ELSEVIER

Diagnostic challenges for multiplexed protein microarrays

Stephen R. Master, Charlene Bierl and Larry J. Kricka

Department of Pathology & Laboratory Medicine, University of Pennsylvania Medical Center, 3400 Spruce Street, Philadelphia, PA, 19104, USA

Multiplexed protein analysis using planar microarrays or microbeads is growing in popularity for simultaneous assays of antibodies, cytokines, allergens, drugs and hormones. However, this new assay format presents several new operational issues for the clinical laboratory, such as the quality control of protein-microarray-based assays, the release of unrequested test data and the use of diagnostic algorithms to transform microarray data into diagnostic results.

Protein microarrays are important analytical tools for the multiplexed analysis of large numbers of substances. Currently, there are two microarray formats: 2D planar microarrays and microbead (liquid or suspension) microarrays [1–10]. The 2D microarray comprises reagents for individual tests immobilized as an ordered array or grid of discrete reagent areas (spots) on a flat surface (e.g. a microscope slide). The bead-microarray format comprises encoded microbeads (e.g. each with a unique fluorescence signature) and each type of bead is coated with a different reagent. Both protein microarray formats are popular in biochemical analysis, and prominent examples include multiplexed (simultaneous) assays for cytokines [11,12], allergens [13–17], drugs [18] and hormones [18]. The emergence of the microarray assay format poses several new operational issues for the clinical laboratory. We critically examine issues related to the quality control of protein-microarray-based assays, the release of unrequested test data and the use of diagnostic algorithms to transform microarray data into diagnostic results. A prime purpose of this article is to stimulate debate and discussion of the challenges and potential of all types of microarray assays in routine testing.

Key issues for multiplexed protein assays in a microarray format

In the clinical laboratory, protein assays are usually performed as discrete tests for known proteins. For example, an immunoassay test for serum thyrotropin (TSH) can be performed in a discrete well of a microwell plate or in a test tube. However, a multiplexed assay (planar or bead format) and a multiplexed assay in association with an algorithm are new types of assay format for the clinical laboratory. A simple multiplexed assay bundles together assays on an array or sets of beads in one tube for economy and the convenience of simultaneous assay. Some or all of the individual test results are then reported from the microarray. Another type of multiplexed assay does not report the individual test results but rather feeds these primary data into an algorithm that produces a single test result.

The types of multiplexed assay can be subdivided further according to the type of result that is expected and the types of control that will need to be considered. The simplest multiplexed assay expects a single answer, such as the typing of a pneumococcal species [19]. In this situation, a single species is expected so a negative result or multiple positives would inherently suggest a test failure. Similarly, although tests for specific antibody responses may have complex expected results (with a deviation from these results suggesting the utility of further evaluation), the overall result is a full vaccination response or the need for revaccination. The next level of complexity involves tests where a pattern of answers would be possible, each with slightly different clinical implications but still within the same disease category. To date, tests for autoimmunity are the most prominent examples of this class [20,21]. Clinical treatment or prognosis on the basis of these tests might differ depending on which antibodies are present, indicating the need for some or all of the individual test results to be reported. The final category is one where individual analytes do not carry significance, but the combination of many analytes using a pattern-recognition algorithm yields a single result. Analogous examples using protein-fragment mass-spectral data rather than protein microarrays *per se* include

Corresponding author: Kricka, L.J. (kricka@mail.med.upenn.edu)

1359-6446/06/$ - see front matter doi:10.1016/j.drudis.2006.09.010

the MammoCheck™ and OvaCheck® tests from Correlogic (http://www.correlogic.com).

This last category presents the most substantial challenge to clinical implementation, although a clear precedent for the idea can be found as long ago as 1975 when Linus Pauling and colleagues proposed detecting urinary biomarker patterns specific to a particular state of health (an approach that he termed 'orthomolecular diagnosis') [22]. One of the first tests to use pattern recognition for clinical diagnostics was the REMEDi HS drug screening system produced by Bio-Rad (http://www.bio-rad.com) in the early 1990s [23]. This technology uses an automated liquid-chromatographic analyzer and UV spectral patterns coupled to a computational rules-based algorithm for identifying individual drugs in biological fluids. The general concept of using simple as opposed to complex combinations of assay measurements to produce a single test result is, of course, much less controversial. The clinical laboratory already uses algorithms to derive test values indirectly (e.g. the Friedwald formula for LDL cholesterol). More significantly, algorithms are used to derive a single parameter that reflects risk {e.g. the triple and quadruple screen [measuring human chorionic gonadotrophin (hCG), unconjugated estriol (E3), α-fetoprotein (AFP), plus inhibin-A for the quadruple test]}. The novelty of microarray-based tests that have associated computational processing lies primarily in the physical assay format (all results from a 'single tube'). Examples of algorithms used in the clinical laboratory as well as commonly used clinical calculations based on individual laboratory results are shown in Table 1.

Quality control of microarray assays

In clinical laboratories the performance of analytical methods is routinely monitored by analyzing quality control (QC) materials with known concentrations of various analytes. The observed results are compared with the range of values determined for the known analytes (control values). Observed values falling within the control range indicate that the method is performing acceptably, whereas values falling outside of this range signify an unacceptable analytical performance. Control materials are manufactured in large quantities and provide a long-term source of specimens with normal and abnormal concentrations of various analytes in a matrix that is the same as the test specimen (e.g. serum).

Despite the importance of routine clinical quality control, its application to microarray-based assays remains underdeveloped. Attention has been paid to the reproducibility of spotting and the activity of reagents on the array as well as to data processing and analysis methods [24]. Also, many array designs include internal control spots to detect nonspecific interactions, verify assay conditions, produce fiduciary (marker) signals for the orientation of the microarray during data analysis, assess the binding of detection reagents, serve as negative and positive assay and scanning controls, and even provide positive controls for protein interactions (http://www.invitrogen.com/protoarray). However, the more complex issues of routine QC have received less attention.

QC protocols developed for the routine clinical laboratory are designed to ensure that analytical methods are performing within desired specifications on a daily, weekly and monthly basis. Results from the analysis of QC materials with high, medium and low levels of a particular protein are assessed using Levy–Jennings plots or in a more structured way by means of a series of interpretive rules that provide the basis for the acceptance or rejection of analytical data [25]. In a typical automated immunoassay for a given protein, the assay is controlled by analyzing the QC materials at the beginning of each day and at the beginning of subsequent 8-h shifts. Results for the QC materials must fall within the target ranges determined by the laboratory for those materials in order for the results on the clinical specimens to be released. The closest analogy to a microarray assay is an immunoassay performed on a microtiter plate (e.g. 96, 384 and 1536 wells). However, this type of assay involves only a single reagent arrayed in the microtiter-plate wells so the QC problem is relatively simple compared with a microarray that has hundreds or thousands of individual reagents in the array. Usually, all three controls (high, medium and low levels of a protein) are included in wells at the start and end of the microtiter plate. Results for the controls are subjected to the same type of analysis described above to decide whether the results are analytically acceptable.

Quality control materials

For a routine microarray method designed to generate data on clinical specimens, there is little guidance on the best way to QC such assays. An initial problem is the availability of controls with high, medium and low levels of each of the proteins that the array measures. Manufacturing such controls would be complex because of the number of different proteins that each control must contain to test each reagent on the array. Routine clinical laboratory QC materials can contain >80 different analytes, but most are stable

TABLE 1

Examples of tests that are combinations of individual tests

Anion gap	The anion-gap calculation (serum sodium; serum chloride; serum bicarbonate); used in the differential diagnosis of acid–base disorders
Corrected white blood cell count	A calculation based on manual count of nucleated red blood cells for a more accurate white blood cell count
Free testosterone	The concentration of bioavailable and free testosterone derived from total serum testosterone, protein-binding constants and the serum concentration of the testosterone-binding proteins; used for assessment of hormonal disorders
LDL	The Friedwald serum LDL calculation (serum total cholesterol; HDL cholesterol; serum triglycerides/5); used for cardiac risk assessment
Nephrogenous cAMP	A calculation based on urinary and plasma cAMP levels; used in the assessment of parathyroid function
'Triple Screen'	The ratio of serum AFP:serum hCG:urine estriol to screen for fetal Down syndrome and trisomy 18
'Quadruple Screen'	The addition of inhibin-A to the estimate of the Triple test; an improved screen for fetal defects

(e.g. potassium ions, uric acid and urea) and are readily obtained in highly purified forms. By contrast, the stability and availability of the protein analytes required to prepare QC materials for protein microarrays is not a trivial issue. The scale of analysis possible with a protein microarray is also the source of potential problems concerning controls. To illustrate this point, we consider an array designed to test for 100 different proteins. If 0.05 g/l represented a high value for these individual protein analytes, then a high control would contain 100 × 0.05 g/l, or 5 g/l, and this total concentration might not be compatible with solubility for the collection of proteins. More significantly, the potential for interactions between controls might complicate the generation of pooled high, medium and low control specimens. The design, generation and distribution of appropriate control material will pose a major challenge to the clinical use of microarray-based assays and such problems are now receiving attention from national organizations such as the National Institute of Standards and Technology (NIST, USA) in the context of biomarker discovery and validation [26]. For assays that utilize algorithmic post-processing to generate a small number of diagnoses from large protein-microarray datasets (see below), the robustness of the resulting classification might also influence the development of such control materials.

Interpreting quality control data

If the hurdles presented by controls can be overcome, the next problem is developing a set of rules for interpreting the combined QC data from the array along with appropriate criteria to determine the acceptability of the analytical results from the array. No problem exists if the array returns results that are within the control limits for the individual proteins in each QC material. However, it is less clear how the laboratory should handle the total assay if some of the analytical results for the QC materials are not within the predetermined control limits. Once again, the scale of microarray analysis produces a correspondingly scaled QC challenge. Figure 1 shows some possible results from the analysis of

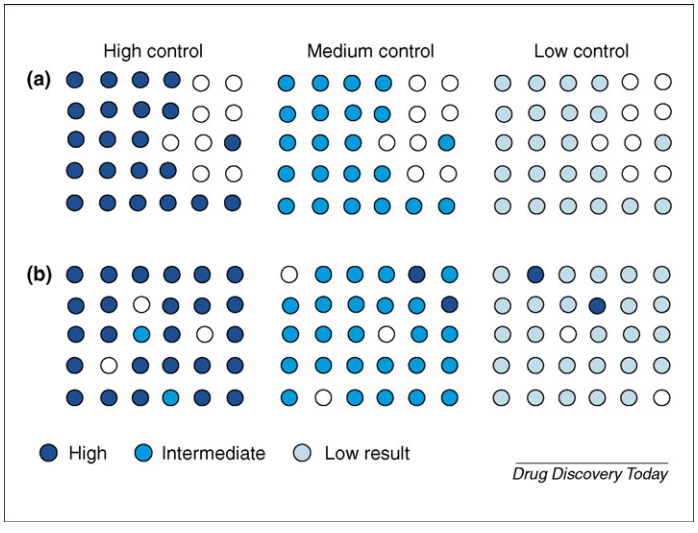

FIGURE 1

Simulated planar microarray results for QC materials with high, medium or low levels of analytes. (a) The systematic failure and **(b)** the random failure of a subset of locations on the array to provide the expected result for the high, medium or low controls. Both sets of results would render results from test arrays questionable at specific locations on the array.

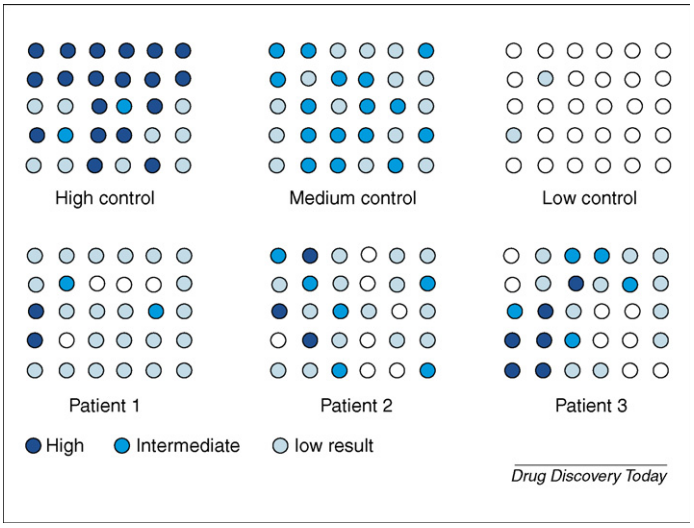

FIGURE 2

Simulated planar microarray results for samples from three controls (high, medium and low) and three patients (Patient 1, 2 and 3). The controls have random failures and this poses a significant problem for the interpretation of the results obtained for the three patients.

three control materials that have high, medium and low values for each analyte that the array is designed to test. If all assays for all analytes in all three controls are either acceptable or unacceptable on the basis of the established QC ranges, the assays either pass or fail the QC. Similarly, if a specified analyte consistently fails (Figure 1a), the result would indicate there is a particular problem with specific reagents on the array. However, given the number of analytes, it is far more likely that a subset of analytes will fail in a subset of controls (Figure 1b). If the controls form part of a batch of assays that includes patient samples, as shown in Figure 2, then the problem is how to interpret the results obtained for the three patients in light of the partially failed controls.

Several options for resolving this issue can be proposed. One course of action would be to reject all of the results, undertake the appropriate trouble-shooting, and then reanalyze the controls and samples on additional microarrays. Another option is to accept data selectively for the three patients from locations on the microarrays that passed for each of the three controls. However, because the goal of this process was to analyze all three patient samples for all of the analytes, no benefit is derived from taking this approach as opposed to simply re-analyzing all samples. It is possible to create a subset of the array when using a microbead format; however, the practical challenges involved (particularly with respect to ensuring there are no significant interferences for assay validation of a given mix of beads) might be substantial. The fundamental issue (needing to rerun entire arrays for subsets of QC failure) is further exacerbated in formats that combine a series of individual microarrays on a single slide (an array of arrays). Should the slide be considered in its entirety, or should control data be generated for each of the individual arrays (e.g. one array for each of the high, medium and low controls)? Even more problematic are devices that combine several arrays on the same substrate and also have an array structure that provides replicate testing for each analyte within each array (see Figure 3 for a hypothetical device that has six arrays on a slide, and each array tests for ten analytes in triplicate). When this is used to analyze a QC material, many

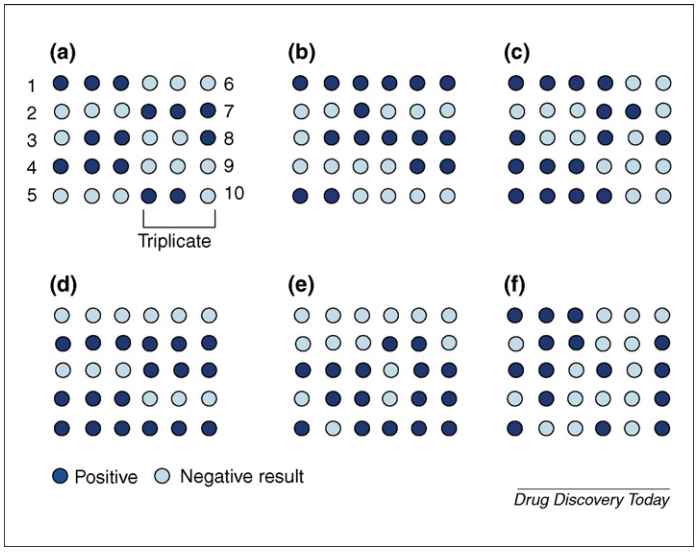

FIGURE 3

A schematic of an array-of-arrays-type device for testing multiple samples. Each sample is tested in triplicate for each of the ten analytes (locations 1–10) on each of the arrays **(a–f)**. Variability in results is illustrated using a simple positive and negative classification of the results. Results can vary from 3/3 to 0/3 of the triplicate test areas (spots) on the array providing the expected result, for a control or for test specimens.

different combinations of results are possible. For individual arrays there might be results for some triplicates that are all within control limits, others that are all outside of the control limits, and yet others that have combinations of results that are inside and outside of the control limits. In a worst-case scenario, the analytical performance of the individual arrays on the slide might also vary. It would therefore be difficult to decide what the appropriate criteria should be for determining whether the entire device was functioning correctly. The above QC considerations are compounded as the number of individual reactions performed on the array increases. We suggest that a definitive answer to these questions for a given assay will only be obtained when the clinical application and the post-assay computational processing (if any) are taken into consideration. The scale of replication on planar microarrays is limited by the size of the substrate and the number of analytes. However, in the case of bead arrays it is a simple matter to perform replicates on a much larger scale (hundreds of beads per analyte), which allows the rejection of outliers in the data and the stabilization of results.

It is clear that it will be important to resolve QC issues for large-scale and ultra-large-scale microarray analysis if this form of analysis is to become established as a routine method for the clinical laboratory. Many current array-based assays include replicate testing (e.g. in triplicate for planar arrays and 50–200 for bead arrays), standard curves or internal controls. Although each of these measures provides additional levels of assurance as to the validity of a result, none provides the same level of oversight of analytical performance as testing QC materials. Recently, several microarray-based assays have been cleared by the FDA. These include a bead-suspension microarray assay from the Bio-Rad Laboratories for ANA testing based on the BioPlex^{TM} 2200 analyzer and the Zeus Scientific (http://www.zeusscientific.com) AtheNA Multi-Lyte ANA Test System for the semi-quantitive detection of IgG class antibodies to specific antigens (SSA, SSB, Sm, RNP, Scl -70, Jo-1, Centromere B and Histone) in human serum. For the Bio-Plex^{TM} 2200 ANA assay, two controls are supplied that contain normal and abnormal levels of all 13 analytes. The user can only report out analytes that pass QC, and because the Medical Decision Support Software requires all test results, there will be no Medical Decision Support Software output if any analyte fails QC [21].

Releasing unrequested results from a microarray

The clinical utility of microarrays and other multiplexed assays is derived primarily from their parallel measurement of large numbers of analytes. However, this advantage might raise the issue of 'unnecessary' clinical testing if the number of measured analytes substantially exceeds the set that would be indicated in a given clinical situation. Conversely, although microarray technology might ultimately prove cost-effective for batch analysis, there is no current consensus as to how unrequested test results should be stored and/or reported. At least two commercial platforms (Randox Biochip, http://www.randox.com; and Athena Multi-Lyte, http://www.zeusscientific.com) explicitly provide capabilities for the retrospective reporting of analyte measurements that were not initially requested. This might provide a substantial advantage in investigatory settings (such as clinical trials support), where the availability of larger datasets for subsequent mining might be beneficial. However, it is unclear how this capability would affect the clinical laboratory because the storage of unordered test results without reporting them might conceivably raise legal questions in cases where those results would substantially alter diagnosis or treatment. It is worth noting that this consideration primarily applies in cases where the individual results are reported, rather than in those cases where a larger pattern of analytes (each, perhaps, measured with lower precision) is used to derive a single, aggregate diagnosis.

Algorithms

As discussed above, the clinical laboratory uses a variety of calculated test values in routine practice, including the so-called 'quadruple screen' for the prenatal risk assessment of trisomy 21 [27]. The number and variety of analytes available for analysis when using protein microarrays substantially exceeds those seen before in a single assay and, as a result, the range of interpretive options for such a device is correspondingly complex. Significantly, high-level pattern recognition approaches are likely to be routinely applied to these results. A variety of interpretive tools has been used for analogous discovery work with DNA microarrays; these include various clustering approaches (hierarchical clustering and self-organizing maps [28,29]) as well as supervised classification algorithms (support vector machines, k-nearest neighbors and neural networks [30–32]). The second group of analytical approaches is likely to be of substantial clinical use after validation. The potential clinical applicability of pattern classification algorithms carries with it a corresponding regulatory burden, and any approach utilized for diagnosis would, of course, be subject to FDA regulation as a 'device' if marketed within USA. Currently, the only FDA-approved pattern-recognition algorithm associated with a protein microarray (or any other highly multiplexed device) is the k-nearest neighbors (kNN) classifier used for ANA interpretation on the BioPlex 2200 (http://www.fda.gov/cdrh/pdf4/K043341.pdf) [21,33]. This algorithm utilizes the 11 closest

reference samples from a database of >1400 patients to determine a final diagnostic category. Interestingly, regulatory approval was obtained by referencing (as a predicate device) an algorithm (non-kNN) previously used to match chemical library patterns for toxicology screening. Currently, it remains unclear whether the FDA will allow this predicate to be used narrowly (e.g. kNN and rules-based classifiers alone) or whether it will be seen as a broader validation of the general concept of pattern-recognition algorithms used for the diagnostic interpretation of protein-microarray data. A further issue is the recent patenting of pattern-recognition approaches to processing data (the so-called 'hidden patterns' patent). This approach facilitates the identification of hidden or non-obvious patterns of changes of constituents in serum, urine, saliva and sweat. These discriminatory biological data patterns can then be used to identify disease and other conditions [34]. The scope and potential impact of this intellectual property on multiplex testing, particularly given the long history of pattern recognition, is currently unclear.

Conclusions

The planar and microbead microarray multiplexed assay formats provide a convenient and efficient means to test samples for many analytes in a parallel manner. So far, most concerns over this technology have related to the pros and cons of the planar versus microbead format, immobilization strategies for the binding agent or sample, cross-reactivity and various aspects of the informatics challenge posed by analytical systems capable of producing millions of assay results per day. However, operational issues relating to quality control, reporting of unrequested test results and the use of algorithms pose potential obstacles to the widespread use of this technology, especially in clinical testing. Indeed, the FDA has commented on the newness of the multiplex type of test, and has provided written comments to the manufacturer of a diagnostic device that incorporates a multiplexed protein test and an algorithm (http://www.fda.gov/cdrh/oivd/letters/071204-correlogic.html) [35], and a device that incorporates a DNA microarray [this was subsequently cleared for marketing in December 2004 (http://www.fda.gov/bbs/topics/news/2004/new01149.html)]. Despite recent progress in the development of planar and microbead arrays, a wider appreciation of the ramifications for these types of assay in routine testing is urgently needed, particularly in view of the emerging interest in large panels of biomarkers for 'biophysicals' (for example, Biophysical250, http://www.biophysicalcorp.com).

References

1 Lipshutz, R.J. *et al.* (1999) High density synthetic oligonucleotide arrays. *Nat. Genet.* 21 (Suppl), 20–24
2 Southern, E.M. (2001) DNA microarrays. History and overview. *Methods Mol. Biol.* 170, 1–15
3 Ekins, R.P. (1998) Ligand assays: from electrophoresis to miniaturized microarrays. *Clin. Chem.* 44, 2015–2030
4 Schena, M., ed. (2000) *Microarray biochip technology*, p. 297, Eaton Publishing
5 Zangar, R.C. *et al.* (2005) Studying cellular processes and detecting disease with protein microarrays. *Drug Metab. Rev.* 37, 473–487
6 Bertone, P. and Snyder, M. (2005) Advances in functional protein microarray technology. *FEBS J.* 272, 5400–5411
7 Stoll, D. *et al.* (2005) Protein microarrays: applications and future challenges. *Curr. Opin. Drug Discov. Devel.* 8, 239–252
8 Lueking, A. *et al.* (2005) Protein biochips: a new and versatile platform technology for molecular medicine. *Drug Discov. Today* 10, 789–794
9 Angenendt, P. (2005) Progress in protein and antibody microarray technology. *Drug Discov. Today* 10, 503–511
10 Nolan, J.P. and Sklar, L.A. (2002) Suspension array technology: evolution of the flat-array paradigm. *Trends Biotechnol.* 20, 9–12
11 Huang, R.P. *et al.* (2001) Simultaneous detection of multiple cytokines from conditioned media and patient's sera by an antibody-based protein array system. *Anal. Biochem.* 294, 55–62
12 Lin, Y. *et al.* (2003) Detection of multiple cytokines by protein arrays from cell lysate and tissue lysate. *Clin. Chem. Lab. Med.* 41, 139–145
13 Harwanegg, C. and Hiller, R. (2005) Protein microarrays for the diagnosis of allergic diseases: state-of-the-art and future development. *Clin. Chem. Lab. Med.* 43, 1321–1326
14 Fall, B.I. *et al.* (2003) Microarrays for the screening of allergen-specific IgE in human serum. *Anal. Chem.* 75, 556–562
15 Harwanegg, C. *et al.* (2003) Microarrayed recombinant allergens for diagnosis of allergy. *Clin. Exp. Allergy* 33, 7–13
16 Jahn-Schmid, B. *et al.* (2003) Allergen microarray: comparison of microarray using recombinant allergens with conventional diagnostic methods to detect allergen-specific serum immunoglobulin E. *Clin. Exp. Allergy* 33, 1443–1449
17 Kim, T.E. *et al.* (2002) Quantitative measurement of serum allergen-specific IgE on protein chip. *Exp. Mol. Med.* 34, 152–158
18 FitzGerald, S.P. *et al.* (2005) Development of a high-throughput automated analyzer using biochip array technology. *Clin. Chem.* 51, 1165–1176
19 Lin, J. *et al.* (2006) Validation of a multiplex pneumococcal serotyping assay with clinical samples. *J. Clin. Microbiol.* 44, 383–388
20 Shovman, O. *et al.* (2005) Multiplexed AtheNA multi-lyte immunoassay for ANA screening in autoimmune diseases. *Autoimmunity* 38, 105–109
21 Binder, S.R. *et al.* (2005) Computer-assisted pattern recognition of autoantibody results. *Clin. Diagn. Lab. Immunol.* 12, 1353–1357
22 Dirren, H. *et al.* (1975) Sex-related patterns in the profiles of human urinary amino acids. *Clin. Chem.* 21, 1970–1975
23 Sadeg, N. *et al.* (1997) Automated liquid-chromatographic analyzer used for toxicology screening in a general hospital: 12 months' experience. *Clin. Chem.* 43, 498–504
24 Kricka, L.J. and Master, S.R. (2005) Validation and quality control of protein microarray-based analytical methods. *Methods Mol. Med.* 114, 233–255
25 Westgard, J.O. *et al.* (1981) A multi-rule Shewhart chart for quality control in clinical chemistry. *Clin. Chem.* 27, 493–501
26 Barker, P.E. (2003) Cancer biomarker validation: standards and process. *Ann. N. Y. Acad. Sci.* 983, 142–150
27 Benn, P.A. (2002) Advances in prenatal screening for Down syndrome: I. general principles and second trimester testing. *Clin. Chim. Acta* 323, 1–16
28 Eisen, M.B. *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A.* 95, 14863–14868
29 Tamayo, P. *et al.* (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. U. S. A.* 96, 2907–2912
30 Brown, M.P. *et al.* (2000) Knowledge-based analysis of microarray gene expression data using support vector machines. *Proc. Natl. Acad. Sci. U. S. A.* 97, 262–267
31 Yeang, C.H. *et al.* (2001) Molecular classification of multiple tumor types. *Bioinformatics* 17, S316–S322
32 Xu, Y. *et al.* (2002) Artificial neural networks and gene filtering distinguish between global gene expression profiles of Barrett's esophagus and esophageal cancer. *Cancer Res.* 62, 3493–3497
33 Binder, S.R. *et al.* (2006) Protein arrays and pattern recognition; new tools to assist in the identification an management of autoimmune disease. *Autoimmun. Rev.* 5, 234–241
34 Hitt, B.A. *et al.* (2005) Process for distinguishing between biological states based on hidden patterns from biological data. *U.S. Patent.* 6925389
35 Wagner, L. (2004) A test before its time? FDA stalls distribution process of proteomic test *J. Natl. Cancer Inst.* 96, 500–501