Excitement and Realities in Microarray Analysis of the Biological Effects of Polyphenols

Stephen Barnes\(^1,4,5\) and David B. Allison\(^2,3,4\)

\(^1\)Department of Pharmacology & Toxicology; \(^2\)Section on Statistical Genetics; \(^3\)Department of Biostatistics, UAB Clinical Nutrition Research Center; \(^4\)Center for Nutrient-Gene Interaction in Cancer Prevention, University of Alabama at Birmingham, Birmingham, Alabama, USA; \(^5\)Purdue University – University of Alabama at Birmingham Botanicals Center for Age-Related Disease, Birmingham, Alabama, USA

Abstract

The mechanisms by which polyphenols and their metabolites alter the functions of cells (proliferation, differentiation, and apoptosis) is of great interest as investigators explore their roles in the prevention of disease. DNA microarray analysis is a widespread technology that provides a glimpse of global gene expression. It may provide important perspectives in polyphenols research, involving new leads for further scientific investigation. However, what are its limitations? Are the statistics used in the analysis of microarray data fundamentally sound? Do we really know how to approach the analysis of a system? These topics are discussed along with some recent applications of microarray analysis to polyphenols research.

Keywords: Bonferroni correction, EGCG, fold change, genistein, replication.

Introduction

The genome landscape

The advent of sequencing of many of the genomes of humans and experimental animals, as well as lower organisms, is leading to a transformation of traditional research paradigms. Gene hunters have searched these genome sequences by computational methods and have extracted both known and predicted genes. Although the total number of genes in the human genome was much less than originally presumed (now these may be only as “few” as 24,000) (Pennisi, 2003), it has allowed the selection of oligonucleotides that can be used to identify specifically and to quantify individual mRNAs (the transcriptome) that encode the proteins expressed in an organism. The oligonucleotides are attached to microarrays (gene chips) that, in the case of the human genome array, include those from transcripts of more than 33,000 different genes. Similar arrays are available for the mouse genome, and it is expected that the rat genome will be fully represented on one chip in the near future. A parallel technology involves the use of cDNAs that can be printed onto glass slides. It is hoped that this latter method may substantially reduce the cost of carrying out analysis of the transcriptome. At the present time, the substantial cost of the gene chip approach (US$400+ per chip) causes investigators to abrogate their experimental design in such a way that the value of this interesting technology is put at risk. There can be no doubt that solving the technical problems and providing the methodology at a nominal cost will lead to the predicted revolution in the understanding of biological systems.

Driving forces in polyphenols research

An ever-increasing community of investigators is studying the effect of polyphenols in biological systems. Polyphenols have a plethora of reported biochemical activities, ranging from inhibition of tyrosine kinases (Akiyama et al., 1987), mimicking of physiological estrogens (Jacobs & Lewis, 2002), action as antioxidants...
High numbers of measured variables and statistical analysis

The sudden shift from science focused on discrete hypotheses to experiments that allow “discovery” has excited many investigators. It has been seductive because apparently whole new areas of research have been opened up for investigation. However, too many times initial exciting reports have not been confirmed, either by other investigators or even by the team that provided the first report. Why is this?

In conventional experiments, there are only a few (e.g., $k = 1$ to 5) variables to consider, and the experiment is replicated at least $n$ times, where $n \gg k$. Statistical methods for analyzing such data are well described and are familiar to most investigators. However, in high-dimensional biology experiments such as DNA microarrays, where $k \gg n$ (e.g., $k = 100$ to $100,000$ and $n = 1$ to 5), standard statistical methods are often not optimal. Consequently, new statistical methods are needed; indeed, many are proffered in the literature. However, few have been rigorously validated (Mehta et al., 2004).

One obvious solution is to decrease the number of variables measured and/or parameters to be estimated—this is already occurring. For instance, manufacturers are putting a much smaller ($n = 20$ to 100) number of DNA or protein targets on arrays that are designed to specialize in biochemical events that are closely related. Some examples of these approaches are presented later in this article. But even here, the number of replicates needed to avoid unacceptable rates of false positives and false negatives is likely to be far greater than most investigators are prepared to afford. For genome-wide analysis, because of the cost of DNA arrays, investigators have typically limited their efforts to 1–3 replicates, and journals have accepted this situation. However, this has to change for the good of science.

A conservative way to take into account the large numbers of variables (parameters) that are studied (estimated) in DNA array analysis is the Bonferroni correction (Wayne & McIntyre, 2002; Bennett et al., 2003). In this method, the value of $p$ needed to reject the null hypothesis is divided by the number of tests conducted. For an array with 10,000 different DNA targets and a separate test on each target, this means that $p$ has to be less than 0.000005 to reach significance at the experiment-wise 0.05 $\alpha$ level. Even for an experiment with 100 DNA targets, $p$ has to be less than 0.0005. This is a harsh correction and often seen as impractical. Statisticians are therefore developing alternative procedures that maintain the expected proportion of “discoveries” (i.e., rejected null hypotheses) that are false below some specified level. This is referred to as the false discovery rate, or FDR, and is increasingly seen as a more sensible alternative to the Bonferroni correction and other techniques for controlling the type 1 error rate (Keselman et al., 2002; Sabatti et al., 2002).

Ensuring that there is a reasonable probability of finding genes that are differentially expressed across conditions while at the same time maintaining FDR rates at practical levels generally involves collecting far more biological replicates than is commonly practiced. There is an additional advantage in using a larger number of replicates in that it allows one to be able to permute or resample from the data a sufficient number of different ways to allow robust non-parametric inference to be useful (Chernick, 1999; Zhang & Zhao, 2000) or, if the sample size is larger still (e.g., $\sim 30$ or more), one can rely on the asymptotic properties of parametric statistics, thereby leading to more valid analyses.
Value of repeated experiments

It is possible to estimate the fraction of true and false positives in an experimental data set (Gadbury et al., 2004). In general, the efficiency of discovering the true effects rises rapidly with the number of experimental replicates. However, the fraction of false positives may decline much more slowly depending on the data and the cut-off used. Array analysts try to limit the propensity of misleading findings by applying a twofold change criterion as a filter—however, as we shall see later, this is a very arbitrary method and may simply be unrealistic in a system. If only a small number of genes are truly changed, investigators may still find themselves with a big haystack to search for the true positives among an excess of false positives. At this point, the extensive degeneracy of the gene array reveals itself. Many combinations of gene changes may become apparent, particularly through the rose-tinted glasses of the investigator searching for confirmation of a previously cherished hypothesis. Repeating the same experiment in different labs will lead to data that are a mixture of real events and a much larger set of false positives. Because these two types of results cannot be distinguished, quite different, but compelling explanations of the data are proffered. One way that helps to resolve the dilemma is to repeat the entire experiment in a lab (many will remember telling a graduate student, “Interesting, but repeat the experiment two more times and then come back and see me”). The true positive results are expected to occur with higher frequency in each repeat experiment, whereas the false positives will appear randomly. By using a Venn diagram, only those results that overlap between experiments should be considered as possible true positives. More sophisticated approaches to searching for such interactions are now being developed (Kim et al., in press). It should be pointed out that even those factors that overlap between repeated experiments have a finite false discovery rate; however, they are a smaller proportion each time the experiment is repeated.

Importance of gene-gene relationships

Often, an implicit assumption has been that each of the genes or proteins that are being measured varies independently of the others. This is palpably not true. Gene and protein families are produced in a coordinated manner during the life of a cell. For instance, the DNA synthesis that precedes cell division requires a carefully orchestrated set of proteins both in content, place, and time. Therefore, some complexity in the data, may be reduced by grouping known associated factors. Similarly, previous and independently observed variance estimates can be included in the analysis using Bayesian approaches (Long et al., 2001; Moloshok et al., 2002; Yang et al., 2004). If associations are not known a priori, then they may become apparent from the data. Investigators use cluster and principal component analyses to distill such associations. It should be noted that these may become hypotheses for future analysis rather than being proved by the current analysis. Investigators may use evidence of such associations in reports of work using non-array approaches to give support for the hypotheses.

There are two very important questions regarding DNA microarray data with respect to how well are they connected to the world outside of the experiment. For example, is a change observed at the mRNA level translated to the protein level? And can we recognize the important changes that occur that make a cell behave differently (presumably this is the purpose of the experiment)? There has been a concern about the poor relationship between mRNA and protein expression (Ideker et al., 2001; Griffin et al., 2002). Although housekeeping genes and proteins (i.e., highly expressed genes) are well-correlated, low-abundance genes are not (Ideker et al., 2001). A possible explanation is that housekeeping genes are continually expressed, whereas low-abundance genes may only be expressed at certain times in the cell life cycle (e.g., the cyclins). Because transcription of genes to mRNA and translation of the mRNA to protein do not occur simultaneously, the time of sampling following the application of a stimulus may critically affect the apparent relationship between gene transcription and translation. Shortly after the stimulus, gene transcription dominates, whereas later translation to protein is the more important event (Fig. 1). To investigate the correlation, one should integrate the levels of the mRNA and the protein over time. A similar approach can be used to apply more critically microarray analysis when examining the effects of both time and concentration of an effector substance. In the one-dimensional world that most of us are more familiar with, collecting multiple data points about a process over time has been the norm and is informative about how changes at the cellular level have come about. Single data images, whether they are of microarray data or photographs, can be very misleading. Multiple data images increase the likelihood that the observed changes in gene expression are true.

Filtering data: Validity of the two-fold change

Investigators have used a two-fold change in expression cutoff as a filter to try to lower the false positive rate. However, is this justified? Almost certainly not. To apply a fixed fold-change cutoff to all experiments would be equivalent to assuming the measurement error and biological variability was the same for all genes and that
Microarray analysis of the biological effects of polyphenols

Figure 1. Effect of a stimulus on the correlation of mRNA and protein synthesis. In (A), a housekeeping gene that is continually expressed (e.g., actin or ATP synthetase) may undergo a small increase of synthesis of its mRNA that later results in increased synthesis of the corresponding protein. The ratio of mRNA to protein for this gene therefore changes only a little (i.e., they are well correlated at all times). In (B), a cell cycle gene is not expressed until the stimulus is given. Its mRNA increases rapidly and then falls off—the corresponding protein appears with some delay. The ratio of mRNA to protein for this gene varies enormously.

sample size had no effect on the variability of fold-change estimates. All of these assumptions are clearly erroneous.

At this point, we have very little knowledge about how a system as complex as a cell really works. In our one-dimensional world, we have traditionally fought to show our colleagues that the specific gene or protein we elected to study is important (compared to everyone else’s selection). In doing so, we’ve done our best to ignore the systems issues that are actually how the cell works. So, what are the properties of a complex system? Can we translate our knowledge of other systems to those that may operate in a cell? Many systems, such as a religion, a democracy, or an economic model, have very distinct core values. These hardly change because they are fundamental to the system. When they do change, so does the system. On the other hand, these systems are successful because elements in them can flourish. However, the latter are not the controlling factors. Thus, the largest changes that are observed may be unrelated to what we need to know.

So, how do we recognize the important factors? One way is to apply a statistical weighting factor to the fold-change (i.e., importance I = fold change/variance). In this scenario, a low variance will amplify a small fold-change and high variance will reduce it. Thus, there is another advantage to analyzing a larger number of arrays—the ability to obtain a good estimate of variance (Mariani et al., 2003). Furthermore, integrating fold-change and variance across associated genes or proteins gives an estimate of the importance of a particular factor in effects space, an important parameter in systems biology.

Microarrays and polyphenols

There are a limited number of publications where microarray analysis has been used to examine the effects of polyphenols on gene expression in cells and tissues (summarized in Table 1). Li and Sarkar (2002) examined the effect of the isoflavone genistein on the human prostate cancer PC3 cell line. The gene expression changes were examined over a concentration range from 0 to 30 μM and at 0, 6, 36, and 72 h. Each sample was analyzed in two separate experiments that served to reduce (although not eliminate) the false discovery rate. Using the two-fold change criterion, a total of 832 genes out of 12,558 genes on the microarray were changed. Among these, 774 were downregulated and 58 were upregulated by genistein. Cluster analysis using the time-dependence of the data revealed nine different types of expression change—one cluster showed an increase throughout the experiment. These genes involved signal transduction, protein dephosphorylation, heat shock response, inactivation of MAP kinase, apoptosis, and cell cycle arrest. An important aspect of this study was that the investigators also verified by using quantitative reverse transcriptase-polymerase chain reaction (RT-PCR) the expression of 26 genes of the 832 genes identified by microarray analysis.
Table 1. Microarray type and number of genes in studies on polyphenols.

<table>
<thead>
<tr>
<th>Polyphenol</th>
<th>Tissue</th>
<th>Array type</th>
<th>Number of genes</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genistein</td>
<td>PC3</td>
<td>Affymetrix human U95</td>
<td>12,588</td>
<td>Li and Sarkar (2002)</td>
</tr>
<tr>
<td>EGCG</td>
<td>CaSki</td>
<td>Microarray chip</td>
<td>384</td>
<td>Ahn et al. (2002)</td>
</tr>
<tr>
<td>Genistein</td>
<td>MCF-7</td>
<td>Glass; cDNA</td>
<td>448</td>
<td>Chen et al. (2003)</td>
</tr>
<tr>
<td>Genistein</td>
<td>Mesenteric artery</td>
<td>Not stated</td>
<td>Not stated</td>
<td>Rodrigo et al. (2003a, 2003b)</td>
</tr>
</tbody>
</table>

In a related study, Wang and Mukhtar (2002) investigated the effects of (-)-epigallocatechin gallate (EGCG) on human prostate cancer LNCaP cells. In this case, a single concentration (12 μM) and a single time point were used. The experiment was repeated at least two times with fresh cells. In this study, the mRNAs from the EGCG-treated and control cells were converted to cDNAs labeled with Cy3 and Cy5 fluorescent dyes, respectively. This enabled the samples to be mixed and then analyzed on the same microarray. This reduces measurement error caused by changes in the sensitivity of the fluorescence scanner when used on two different occasions. The array contained a total of 250 genes that were members of the kinase and phosphatase families. Once again, a two-fold change cutoff was used to establish significance. This led to the identification of 25 genes of which 16 were upregulated and 9 downregulated.

When independent experiments were considered, only three downregulated genes and one upregulated gene were consistently changed (Table 2). Of those not deemed to have changed, one is the phosphatase PTEN, a tumor suppressor gene. However, the expression of this gene is tightly regulated, and the difference between proliferation and apoptosis is less than a two-fold change. Such changes would not be detected using current microarray protocols and conventions.

The effects of EGCG were also studied in a human cervical cancer CaSki cell line (Ahn et al., 2003). These cells undergo cell cycle G1 arrest at 35 μM and apoptosis at 100 μM. RNA was converted to cDNA with Cy3jCy5 labeling and the cDNA hybridized to a 384-gene microarray containing mostly cancer-related genes. There were 16 genes that were downregulated after incubation with EGCG for 12 and 24 h and 4 genes that were upregulated. Because this study used multiple concentrations of EGCG (35 and 100 μM), the individual microarray data sets represent a form of replication. Therefore, genes that appear in all the data sets are more likely to be true positives. These (seven downregulated and three upregulated) are shown in Table 3. Interestingly, three of the downregulated genes do not have a known function at this time. A similar method of replication analysis can be carried out between data sets obtained at the same concentration (35 μM) of EGCG, but at different times. In that case, seven genes are downregulated and four are upregulated (Table 4).

Chen et al. (2003) examined the effect of genistein (50 and 100 μM) on human breast cancer MCF-7 cells. The microarray they used contained 448 human genes. Using a twofold change criterion, eight genes were downregulated, but only one, ERα, was changed at both concentrations. Three genes were upregulated, but only one (heat shock protein 105) at both concentrations. Because genistein inhibited cell growth and caused apoptosis at these concentrations, it is surprising that more genes were not changed in expression. It is probable, therefore, the changes were more subtle than could be assessed by the two-fold change cutoff.

Ramanathan and Gray (2003) used differential display reverse transcriptase PCR to identify 10–12 transcripts in MCF-7 cells that were upregulated by genistein. One of these, PE-13.1, was activated by genistein, but not 17β-estradiol. However, its sequence did not match any previously reported gene, although it had a good match to the BAC clone RP11-58052. It also did not match any predicted protein in the human EST database. In this respect, it is analogous to the DING protein recovered from MCF-7 cells using a genistein affinity procedure—DING has no corresponding gene (Belenky et al., 2003).

Microarrays were also used to examine the effects of genistein in the vasculature. Rodrigo et al. (2003a) found that in mesenteric artery endothelial cells, genistein downregulated endothelin converting enzyme-1 (ECE-1) that catalyzes the conversion of pro-endothelin...
**Table 3.** Genes whose expression in human cervical cancer CaSki cells were consistently downregulated according to EGCG concentration or time of incubation.

<table>
<thead>
<tr>
<th>Genes/proteins</th>
<th>Fold decrease due to time</th>
<th>Fold decrease due to concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td>NADH dehydrogenase (ubiquinone) 1 β</td>
<td>2.2</td>
<td>—</td>
</tr>
<tr>
<td>CD83 antigen (activated B lymphocytes)</td>
<td>2.7</td>
<td>3.1</td>
</tr>
<tr>
<td>Dual specificity phosphatase 1</td>
<td>4.1</td>
<td>2.9</td>
</tr>
<tr>
<td>Ras homologue gene family (rho G)</td>
<td>2.8</td>
<td>2.8</td>
</tr>
<tr>
<td>HSPC135 protein</td>
<td>2.2</td>
<td>—</td>
</tr>
<tr>
<td>DGFZP58600120 protein</td>
<td>2.7</td>
<td>3.1</td>
</tr>
<tr>
<td>DGFZP5860917 protein</td>
<td>3.0</td>
<td>2.7</td>
</tr>
<tr>
<td>KIAA0210 gene product</td>
<td>3.3</td>
<td></td>
</tr>
<tr>
<td>KIAA0250 gene product</td>
<td>2.8</td>
<td>2.9</td>
</tr>
<tr>
<td>In activated T/LAK lymphocytes</td>
<td>—</td>
<td></td>
</tr>
</tbody>
</table>

*Adapted from Ahn et al. (2003).

**Table 4.** Genes whose expression in human cervical cancer CaSki cells were consistently up regulated according to EGCG concentration or time of incubation.

<table>
<thead>
<tr>
<th>Genes/proteins</th>
<th>Fold increase due to time</th>
<th>Fold increase due to concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Solute carrier family 25</td>
<td>2.1</td>
<td>—</td>
</tr>
<tr>
<td>Vimentin</td>
<td>2.4</td>
<td>2.9</td>
</tr>
<tr>
<td>Ribosomal protein L19</td>
<td>2.2</td>
<td>2.7</td>
</tr>
<tr>
<td>Mitogen-activated protein kinase kinase kinase 3</td>
<td>2.3</td>
<td>2.4</td>
</tr>
</tbody>
</table>

*Adapted from Ahn et al. (2003).

There is only one study where application of DNA microarray analysis has been applied to tissues. Naciff et al. (2002) investigated the effects of 17α-ethinylestradiol, bisphenol A, and genistein (three different doses for each compound) on gene expression in the uterus of developing female rats. Although 17α-ethinylestradiol and bisphenol A upregulated the expression of a similar group of genes (26 and 35, respectively), genistein altered the expression of 227 genes, most of which were downregulated.

**Conclusions**

Microarray analysis is an exotic scientific tool that could revolutionize polyphenols research. At this time, its application to the study of polyphenols has been limited. Improvements in experimental design and statistical methodology will be needed in order to realize its full potential.

**Acknowledgments**

Research on the work on polyphenols at UAB is supported by funds from federal grants to the UAB Center for Nutrient-Gene Interaction in Cancer Prevention (U54 CA100949), the Purdue-UAB Botanicals Center.
for Age-Related Research (P50 AT-00477, Connie Weaver, PI), NIH grants R01 DK56366, P30 DK56336, P20 CA093753, U24 DK058776, and R01ES09912, NSF grants 0090286 and 0217651, and a grant from the UAB-HSF-GEF.

References


Microarray analysis of the biological effects of polyphenols


