Volume 365, Issue 9458 , 5 February 2005-11 February 2005, Pages 454-455

Microarrays and molecular research: noise discovery?

John PA Ioannidis[a, b, c]

[a]Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina 45110, Greece

[b]Biomedical Research Institute, Foundation for Research and Technology–Hellas, Ioannina, Greece

[c]Institute for Clinical Research and Health Policy Studies, Tufts-New England Medical Center, Tufts University School of Medicine, Boston, Massachusetts, USA

Comments

The promise of microarrays has been of apocalyptic dimensions. As put forth by one of their inventors, “all human illness can be studied by microarray analysis, and the ultimate goal of this work is to develop effective treatments or cures for every human disease by 2050”.[1] All diseases are to be redefined, all human suffering reduced to gene-expression profiles. Cancer has been the most common early target of this revolution[2] and publications in the most prestigious journals have heralded the discovery of molecular signatures conferring different outcomes and requiring different treatments. Yet, in today's *Lancet*, Stefan Michiels and colleagues show that, on close scrutiny, in five of the seven largest studies on cancer prognosis, this technology performs no better than flipping a coin. The other two studies barely beat horoscopes. Why such failure?
Give me information on a single gene and 200 patients, half of them dead, please. I bet I can show that this gene affects survival (p<0·05), even if it does not. One can do analyses: counting or ignoring exact follow-up; censoring at different timepoints; excluding specific causes of death; exploiting subgroup analyses; using dozens of different cut-offs to decide what constitutes inappropriate gene expression; and so forth.[3] Without highly specified a-priori hypotheses, there are hundreds of ways to analyse the dullest dataset. Thus, no matter what my discovery eventually is, it should not be taken seriously, unless it can be shown that the same exact mode of analysis gets similar results in a different dataset. Validation becomes even more important when datasets become complex and analytical options increase exponentially. Typically, patients are split into separate training and validation sets. In another common approach, each patient is left out in turn, a model is built, and then checked against the excluded patient.[4]

Validation is still an analysis and can be manipulated as can any analysis. Several variants of inadequate or incomplete validation have been described.[2 and 5] Furthermore, when the same team does both the original analysis and validation thereof, one might consciously or unconsciously select the best-performing pair of training-validation data and analytical mode. Against this licence-to-analyse, one can use always and strictly the same method, generate randomly many training and validation sets, and examine whether results are stable. But then, as Michiels and colleagues show, the selected "important" genes rarely coincide across random replicates. Published estimates often seem excessively optimistic, probably due to serendipitous selection bias either in the analysis mode or in the validation process.

Microarrays produce information of unparalleled wealth. This information is their great, fascinating advantage—and their downfall. Let us suppose for a moment that no gene is important for any disease outcome and that it is all random noise. That scenario is scary: this noise is so data-rich that minimum, subtle, and unconscious manipulation can generate spurious "significant" biological findings that withstand validations by the best scientists, in the best journals. Biomedical science would then be entrenched in some ultramodern middle ages, where tons of noise is accepted as "knowledge". However, hopefully, some biological variables must indeed be important—but how do we suppress surrounding noise? If 30 genes determine the outcome of a specific cancer, we expect upfront that each gene (of 30 000 tested) has a 1:1000 chance on average to be truly important. The same caveat applies not only in gene-related applications, but also in proteomics,[6] and all discovery-oriented molecular research where enormous databases can be rapidly generated from just a handful of patients.[7] With such massive information, usually there cannot be any strong a-priori hypothesis that specific biological factors are more important than others. Any confident claims of "biological plausibility" sit on very slippery ground.[8]

True discovery remains a challenge in the molecular era. Routinely repeated random sampling for multiple validations is useful. Perhaps more importantly, validations should be done by several completely independent teams. I cannot stress "completely" enough here. Some journals, dismayed at the questionable replication of some molecular research,[9] propose that papers should also contain independent replications.[10] Yet do same-team approaches ensure independence? Any intermingling of the process for generating and replicating the hypothesis entails the danger of somehow diluting the independence of the replication.

Sample size is also essential. A recent editorial hailed the advent of "small studies with high density of data".[11] Well, I think there is no free lunch in good research. Microarrays need evidence and this cannot be obtained from a couple of small studies, no matter how high-tech. Small sample sizes might actually hinder the identification of truly important genes.[2] Molecular medicine may eventually fulfil its arrays of promises.[12] However, we should aim for many independent studies with a total of several thousand patients, a hundred-fold more than the current standard. If we truly believe that microarrays and molecular research in general are important, we should not settle for less.

I declare that I have no conflict of interest.

Volume 365, Issue 9472 , 14 May 2005-20 May 2005, Pages 1683-1684

Prediction of cancer outcome with microarrays

Norio Iizuka[a], Yoshihiko Hamamoto[b] and Masaaki Oka[a]

[a]Department of Surgery II, Yamaguchi University School of Medicine, 1-1-1 Minami-Kogushi, Ube, Yamaguchi 755-8505, Japan

[b]Department of Computer Science and Systems Engineering, Faculty of Engineering, Yamaguchi University, Ube, Yamaguchi, Japan

Comments

Stefan Michiels and colleagues[1] showed that the prognostic value of published microarray results should be considered with caution. However, their approach is much less effective in assessing the combination-based predictor developed by us.[2]

As shown in their figure 1,[1] four of our 12 genes[2] barely passed in their artificial samples at a frequency of more than 50%; none of the four genes would pass because the training sample size is larger in their resampling approach. This is not surprising, because even *P53*, a hallmark gene for cancer, could not account for 50% of cancer patients if it is examined in hundreds of samples. Thus, there would be no gene that could account for more than half the cancer patients in a large cohort. It is more important to make an association between a disease and a pathway—not a single gene. As such, Michiels and colleagues' approach, which focuses on expression of a single gene, is incorrect. Which of the many genes involved in pathways that could be commonly present in a large cohort (eg, 1000 samples) is the most suitable for constructing a robust predictor? This was the question we bore in mind when we constructed the 12-gene predictor[2] according to the pattern of a combination of genes. We believe that the selection of genes used in a classifier is more important than the design of a classifier in constructing a robust predictor. In fact, our accuracy rate of 93% is attributable to searching for all combinations of 12 out of 50 through training subsets.

Michiels and colleagues showed that *HLA-DQA1* (*HLA-DQA*) was the most robust feature among 4861 genes in our cohort. Our recent study[3] identified four HLA family genes (*HLA-DRA*, *HLA-DRB1*, *HLA-DQA1*, and *HLA-DG*) associated with cancer recurrence in the same cohort used in our previous work, all of which are expressed coordinately with each other (Pearson's correlation coefficient 0·708–0·888). Thus, HLA family genes and the relevant pathway (ie, the immune system) have an important role in recurrence of hepatocellular carcinoma. In our previous work,[2] an *HLA-DRA*-integrated predictor indeed classified our patients more accurately than an *HLA-DQA1*-integrated predictor. This is a clear example of the fact that the most robust feature gene does not necessarily give the most robust predictor when taking a combination into consideration. In other words, the N best genes are not always the best N genes. Our work,[3] in which the 12 genes[2] used in our predictor are not always located within the top 12 in ranking by the Fisher ratio supports this work. Thus, Michiels and colleagues' approach of selecting genes on the basis of their individual effectiveness cannot yield any optimum gene subsets.

We agree with Michiels and colleagues that one-time gene selection with a training set of small size is insufficient for constructing a robust predictive system. When the number of available training samples is small, the gene subset is strongly affected by the variability of the samples. However, this finding is not a new one—it was proposed in the 1970s and confirmed in the 1990s in the field of pattern recognition (references available from the authors). To resolve this issue and to obtain a reliable gene subset, it is very important to establish the best way to cope with this variability. Of course, we must assess the constructed predictor in an independent large cohort, as proposed by Ntzani and Ioannidis.[4]

Since an unstable sample label can affect gene selection and classifier design, we should reconsider the suitability of sample labels in each cancer type.[5]

We declare that we have no conflict of interest.

Volume 365, Issue 9472 , 14 May 2005-20 May 2005, Page 1683

Predictionofcanceroutcomewithmicroarrays

Elia Biganzoli[a], , Nicola Lama[b], Federico Ambrogi[a], Laura Antolini[a] and Patrizia Boracchi[b]

[a]Unit of Medical Statistics and Biometry, Istituto Nazionale per lo Studio e la Cura dei Tumori, Via Venezian 1, 20133, Milano, Italy

[b]Institute of Medical Statistics and Biometry, Universita degli Studi di Milano, Via Venezian 1, 20133, Milano, Italy

Comments

Stefan Michiels and colleagues (Feb 5, p 488)[1] raise important issues about the relevance of current genomic research for outcome prediction.

The need to assess the real gain provided by genomics has been already pointed out.[2] However, Michiels and colleagues stress careful assessment for random sampling variability in results from high-throughput analyses on small samples. Advocating validation by repeated random sampling is understandable; however, splitting the sample into training and validation sets leads to a loss of information in the classification-building process.

If sample size is limited, the removal of training observations for validation purposes decreases the model's precision. Moreover, if the size of the validation set is too low, performance estimation will be uncertain. In fact, the patterns of the confidence intervals in Michiels and colleagues' figure 2 show a decrease in the estimated misclassification rate, but also in its precision (wider confidence intervals) with the increased training sample size. Bootstrap resampling methods[3] might be better suited to addressing this issue.

Resampling should also be done in the prefiltering of the whole gene set to account for the variability of this process. We adopted this strategy in the reanalysis of van't Veer and colleagues' data,[4] and reported the type of error rate controlled in prefiltering and gene selection.[5] The classifier found considering the whole 78 cases included 119 genes, sharing only 28 (40%) of the 70 genes of the original paper.[4] Throughout our cross-validation procedure, 91% of them were included at least once in the different gene sets.

To assess a model's accuracy, sensitivity and specificity should be reported instead of the misclassification error rate. However, these measures do not directly reflect the clinical relevance of the classifier, which is rather quantified by the negative and positive predictive values.[2]

A key issue, if not the priority, is that of study design. Are we sure that we have so far applied the best one for outcome prediction with microarray data? The unmatched case-control design to follow-up data seems widespread in microarray studies. However, an artificially balanced composition of the study set to allow maximisation of the power of the comparisons between cases (patients whose disease recurred within a reference follow-up time) versus controls (whose disease did not recur) does not reflect the target cohort, thus preventing the possibility of correctly estimating the incidence of class conditional failure. Nevertheless, the misleading naive application of the Kaplan-Meier estimator to such studies is widespread.[2]

Therefore, before aiming “for many independent studies with a total of several thousand patients”, we must first consider better designs for them. This is not a mere sample size issue—the best internal validation procedure will not solve a problem caused by poor design.

Overall, the underlying issue in microarray studies is the lack of standard methods for design, data analysis, and performance assessment according to clinical aims. Achieving these goals requires cooperative efforts between multidisciplinary research networks.

Partly supported by EU Biopattern project FP6-2002-IST-1 N° 508803. We declare that we have no conflict of interest.

Volume 365, Issue 9472 , 14 May 2005-20 May 2005, Page 1685

Prediction of cancer outcome with microarrays

Shea N Gardner[a], and Michael Fernandes[a]

[a]Lawrence Livermore National Laboratory, Livermore, CA 94551, USA

Comments

Stefan Michiels and colleagues[1] list the potential and limitations of microarrays for the prediction of cancer outcome. Cancer is a disease characterised by complex, heterogeneous mechanisms, and studies to define factors that can direct new drug discovery and use should be encouraged. However, this is easier said than done. Casti[2] teaches that a better understanding does not necessarily extrapolate to better prediction, and that useful prediction is possible without complete understanding. To attempt both explanation and prediction in a single non-mathematical construct is a tall order.

At this stage of incomplete knowledge, predictive ability can be enhanced by considering hybrid approaches: computational methods based on known mechanisms of the disease,[3 and 4] together with microarrays. Multiple initiatives directed to the same objective—ie, prediction of cancer outcome—may provide validation. Biomarkers serve to predict, not explain, and the guidance provided by Ransohoff[5] is relevant to outcome validation in cancer. But the first step is to assure a robust method that can allow for replication. A large-scale clinical study with a complementary approach is an appealing possibility.

We declare that we have no conflict of interest.