MICROARRAY DATA ANALYSIS: ACQUIRING A SYSTEMIC VIEW IN BIOLOGY

Alessandro Giuliani
Environment and Health Department, Istituto Superiore di Sanità, Viale Regina Elena 29900161 Roma, Italy

Keywords: Omics Sciences, Systems Biology, Multidimensional Statistics, Dynamics, Attractors, Reductionism vs. Holism.

Contents

1. Units and Variables: The Basic Nature of the Problem
2. The Pessimistic Way (The Curse of Dimensionality)
3. The Optimistic Way (The Blessing of Dimensionality)
4. Conclusion: Where We Go From Here
Glossary
Bibliography
Biographical Sketch

Summary

The last two decades witnessed a wide diffusion of the so-called high-throughput techniques in all the fields of biological research. Genome-wide expression platforms (microarray) are the by far most common high-throughput technologies and it became a standard in the biomedical research. Passing from the analysis of single gene expression levels, like in traditional molecular genetics, to the simultaneous analysis of more than twenty-thousands is provoking a complete re-shaping of the perspective we look at biology forcing scientists to acquire a systemic view. The description of different statistical analysis perspectives on microarray data is a privileged observatory for getting the sense of this change. The evolution leading from the efforts to get rid of the 'highly pathological' situation (in terms of classical statistical methods) of having much more statistical units than random variables remaining in the realm of statistical orthodoxy to the consideration of the high dimensionality of gene expression data as a blessing instead of a curse and the consequent development of a 'statistical-mechanics' like approach to biology is the theme of this work.

1. Units and Variables: The Basic Nature of the Problem

A DNA microarray consists in an arrayed series of thousands of microscopic spots of DNA oligonucleotides, called features, each containing picomoles of a specific DNA sequence, known as *probes* (or *reporters*). This can be a short section of a gene or other DNA element that are used to hybridize a cDNA or cRNA sample (called *target*) under high-stringency conditions. A cell culture is grown over the microarray and, thanks to the Watson and Crick base pairing; each RNA molecule (cRNA) present in the culture binds to its correspondent probe on the array. Probe-target hybridization is usually detected and quantified by detection of fluorophore-, or chemiluminescence-labeled targets to determine relative abundance of the specific RNA molecule in the target. The

amount of fluorescence measured at each specific spot is thus a proxy of the amount of expression of the correspondent gene in the culture. Since an array can contain tens of thousands of probes, a microarray experiment can accomplish many tests in parallel.

The aim of a typical microarray experiment is to give a characterization of the differences between k different biological samples (cell cultures, tissue specimens..) in terms of n gene expression profiles as measured by microarray technology. The above question suggests a natural statistical formalization in terms of biological samples as statistical units and single gene expression levels as random variables. An eventual hypothesis testing or discrimination task can be guided by the pertaining of biological samples to different classes (e.g. control and drug-treated samples, normal and pathological samples, different tissues samples….), this task is accomplished by the detection of statistically significant differences between the groups in the $n-$ dimensional space spanned by the n microarray probes. A typical microarray experiment involves a value of k going from 10 to 100, while n is in the order of 10000-30000. This poses a hard problem to classical inferential statistics that is based on the reverse situation of a much higher number of statistical units with respect to variables. To grasp the relevance of this point is important to go back to the birth of modern inferential statistics.

Student, from where the world-wide famous Student's t distribution derives its name, is the pseudonym of the famous English statistician William Gosset (1867-1932) that in 1899 joined the Dublin Guinness brewery as responsible of the quality control. In 1906 he had to solve the task of finding the optimal strategy for checking the degree of invariance of beer composition along the production process. Keeping the beer composition inside a relatively strict interval was essential for the maintenance of a good and recognizable taste for Guinness production, the check had to be based on the selection of k bottles for each lot and their chemical analysis. The statistical problem can be formulated as the choice of an acceptance threshold neither too strict so not to stop the production too often so causing relevant problems to the production line (with consequent money losses) nor too relaxed so not to put on the market a low quality product with bad consequences for Guinness brand (and again with relevant money losses).

This is a classical operational research problem in which there is a trade-off between two competing optimization tasks and the best solution consists in finding a global minimum (maximum) of a parameter influenced in opposite directions by the two tasks. Gosset solved the problem by empirically building a control chart in which he registered how many times a given observed malt concentration in the k bottles sample went together with a real fault somewhere along beer production line that asked for a corrective intervention (real positive cases) and how many times the same value was correspondent to a false alarm (false positive cases). In the first case stopping the production and checking for the presence of a fault along the process was the right choice, in the second case the correct (most convenient) choice was to let things go as usual. Do not stop the production when in presence of a real positive result is what we call '*Error of the I type'*, stopping the production when unnecessary is the '*Error of the II type'*, the two errors are clearly strictly related and correspond to the two competing tasks. Gosset discovered that the best compromise between these two different error

sources (weighted for their respective financial relevance) was to set a threshold of $p = 0.05$, i.e. to accept to be wrong (in the I type way) in the 5% of cases, thus taking as threshold limit for stopping the production a displacement from the optimal malt concentration that (in the previous experimentation) was discovered to correspond to a false positive in the 5% of cases. This very elegant result (Gossett used the Student's pseudonym in the publication so to avoid problems with Guinness brewery for disseminating industrial secrets) was taken as such by experimental science that denominated H_0 (null hypothesis) the absence of a relevant effect (the production has no problem, in beer terms) and consequently the observed experimental result is due to chance, and H_1 (alternative hypothesis) the presence of a real cause at the basis of observed results (the production line presents a specific malfunctioning, in beer terms) respectively. Consequently the scoring of a $p < 0.05$ was considered (more or less a-critically) as the signature of a 'statistically significant' result and thus as the reach of an acceptable factual basis for a given scientific model of explanation [8]. This line of thought was adapted to other statistical tests and procedures (Analysis of Variance, Regression models, Non-parametric tests..) remaining substantially invariant in its basic philosophy, so we must not to be surprised by the presence of paradoxical effects deriving from the application of the Student result very far from its place of birth. In the case of microarrays the paradox comes from the fact that, at odds with the original Gosset's formulation, we are not checking for the statistical significance of only one (or very few) variables but of thousands of them. To accept a type I error (or False Discovery Rate, FDR in the microarray jargon) the 5% of times is very reasonable as for the between groups difference for the amount of transcription of a single gene, but let's imagine to repeat this operation 20000 times (the number of genes normally present in a microarray chips), it is immediate to understand how the 20000 times repetition of an operation that **each time** has a 5% probability of error implies a huge number of false positives (high FDR), with around 1000 genes expected to be 'significant' for the pure effect of chance [1]. In microarray experiments we are very far outside the realm of a rational use of the elegant Gosset's solution, and we need to seriously face the problem of test multiplicity. In microarray literature there are many proposals for facing the test multiplicity problem, these efforts coarsely pertain to two main lines of thought we can define as the 'pessimistic' and the 'optimistic' ways. In the pessimistic approach high dimensionality is considered as a curse to be bravely faced by limiting the connected risks of false positives by a smart use of statistical tools. In the optimistic way, on the contrary, high dimensionality is considered as a blessing for the possibility offered to turn upside-down the usual approach of molecular genetics and thus strating to explore an organizational analysis more general (and much more promising in terms of realism) than the single genes [2,3].

2. The Pessimistic Way (The Curse of Dimensionality)

The most conservative reaction to the challenge high dimensionality poses to the common way of reasoning in biomedical sciences (look for the single genes affected and build an explanation on them) is to concentrate only on a small sub-set of the information present in the microarrays. This can be done by adopting a 'hypothesis driven' approach (low dimensional arrays made of only few genes selected on the basis of a given mechanism of action) or by looking for a statistically sound strategy for

setting to a minimum the FDR. The first approach needs no comment: it is coincident with the simple coming back to the 'old mode' , this is a completely acceptable choice but it is not pertinent here. Historically, the first' pessimistic' way to face the curse of dimensionality was to skip any concern about inferential statistics and using the simple Fold Change (FC) as mark of a relevant result: the genes relevant for the phenomenon under study are those that present at least a doubling with respect to their baseline activity ($FC = 2$) or a triplicate (FC = 2) or a ten-fold increase ($FC = 10$). This shortcut in the beginning seemed very effective even for its extreme simplicity, but soon appeared that the genes that were 'significant' in an experiment were completely different from the genes significant in a replica. In other words the scientists recognized they made the very bad affair of exchanging a well known error probability (classical statistics) for an unknown one [1].

The next move of this frontal attack to high dimensionality, was the application of 'Bonferroni-like' strategies following a classical method to deal with multiple tests developed in the thirties by the Italian statistician Carlo Emilio Bonferroni (1892-1960) in the realm of insurances. According to Bonferroni, the significance level used as threshold in a test must be divided by the number of independent tests we perform on the same data set. Consequently, if we measure the activity of 10 genes, the 'corrected p' to replace the usual 0.05 threshold, should be $p(\text{Bonferroni}) = 0.05/10 = 0.005$, while for 100 genes the threshold becomes $p(\text{Bonferroni}) = 0.05/100 = 0.0005$ and so forth.

While this procedure clearly limits the number of false positives, it is very depressing for biologists because only very few outlier activities survive the correction, so eliminating a lot of potentially relevant information. Moreover, at the basis of the Bonferroni strategy is the assumption each gene 'plays its game' independently from the others, that is clearly absurd in biology. The lesson emerging from these first efforts, is the need to consider statistics no more as a standardized recipe to be applied to the data, in the most rigorous as possible way, but a as a way to sketch a quantitative picture of the studied system. This implies the choice of a statistical tool cannot be considered as neutral but, to be effective, must reflect some relevant intrinsic features of the biological system at hand.

From this point onward, the response to the dimensionality challenge followed a less 'automatic' path and tried to pass around the obstacle instead of facing it directly or, worse, completely refusing its existence. Still in the realm of the pessimistic way are located the so called GCT (gene-class testing) strategies. The philosophy of a GCT strategy is the explicit use of *a priori* knowledge about the presence of functional classes made of genes having a similar physiological role. In GCT, the first step of collection of all the genes 'significantly' related to the particular discrimination task into a list is followed by the analysis of such a list in terms of 'differential enrichment' of genes pertaining to certain functional classes with respect to the whole genome list. Thus if the 'significant genes' display a much higher relative frequency of 'genes involved in immune response' with respect to the whole genome, independently of the reliability of the observed significance of a given gene, we can safely affirm that the studied phenomenon involves 'the immune response' as such. The statistical

significance of the functional distribution of the genes in the list with respect to a chance assortment is a much better defined statistical problem with respect to the statistical significance of each single gene as for the discrimination task, and can be solved by a straightforward chi-square like approach. Gene functional classes are based on the so called GO (Gene Ontologies): complex hierarchical classification of genes going from very general (immunity, metabolism, DNA repair, membrane structure..) to very detailed functional definitions (Natural Killer activity, pro-apoptotic, MAP Kinase pathway..). At this point the subject of the analysis is no more the single gene but the entire list that is checked for the tenability of its characterization as a specific non-random choice of functions. The GST approach has a much more realistic attitude in biological terms (genes are considered as *proxyes* of global functions and not *per se*) but nevertheless remains very unpractical and affected by many ambiguities for the polysemic character of biomedical information: as a matter of fact the same gene can be allocated in many different classes of biological activity. Moreover we could think of completely independent classifications based, let's say on cellular location of the correspondent protein or its sequence homology with other products.

Thus the generic question 'gene X is more similar to gene Y or gene Z?' does not allow for any context independent answer, being strictly dependent on the organization level we are considering or the features we think are more important for the particular case. Thus, while looking for a 'biological meaning' in addition to the brute *a posteriori* statistical significance, so giving rise to a kind of Bayesian approach, while in principle is surely correct can be very difficult to attain and open to a lot of alternative interpretations. The Bayesian character of statistics application to biology is a very important point to discuss: in the case of GST a given *a posteriori* statistical result (in the form of a list of genes significantly different for the discrimination task) is accepted as 'biologically meaningful' if (and only if) it can be interpreted in the light of accepted biological theory. That is to say our results (*a posteriori* knowledge) are asked to 'make sense' in terms of a specific and recognizable suite of already known biological functional modules (*a priori* knowledge).

The work of Rev. Thomas Bayes (1702-1761) is still more evident in the so called network-based analysis of microarray data that acquired a great popularity following the development of Systems Biology studies [4].

Figure1 reports the accepted regulation frame for apoptosis, the so called programmed cell death.

In this website are collected thousands of such regulation pathways described by means of the box-and-arrow formalism, each box represents a cellular constituent (gene, protein, organic metabolite), each arrow a regulation link (is increased/decreased by, binds to, is modulated by, interacts with…) between two nodes of the network.

The network formalization in which each regulation module is presented as an integrated system of mutual regulation among the elementary constituents is very common in biology given it allows for a convenient and vivid display of very intermingled pathways. What is new of systems biology approaches is the effort of deriving quantitative general consequences from the wiring structure of such networks

by both the application of topological graph theory derived descriptors and novel application of some intriguing qualitative dynamics results coming from systems analysis and automatic control fields.

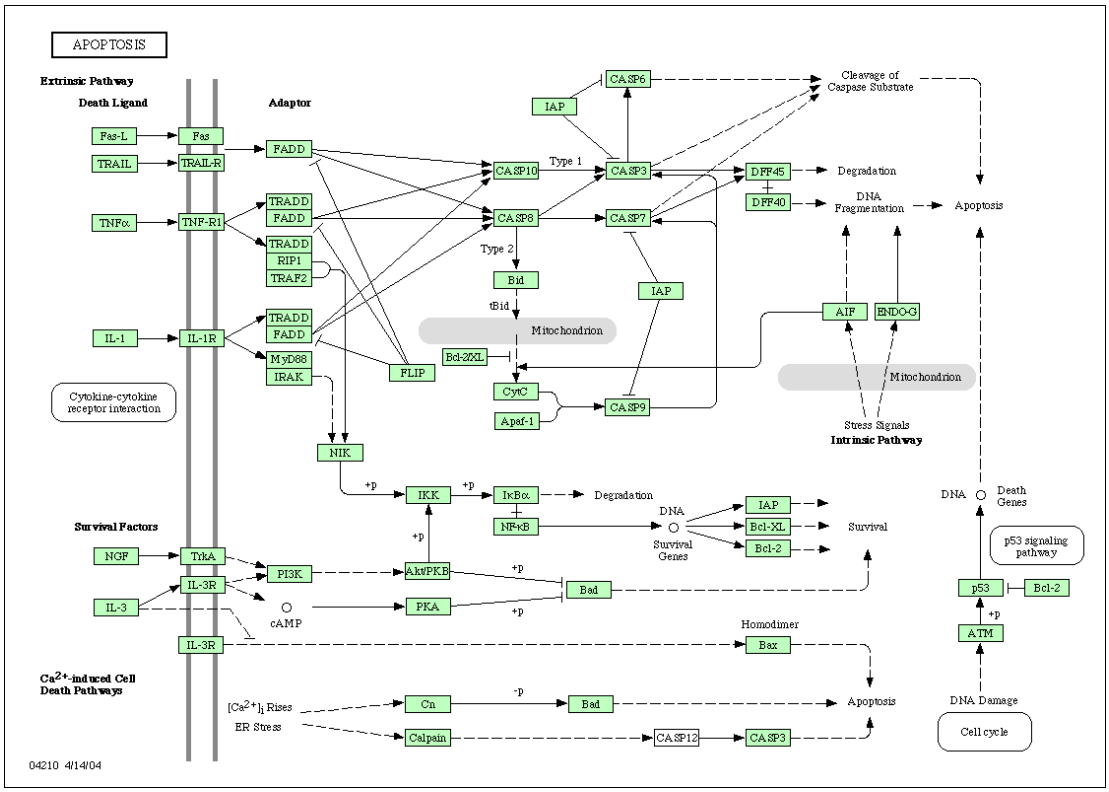

Figure 1. The reported cartoon comes from KEGG (Kyoto Encyclopedia of Genes and Genomes) freely available at the website: http://www.genome.ad.jp/kegg/.

-

-

Bibliography

1. Allison DB, Cui X., Page GP, Sabripur M. (2006) Microarray data analysis: from disarray to consolidation and consensus. *Nature Rev. Genet*. 7(1): 55-67.

2. Alter O., Brown PO, Botstein D. (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc.Natl.Acad. Sci. USA* 97(18): 10101-10106.

3. Alter O. (2006) Discovery of principles of nature from mathematical modeling of DNA microarray data. *Proc.Natl.Acad. Sci. USA* 103 (44): 16063-16064.

4. Bar-Yam Y., Harmon D., de Bivort B (2009) Attractors and democratic dynamics. *Science* (323): 1016-1017.

5. Benigni R. and Giuliani A. (1994) Quantitative modeling and biology: The multivariate approach. *American Journ. of Physiol.* (266) (35), R1697-R1704

6. Brock A, Chang H, Huang S (2009) Non-genetic heterogeneity- a mutation-independent driving force for the somatic evolution of tumours. *Nat. Rev. Genet*. (10): 336-342.

7. Chang H.H., Hemberg M., Barahona M., Ingber D.E., Huang S. (2008) Transcriptome wide noise controls lineage choice in mammalian progenitor cells. *Nature*; 453:544–7.

8. Dale A. (1999) A History of Inverse Probability: from Thomas Bayes to Karl Pearson. Springer, Berlin.

9. Gallavotti G. (1999) Statistical Mechanics: A Short Treatise, Springer, Berlin.

10. Huang S, Eichler G, Bar-Yam Y, Ingber DE (2005) Cell fates as high-dimensional attractor states of a complex gene regulatory network. *Phys.Rev.Lett.* (94): 128701-1-6.

11. Laughlin, R.B., Pines D., Schmalian J., Stojkovic B.P., Wolynes P. (2000) The middle way. *Proc. Natl. Acad. Sci. USA* 97: 32–37.

12. Overington JP, Al Lazikani B, Hopkins AL (2006) How many drug targets are there? *Nat.Rev. Drug. Discovery* 5: 993-996.

13. Roden JC, King BW, Trout D., Mortazavi A., Wold BJ, Hart CE (2006) Mining gene expression data by interpreting principal components. *BMC Bioinformatics* (7): 194.

14. Tsuchyia M., S.T. Wong, Z. X. Yeo, A. Colosimo, M.C. Palumbo, L. Farina, M. Crescenzi, A. Mazzola,

R. Negri, M.M. Bianchi, K. Selvarajoo, M. Tomita and A.Giuliani (2007) Gene expression waves: cell cycle independent collective dynamics in cultured cells *FEBS J.* 274: 2874-2886.

15. Webber C.L., Marwan N., Facchini A., Giuliani A. (2009) Simpler methods do it better: Success of Recurrence Quantification Analysis as a general purpose data analysis tool. *Physics Letters A* **373**: 3753-3756.

Biographical Sketch

Alessandro Giuliani was born in Roma in 1959, in 1981 he graduated in Biology at University 'La Sapienza' of Roma where he further specialized in Statistics. Alessandro Giuliani is involved since more than 25 years in the development of soft mathematical modeling for biology, along this research line he authored more than 170 papers in peer-review journals going from computational physics to chemistry, molecular biology, genetics, medicine and ecology. Since 1997 he is senior scientists at the Istituto Superiore di Sanità (Italian NIH) and collaborates with many different research groups both in Italy and abroad. Dr. Giuliani is currently the Editor –in-chief of Systems and Synthetic Biology (Springer).