Composite Probe and Signal Recovery of Compressed Sensing Microarray

Zhenhua Gan[1,2,4], Baoping Xiong[1,3], Fumin Zou[1,4(✉)], Yueming Gao[3], and Min Du[2,3]

[1] College of Information Science and Engineering, Fujian University of Technology, No. 3 Xueyuan Road, University Town, Minhou, Fuzhou, Fujian, China
{ganzh, fmzou}@fjut.edu.cn, xiongbp@qq.com
[2] College of Electrical Engineering and Automation, Fuzhou University, No. 2 Xueyuan Road, University Town, Minhou, Fuzhou, Fujian, China
dm_dj90@163.com
[3] College of Physics and Information Engineering, Fuzhou University, No. 2 Xueyuan Road, University Town, Minhou, Fuzhou, Fujian, China
fzugym@yahoo.com.cn
[4] Key Lab of Automotive Electronics and Electric Drive Technology of Fujian Province, No. 3 Xueyuan Road, University Town, Minhou, Fuzhou, Fujian, China

Abstract. Due to the large number of uncertain factors in hybridization, image capture and processing of the microarray, multiple probes were generally arranged to improve the reliability of the measurement. However, the small area limited the number of probes that were allowed to be added on, so a composite probe would be the better choice. A composite probe contained the linear combination of a variety of gene fragments. It was used so that the microarray could easily realize the repeated gene fragments within a limited region. The number of composite probes would rapidly dwindle when it compared to a traditional microarray. At the same time, since the sparse characteristics of biological gene mutation, the compressed sensing idea is adopted to recovery the gene variation in the composite probes. The 96 fragments can be used with the 48 × 96 sparse random matrix to construct the 48 composite probes when the sparsest level K is no more than 12. Simulation results show that compressed sensing can accurately recover the gene mutation by using the Orthogonal Matching Pursuit (OMP) algorithm.

Keywords: Compressed sensing · Microarray · Composite probe · Sparse random matrix · OMP

1 Introductions

Microarray is a newly technology for high-throughput and quantitative detection in the biology science area. The abilities of microarray to express of thousands of genes simultaneously in a single detection have allowed the application in wide variety of fields, such as molecular biology, genetics, agriculture, disease diagnosis, medical treatment,

J. Pan et al. (eds.), *Genetic and Evolutionary Computing*, Advances in Intelligent Systems and Computing 536, DOI 10.1007/978-3-319-48490-7_2

food safety supervision, and judicial identification [1]. In a traditional microarray, each of the probe represents a complementary gene segment to be used to detect the corresponding gene information [2].

For the measurement noises, multiple probes were usually arranged to improve the reliability of the determination. The same probe was an effective way to avoid the information losses due to the interference of noises, but the repeated arrangement of probes resulted in an increase in the number of probes on the microarray. Thereupon, the weak fluorescence and the small size of probe were producing adverse effects while the density was increased, which also had caused serious irreparable damage for the ability to obtain reliable expression of the probes.

A more efficient method for solving the above problems was to use the composite probes. In this way each composite probe located in a spot was designed to detect the expression of multiple gene fragments simultaneously. The microarray scanner read the intensity of linear combination information from the composite probe, and the message of each gene probe would be obtained via the appropriate recovery algorithm [3].

Traditional cDNA gene sequencing probes produced a large number of mostly useless information, due to the fact that differences in the sequence between the reference sample and test sample were sparse. Because of the sparse characteristics of biological gene mutation, the compressed sensing idea was adopted to recover the gene variation in the composite probes. The compressed sensing theory had provided a strong support for the accurate recovery of the sparse signals, and it had been widely used in biological sensing, radar detection, data compression, image processing, and pattern recognition [4]. The compressed composite probes were constructed based on the compressed sensing ideas. The difference gene sequencing signals could be recovered by observing a small amount of the composite probes [5, 6].

The application of the composite probes on microarray was confirmed by [3]. And a composite probe method for constructing the compressed sensing microarray was proposed in [6]. A sparse low density parity check code (LDPC) as the measurement matrix to construct a compressed sensing microarray, and the recovery algorithm for the gene difference information were also proposed in [6]. For more information, the sparse random matrix in the recovery algorithm had the advantage of being a simple structure, low computational complexity, and easy to update and store in [7].

2 Design of Composite Probe for Compressed Sensing Microarray

2.1 Compressed Sensing

Compressed sensing is a sampling and reconstruction theory for sparse singles. Signal or the signal after a special transformation, with sparse or compressible characteristics, is the premise of compressed sensing [8, 9]. Considering a discrete digital signal $x \in R^N$ that has $K << N$ non-zero elements, the signal x is K sparse and N-K elements in the signal x will be 0 or close to 0. Since the signal x is generally not directly measured, we could design an $M \times N$ measurement matrix A to observe M linear combinations of the x, where $K << M << N$.

$$y_{M\times 1} = A_{M\times N} x_{N\times 1} \qquad K << M << N \tag{1}$$

Although the Eq. (1) is a underdetermined system, we also could reconstruct the signal x for K sparse by solving the constrained l_0 minimization,

$$\hat{x} = \arg\ \min \|x\|_0 \quad s.t. \quad y = Ax \tag{2}$$

where $\|x\|_0$ denotes the l_0-norm.

Unfortunately, solving the l_0 minimization is known as NP-hard. In order to solve this problem, it is usually converted into minimizing the l_1 with the optimization constraints. As long as the measure matrix A satisfies the restricted isometric property (RIP), the Eq. (1) agree with the following constraints,

$$\hat{x} = \arg\ \min \|x\|_1 \qquad s.t. \quad y = Ax \tag{3}$$

where $\|x\|_1$ denotes the l_1-norm [10, 11].

2.2 Composite Probe for Compressed Sensing Biological Microarray

The biological microarray uses the principle of molecular hybridization, which the gene to bind specific complementary sequences in the microarray probes. Since fluorescent labeling has been achieved already, we can get the fluorescent signal by light excitation. The information of the corresponding gene fragments from the resulting fluorescence signals can also be analyzed.

A typical cDNA microarray is fixed with a large number of probe spots located on the surface, but each probe consists of the single gene fragment, which can only detect specific complementary sequence segments. The detection principle of traditional cDNA is shown in Fig. 1.

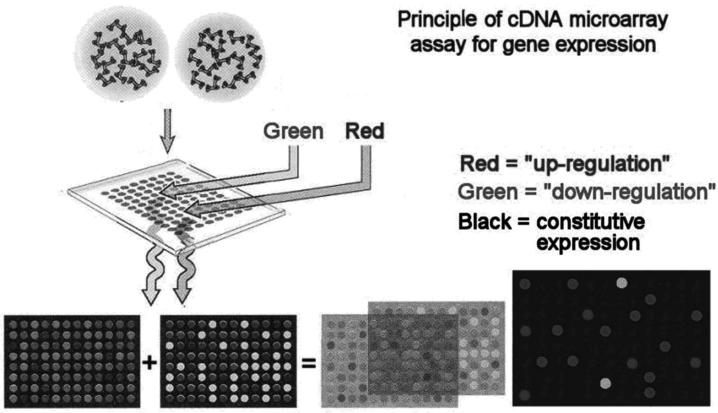

Fig. 1. The principle of traditional cDNA detection

The fluorescence intensity is at its most when the probes are matched normally on the microarray, and the intensity is at its weakest when the probes are mismatched. When the probes are not paired, there is little to no fluorescence intensity. The fluorescence intensity generated by match pairs is 5 times to 35 times more intense than that of a single or two bases mismatch in the probe's sequence. So the accurate determination of the fluorescent intensity is the basis of the specific detection of the biological sequence of microarray probe [12].

The composite probe fluorescence intensity is reflected the cumulative number of fluorescent molecules in various biological fragments fixed in the probe's spot. Literature [7] uses similar techniques as literature [3], which the design of the composite probe is realized by mixing the existing probe molecules according to the linear relationship of the measurement matrix A. This method can be used concurrently with the existing cDNA processing technology.

In particular, there are only a small fraction of the genes to be in a state of mutation. We are considering the difference that the gene expression of test sample is compared with the reference sample. And the difference of the signals which produced by two samples is nature sparse.

In order to construct a compressed sensing microarray with M composite probes, an $M \times N$ measurement matrix A with $M << N$ must be designed for N gene fragments. And we design the measurement matrix A with binary 0/1 elements only to simplify the construction difficulty of the compressed probes.

In two-color microarray of cDNA experiments, the reference sample is labeled by Cy3 while the test sample is labeled by Cy5 [13]. We are comparing two channel's sample by data vectors x_{cy3} and x_{cy5}, and interesting the difference expression of $x = x_{cy3} - x_{cy5}$.

Since there are differences in the small number of gene segments, the distribution of the x is sparse. The compressed sensing idea is relevant to the applications of DNA microarrays in the gene variation. Figure 2 illustrates the structure of the composite probe.

Each row of the matrix A represents a linear combination of the gene fragments. The m-th composite probe is determined by the positions of the gene fragment in the m-th row of matrix A. The combination structure of a composite probe is shown as the following,

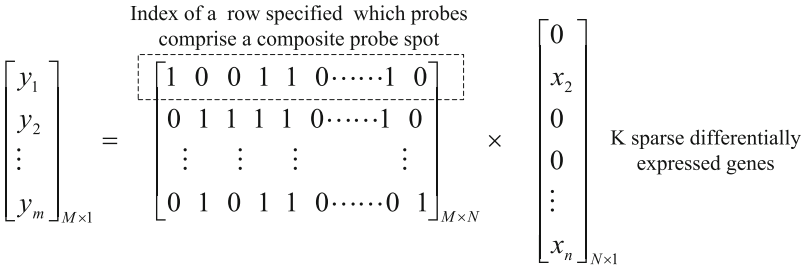

Fig. 2. Illustration of the compressed microarray

$$y_j = \sum_{i=1}^{N} a_{ji} x_i, \qquad j = 1, 2, \cdots M \tag{4}$$

where M < < N. Additionally, if the number of nonzero elements is different in each row, the actual mixed solution of probes should be diluted to the specified volume to ensure the consistency of the dilution.

3 Composite Probe Recovery Using Compressed Sensing

3.1 Sparse Random Measurement Matrices

Each column of the random sparse M × N matrix contains only uM non-zero elements with independent and identical distribution [14]. Literature [15, 16] also have pointed out that the recovery effect of sparse random measurement matrix is consistent with the gauss random measurement matrix. Moreover, the literature [14] have further proved that the sparse random matrix satisfies the RIP.

Due to the each row of the matrix represents a linear combination of a probe spot. We limit the elements of the random sparse matrix into binary 1/0 for the sake of constructing simplicity. The configuration process for sparse random matrix is as follows,

(1) Production M × N matrix of zeros;
(2) The position of each column elements is randomly selected according to the sparse coefficient u of the matrix, and these elements would be set to 1.

3.2 Recovery of Variation Gene from Composite Probe

In two-color microarry of cDNA, we are comparing two channel sample by x_{cy3} and x_{cy5}, and interesting in the difference expression of $x = x_{cy3} - x_{cy5}$. By sparse random matrix, the normalized observation value of the composite probe is defined as $y = y_{cy3} - y_{cy5}$.

If the compressed sensing recovery x is obtained directly by the combination method, which is a NP-hard as well known. Formula (3) is an l_1-norm optimization problem, compared to time-consuming convex optimization, the classical sparse approximation methods, such as the Orthogonal Matching Pursuit (OMP) algorithm, would be very suitable.

In the OMP algorithm, the residual vector r, which is the error of approximation vector y, is smaller and smaller after several iterations [17].

Let $x_k = \arg\min_x \|y - A_k x\|_2$, $r_k = y - A_k x$, $A_k = [A_{k-1}\ a_k]$ be a sub-matrix which selected in step k. Then the OMP algorithm process as follows [18, 19].

Input: compressed sampling matrix A, measured value y, the sparsity level K.

Output: reconstruction of the signal ^x, estimated support I.

Initialization: $x_0 = 0$, $r_0 = y$, $k = 0$, estimated support $I = \emptyset$.

(1) $k \leftarrow k + 1$;
(2) the index that is the best match with the residual vector r_{k-1}, and $\lambda k \leftarrow \text{argmax}_j\{| < r_{k-1}, a_j >|\}$;
(3) update the index $I_k = [I_{k-1}\ \lambda_k]$, and $A_k = [A_{k-1}\ a_k]$;
(4) reconstruction $\hat{x} \leftarrow [A_k]^{-1}\ y$;
(5) update the residual vector as $r_k \leftarrow y - A_k\ (\hat{x})$;
(6) If $k \leq K$, then execute step (1), otherwise stop at $k > K$.

4 Simulation Results and Analysis

We have designed N = 96 cDNA microarray simulation probes with the idea of array-based comparative genomic hybridization (aCGH). The difference between the reference probes and the test probes, i.e., the sparsity level is K = 12. In the simulation experiments, the differences between the reference probes and the test probes have subjected to random distribution, and the locations of these different composite probes are also subjected to random distribution.

Figure 3a illustrates the reference probe x_{cy3}, and Fig. 3b demonstrates the probe x_{cy5}. Then, the differences between them, i.e., $x = x_{cy3} - x_{cy5}$ are shown in Fig. 3c.

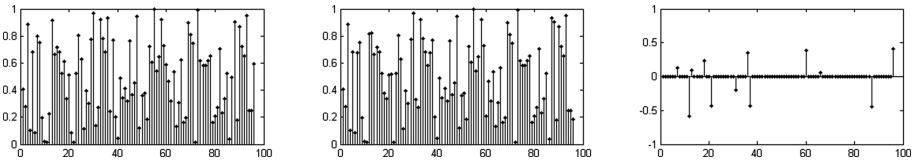

Fig. 3. **a**. The probe *xcy3,* **b**. The probe *xcy5,* **c**. $x = x_{cy3} - x_{cy5}$

We also have designed the sparse random matrix as compressed sensing measurement matrix A and let the elements sparsity coefficient $u = 0.25$. And M = 48 composite probes of compressed sensing microarray are constructed from N = 96 gene fragments by matrix A in the mixed method.

The observations of the composite probes are shown in Fig. 4a and Fig. 4b, while the differences between them, i.e., $y = y_{cy3} - y_{cy5}$ are shown in Fig. 4c.

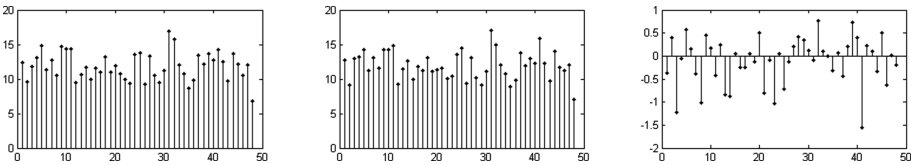

Fig. 4. **a**. The composite probes ycy3, **b**. The composite probes ycy5, **c**. $y = y_{cy3} - y_{cy5}$

We have used the OMP recovery algorithm to successfully reconstruct the gene different vector x, at N = 96, M = 48, K = 24, u = 0.25. As shown in Fig. 5, the recovery is so accurate that the relative error is $e = 4.4016 \times 10^{-15}$.

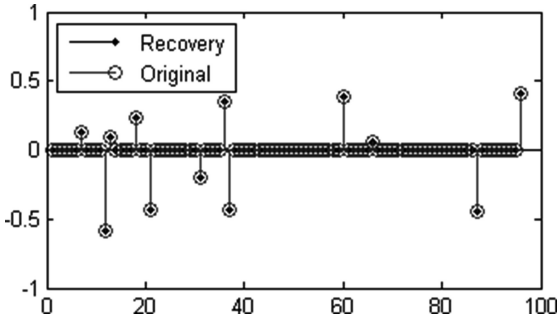

Fig. 5. The recovery of $\hat{x} = x_{cy3} - x_{cy5}$ with $e = 4.4016 \times 10^{-15}$

The structural parameters of cDNA simulation microarray have remained unchanged at N = 96, M = 0.5 N and the sparsity coefficient u = 0.25 for matrix A, and sparse K has been changed from zero to M. We still have used the OMP algorithm to recover the vector x. The accurate reconstruction ratios of the simulation signals are shown in Fig. 6.

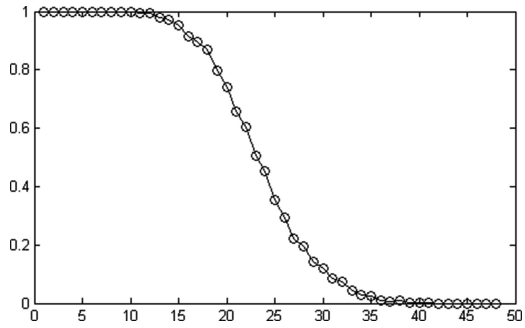

Fig. 6. The accurate reconstruction ratio of $\hat{x} = x_{cy3} - x_{cy5}$ for M = 48

As shown in Fig. 6, the compressed sensing algorithm recovers the probe's difference signals with high accuracy, at N = 96, M = 0.5 N, u = 0.25 and K ≤ 12.

Figure 7 demonstrates the accurate reconstruction ratio of simulation probes under the OMP recovery algorithm, when only the number of composite probe, i.e., M has been changed from zero to N.

It is shown in Fig. 7, compressed sensing algorithm achieves high accurate recovery for difference signals between the reference probes and the sample probes when the sparse random measurement matrix A is used at N = 96, K = 12, M ≥ 48, u = 0.25.

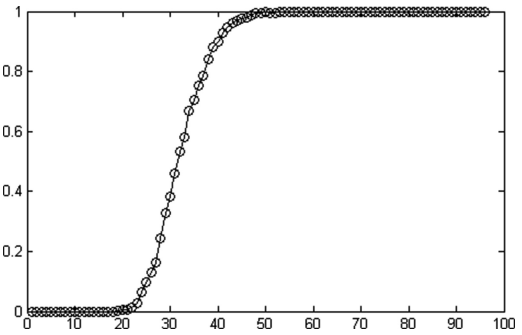

Fig. 7. The accurate reconstruction ratio of $\hat{x} = x_{cy3} - x_{cy5}$ for K = 12

5 Summary and Conclusions

There are a large number of uncertain factors in hybridization, image capture and processing of the microarray. In order to improve the reliability of the measurement, multiple probes are generally arranged to carry out repeated measurements. With a composite probe, a single spot of the compressed sensing microarray can easily and simultaneously measures many gene fragments, so that the repeated measurements of gene fragments can be realized with a limited number of spots. Considering the randomness and sparsity of genetic mutation, the total number of the composite probes installed in the compressed sensing microarray can be sharply reduced compared to that in the traditional microarray. Simulation experiment results show that, by using composite probes with gene fragment at N = 96, M = 0.5 N, and sparse random measurement matrix sparsity coefficient $u = 0.25$, when difference of cDNA probes $K \leq 12$, based on OMP algorithm for compressed sensing, the high accuracy recovery of the difference signal of cDNA can be realized.

Acknowledgments. This work is partially supported by the National Natural Foundation Project (61304199), the Ministry of Science and Technology projects for TaiWan, HongKong and Maco (2012DFM30040), the Major projects in Fujian Province (2013HZ0002-1,2013YZ0002, 2014YZ0001), the Science and Technology project in Fujian Province Education Department (JB13140/GY-Z13088), and the Scientific Fund project in Fujian University of Technology (GY-Z13005,GY-Z13125).

References

1. Ping, L.Y.: Biological sensors and biological chips: the field of biological macromolecules. Chin. J. Lab. Diagn. **9**(4), 645–648 (2005)
2. Guolian, H., Chen, D., Shukuanl, X., et al.: Novel detection system of microbe chip and its application. Acta Optica Sinica **27**(3), 499–504 (2007)
3. Shmulevich, I., Astola, J., Cogdell, D., et al.: Data extraction from composite oligonucleotide microarrays. Nucleic Acids Res. **31**(7), 431–439 (2003)

4. Jiao, L.C., Yang, S.Y., Liu, F., et al.: Development and prospect of compressive sensing. Acta Electronica Sinica **39**(7), 1651–1662 (2011)
5. Sheikh, M.A., Sarvotham, S., Milenkovic, O., et al.: DNA array decoding from nonlinear measurements by belief propagation. In: 2007 IEEE/SP Workshop on Statistical Signal Processing, SSP 2007, pp. 215–219. IEEE (2007)
6. Parvaresh, F., Vikalo, H., Misra, S., et al.: Recovering sparse signals using sparse measurement matrices in compressed DNA microarrays. IEEE J. Sel. Top. Signal Process. **2** (3), 275–285 (2008)
7. Gilbert, A., Indyk, P.: Sparse recovery using sparse matrices. Proc. IEEE **98**(6), 937–947 (2008)
8. Wang, J.-W., Wang, X.: Image reconstruction method based on compressed sensing for magnetic induction tomography. J. Northeast. Univ. (Nat. Sci.) 36(12), 1687–1690 (2015)
9. Dai, Q.-H., Fu, C.-J., Ji, X.-Y.: Research on compressed sensing. Chin. J. Comput. **34**(3), 425–434 (2011)
10. Shi, G.M., Liu, D.H., Gao, D.H., Liu, Z., Lin, J., Wang, L.J.: Advances in theory and application of compressed sensing. Acta Electronica Sinica **37**(5), 1070–1081 (2009)
11. Fei, X., Qingshan, Y.: Neurodynamic optimization method for recovery of compressive sensed signals. Appl. Res. Comput. **32**(8), 2551–2553 (2015)
12. Shen, B., Tu, D.-W., Zeng, A.-H.: DNA chip fluorescence detection by CCD. Opt. Instrum. **27**(5), 16–20 (2005)
13. Xu, Y., Ruan, Q.-F., Li, Y.-P.: Analysis methods of expression genes. J. Food Sci. Biotechnol. **27**(1), 122–126 (2008)
14. Bo, Z., Yu-lin, L., Kai, W.: Restricted isometry property analysis for sparse random matrices. J. Electron. Inf. Technol. **1**, 169–174 (2014)
15. Candes, E.J., Tao, T.: Decoding by linear programming. IEEE Trans. Inf. Theory **51**(12), 4203–4215 (2005)
16. Xiaobo, L.: Research on measurement matrix based on compressed sensing. Beijing Jiaotong University, pp. 16–19 (2010)
17. Tropp, J.A., Gilbert, A.C.: Signal recovery from random measurements via orthogonal matching pursuit. IEEE Trans. Inf. Theory **53**(12), 4655–4666 (2008)
18. Feng, L., Yi, G.: Compressed sensing analysis. Science Press, Beijing, pp. 66–69 (2015)
19. Wang, J.: Support recovery with orthogonal matching pursuit in the presence of noise. IEEE Trans. Signal Process. **63**(21), 5868–5877 (2015)

http://www.springer.com/978-3-319-48489-1

Genetic and Evolutionary Computing
Proceedings of the Tenth International Conference on Genetic and Evolutionary Computing, November 7-9, 2016 Fuzhou City, Fujian Province, China

Pan, J.-S.; Lin, J.C.-W.; Wang, C.-H.; Jiang, X.H. (Eds.)

2017, XIV, 316 p. 128 illus., Softcover
ISBN: 978-3-319-48489-1