New experimental techniques in genomics: challenges for data processing and analysis

Burzykowski T

Center for Statistics, Limburgs Universitair Centrum, Diepenbeek, Belgium

Abstract

A lot of current research in biology and medicine is aimed at understanding, what is the role of particular fragments of a living's organism genome from a point of view of different biological process taking place in a cell. All cells in an organism contain the same DNA, but despite that, they actually differ. The differences are due to the fact that, stimulated by cell regulatory mechanisms or environmental factors, fragments of DNA (genes) express their code and provide the instructions when and in what quantity to produce specific proteins. This process is called **gene expression**. Differential gene expression implies differential protein abundance, and thus induces different cell functions. Gene expression level is a measure that aims at a quantitative description of the gene expression.

As specified by the **central dogma of molecular biology** (Crick, 1970) [1], in the first step, called **transcription**, the genetic information carried by a gene is transcribed into mRNA (**messenger RNA**). In the second step, called **translation**, an appropriately modified copy of mRNA migrates to the cytoplasm where it serves as a template for protein synthesis. Because the synthesis of the protein associated with a particular gene involves the transcription of DNA into mRNA, one might assume the abundance of mRNA produced during the transcription as a measure of the gene expression.

Analysis of mRNA-based measures of gene expression levels is one of the best practical solution available at this moment. However, due to a variety of reasons, mRNA and protein levels and their alterations often poorly correlate. To study the functional and biochemical features of specific cell types, one should actually investigate both the gene expression levels and the type and abundance of produced proteins. The latter is the aim of **proteomics**. The investigation of **protein expression levels** may be even more important in situations when gene expression analysis is not feasible (e.g. for the complex protein mixtures present in body fluids such as plasma, synovial and cerebrospinal fluid).

Rapid developments in molecular biology technology have led to the development of various experimental methods making possible investigation of gene- or protein-expression levels. These methods include, e.g. microarrays (cDNA, Schena et al. [2]; or oligonucleotide, Lockhart et al. [3]); Serial Analysis of Gene Expression (SAGE; Velculescu et al. [4]); 2-dimensional electrophoresis with mass spectrometric identification; combined fractional diagonal chromatography (COFRADIC; Gevaert and Vandekerckhove, 2004, [5]).

All these techniques share several common features. For instance:

- they use sophisticated instrumentation;
- they are very sensitive, also to systematic effects due to time, place, reagents, personnel, etc.;
- they yield complex data, in terms of correlation, variability, etc.;
- they generate many (10^2-10^5) measurements per biological sample;
- their reproducibility can easily be compromised.

Because of these features, processing and analyzing data produced by these methods is still a challenging task. More specifically, several problems can be listed:

- the need for preliminary preprocessing, aimed at removing of artifacts, normalization, summarizing signals, assessing quality etc.;
- the data require novel methods of analysis, as they do not fit into the classical framework where the number of observations (samples) is greater than the number of variables (measurements);
- taking into account the complexity of data requires advanced methods of analysis;
- the large amount of measurements creates a need for tools allowing automated analyses; on the other hand, it results in computational problems when advanced techniques are used.

ADDRESS FOR CORRESPONDENCE:
Tomasz Burzykowski
Center for Statistics, Limburgs Universitair Centrum,
3590 Diepenbeek, Belgium
e-mail: tomasz.burzykowski@luc.ac.be

In general, one can note that the new experimental technologies are developed at a quicker pace than the methods that can address the problems mentioned above. As a result, sometimes the basic issues regarding repeatability of measurements, susceptibility to systematic effects, generalizability of findings etc., are not well understood and/or resolved before the technology is put into practice. Obviously, this can lead to potentially serious problems.

Potential promises and pitfalls related to the use of the novel genomic technologies are very well illustrated by the following example related to the use of protein mass spectra to discriminate between cancerous and normal samples.

In brief, in surface-enhanced laser desorption and ionization time of flight (SELDI-TOF) mass spectrometry a biological sample (such as serum) is apllied to a precoated stainless steel slide, which binds preferentially a particular class of proteins based on their physiochemical properties. The sample is further mixed with an energy absorbing matrix, which causes the entire mixture to crystallize as it dries. The crystal is put into a vacuum chamber and is hit with a laser, what produces ionized protein molecules in the gas phase. A brief electric field is then applied to accelerate the ions down a flight tube, and a detector at the end of the tube records the time of flight, from which the mass-to-charge ratio (m/z value) of the protein can be derived. A typical spectrum consists of the sequentially recorded numbers of ions arriving at the detector (the intensity) coupled with the corresponding m/z value. Peaks (local maxima) in the intensity plot ideally correspond to individual proteins. One can distinguish them from features (the observed intensities at a particular m/z values). A set of spectra will have thousands of features, but only a small fraction of these would correspond to peaks.

Based on a spectrum, one can attempt to build a proteomic pattern, that is, a pattern discriminating between spectra coming from different biological samples. It can be formed by a small key subset of proteins or peptides buried among the entire repertoire of thousands of proteins represented in the sample spectrum. The pattern can be defined by peaks (or features) at key m/z positions.

Petricoin et al. [6] used SELDI-TOF mass spectra to discriminate between ovarian cancer and normal samples. They used samples from 100 ovarian cancer patients, 100 normal controls and 16 pts. with "benign disease" (216 in total). They constructed a proteomic pattern based on 50 cancer and 50 normal spectra ("training set"), and then tested it on the remaining 116 samples ("test set"). As a result, they were able to correctly classified 50 out 50 of the "test" ovarian cancer cases (100% sensitivity) and 63 out of 66 of the "test" non-malignant cases (95% specificity). The estimated values of sensitivity and specificity are impressive and the results deservedly attracted a lot of attention.

In 2004 the same team published results of an additional analysis of the data, using a higher-resolution technique called the hybrid quadrupole time-of-flight (QqTOF) mass spectrometry (Conrads et al. [7]). Using the same biological samples as Petricoin et al. [6], they constructed a proteomic pattern capable of achieving a 100% sensitivity and 100% specificity for identifying cancer from normal.

The paper by Conrads et al. [7] does suggest some issues, though. For instance, the authors acknowledge that their quality assurrance and control (QA/QC) measurements "indicated 32 spectra that were of lesser quality (…). These mass spectra were all generated at the end of the experimental run, suggesting that a deviation in the process had occurred". It appears that these 32 spectra were removed from both the Petricoin et al. [6] and Conrads et al. [7] analyses. More importantly, however, from *Fig. 7* of Conrads et al. [7] one can infer that samples were processed in batches, with normal samples processed first, and control samples next. Thus, a part of the normal samples was processed at the time when the quality of the measurements was deteriorating. This raises a question whether the obtained results are due to confounding of bad quality samples with normal samples.

More insight into the results of Petricoin et al. [6] and Conrads et al. [7] was provided by Baggerly et al. [8]. They reanalyzed the three following datasets:

1. the one described in Petricoin et al. [6] (216 spectra), with spectra obtained using the Ciphergen H4 ProteinChip array;
2. the same 216 samples run on the Ciphergen WCX2 ProteinChip array (corresponding to the paper of Conrads et al. [7]);
3. a new set of 253 spectra (91 normal and 162 cancer samples), run on the WCX2 array.

Based on their analysis, Baggerly et al. [8] reported the following:

- There was an apparent change in protocol in the middle of dataset 1, which might be due to, e.g. a shift between chip types. As the authors comment: "Such technological differences can give rise to real differences in the spectra, but these differences are not biologically interesting".
- There was an offset (a shift along the horizontal m/z scale) between datasets 2 and 3, that was substantially larger than the nominal precision of the procedure. As the authors note: "A shift of this magnitude could cause the same protein to be identified differently in the two different experiments, obscuring the biology".
- They were unable to separate normals from cancers in dataset 3 using the proteomic pattern developed from dataset 2. This questions the generalizability of the discrimination procedure developed by Petricoin et al. [6] and Conrads et al. [7].
- They were able to perfectly classify the samples in dataset 3 using a set of features lying wholly in the noise region (low m/z) of the spectra. However, as there can be no biological reason for the differences between samples in this region, this would suggest a systematic difference in the way the groups of samples were processed. The authors note that "the features supplied for dataset 1 are also in the lower end of the mass range".

The above example clearly illustrates challenges related to the practical use of the novel techniques aimed at quantification of gene- or prtoein-expression levels. Undoubtedly, the techniques offer a great potential for getting more insight into interesting biological processes. However, before implementing them in practice, one should very carefully assess their properties. For instance, one should ensure repeatability of results. To this aim, potential sources of variability should be investigated and methods to control them developed. One should also develop

methods of calibration and normalization of measurements, which would ensure comparability of the results obtained, e.g. in different experiments. Experiments aiming at development of methods of practical application of the techniques should be carefully designed. "Classical" principles of experimental design – e.g. randomizing the order of processing the samples, "blinding" of procedures of sample-processing, balancing the distribution of important experimental factors – might be here more important than ever, given the high susceptibility of the novel techniques to systematic effects. Finally, appropriate methods of data analysis, taking into account potentially complex structure of the data, should be used.

References

1. Crick FHC. The central dogma of molecular biology. Nature, 1970; 227: 561-3.
2. Schena M, Shalon D, Davis RW et al. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science, 1995; 270: 467-70.
3. Lockhart DJ, Dong H, and Byrne MC. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nature Biotechnology, 1996; 14: 1675-80.
4. Velculescu VE, Zhang L, Vogelstein B et al. Serial analysis of gene expression. Science, 1995; 270: 484-7.
5. Gevaert K, Vandekerckhove J. COFRADIC: the Hubble telescope of proteomics. Drug Discovery Today: Targets, 2004; 3 (Suppl.): 16-22.
6. Petricoin EF, Ardekani AM, Hitt BA, et al. Use of proteomic patterns in serum to identify ovarian cancer. Lancet, 2002; 359: 572-7.
7. Conrads TP, Fusaro VA, Ross S et al. High-resolution serum proteomic features for ovarian cancer detection. Endocrine-Related Cancer, 2004; 11: 163-78.
8. Baggerly KA, Morris JS, Coombes KR. (2004) Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. Bioinformatics, 2004; 20: 777-85.