MATRIX FORMULATION OF A UNIVERSAL MICROBIAL TRANSCRIPT PROFILING SYSTEM

J. Patrick Fitch, Jefferson Ng, and Bahrad A. Sokhansanj

Biology & Biotechnology Research Program
Lawrence Livermore National Laboratory; University of California
L-452, 7000 East Avenue; Livermore, CA 94550

ABSTRACT

DNA chips and microarrays are used to profile gene transcription. Unfortunately, the initial fabrication cost for a chip and the reagent costs to amplify thousands of open reading frames for a microarray are over $100K for a typical 4 Mbase bacterial genome. To avoid these expensive steps, a matrix formulation of a universal hybrid chip-microarray approach to transcript profiling is demonstrated for synthetic data. Initial considerations for application to the 4.3 Mbase bacterium *Yersinia pestis* are also presented. This approach can be applied to arbitrary bacteria by recalculating a matrix and pseudoinverse. This approach avoids the large upfront expenses associated with DNA chips and microarrays.

1. INTRODUCTION

The adenine (A), cytosine (C), guanine (G), and thymine (T) nitrogenous bases of DNA preferentially bind A-to-T and G-to-C. The complementary base pairing property can be used as a powerful detection method by the hybridizing (binding) of a reference strand of DNA to a test strand. DNA sequencing, transcript profiling, and clinical diagnostics are a few of the applications that exploit DNA hybridization. This paper focuses on transcript profiling applications to detect the level of mRNA (gene expression). The hybridization approaches are derived from DNA chips [1], DNA microarrays [2], and sequencing by hybridization methods [3].

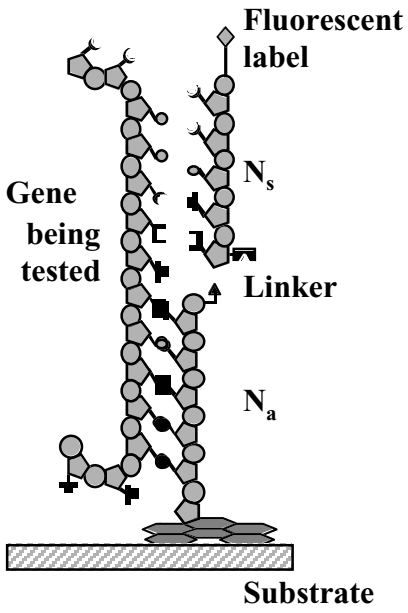

Fig. 1. Description of hybridization experiment for transcript profiling.

The matrix approach that we propose is based on a generalized experiment described graphically in Fig. 1. Detection of a test strand of DNA is accomplished by hybridization with a complementary reference strand of length $N=N_a+N_s$. The N-mer is composed of an array of N_a-mers attached to a glass substrate in a manner similar to microarrays. The additional N_s bases for the detection are introduced in solution. Using extension enzymes or special linker chemistries, the discrimination power after hybridization is the same as with a whole N-mer. Even when multiple N_s-mers hybridize

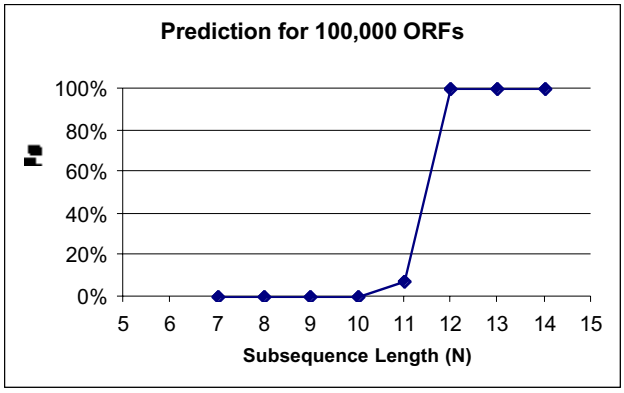

Fig. 2. Model prediction that 12 bases are needed to discriminate 100,000 genes (ORFs).

to the same spot, we want to identify the source genes. One solution is to place the competing N_s-mers into separate pools or experiments.

Our goal is to monitor message RNA level of most genes (over 90%) in a prokaryotic organism (bacterium) without custom microarrays. Nothing in the described method limits the approach to prokaryotes. The hybridization, however, is currently restricted in oligonucleotide length to about 10 bases. As the length increases from 10-mers to 12-mers and longer (see Fig. 2), eukaryotic profiling including human gene profiling will be possible with the same technique [4].

2. MATRIX FORMULATION

A matrix formulation of the experiment is **M** = **P D G**, where

G = gene expression vector to be estimated (g x 1 column vector) representing the level of activity for g genes.

D = DNA matrix (4^N x g) that maps individual genes into constitutive N-mers. Let **D**(i,j) be the value of **D** at position (i,j) where (i,j) run (1,1) to (4^N, g). **D**(i,j) is defined as the number of times that the i th N-mer occurs in gene j.

N = $N_a + N_s$ is the number of bases in the DNA reference strand.

N_a = number of bases in the oligos attached to the substrate.

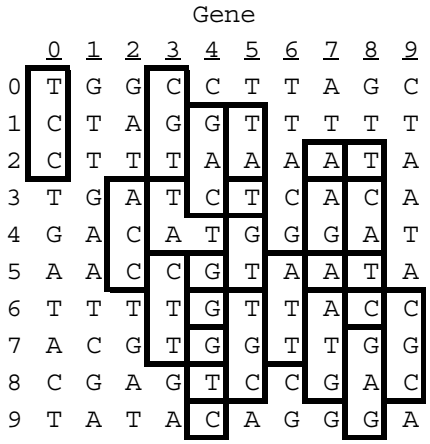

Fig. 3. Ten genes of length 10 are used as a synthetic genome for demonstrating the technique. Unique 3-mers indicating the discrimination power of the detection are highlighted. Note that there is no 3-mer unique to gene #1.

N_s = number of bases in the oligo extension introduced in solution during hybridization.

P = pooling matrix (p 4^{Na} x 4^N) where p is the number of experimental pools.

M = measurement vector (p 4^{Na} x 1 column vector) representing the outcome of an experiment.

We have assumed that all possible N_a-mers are attached to the substrate. The matrix formulation is more general than this and any set of oligos (even of varying lengths) could have been used. However, we anticipate using the HyChip™ by HySeq for our transcript profiling and it uses all possible 5-mers at this time [5].

With the experiment constructed as a matrix multiply, the same approach as in the classic **Ax**=**b** matrix formulation can be used. Recall that if $\mathbf{A}\mathbf{A}^T$ is invertible, then the optimal least squares solution or pseudoinverse is

$$\mathbf{x}_{LS} = \mathbf{A}^T (\mathbf{A}\,\mathbf{A}^T)^{-1}\,\mathbf{b}. \qquad (1)$$

3. EXAMPLE

To demonstrate the matrix approach, consider an artificial genome with 10 genes of 10 randomly assigned bases. In Fig. 3, we show the genes as

	Gene									
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
AAA	0	0	0	0	0	0	0	0	0	0
AAC	0	0	0	0	0	0	0	0	0	0
AAG	0	0	0	0	0	0	0	1	0	0
AAT	0	1	0	0	0	0	0	1	0	1
ACA	0	0	0	0	0	0	0	0	0	0
ACC	0	0	1	0	0	0	0	0	0	0
ACG	0	0	0	0	0	0	1	0	0	1
ACT	1	0	0	1	1	0	0	0	0	0
AGA	0	0	0	0	0	0	0	1	0	0
AGC	0	0	0	0	0	0	0	0	0	0

Fig. 4. The first 10 rows of the 64 row by 10 column **D** matrix. Rows with a single nonzero element are unique identifiers of a gene. For instance, AAG is 0 except for the 1 in the gene 7 column. Compare these entries with the highlighted 3-mers in Fig. 2.

well as highlighting the unique 3-mers i.e., a 3-mer that appears in only one gene.

For this example, we use a 3-mer reference strand (N=3) comprised of a single base attached to a substrate (N_a=1) and a 2-mer in solution (N_s=2). There are four spots on the slide (A, C, G, and T). A few of the rows of the **D** matrix generated from the genes in Fig. 3 are shown in Fig. 4. The first N_a bases in the index column (the first A for the rows in Fig. 4) correspond to the attached N_a—mers so that the pooling and measurement matrices are easier to interpret. Rows that contain a single non-zero element represent 3-mers that uniquely identify a gene e.g., AAG, ACC, AGA for genes 7, 2, and 7, respectively. If each gene had many unique identifiers, the detection problem would be simple. For a short oligo detection system, however, there are many shared N-mers and so multiple hits must be utilized in the algorithm. The design issues are to implement a least squares estimation of the level of gene expression while introducing a pooling strategy (the **P** matrix) that minimizes the number of experiments needed.

For the synthesized data we have used N_a=1 and N_s=2. The experimental constraints result in a **P** matrix that is sparse and with repeating subunits. Denoting each pool by $\mathbf{P_i}$, the pooling matrix is of the form:

$$\begin{vmatrix} \mathbf{P_1} & \mathbf{O} & \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{P_1} & \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} & \mathbf{P_1} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} & \mathbf{O} & \mathbf{P_1} \\ \mathbf{P_2} & \mathbf{O} & \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{P_2} & \mathbf{O} & \mathbf{O} \\ & \cdots & & \\ \mathbf{O} & \mathbf{O} & \mathbf{P_p} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} & \mathbf{O} & \mathbf{P_p} \end{vmatrix}$$

Each of the $\mathbf{P_i}$ are repeated 4^1=4 times because N_a is 1 in our example. Each $\mathbf{P_i}$ row vector is of length 4^2=16 (N_s is 2) and comprised of 1s and 0s. A 1 indicates that a particular N_s-mer is used in the pool. **O** is a zero row vector of length N_s. For this example, we designed a set of p=5 pools shown in Fig. 5.

	P1	P2	P3	P4	P5
AA	0	0	0	0	1
AC	1	0	0	0	0
AG	1	0	0	0	0
AT	0	0	0	0	0
CA	0	0	0	0	0
CC	0	0	1	0	0
CG	0	0	0	0	0
CT	0	0	0	0	0
GA	0	0	0	0	0
GC	0	0	0	0	0
GG	0	0	0	0	0
GT	0	0	0	1	0
TA	0	0	0	0	0
TC	0	0	0	0	1
TG	0	0	0	0	0
TT	0	1	0	0	0

Fig. 5. The five pools used for the synthetic genome example. **PD** has matrix rank 10.

Both **D** and **P D** are of matrix rank 10. Therefore the pseudoinverse exists and can be determined using singular value decomposition or other techniques. The pseudoinverse was verified using MATLAB™[6]. The calculation implements Eq. (1) and estimates

$$\mathbf{G_{LS}} = \mathbf{(PD)}^T \mathbf{(PD\ (PD)}^T)^{-1}\ \mathbf{M}. \qquad (2)$$

4. APPLICATION TO BACTERIA

To apply this approach to a specific bacterium like *Yersinia pestis* requires development of several tools and algorithms. *Y. pestis* is roughly a 4.3Mbase genome and has over 4,000 genes. Using an experimental system based on two 5-mers (N_a=N_s=5), we use the matrix approach to design a pooling strategy. The algorithm is

1. Determine the occurrence rates of every N-mer (0, 1, 2, etc.)
2. Starting with unique occurrences, then doublets, etc. determine which attached N_a-mers are contaminated by the proposed N-mer and are not already in a pool.
3. Find the largest set of N_s-mers meeting criterion 2 with non-intersecting contamination sets. Place in one pool and then return to step 2.

This approach is basically a Gram-Schmidt orthogonalization. We have calculated the **D** matrix for *Y. pestis* and are now designing the pooling strategy. Fig. 6 shows the percent of ORFs that have a unique N-mer for N ranging from 5 to 14. Both model and data are presented showing that a 10-mer (N=10) is sufficient to transcript profile *Y. pestis*.

5. SUMMARY

In addition to designing the minimal pooling strategy, the matrix formulation should facilitate the design of calibration oligos and other strategies for making microarrays more quantitative and repeatable. We are also introducing design constraints in the pool specification to accommodate ribosomal and other contaminating signals.

6. ACKNOWLEDGEMENTS

This work was performed under the auspices of the U.S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48. Publication UCRL-JC-141175.

The authors would also like to thank T. Slezak and the LLNL informatics team, T. Carrano (LLNL), G. Andersen and the LLNL microbiology team, and R. Drmanac (HySeq) for discussions of this R&D.

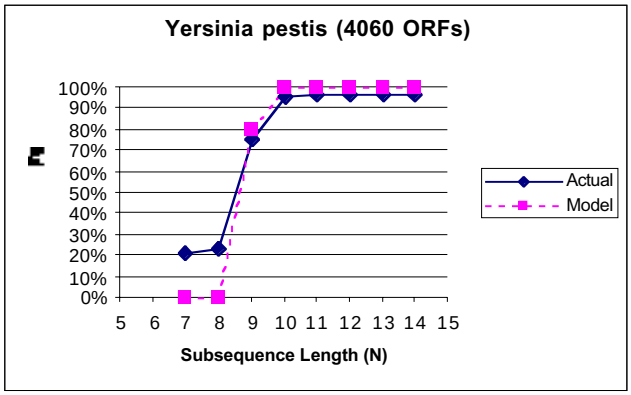

Fig. 6. Model and data overlaid showing that over 90% of the *Yersinia pestis* genes have at least one unique 10-mer.

7. REFERENCES

[1] Affymetrix Products: GeneChip™, [online] http://www.affymetrix.com/products/tech_probe.html.

[2] J. L. DeRisi, V. R. Iyer, and P. O. Brown, Exploring the metabolic and genetic control of gene expression on a genomic scale, *Science*, vol. 278, pp. 680-686, Oct. 1997. http://cmgm.standford.edu/pbrown/mguide/index.html

[3] S. Drmanac, D. Kita, I. Labat, B. Hauser, J. Burczak, R. Drmanac, Accurate sequencing by hybridization for DNA diagnostics and individual genomics, *Nature Biotechnology*, 16, 54-58, 1998.

[4] J. P. Fitch, B. A. Sokhansanj, L. L. Ott, and T. R. Slezak, Informatics and simulation in a genomic approach to understanding virulence, *Proc. IEEE Information Technology Applications in Biomedicine (ITAB 2000)*, Arlington, VA, pp. 302-307, Nov. 9-10, 2000.

[5] Universal DNA sequencing chip, [online] http://www.hyseq.com/products/puniv.html.

[6] MATLAB™ version 5.2.0, The MathWorks, Inc., January 1998.