ACADEMIC PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

GENOMICS

Genomics 81 (2003) 98–104

www.elsevier.com/locate/ygeno

Short Communication

Global functional profiling of gene expression☆

Sorin Drăghici,[a,*] Purvesh Khatri,[a] Rui P. Martins,[b] G. Charles Ostermeier,[b,c] and Stephen A. Krawetz[b,c]

[a] *Department of Computer Science, Wayne State University, 5143 Cass Avenue, Detroit, MI 48202, USA*

[b] *Center for Molecular Medicine and Genetics, Wayne State University, 253 C.S. Mott Center, 257 E. Hancock, Detroit, MI 48202, USA*

[c] *Department of Obstetrics and Gynecology, Wayne State University, 253 C.S. Mott Center, 257 E. Hancock, Detroit, MI 48202, USA*

Received 4 October 2002; accepted 5 November 2002

Abstract

The typical result of a microarray experiment is a list of tens or hundreds of genes found to be differentially regulated in the condition under study. Independent of the methods used to select these genes, the common task faced by any researcher is to translate these lists of genes into a better understanding of the biological phenomena involved. Currently, this is done through a tedious combination of searches through the literature and a number of public databases. We developed Onto-Express (OE) as a novel tool able to automatically translate such lists of differentially regulated genes into functional profiles characterizing the impact of the condition studied. OE constructs functional profiles (using Gene Ontology terms) for the following categories: biochemical function, biological process, cellular role, cellular component, molecular function, and chromosome location. Statistical significance values are calculated for each category. We demonstrate the validity and the utility of this comprehensive global analysis of gene function by analyzing two breast cancer datasets from two separate laboratories. OE was able to identify correctly all biological processes postulated by the original authors, as well as discover novel relevant mechanisms.

Keywords: Functional analysis; Microarrays; Biological process; Pathway analysis; Breast cancer; Onto-Express; Gene Ontology; GenBank; UniGene

Introduction

Sporadic and familial breast cancer accounts for one-third of cancer diagnoses and 15% of cancer deaths among U.S. women [1]. It is estimated that 205,000 new cases will be diagnosed and 39,600 deaths will occur in 2002, making it the most common noncutaneous cancer and second leading cause of cancer death [2]. The clinical outcome of sporadic breast cancer relies heavily on histological evaluation of biopsied primary and metastasized tumors [3,4]. Information concerning the expression of a number of molecular markers such as estrogen and progesterone receptor (ER, PgR), p53, Ki67, Bcl-29, and HER2/neu [5,6], as well as the penetrance of BRCA1 and BRCA2 mutations in the family germ line [7], can provide a more concise prognostic picture. The trend of using marker expression information has prompted a concerted effort to investigate the diagnostic and prognostic efficacy of generalized patterns of gene expression in breast cancers using microarray technologies. This has led to the characterization of molecular signatures specific to certain types of tumors [8–10] and cell types within those tumors [11,12]. Just as information on markers like ER and PgR can add to a more accurate prognosis, examination of thousands of established and putative markers could provide the much needed information to tailor specifically a patient's treatment regimen.

Various technologies such as cDNA and oligonucleotide arrays are now available to meet this challenge. Independent of the platform and the analysis methods used, the result of a microarray experiment is a list of differentially expressed

☆ This work was funded in part by a Sun Microsystems grant awarded to S.D., NIH Grant HD36512 to S.A.K., a Wayne State University SOM Dean's Post-Doctoral Fellowship, and an NICHD Contraception and Infertility Loan to G.C.O. Support from the WSU MCBI mode is gratefully appreciated.

* Corresponding author.

E-mail address: sod@cs.wayne.edu (S. Drăghici).

0888-7543/03/$ – see front matter © 2003 Elsevier Science (USA). All rights reserved.
doi:10.1016/S0888-7543(02)00021-6

genes. Most data analysis methods available concentrate on this aspect [13]. However, a major challenge is to translate these lists of differentially regulated genes into a better understanding of the underlying biological phenomena. Currently, there are no tools available to support the researcher in this arena. Many a researcher parses such lists of genes manually, using literature searches and browsing public databases, in an attempt to extract the relevant biological processes and pathways. This is an extremely tedious and error-prone process that usually takes many months. Furthermore, even if this manual processing could be done in a systematic and complete manner, we show that the simple frequency of a given biological process among the differentially regulated genes may be misleading.

Onto-Express (OE) is a tool designed to mine the available functional annotation data and help the researcher find relevant biological processes [14]. Many months of tedious and inexact manual searches are substituted by a few minutes of fully automated analyses. The result of these analyses is a functional profile of the condition studied. In the latest version, this functional profile is accompanied by the computation of significance values for each functional category. Such values allow the user to distinguish between significant biological processes and random events. OE's utility is demonstrated by analyzing data from two recent breast cancer studies. This tool is available online at http://vortex.cs.wayne.edu/Projects.html.

Results and discussion

OE's input is a list of genes specified by either accession number, Affymetrix probe ID, or UniGene cluster ID. A functional category can be assigned to a gene based on specific experimental evidence or by theoretical inference (e.g., similarity with a protein having a known function). OE shows explicitly how many genes in a category are supported by experimental evidence (labeled "experimented") and how many are inferred ("inferred"). Those genes for which this information is not available are labeled "non-recorded." The results are provided in graphical form and e-mailed to the user on request. OE constructs a functional profile for each of the Gene Ontology (GO) categories [15]: cellular component, biological process, and molecular function as well as biochemical function and cellular role, as defined by Proteome [16]. As biological processes can be regulated within a local chromosomal region (e.g., imprinting), an additional profile is constructed for the chromosome location.

The following example illustrates OE's functionality. Let us consider that we are using an array containing 2000 genes to investigate the effect of ingesting a certain substance X. Using classical statistical and data analysis methods we decide that 200 of these genes are differentially regulated by substance X. For each of these 200 genes, OE uses the available public data containing information about the biochemical function, biological process, cellular role, cellular component, molecular function, and chromosome location. Let us focus on the biological process, and assume that the results for the 200 differentially regulated genes are as follows: 160 of the 200 genes are involved in mitosis, 80 in oncogenesis, 60 in the positive control of cell proliferation, and 40 in glucose transport. As demonstrated in [17], these results are tremendously useful as such, since they save the researcher the inordinate amount of effort involved in going through each of the 200 genes, compiling lists with all biological processes each gene is involved in, and then cross-referencing all those biological processes to determine how many genes are in each process.

Table 1
The statistical significance of the data mining results

Biological process	Genes found	Genes expected	
Mitosis	160	160	Not better than chance
Oncogenesis	80	80	Not better than chance
Positive control of cell proliferation	60	20	Better than chance
Glucose transport	40	10	Much better than chance

Note. The number of genes that are involved in a given biological process can be misleading. Mitosis may appear to be the most important process affected since 160 of the 200 differentially regulated genes are involved in mitosis. In fact, this is no better than chance alone.

If we now look at the functional profile described above, we might conclude that substance X may be related to cancer since mitosis, oncogenesis, and cell proliferation would all make sense in that context. However, a reasonable question is: what would happen if all the genes on the array used were part of the mitotic pathway? Would mitosis continue to be significant? Clearly, the answer is no. Therefore, it is necessary to compare the actual number of occurrences with the expected number of occurrences for each individual category.

This comparison is shown in Table 1 for the example considered. Now, the functional profile appears to be completely different. There are indeed 160 mitotic genes but, despite this being the largest number, we actually expected to observe 160 such genes so this is not better than chance alone. The same is true for oncogenesis. The positive control of cell proliferation starts to be interesting because we expected 20 and observed 60. This is three times more than expected. However, the most interesting is the glucose transport. We expected to observe only 10 such genes and we observed 40, which is four times more than expected. Considering the expected numbers of genes changed radically the interpretation of the data. Now, we may want to consider the correlation of X with diabetes instead of cancer.

The example above serves only an illustrative purpose. In practice, several other factors need to be considered. First, several data analysis methods have different error

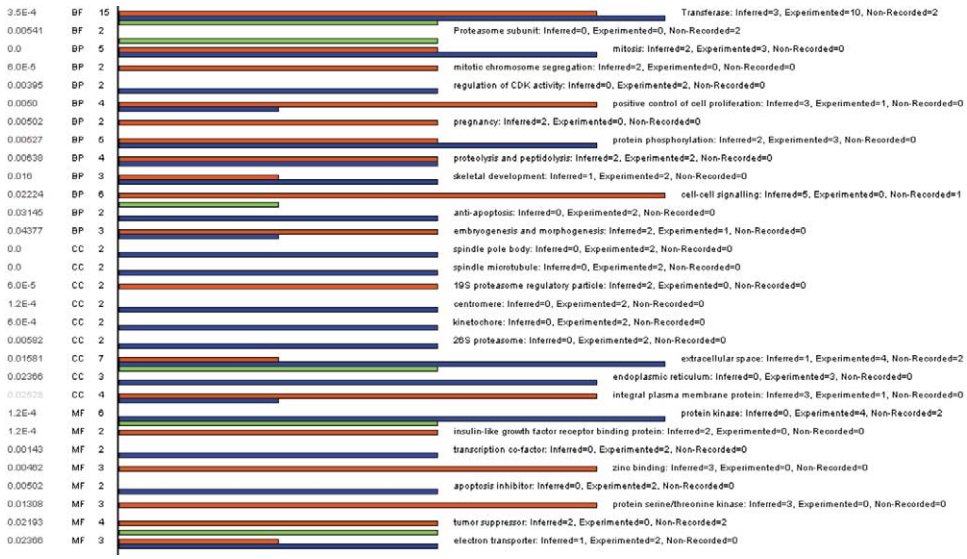

Fig. 1. Significant correlations were observed between the expression level and poor breast cancer outcome for 231 genes [18]. This subset of genes was processed by Onto-Express to categorize the genes into functional groups as follows: **BF,** biochemical function; **BP,** biological process; **CC,** cellular component; **MF,** molecular function. The 30 different functional groups associated with poor disease outcome in a significant way ($p < 0.05$ in left column) are shown. Red bar graphs represent genes for which the function was inferred, blue graphs represent genes for which the function was proved experimentally, and green graphs represent genes for which this type of information was not recorded in the source database.

rates. Thus, OE's input can contain false positives (genes reported as being differentially regulated when they are not). Since the presence and number of such false positives can influence the results, it is important to take this into consideration when interpreting the results. Second, if a custom array is purposefully enriched with a certain type of genes, the significance of those specific genes will appear to be artificially lower. This biological bias has to be considered when interpreting OE's results. Finally, microarray data are typically obtained from several repeated experiments. If a certain biological process is found to be affected in repeated, independent experiments, it is likely that the process is indeed so, independent of the number of genes representing that process on the array.

To illustrate OE's capabilities, we have applied it to a number of breast cancer datasets. As will be shown, our approach has revealed several novel insights. A microarray strategy was recently used to identify 231 genes (from an initial set of 25,000) that can be used as a predictor of clinical outcome for breast cancer [18]. Using a classical approach based on putative gene functions and known pathways, Van't Veer et al. identified several key mechanisms such as cell cycle, cell invasion, metastasis, angiogenesis, and signal transduction as being implicated in cases of breast cancer with poor prognosis. The 231 genes found to be good predictors of poor prognosis were submitted to OE using the initial pool of 25,000 genes as the reference set. We concentrated on those functional categories significant at 5% ($p < 0.05$) and represented by two or more genes (Fig. 1). Our approach was validated by the fact that the results included most of the biological processes postulated to be associated with cancer including the positive control of cell proliferation and antiapoptosis. Oncogenesis, cell cycle control, and cell growth and maintenance are not significant at 5% but do become significant if the threshold is lowered to 10% (see Fig. 2).

OE also identified a host of novel mechanisms. Protein phosphorylation was one of these additional categories significantly correlated with poor prognostic outcome. Apart from its involvement in a number of mitogenic response pathways, protein phosphorylation is a common regulatory tactic employed in cell cycle progression. PCTK1 [19] and STK6 [20] are among the cell cycle regulatory kinases identified as corollaries to prognostic outcome. Similarly, antiapoptotic factors surivin [21,22] and BNIP3 [23] were identified. Both mechanisms are believed to be intimately linked and active in regulating cell homeostasis and cell cycle progression.

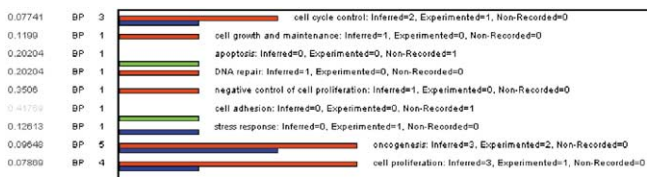

Fig. 2. Some interesting biological processes that are not significant at the 5% significance level. Note that processes commonly associated with cancer such as cell proliferation, cell cycle control, and oncogenesis *are* significant at the 10% significance level. Furthermore, the statistical analysis for apoptosis, cell growth, and maintenance, etc., should be interpreted cautiously since they are represented by a single gene. Red bar graphs represent genes for which the function was inferred, blue graphs represent genes for which the function was proved experimentally, and green graphs represent genes for which this type of information was not recorded in the source database.

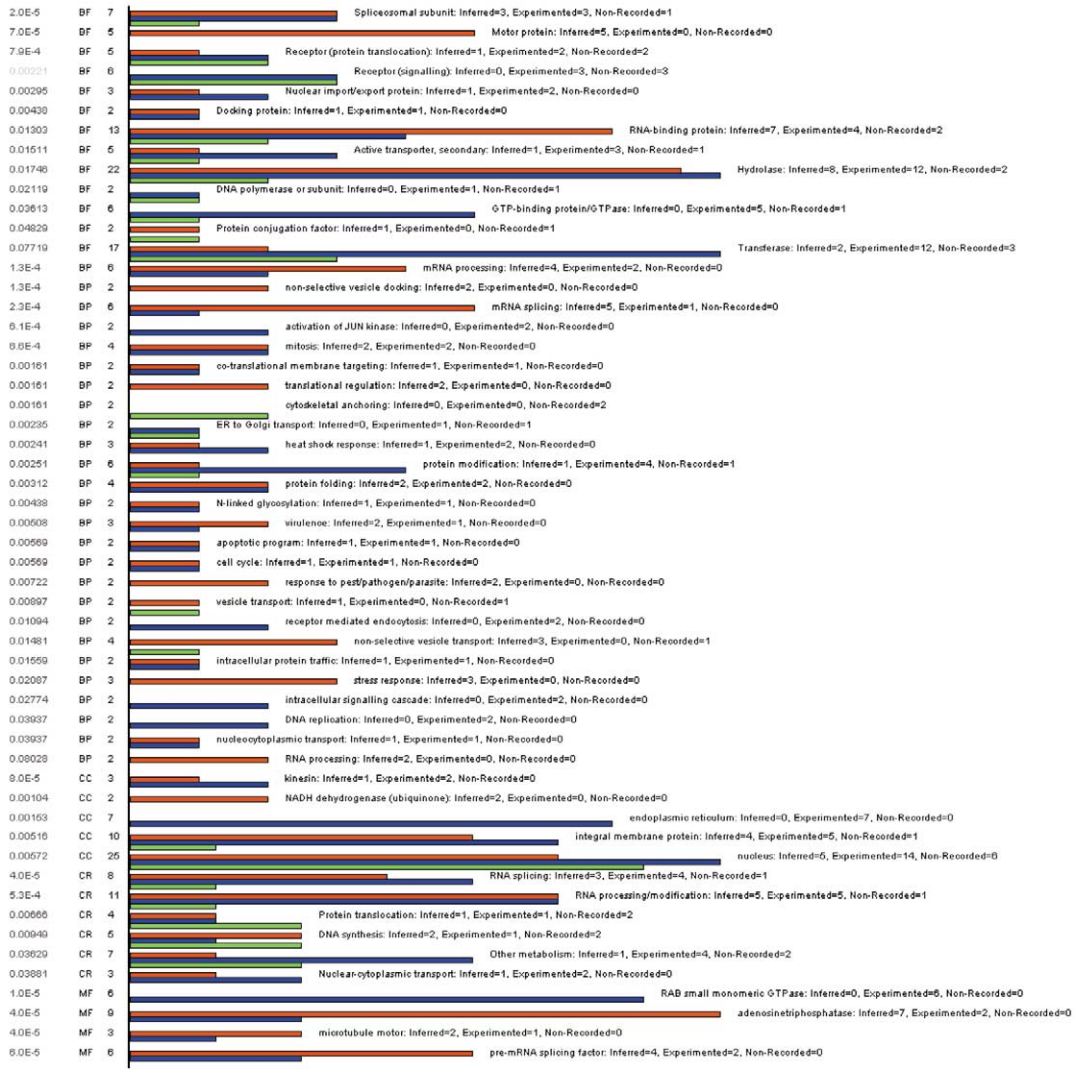

Fig. 3. Functional categories significantly ($p < 0.05$) stimulated by BRCA1 overexpression in breast cancer [24]: BF, biochemical function; BP, biological process; CC, cellular component; CR, cellular role; MF, molecular function. Red bar graphs represent genes for which the function was inferred, blue graphs represent genes for which the function was proved experimentally, and green graphs represent genes for which this type of information was not recorded in the source database.

The second dataset used to validate our methods focused around the link between BRCA1 mutations and tumor suppression in breast cancer. The expression of 373 genes was found to be significantly and consistently altered by BRCA1 induction [24]. We submitted this set to OE using the genes represented on the HuGeneFL microarray (aka HU6800; Affymetrix, Santa Clara, CA, USA) as the reference set. This array contains approximately 6800 human ESTs. We divided the genes into up-regulated and down-regulated. The functional categories significantly represented in the set of up-regulated genes are stimulated by BRCA1 overexpression (Fig. 3). Functional categories significantly represented in the set of down-regulated genes are inhibited by BRCA1 overexpression (Fig. 4). Once again, our approach was validated by the fact that the biological processes found to be significantly affected included several processes known to be associated with cancer: mitosis, cell cycle control, and the control of the apoptotic program. This analysis also showed that BRCA1 had somewhat of a homeostatic effect on the cells, promoting many cell survival and maintenance pathways (e.g., mRNA processing, splicing, protein modification, and folding). BRCA1 is known to be involved in cell cycle checkpoint control (i.e., acting as a tumor suppressor [25]) and significantly down-regulates several genes that normally promote transition through the cell cycle, including CDC2, CDC25B, and the c-Ha-ras1 proto-oncogene.

Methods

Several different statistical approaches can be used to calculate a p value for each functional category F. Let us consider there are N genes on the microarray used. Any

Fig. 4. Functional categories significantly ($p < 0.05$) inhibited by BRCA1 overexpression in breast cancer [24]: BF, biochemical function; BP, biological process; CC, cellular component; CR, cellular role; MF, molecular function. Red bar graphs represent genes for which the function was inferred, blue graphs represent genes for which the function was proved experimentally, and green graphs represent genes for which this type of information was not recorded in the source database.

given gene is either in category F or not. In other words, the N genes are of two categories: F and non-F (NF). The researcher uses his or her choice of data analysis methods to select which genes are regulated in his or her experiments. Let us assume that they picked a subset of K genes. We observe that x of these K genes are F and we want to find out what is the probability of this happening by chance. So, our question is: given N genes of which M are F and $N - M$ are NF_1 we pick randomly K genes and we ask what is the probability of having exactly x genes of type F. Once we pick a gene from the array, we cannot pick it again so this is clearly sampling without replacement.

The probability that a certain category occurs x times just by chance in the list of differentially regulated genes is appropriately modeled by a hypergeometric distribution with parameters (N, M, K) [26]:

$$P(X = x|N, M, K) = \frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}} \tag{1}$$

Based on this, the p value of having x genes or fewer in F can be calculated by summing the probabilities of a random list of K genes having 1, 2, . . . , x genes of category F [27,26]:

$$p = \sum_{i=0}^{x} \frac{\binom{M}{i}\binom{N-M}{K-i}}{\binom{N}{K}} \tag{2}$$

This corresponds to a one-sided test in which small p values correspond to underrepresented categories. The p value for overrepresented categories can be calculated as

$$p = 1 - \sum_{i=0}^{x} \frac{\binom{M}{i}\binom{N-M}{K-i}}{\binom{N}{K}} \tag{3}$$

when the sum is larger than 0.5.

The hypergeometric distribution is rather difficult to calculate when arrays such as Affymetrix HGU133A (22,283 genes) are assayed. However, it is well known that the hypergeometric tends to the binomial when N is large. If a binomial is used, the probability of having x genes in F in a set of K randomly picked genes is given by the classical formula of the binomial probability in which the probability of extracting a gene from F is estimated by the ratio of F genes present on the chip M/N,

$$P(X - x|K, M/N) = \binom{K}{x}\left(\frac{M}{N}\right)^{x}\left(1 - \frac{M}{N}\right)^{K-x}, \tag{4}$$

and the p-value can be calculated as

$$p = 1 - \sum_{i=0}^{x-1} \binom{K}{i}\left(\frac{M}{N}\right)^{i}\left(1 - \frac{M}{N}\right)^{K-i} \tag{5}$$

Alternative approaches include a χ^2 test for equality of proportions [28] and Fisher's Exact test [29]. For the purpose of applying these tests, the data can be organized as shown in Table 2. The dot notation for an index is used to represent the summation on that index. In this notation, the number of genes on the chip is $N = N_{.1}$, the number of genes in functional category F is $M = n_{11}$, the number of genes selected as differentially regulated is $K = N_{.2}$, and the number of differentially regulated genes in F is $x = n_{12}$. Using this notation, the χ^2 test involves calculating the value of the χ^2 statistic,

Table 2

	Genes on chip	Differentially regulated genes	
Having function F	n_{11}	n_{12}	$N_{1.} = \Sigma_{j=1}^{2} n_{1j}$
Not having F	n_{21}	n_{22}	$N_{2.} = \Sigma_{j=1}^{2} n_{2j}$
	$N_{.1} = \Sigma_{i=1}^{2} n_{i1}$	$N_{.2} = \Sigma_{i=1}^{2} n_{i2}$	$N_{..} = \Sigma_{i,j} n_{ij}$

Note. The significance of a particular functional category F can be calculated using a 2×2 contingency table and a χ^2 or Fisher's Exact test for equality of proportions. The N genes on a chip can be divided into genes that are involved in the functional category of interest F ($n_{11} = M$) and genes that are not involved in F (n_{21}). The K genes found to be differentially regulated can also be divided into genes involved ($n_{21} = x$) and not involved (n_{22}) in F.

$$\chi^2 = \frac{N_{..}\left(|n_{11}n_{22} - n_{12}n_{21}| - \frac{N_{..}}{2}\right)^2}{N_{1.}N_{2.}N_{.1}N_{.2}} \tag{6}$$

where $N_{..}/2$ in the numerator is a continuity correction term that can be omitted for large samples [30]. The value thus calculated can be compared with critical values obtained from a χ^2 distribution with $df = (2 - 1) \cdot (2 - 1) = 1$.

However, the χ^2 test for equality of proportion cannot be used for small samples. The rule of thumb is that all expected frequencies $E_{ij} = (N_{i.}N_{.j}/N_{..})$ should be greater than or equal to 5 for the test to provide valid conclusions. If this is not the case, Fisher's Exact test can be used [28,31,32]. Fisher's Exact test considers the row and column totals $N_{1.}$, $N_{2.}$, $N_{.1}$, $N_{.2}$ fixed and uses the hypergeometric distribution to calculate the probability of observing each individual table combination as follows:

$$P = \frac{N_{1.}! \cdot N_{2.}! \cdot N_{.1}! \cdot N_{.2}!}{N_{..}! \cdot n_{11}! \cdot n_{12}! \cdot n_{21}! \cdot n_{22}!} \tag{7}$$

Using this formula, one can calculate a table containing all the possible combinations of $n_{11}n_{12}n_{21}n_{22}$. The p value corresponding to a particular occurrence is calculated as the sum of all probabilities in the table lower than the observed probability corresponding to the observed combination [29].

Related work by Audic and Claverie on EST data [33] used a Poisson distribution and a Bayesian approach to calculate the probability of observing a given number of tags. This can also be used as an alternative to our proposed approach using the combination χ^2–binomial–Fisher's Exact test. However, extensive simulations performed by Man et al. showed that the χ^2 test has the best power and robustness [29]. Accordingly, Audic and Claverie's test was not implemented.

OE provides implementations of the χ^2 test, Fisher's Exact test, and the binomial test. For a typical microarray experiment when the number of genes on the chip $N \simeq 10{,}000$ and the number of selected genes is $K \simeq 100 = 1\%N$, the binomial approximates well the hypergeometric, and therefore, the hypergeometric was not implemented. Fisher's Exact test is required when the sample size is small and the χ^2 test cannot be used. The user can select between the binomial and the χ^2 test. If χ^2 is chosen, the program automatically calculates the expected values and uses Fisher's Exact test when χ^2 becomes unreliable (expected values less than 5).

The exact biological meaning of the calculated p values depends on the list of genes submitted as input. For example, if the list contains genes that are up-regulated and mitosis appears more often than expected, the conclusion might be that the condition under study stimulates mitosis (or more generally, cell proliferation) in a significant way. If the list contains genes that are down-regulated and mitosis appears more often than expected, exactly as before, the conclusion might be that the condition significantly inhibits mitosis. This can be aided by calculating the enrichment of various clusters in hierarchical cluster analysis [34].

In many experiments there are several probe sets that are flagged as absent on the array. A researcher may wish to exclude such genes from the analysis. If a reference chip is selected, all genes on the chip will be included in the analysis. If the user desires to exclude such genes from analysis, they can prepare a list of present genes and use it as the reference. This will adjust the probabilities correspondingly.

A correction for multiple experiments may be useful since repeated tests are conducted to determine the significance of a given GO term. We are currently working on implementations of several corrections for multiple experiments including Holm, Bonferroni, and bootstrapping.

Conclusions

In contrast to the approach of looking for key genes of known specific pathways or mechanisms, global functional profiling can reveal novel biological mechanisms. We presented a method for constructing profiles based on public data and GO categories and terms. The method also provides information about the statistical significance of each of the pathways and categories used in the profiles. We validated our approach by analyzing two public cancer datasets. In both cases, the biological processes reported as significantly impacted included well-known cancer-related processes, thus confirming the validity of the technique. Furthermore, our analysis revealed several novel insights into the mechanisms of breast cancer. Onto-Express is available online at http://vortex.cs.wayne.edu/Projects.html.

References

[1] J. Lacey Jr., S. Devesa, L. Brinton, Recent trends in breast cancer incidence and mortality, Environ. Mol. Mutag. 39 (23) (2002) 82–88.
[2] American Cancer Society: Facts and figures, Tech. rep., American Cancer Society, Atlanta, GA, 2002.
[3] L. Latinovic, G. Heinze, P. Birner, H. Samonigg, H. Hausmaninger, E. Kubista, W. Kwasny, M. Gnant, R. Jakesz, G. Oberhuber, Prog-

nostic relevance of three histological grading methods in breast cancer, Int. J. Oncol. 19 (6) (2001) 1271–1277.
[4] S. Frkovic-Grazio, M. Bracko, Long term prognostic value of Nottingham histological grade and its components in early (pt1n0m0) breast carcinoma, J. Clin. Pathol. 55 (2) (2002) 86–87.
[5] A. Makris, D. Allred, T. Powles, M. Dowsett, I. Fernando, P. Trott, A. Ashley, M. Ormerod, J. Titley, C. Osborne, Cytological evaluation of biological prognostic markers from primary breast carcinomas, Breast Cancer Res. Treat. 44 (1) (1997) 65–74.
[6] P. Birner, G. Oberhuber, J. Stani, C. Reithofer, H. Samoniggand, H. Hausmaninger, E. Kubista, W. Kwasny, D. Kandioler-Eckersberger, R. Jakesz, Evaluation of the United States Food and Drug Administration-approved scoring and test system of her-2 protein expression in breast cancer, Clin. Cancer Res. 7 (6) (2001) 1669–1675.
[7] K. Nathanson, Breast cancer genetics: what we know and what we need, Nat. Med. 7 (5) (2001) 552–556.
[8] C. Perou, T. Sørlie, M. Eisen, M. van de Rijn, S. Jeffrey, C. Rees, J. Pollack, D. Ross, H. Johnsen, et al, Molecular portraits of human breast tumours, Nature 406 (6797) (2000) 747–752.
[9] P. Lonning, T. Sørlie, C. Perou, P. Brown, D. Botstein, A. Borresen-Dale, Microarrays in primary breast cancer-lessons from chemotherapy studies, Endocr. Relat. Cancers 8 (3) (2001) 259–263.
[10] T. Sørlie, C. M. Perou, R. Tibshirani, T. Aasf, S. Geislerg, H. Johnsenb, T. Hastiee, M.B. Eisen, M. van de Rijni, et al., Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications, Proc. Natl. Acad. Sci. USA 98 (19) (2001) 10869–10874.
[11] D. Sgroi, S. Teng, G. Robinson, R. LeVangie, J. Hudson, A. Elkahloun, In vivo gene expression profile analysis of human breast cancer progression, Cancer Res. 59 (22) (1999) 5656–5661.
[12] C. Perou, S. Jeffrey, M. van der Rijni, C. Rees, M. Eisen, D. Ross, A. Pergamenschikov, C. Williams, S. Zhu, et al., Distinctive gene expression patterns in human mammary epithelial cells and breast cancers, Proc. Natl. Acad. Sci. USA 96 (16) (1999) 9212–9217.
[13] S. Drăghici, Statistical intelligence: effective analysis of high-density microarray data, Drug Discovery Today 7 (11) (2002) S55–S63.
[14] P. Khatri, S. Drăghici, G.C. Ostermeier, S.A. Krawetz, Profiling gene expression utilizing Onto-Express, Genomics 79 (2) (2002) 266–270.
[15] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, et al., Gene ontology: tool for the unification of biology, Nat. Genet. 25 (2000) 25–29.
[16] Proteome, Proteome BioKnowledge Library, Tech. rep., Incyte Genomics http://www.incyte.com/sequence/proteome (2002).
[17] G.C. Ostermeier, D. Dix, D. Miller, P. Khatri, S.A. Krawetz, Spermatozoal RNA profiles of normal fertile men, Lancet 360 (9335) (2002) 773–777.
[18] L.J. van't Veer, H. Dai, M.J. van de Vijver, Y.D. He, et al., Gene expression profiling predicts clinical outcome of breast cancer, Nature 415 (2002) 530–536.
[19] M. Meyerson, G.H. Enders, C.L. Wu, L.K. Su, C. Gorka, C. Nelson, E. Harlow, L.H. Tsai, A family of human cdc2-related protein kinases, EMBO J. 11 (8) (1992) 2909–2917.
[20] M. Kimura, S. Kotani, T. Hattori, N. Sumi, T. Yoshioka, K. Todokoro, Y. Okano, Cell cycle-dependent expression and spindle pole localization of a novel human protein kinase, J. Biol. Chem. 272 (21) (1997) 13766–13771.
[21] A. Giodini, M. Kallio, N. Wall, G. Gorbsky, S. Tognin, P. Marchisio, M. Symons, D. Altieri, Regulation of microtubule stability and mitotic progression by survivin, Cancer Res. 62 (9) (2002) 2462–2467.
[22] F. Li, G. Ambrosini, E. Chu, J. Plescia, S. Tognin, P. Marchisio, D. Altieri, Control of apoptosis and mitotic spindle checkpoint by survivin, Nature 396 (6711) (1998) 580–584.
[23] J. Boyd, Adenovirus E1B 19 kDa and Bcl-2 proteins interact with a common set of cellular proteins, Cell 79 (2) (1994) 341–351.
[24] P.L. Welcsh, M.K. Lee, R. M. Gonzalez-Hernandez, D.J. Black, M. Mahadevappa, E.M. Swisher, J.A. Warrington, M.-C. King, BRCA1 transcriptionally regulates genes involved in breast tumorigenesis, Proc. Natl. Acad. Sci. USA 99 (11) (2002) 7560–7565.
[25] A.R. Venkitaraman, Cancer susceptibility and the functions of BRCA1 and BRCA2, Cell 108 (2) (2002) 171–182.
[26] S. Tavazoie, J.D. Hughes, M.J. Campbell, R.J. Cho, G.M. Church, Systematic determination of genetic network architecture, Nat. Genet. 22 (1999) 281–285.
[27] G. Casella, Statistical inference, Duxbury, 2002.
[28] L.D., Fisher, G. van Belle, Biostatistics: a methodology for health sciences, John Wiley and Sons, New York, 1993.
[29] M.Z. Man, Z. Wang, Y. Wang, POWER SAGE: comparing statistical tests for SAGE experiments, Bioinformatics 16 (11) (2000) 953–959.
[30] T. Glover, K. Mitchell, An introduction to biostatistics, McGraw-Hill, New York, 2002.
[31] J.W. Kennedy, G.W. Kaiser, L.D. Fisher, J.K. Fritz, W. Myers, J. Mudd, T. Ryan, Clinical and angiographic predictors of operative mortality from the collaborative study in coronary artery surgery (CASS), Circulation 63 (4) (1981) 793–802.
[32] M.E. Stokes, C.S. Davis, G.G. Koch, Categorical Data Analysis Usingthe SAS System, SAS Institute, Carry, NC, 2002.
[33] S. Audic, J.-M. Claverie, The significance of digital gene expression profiles, Genome Res. 10 (7) (1997) 986–995.
[34] R. Cho, M. Huang, M. Campbell, H. Dong, et al., Transcriptional regulation and function during the human cell cycle, Nat. Genet. 27 (2001) 48–54.