An Overview of DNA Microarray Image Requirements for Automated Processing

Peter Bajcsy
National Center for Supercomputing Applications (NCSA), University of Illinois at Urbana-Champaign (UIUC)
pbajcsy@ncsa.uiuc.edu

Abstract

We present an overview of DNA microarray image requirements for automated processing and information extraction from spotted glass slides. Motivation of our review comes from the need to automate high-throughput microarray data processing due to exponentially growing amounts of microarray data. In order to automate microarray image processing and draw biologically meaningful conclusions from experiments, one has to understand the processing flow, modeling assumptions, uncertainties involved, and the computational tradeoffs of multiple approaches. We present a model of an ideal microarray image and microarray deviations from the model in real experiments. In the summary, we discuss several open problems and the current challenges of high-throughput microarray image processing.

1. Introduction

DNA microarray image processing is one of the information extraction problems occurring in molecular biology and bioinformatics [12]. Molecular biologists and bioinformaticians are using microarray technology for identifying a gene in a biological sequence and predicting the function of the identified gene within a larger system [6] (although there is still an active debate about how to define the bioinformatics discipline [8]). Microarray technology is based on creating DNA microarrays that are typically composed of thousands of DNA sequences, called probes, fixed to a glass or silicon substrate. Usually, samples from two sources are labeled with different fluorescent molecules (emitting at red and green wavelengths) and hybridized together on the same array. The array is then scanned by activation with lasers at the appropriate wavelength to excite each dye. The relative fluorescence between each dye on each spot is then recorded and a composite image may be produced. The relative intensities of each channel represent the relative abundance of the RNA or DNA product in each of the two samples.

Since the invention of microarray technology in 1995, researchers developed several microarray image processing methods, statistical models and data mining techniques that are specific to DNA microarray analysis [13]. These analyses are usually part of a microarray data processing workflow that includes, grid alignment, spot segmentation, quality assurance, data quantification and normalization, identification of differentially expressed genes and their significance testing, and data mining. An example of microarray data processing workflow is illustrated in Figure 1. The subset of image processing steps is enclosed with a dashed line in Figure 1.

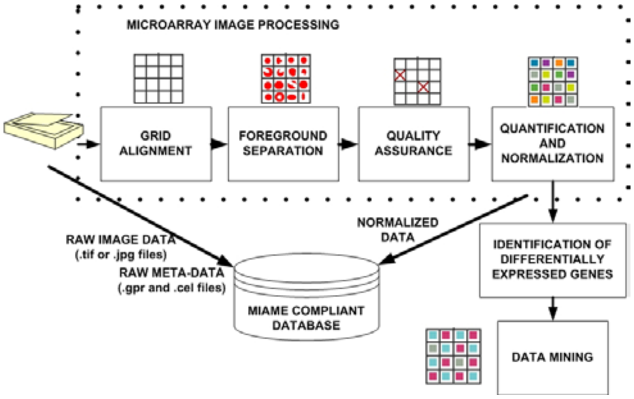
MICROARRAY IMAGE PROCESSING
GRID ALIGNMENT
FOREGROUND SEPARATION
QUALITY ASSURANCE
QUANTIFICATION AND NORMALIZATION
RAW IMAGE DATA (.tif or .jpg files)
RAW META-DATA (.gpr and .cel files)
NORMALIZED DATA
MIAME COMPLIANT DATABASE
IDENTIFICATION OF DIFFERENTIALLY EXPRESSED GENES
DATA MINING

Figure 1: Microarray data processing workflow. The diagram stresses the requirement to archive both raw and processed data.

The major tasks of DNA microarray image processing are (1) to identify the array format including the array layout, spot size and shape, spot intensities, distances between spots, and background fluorescence, and (2) to extract spot descriptors, as well as the uncertainty of the descriptors that represent the underlying microarray experiment. Biological conclusions are then drawn based on the results from data mining and statistical analysis of all extracted descriptors. The reliability of spot descriptors depends on many different factors [18]. For example, one could list basic factors, such as microarray technology components, and protocols for array production, sample labeling, hybridization and image acquisition. Printing parameters, such as pin size and shape, printing speed, temperature and humidity, printing buffers and deposition surface, will all affect the size and morphology of the individual spots. The type of glass and coating, blocking agents, hybridization and wash buffers will affect background fluorescence. DNA microarray image analysis programs must be easily adapted to these varying parameters.

In this paper, we will overview microarray image processing requirements since their understanding is critical for automation in high throughput microarray environments. We present a model of an ideal microarray image and overview microarray variations in real experiments. In the summary, we discuss several open problems and the current challenges of high-throughput microarray image processing.

2. Microarray Image Processing Requirements

In order to choose an appropriate image processing approach and automate DNA microarray image analysis, one has to understand variations of input microarray images in terms of (1) the image content including foreground and background morphology (e.g., grid layout, spot location, shape and size), and intensity information (e.g., spot descriptors derived from foreground and background intensities), (2) the computer characteristics of input digital images (e.g., number of channels, number of bytes per pixel, file format). Figure 2 shows two examples of microarray images and their very different appearance. These variations have to be compensated by microarray image processing algorithms so that the processing performance meets expected accuracy and speed requirements.

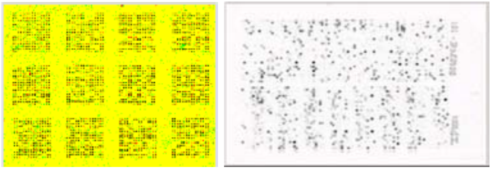

Figure 2: Examples of microarray images with double-fluorescent (left) and radioactive (right) labeled samples that differ in terms of the content (spot geometry, spot size and intensity meaning) and computer characteristics (number of channels and number of bytes per pixel).

What are our expected accuracy and speed requirements on microarray image processing? To answer this question, we consider an ideal microarray image first. Next, we describe an overview of the current understanding of image variations and their sources. Finally, we present the image processing requirements that one should strive to meet.

2.1. Ideal Microarray Image

First, let us define an "ideal" cDNA microarray image in terms of its image content. The image content would be characterized by deterministic grid geometry, known background intensity with zero uncertainty, pre-defined spot shape (morphology), and constant spot intensity that (a) is different from the background, (b) is directly proportional to the biological phenomenon (up- or –down-regulation), and (c) has zero uncertainty for all spots. Figure 3 shows an example of such an ideal microarray image. While finding such an ideal cDNA image is probably a pure utopia, it is a good starting point for understanding image variations and possibly simulating them [3].

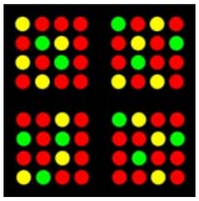

Figure 3: Illustration of an "ideal" microarray image.

Another aspect of an "ideal" cDNA microarray image can be expressed in terms of statistical confidence. If one could not possibly acquire an ideal microarray image, then a high statistical confidence in microarray measurements would be obtained with a very large number of pixels per spot (theoretically it would reach infinity). However, the cost of experiments, the limitations of laser scanners in terms of image resolution, storage of extremely high resolution images and other specimen preparation issues are the real world constraints that have to be taken into account.

The above considerations about an "ideal" microarray image can be used for simulations [3]. Simulations of cDNA microarray images can generate data sets for testing multiple microarray processing algorithms. Motivation for developing microarray image simulation programs also comes from the fact that it is difficult to obtain (a) physical ground truth as an image valuation standard because of the image preparation complexity, and (b) large number of replicates of biological samples as a statistically significant standard because of the cost. In addition, simulations can provide scientific insights about various impacts of microarray preparation fluctuations on the accuracy of final biological conclusions. However, while simulations improve our understanding, they have to be verified by processing real microarray images. Another challenge with simulations is related to setting input simulation parameters since they might depend on individual laboratory procedures and on each microarray acquisition apparatus.

2.2. Sources of Microarray Image Variations

Let us investigate sources of DNA microarray image variations. The cDNA technology is a complex electrical-optical-chemical process that spans (a) cDNA slide fabrication, (b) mRNA preparation, (c) fluorescence dye labeling, (d) gene hybridization, (e) robotic spotting, (f) green and red fluorophores excitation by lasers, (g) imaging using optics, (h) slide scanning, (i) analog to digital conversion using either charge-coupled devices (CCD) or photomultiplier tubes (PMT), and (j) finally image storage and archiving. It is hard to estimate the number of random factors in this complex electrical-optical-chemical process and hence we will focus only a few major factors.

We should perhaps mention that some of the variations are temporally varying, some are ergodic (no sample helps meaningfully predict values that are very far away in time from that sample), and some appear as systematic errors more than as random errors. We overview a few sources of image variations observed in foreground, background and intensity information.

Variations of microarray image channels: Based on the cDNA labeling type used during microarray slide preparation (hybridization), one can obtain, for instance, single-, double- or multi-fluorescent images. Most microarray data contain double-fluorescent images from scanners that operate at two wavelengths, e.g., 532nm (red) and 632nm (green) wavelengths forming two channels shown in Figure 2 left. In general, microarray image data can consist of any number of channels. It is possible to foresee the use of more than two or three channels in future, for example, by using hyperspectral imaging [2].

Another variation of microarray image channels is the storage file format, data compression and data accuracy (number of bytes per pixel). For example, a storage file format with lossy data compression introduces undesirable spatial blur of spots and the microarray image analysis becomes less accurate. Similarly, the number of bits per pixel has to accommodate the dynamic range of an analog signal produced by the red or green fluorophores excitation due to laser illumination. Dynamic range corresponds to the maximum minus minimum measured amplitude, and any value outside of the range [min, max] will be mapped to either min or max. For a fixed number of bits and increasing dynamic range, the uncertainty of each intensity measurement increases. In other words, the bins for all analog values converted to the same digital number are becoming wider. Figure 4 illustrates this concept for a digital number represented by two bits.

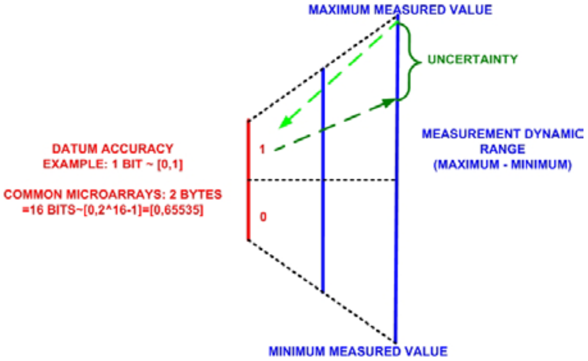
MAXIMUM MEASURED VALUE
UNCERTAINTY
MEASUREMENT DYNAMIC RANGE (MAXIMUM - MINIMUM)
DATUM ACCURACY EXAMPLE: 1 BIT ~ [0,1]
COMMON MICROARRAYS: 2 BYTES =16 BITS~[0,2^16-1]=[0,65535]
1
0
MINIMUM MEASURED VALUE

Figure 4: Illustration of data accuracy, uncertainty and dynamic range dependencies.

In general, microarray image processing algorithms should be able to handle any number of input channels, file format and data accuracy. It should be understood that image analysis results will contain some uncertainty due to file storage and datum accuracy constraints.

Variations of grid geometry: A microarray slide preparation should be considered as one source of variation in grid geometry [4], [10], and [17]. For example, it is important to know that if a spotting machine with several dipping pins prints multiple 2D arrays of spots, then the dipping pins might bend over time and cause irregularity in a 2D arrangement of the printed spots [4]. If measured spot grids are unpredictably irregular then template-based approach to finding spots [14] leads to (a) inaccurate results or (b) unacceptable costs for creating grid templates that would be custom-tuned to each batch of observed grid geometries. An example of alignment inaccuracies is shown in Figure 5.

Similarly, any rotational offset of a slide or dipping pins will cause a rotated 2D grid in a microarray image with respect to the image edge. Figure 6 shows an example of a rotated sub-grid with irregularly spaced rows and columns. Other sources of variations in spot locations are the slide material, such as nylon filters, glass slides, and probe types, such as radioactively labeled probes and fluorescently labeled probes [16]. These variations can be caused (a) by mechanical strain (nylon filters), or (b) by low discrimination power for small surface areas (glass slides), strong background signal (fluorescently labeled probes) or strong signal interference of neighboring spots (radioactively labeled spots). The variations due to mechanical strain introduce warping into the grid geometry. It is important to understand the strain extreme cases in order to limit the search space of grid geometry.

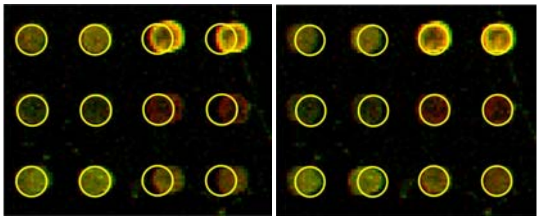

Figure 5: Template-based alignment results obtained by visually aligning the left two columns (left) or the right two columns (right) of microarray spots.

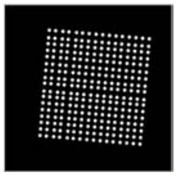

Figure 6: Irregularly spaced and rotated grid geometry of microarray spots.

Due to a small discrimination power, many spots might not be detected [4]. Figure 2 illustrates that many spots might be missing from a 2D array because spot signals are undistinguishable from the background. The absence of spots in a grid poses a challenge for automated grid alignment in addition to other spot location variations. Clearly, missing spots decrease the likelihood of successfully identifying grid configurations by data driven approaches because of a smaller amount of grid evidence. For example, a fully automated grid alignment method would fail to detect correctly a grid if one row of spots from the grid along its border would be completely missing (no evidence about the row existence as illustrated in Figure 7).

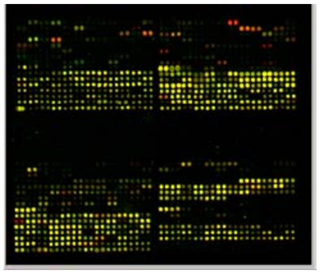

Figure 7: Four sub-grids on one microarray slide. The lower right sub-grid has one less row than other sub-grids.

Variations of background: Background variations occur due to (a) microarray slide preparation (hybridization and spotting errors), (b) inappropriate acquisition procedures (presence of dust or dirt), and (c) image acquisition instruments (non-linearity of imaging components). While the (a) and (b) types of background variations should be detected by microarray quality assurance (see example in Figure 8), the variation due to image acquisition instruments cannot be removed by a user. Thus, many image processing algorithms compensate for background variations by modeling its probability distribution function (PDF). The most frequent model is the Gaussian PDF [3]. Other statistical models to consider would be a uniform PDF or a functional PDF depending on the observed properties of acquired images. For instance, a functional PDF could simulate a positive or negative slant surface function (background intensity shading) that would be combined with spike noise, where spike noise intensities follow an exponential distribution. Figure 9 shows background examples that could be modeled by Normal or Student's t PDF models. It is also necessary to mention that while all channels of microarray images might follow the same PDF, each channel would likely have different parameters for the chosen PDF model.

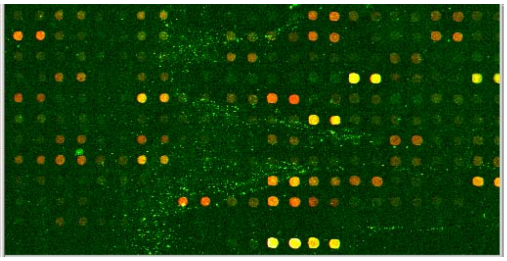

Figure 8: Background variation due to slide washing that should be detected by quality assurance.

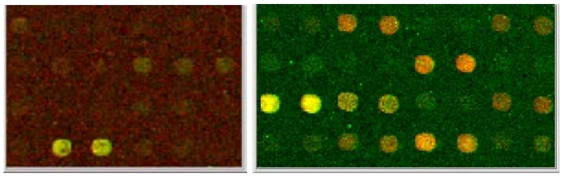

Figure 9: Examples of background noise that could be modeled with PDF models of noise. (Gaussian PDF – left and Student's t PDF – right).

Variations of spot morphology: Another issue to mention is the shape of microarray grid elements (or grid shape primitives). Although the majority of current cDNA microarray imagery is produced with circular spots as shape primitives, one can find the use of other primitive shapes, e.g., lines or rectangles (see the CLONDIAG chip [5]). It is very likely that other primitive shapes than a round spot shape will be used in microarray technology in the future. Figure 10 shows examples of rectangular and triangular shapes.

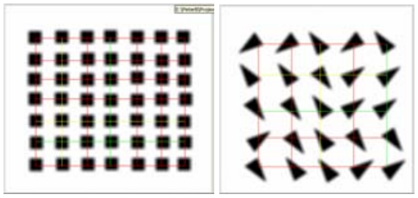

Figure 10: Examples of spot morphologies other than circular.

For the currently most common circular spots, there exist a large number of shape deviations (equals to the total number of foreground and background pixel combinations inside of a grid cell). Figure 11 shows a few classes of morphological deviations as found in microarray images. There are many more spot deviations that have to be analyzed during spot quality assessment. The goal of assessment is to determine a validity of measured spot information and our confidence in deriving any conclusions based on the spot measurement.

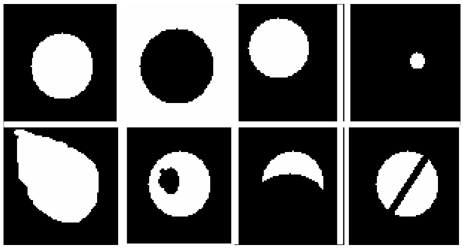

Figure 11: Spatial and morphological variations of spots (from left to right, top row first): (a) a regular spot, (b) an inverse spot or a ghost shape, (c) a spatially deviating spot inside of a grid cell, (d) a spot radius deviation, (e) a tapering spot or a comet shape, (f) a spot with a hole or a doughnut shape, (g) a partially missing spot and (h) a scratched spot.

Variations of foreground and background intensities: Foreground and background intensity variations are also present in microarray image analysis due to slide materials and several labeling techniques. For example, while the fluorescent labeling type leads to microarray images with dark background and bright spots (signal), other labeling types with or without radio-isotopic labels lead to images with bright background and dark spots (see Figure 2 right). A slide material introduces another intensity variation, for example, coated glass slides or nylon membrane or silicon chips. One should understand that it is the background and foreground intensity difference that is relevant to the biological meaning. However, the range of the intensity difference (max – min) and the amplitude of background and foreground variations affect the discrimination of these two classes, as well as our confidence in accurate separation of background and foreground.

Although we described variations of background and the dark-bright schemes for background and foreground, we did not address the issue of foreground spot intensity variations. The reason for this is that microarray images often represent experiments of a discovery type. When discovering biological properties, one cannot predict measurement outcomes such as spot intensity profiles. Thus, one should only adjust parameters of measurement instruments to fully cover the dynamic range of spot intensities so that intensity values are not saturated and possibly discernable from others. As of now, intensities of each spot are modeled according to our previously described ideal microarray image but future research might reveal additional information in the intensity profiles of individual spots.

3. Summary

After reviewing variations of microarray images, one would like to design automated microarray image processing algorithms that are robust to all variations. The robustness would include (1) any number of channels, (2) any storage and computer representation, (3) variable grid and spot locations, (4) unknown background noise, (5) variable background and foreground dark-bright schemes, (6) deviations from spot shapes and (7) deviations from expected spot intensity profiles. Furthermore, the processing algorithms should recognize those cases when missing spots disable automation (or accurate automated image processing) because of the lack of grid evidence.

For anyone who performs scientific experiments with microarray technology, it is important to guarantee microarray image processing repeatability. Assuming that an algorithm is executed with the same data, we expect to obtain the same results every time we perform an image processing step. In order to achieve this goal, algorithms should be “parameter free” so that the same algorithm can be applied repeatedly without any bias with respect to a user’s parameter selection. Thus, for instance, any manual positioning of a grid template is not only tedious and time-consuming but also undesirable since the grid alignment step cannot then be repeated easily. A concrete example of the repeatability issues is presented in [11], where authors compared results obtained by two different users from the same slide (optic primordial dissected from E11.5 wild-type and aphakia mouse embryos) while using the ScanAlyze software package [7]. Each user provided the same input about grid layout first, and then placed multiple grids independently and refined the spot size and position. The outcome of the comparison led up to two-fold variations in the ratios arising from the grid placement differences.

Finally, the amount of microarray image data is growing exponentially and so one is concerned about preparing sufficient storage and computational resources to meet the requirements of end users. For example, finding a grid of spots can be achieved much faster from a sub-sampled microarray image (e.g., processing one out of 5x5 pixels), but the grid alignment accuracy would be less than if the original microarray image had been processed. There are clearly tradeoffs between computational resources (memory and speed/time) and alignment accuracy given a large number of microarray images [1]. While this issue might be resolved without any accuracy loss by using either supercomputers or distributed parallel computing with grid-based technology [9], it might still be beneficial to design image processing algorithms that could incorporate such resource limitations.

DNA microarray image processing is a basic component of learning about gene expression. We have overviewed several DNA microarray processing steps and their requirements for achieving automation in high throughput environments. In the future, researchers will have to address a few additional challenging issues in extracting reliable information about microarray experiments. One of the future challenges of image processing will be the optimization of information extraction and the fine play between over saturation of an image and signals below detection level. A series of questions arises in this context. How can we increase the dynamic range? Can we use partially saturated spots? Shall we reject low quality spot data or attempt to extract whatever useful data can be saved? Can individual spots that are saturated be flagged and rescanned at lower PMT values in an automated fashion until relevant ratios are obtained? Can we construct composite images from different scanning intensities to maximize the number of spots that (a) fall into detectable ranges with good ratios and (b) are not biased by pixels that are too high or too low in intensity?

Other challenges are related to microarray image storage and archival, standardization, automation and fully automated high-throughput processing requirements. There is also a lack of understanding of microarray images at pixel level and uncertainty propagation. The integration of gene expression information with other biological measurements and prior knowledge is also an open area of research. The above questions and challenges have to be answered by additional research and development.

4. References

[1] Bajcsy P., “Gridline: Automatic Grid Alignment in DNA Microarray Scans,” IEEE Transactions on Image Processing, VOL 13, NO 1, pp.15-25, January 2004.

[2] Bajcsy P and R. Kooper, “Prediction Accuracy of Color Imagery from Hyperspectral Imagery,” SPIE 2005, Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XI, Vol. 5806-3428, Orlando, FL, USA.

[3] Balagurunathan Y., E. R. Dougherty, Y. Chen, M. L. Bittner, and J. M. Trent, “Simulation of cDNA Microarrays via a Parameterized Random Signal Model,” Journal of Biomedical Optics, 7(3), 2002.

[4] Buhler J., T. Ideker, D. Haynor, “Dapple: Improved Techniques for Finding Spots on DNA Microarrays,” UV CSE Technical Report UWTR 2000-08-05.

[5] CLONDIAG Chip Technologies,” FluorIS: Array Standardization Tool,” Product Description at http://www.clondiag.com/products/dispo/fluoris/index.php

[6] Chee M., R. Yang and E. Hubbell *et al.*, Accessing genetic information with high-density DNA arrays, *Science* 274 (1996) (5287), pp. 610–614.

[7] Eisen M., “ScanAlyze,” Product Description at. http://rana.lbl.gov/EisenSoftware.htm

[8] Fenstermacher D., “Introduction to Bioinformatics,” Journal of the American Society for Information Science and Technology, 56(5):440-446, 2005.

[9] Foster I. and C. Kesselman. “Computational Grids,” *Chapter 2 of "The Grid: Blueprint for a New Computing Infrastructure"*, Morgan-Kaufman, 1999.

[10] Goryachev A. B., P. F. MacGregor and A. M. Edwards, “Unfolding Microarray Data,” Journal of Computational Biology, Volume 8, Number 4, 2001, pp. 443-461.

[11] Lawrence N. D., M. Milo, M. Niranjan, P. Rashbass, and S. Soullier, “Reducing the variability in cDNA microarray image processing by Bayesian inference,” Bioinformatics, Vol 20., NO. 4, 2004, 518-526.

[12] MacMullen W. J. and S. Denn, “Information Problems in Molecular Biology and Bioinformatics,” Journal of the American Society for Information Science and Technology, 56(5):447-456, 2005.

[13] Quackenbush J: Computational analysis of microarray data. *Nat. Rev. Genet.* 2001, 2(6):418-427.

[14] Scanalytics Inc., “MicroArray Suite,” Product Description at http://www.scanalytics.com/product/hts/microarray.html

[15] Schena M, Shalon D, Davis RW, and Brown PO: Quantitative monitoring of gene expression patterns with complementary DNA microarray. *Science* 1995, 270: 467-470.

[16] Steinfath M., W. Wruck, H. Seidel, H. Lehrach, U. Radelof, and J. O’Brien, “Automated image analysis for array hybridization experiments,” Bioinformatics 2001 17: 634-641.

[17] Whitfield CW, Cziko AM, Robinson GE. 2003. Gene expression profiles in the brain predict behavior in individual honey bees. Science. 302:296-9.

[18] Yue H, Eastman PS, Wang BB, Minor J, Doctolero MH, Nuttall RL, Stack R, Becker JW, Montgomery JR, Vainer M, Johnston R**.** 2001. An evaluation of the performance of cDNA microarrays for detecting changes in global mRNA expression. Nucleic Acids Research, 29, No. 8 e41-1.