-neighbor parameters for predicting DNA duplex stability. In *Biochemistry*, volume 35, pages 3555–3562, 1996.
 [15] S.-Y. Shin, I.-H. Lee, and B.-T. Zhang. Microarray probe design using epsilon-multi-objective evolutionary algorithms with thermodynamic criteria. *EvoWorkshops*, pages 184–195, 2006.
 [16] R. Singh, S. Kumar, and L. Gupta. PrimerDB: A synthetic database for primer/ oligonucleotide hybridization and efficiency prediction. In *Proceedings of the Sixth IASTED International Conference on Biomedical Engineering*, pages 1–6, 2008.
 [17] N. Sugimoto, S. Nakano, M. Yoneyama, and K. Honda. Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes. *Nucleic Acids Res.*, 24(22):4501–5, Nov. 1996.
 [18] J. B. Tobler, M. N. Molla, E. F. Nuwaysir, R. D. Green, and J. W. Shavlik. Evaluating machine learning approaches for aiding probe selection for gene-expression arrays. *Bioinformatics*, 18(1):1645–1715, 2002.