Published in final edited form as:
Expert Opin Med Diagn. 2013 January ; 7(1): 37–51. doi:10.1517/17530059.2012.718329.

Challenges in Biomarker Discovery: Combining Expert Insights with Statistical Analysis of Complex Omics Data

Jason E. McDermott, **Jing Wang**, **Hugh Mitchell**, **Bobbie-Jo Webb-Robertson**, **Ryan Hafen**, **John Ramey**, and **Karin D. Rodland**
Pacific Northwest National Laboratory, Richland WA 99352 USA

Abstract

Introduction—The advent of high throughput technologies capable of comprehensive analysis of genes, transcripts, proteins and other significant biological molecules has provided an unprecedented opportunity for the identification of molecular markers of disease processes. However, it has simultaneously complicated the problem of extracting meaningful molecular signatures of biological processes from these complex datasets. The process of biomarker discovery and characterization provides opportunities for more sophisticated approaches to integrating purely statistical and expert knowledge-based approaches.

Areas covered—In this review we will present examples of current practices for biomarker discovery from complex omic datasets and the challenges that have been encountered in deriving valid and useful signatures of disease. We will then present a high-level review of data-driven (statistical) and knowledge-based methods applied to biomarker discovery, highlighting some current efforts to combine the two distinct approaches.

Expert opinion—Effective, reproducible and objective tools for combining data-driven and knowledge-based approaches to identify predictive signatures of disease are key to future success in the biomarker field. We will describe our recommendations for possible approaches to this problem including metrics for the evaluation of biomarkers.

Keywords

biomarkers; molecular signatures; classification; data integration; expert knowledge

1. Introduction

In one sense, the practice of medicine has been dependent on biomarkers since its inception, even though historic biomarkers have been externally visible indicators of fundamental physiological processes, such as fever, swelling, tenderness or rash. With the advent of molecular biology and the fundamental understanding of information flow from gene to transcript to protein, biomedical scientists have been searching for the unique molecular markers associated with disease processes with the goal of improving early detection, determining prognosis, monitoring the response to therapy or selecting those treatments most likely to be efficacious. The molecular species targeted have included genes, transcripts, proteins, metabolites, and recently non-coding, regulatory RNAs. Yet despite a decade of intense effort and significant investment of labor and funds, the number of clinically validated biomarkers approved by the FDA is embarrassingly modest: fewer than 30 in the most recent published compilation [2]

BOX

DEFINITION OF BIOMARKER TYPES [1]

Feature

A measurable biological component (e.g. protein, gene, metabolite) or state of a component (e.g. proteolytically cleaved protein, methylated gene) that can be analyzed as a candidate biomarker

Biomarker

A feature that is indicative of a disease state, response to therapeutic treatment, or other relevant biological state.

Biosignature

A collection of features, which together defines a biomarker.

Risk biomarkers

- Identify patients who are likely to develop disease

Diagnostic biomarkers

- Detection of early disease state
- Classification into disease subtypes
- Characterization of response to treatment

Prognostic biomarkers

- Prediction of disease progression
- Prediction of disease recurrence
- Identification of patients who are likely to respond to a treatment

The search for biomarkers has always tended to focus on one of two basic styles: hypothesis-based or discovery-based. Biomarker identification through hypothesis-based methods is essentially a by-product of the ever-increasing mechanistic understanding of disease processes. For example, knowledge that diabetes mellitus produces a sustained elevation of blood glucose levels led to the identification of glycosylated hemoglobin as a biomarker for diagnosis of diabetes [3]. Similarly, understanding of the mechanisms of growth regulation in mammary epithelial cells led to the use of estrogen receptor status [4] and HER2/neu amplification [5] as independent prognostic markers in breast cancer. In contrast, discovery-based approaches have focused on identifying changes in the presence or relative abundance of molecular species that are tightly associated statistically with the disease state of interest. This type of study can also be hypothesis-generating, in that observations of differential expression that are closely tied to a disease outcome often lead to intense investigations of the function of the candidate biomarker. A case in point is the breast and ovarian cancer-associated gene BRCA1, which was initially identified by positional cloning of a region on chromosome 17 that is frequently deleted in breast cancer [6]. Subsequent studies of BRCA1 function led to the current detailed understanding of BRCA1's role in facilitating DNA repair [7, 8].

BOX

CURRENT FDA APPROVED BIOMARKERS

[2]

Biomarker	Type	Source	Cancer type	Clinical use
α-Fetoprotein	Glycoprotein	Serum	Nonseminomatous testicular	Staging
Human chorionic gonadotropin-β	Glycoprotein	Serum	Testicular	Staging
CA19-9	Carbohydrate	Serum	Pancreatic	Monitoring
CA125	Glycoprotein	Serum	Ovarian	Monitoring
Pap smear	Cervical smear	Cervix	Cervical	Screening
CEA	Protein	Serum	Colon	Monitoring
Epidermal growth factor receptor	Protein	Colon	Colon	Selection of therapy
KIT	Protein (IHC)	Gastrointestinal tumour	GIST	Diagnosis and selection of therapy
Thyroglobulin	Protein	Serum	Thyroid	Monitoring
PSA (total)	Protein	Serum	Prostate	Screening and monitoring
PSA (complex)	Protein	Serum	Prostate	Screening and monitoring
PSA (free PSA%)	Protein	Serum	Prostate	Benign prostatic hyperplasia versus cancer diagnosis
CA15-3	Glycoprotein	Serum	Breast	Monitoring
CA27-29	Glycoprotein	Serum	Breast	Monitoring
Cytokeratins	Protein (IHC)	Breast tumour	Breast	Prognosis
Oestrogen receptor and progesterone receptor	Protein (IHC)	Breast tumour	Breast	Selection for hormonal therapy
HER2/NEU	Protein (IHC)	Breast tumour	Breast	Prognosis and selection of therapy
HER2/NEU	Protein	Serum	Breast	Monitoring
HER2/NEU	DNA (FISH)	Breast tumour	Breast	Prognosis and selection of therapy
Chromosomes 3, 7, 9 and 17	DNA (FISH)	Urine	Bladder	Screening and monitoring
NMP22	Protein	Urine	Bladder	Screening and monitoring
Fibrin/FDP	Protein	Urine	Bladder	Monitoring
BTA	Protein	Urine	Bladder	Monitoring
High molecular weight CEA and mucin	Protein (Immunofluorescence)	Urine	Bladder	Monitoring

BTA, bladder tumour-associated antigen; CA, cancer antigen; CEA, carcinoembryonic antigen; FDP, fibrin degradation protein; FISH, fluorescent *in-situ* hybridization; GIST, gastrointestinal stromal tumour; IHC, immunohistochemistry; NMP22, nuclear matrix protein 22; PSA, prostate-specific antigen.

Source: Nat Rev cancer © 2005 Nature publishing Group

The advent of high throughput omic technologies in the past decade has fueled the discovery-based approach by providing ready access to large, relatively quantitative datasets of differentially expressed mRNAs, microRNAs, and proteins from case control studies. Despite the application of a wide variety of sophisticated approaches for statistical analysis of these large datasets, the results have been disappointing overall. To date, the FDA has approved only two biomarker panels for breast cancer prognosis (OncoType Dx and MammaPrint) and one for ovarian cancer (Ova1).

As a case in point, the Ova1 In Vitro Diagnostic Multivariate Index Assay was derived from a comparison of differentially abundant plasma proteins from women with ovarian cancer, compared to women with benign gynecological diseases, using artificial neural network (ANN) modeling to derive a panel of 5 biomarkers that surpassed the currently available ovarian cancer biomarker, CA125 [9], in the ability to discriminate between invasive ovarian cancer and benign lesions [10][11]. It should be noted that one factor contributing to the successful FDA approval of the Ova1 panel was its restriction to a very narrowly defined, but clinically important diagnostic application, that of triage of women at high risk of ovarian cancer for referral to a gynecological oncologist for primary surgery [12]. This highlights the importance of matching the experimental strategy for biomarker discovery and identification to the intended use of the biomarker. Biomarkers for early detection of disease must possess the specificity to distinguish between clinically significant cancer and related but benign conditions, with the sensitivity to detect very small tumor masses, ideally before clinical symptoms. This is an extremely high bar. Single biomarkers with sufficient sensitivity [e.g., CA125 or prostate specific antigen (PSA)] often lack the specificity required for clinical utility[13, 14]. Thus many investigators have turned to combinations of biomarkers in hopes of attaining both sensitivity and specificity sufficient for true clinical utility.

The most general approach for assembling biomarker panels has centered on the use of sophisticated statistical models on large quantitative datasets, as exemplified by Correlogic's OvaCheck and LabCorp's OvaSure assays for early detection of ovarian cancer. Although both groups started with well-defined, appropriately sized sample sets comprising patients with ovarian cancer, as well as healthy controls, the two groups differed substantially in their approach to statistical analysis [15] [16]). The Mor group (LabCorps/OvaSure) relied on fairly standard classification tools –support vector machines (SVMs), k-nearest neighbor, and classification trees, resulting in a panel of four markers: leptin, prolactin, osteopontin, and IGF-II [17]. A secondary analysis of these markers on an independent and larger sample, which included women with stage I/II ovarian cancer, was reported to achieve a sensitivity of 95.3% and a specificity of 99.4% when CA-125 and macrophage inhibitory factor (MIF) were added to the panel.

In contrast, Correlogics developed their own algorithm, Knowledge Discovery Engine-VS, a refinement of the random forest approach, to analyze their high-dimensional data [15]. Candidate biomarkers were not evaluated independently, but as part of a pattern, resulting in an 11 analyte panel providing sensitivities and specificities approaching 90% [18]. Interestingly, only CA-125 and C-reactive protein had significant discriminatory power when used alone, and several of the best–performing individual markers did not make it into the final multi-analyte panel [18]. In yet a third example, Lokshin and colleagues applied Metropolis algorithms with Monte Carlo simulation to arrive at a candidate panel for early detection of ovarian cancer, down-selecting from an original set of 96 candidates to a four member panel of CA-125, HE4, CEA, and VCAM-1 [19]. The emphasis in their training set was on discriminating early stage ovarian cancer from benign pelvic diseases and other common cancers (breast, colorectal, and lung). Despite marked differences in statistical approach among those searching for early detection markers in ovarian cancer, CA-125 and

HE-4 consistently emerge as the two most discriminating markers [18–21], implying some underlying biological commonality.

A similar comparison of the two FDA-approved transcriptome signatures for prognosis of node negative breast cancer, the 70 gene MammaPrint [22] and 21 gene OncoType Dx [23] assays, reveals no overlap at the level of specific genes [24]. However, a number of meta-analyses of these two classifiers have identified proliferation-associated, cell-cycle regulated genes as the predominant source of discriminatory power in both of these prognostic assays [25–27]. In fact, when proliferation-associated genes are removed from these prognostic gene signatures, the remaining genes are no better at predicting outcome than are random gene signatures mechanistically unrelated to cancer [27].

These observations raise some very fundamental questions about the ability of purely statistical approaches to identify specific and predictive signatures of outcome in the absence of a weighting or evaluative process provided by relevant domain experts. Developing an approach to biomarker identification and characterization that successfully infuses expert knowledge into data-driven statistical analyses requires a solid foundation in the current practice of each approach in isolation.

2. Data-driven approaches to biomarker identification

Data-driven approaches make use of now-prevalent high throughput datasets that facilitate the elucidation of underlying structure. We discuss a sampling of these techniques under three categories: data reduction, classification and visualization. We note that many of these approaches fall into more than one of our categories (Table 1), and discuss this for specific examples.

2.1. Data Reduction

There is often a need in bioinformatics analysis to reduce the complexity of datasets to a manageable size. Data reduction can make complex biological datasets easier to understand by the expert but can also help eliminate the noise inherent in these kinds of measurements. The increased understanding of the underlying trends in data that reduction can provide can assist researchers in focusing biomarker studies on the most relevant biological areas.

Trend analysis involves implementation of simple regression techniques to reveal statistically relevant patterns in data such as gene expression profiles. For example, trend analysis can be used to identify a positive association between expression levels of a particular gene, such as the AD1 and AD2 isoforms of tenascin C, and a separate variable that trends with tumor aggressiveness, e.g., tumor grade [28]. In addition, trend analysis can be applied to population data such as disease incidence to discover underlying patterns, as in a published trend analysis of the WHO database for breast cancer mortality in paired European countries over time; this study indicated that declining breast cancer mortality preceded, and was independent of, widespread implementation of population-based screening by mammography [29].

Clustering is a broad term referring to any attempt to group data according to interactions among elements in the data. Clustering efforts may focus on any of multiple goals, such as grouping patients with similar cancer subtypes, finding genes with similar expression patterns to identify regulatory units and pathways, or inferring function of unannotated genes [30, 31].

For biomarker discovery, clustering can be used as an initial screen to determine if the data sufficiently separates the samples into appropriate classes [32]. If clustering the data shows

good class segregation, then further analysis can be reasonably undertaken to identify the data elements that best characterize the separation. Additionally, clustering can be used after biomarker discovery to verify that the markers identified can separate the samples effectively into distinct clusters [33].

Clustering methods generally fall into two categories: hierarchical, in which a dendrogram is built that represents the partitions of the data at all possible cluster levels, or partitional, where optimal clusters are found for a single k number of clusters [34].

Variations on classical concepts of clustering can be effectively utilized in certain contexts. Fuzzy clustering, in which genes are assigned multiple weighted cluster memberships, is one approach that may be ideal for gene expression clustering, since some genes likely belonging to multiple functional groups [35]. Biclustering is a beneficial approach for situations where some genes exhibit relationships over only a subset of conditions, such as is the case with tumor datasets with heterogeneous tumor subtypes [36].

2.3. Classification

One important application of bioinformatics is to predict class membership of unknown samples based on data gathered from previously characterized samples. For example, a well-studied problem is the classification of tumor samples by malignant potential using gene expression data. A simplistic approach to this problem is to select expression and significance thresholds for determining whether a gene is changed, then use those thresholds to identify genes that are differentially expressed between tumor types [37]. Another basic approach is to look for genes whose expression correlates with tumor-related phenotypes, such as tumor size, serological markers, or metastasis; the selection of candidate biomarkers based on correlated gene expression is very common [38].

Here we discuss a few of the many classification tools that address these problems, but do so in more sophisticated ways.

2.3.1. Regression—Selection of a small, highly predictive set of markers is a universal problem in classification studies. Regression is a common method for selecting the best set of markers. For univariate regression, markers are selected according to their predictive power regardless of the influence of other markers, while multivariate regression estimates a small set of markers that in concert with one another can effectively classify new samples [39]. Methods that include a penalty component in the selection process, resulting in smaller selection sets, have demonstrated superior performance [39]. Many instances of successful application of this technique exist in the literature. Two recent examples include the use of multivariate regression on 953 proteins identified in nasal fluids to identify 3 proteins that discriminated between high and low responders to glucocorticoid treatment for allergic rhinitis [40]. Similar approaches were used to identify a gene expression signature for lymphoid stem cells, and determine its association with adverse outcomes in acute myeloid leukemia [41].

2.3.2. Support Vector Machine (SVM)—This approach is generally focused on binary classification where training data are mapped onto multidimensional space via a kernel function, and a hyperplane is chosen that separates the two classes. The major benefit of a SVM is that the kernel function allows separation even under non-linear circumstances by choice of the kernel function. The algorithm maximize the margins between the hyperplane (a dividing line in high-dimensional space) and the closest correct data point on each side, while minimizing the distance to any misclassified elements. With this hyperplane in place, new data can be classified simply by determining on which side of the division they fall [42].

SVM has become a very popular tool for selection of candidate biomarker panels from high density datasets, including both proteomic, transcriptomic, and microRNA data sets. Recent examples include an SVM-driven classification for ranking ovarian cancer proteomic data, to select the most discriminating spectra for candidate biomarkers [43], use of serum microRNA profiles to train an SVM for classification of early stage breast cancer [44], and use of SVMs to confirm the classification utility of biomarkers identified by other means, such as ANOVA and PCA [45].

2.3.3. Decision trees and random forest—This learning approach builds a tree of questions such that the terminal leaves represent the correct classification of the test samples. The tree is built (training stage) as questions are chosen that minimize the impurity (i.e. heterogeneity) of the resulting subgroups at each level of the tree. Ideal trees will correctly classify all the data using a minimal structure. While a single decision tree can provide accurate classification, compiling an ensemble of related trees using a random sampling of training data and data features can confer additional accuracy to the classification.

For approaches of this kind, termed random forest methods, each tree is built separately as only a subset of the available features is considered when forming the question at each split of each random tree as it is grown [46]. Marker selection is accomplished as the impact of each feature (i.e. gene profile, for gene expression data) is assessed in the forest as a whole for its power to effectively discriminate samples [47]. Recently, a random forest algorithm was applied to serum proteomic data to identify clusters of proteins significantly associated with Alzheimer's disease [48], and elevated body mass index [49] in an effort to identify early markers of these two diseases.

2.3.4. Artificial neural networks—Artificial neural networks (ANNs) are machine learning approaches inspired by biological neural networks. They consist of nodes (neurons) that have associated values, and links between the nodes that have weights. In the standard implementation, ANNs are trained on data and the error that arises from comparing the output of the ANN with the known value for the observation is fed back to modify the weights of the relationships. ANNs have been used to identify a panel of ovarian cancer tumor markers from serum that could significantly outperform the best single biomarker [11]. An advantage of ANNs is that they can identify nonlinear relationships between variables, but they are prone to overfitting of data and produce 'black-box' models with minimal interpretability.

2.3.5. Relative gene expression analysis—One notable issue in identification of robust biomarkers is that the genetic and regulatory networks for individual patients will differ significantly. Thus consistent changes in particular genes, proteins or pathways may not be evident if compared against a non-specific background. For many diseases patient-specific control samples (from non-diseased tissue, for example) may not exist. One way to address the patient-to-patient background variability is to use relative expression analysis methods [50]. These methods examine the differences in expression levels between pairs of genes or proteins in diseased and non-diseased patients, identifying consistent relationships between features in diseased samples that reverse in non-diseased samples. One such approach, the top-scoring pairs (TSP) method, was used to identify gene signatures that could classify breast, prostate, and leukemia cancer patients using a minimal number of features [51]. Accuracy of this approach was comparable to other approaches using a much larger number of features.

2.4. Visualization

Bioinformatics studies often require a way to summarize data and analysis results in a way that is easily interpreted. This generally takes the form of visualization, thus allowing a user to look at an image and extract conclusions that would not be evident otherwise. An additional benefit is that the process involved in summarizing data and generating the images often becomes an analysis in itself, thus yielding novel results.

2.4.1. Principal component analysis (PCA)—This method is used to reduce the dimensionality of complex datasets, so that the important influences can be identified. As such, PCA is both a data reduction and a visualization method. In PCA, data is plotted along axes that represent orthogonal linear combinations of the original variables. In this way, components of the data that represent as much of the variance as possible are brought to light [52]. PCA is commonly used to determine if data classes are well separated when plotted on two to three principal components, thus demonstrating plausibility of successful biomarker identification. The resulting graphical representation often assists in visualizing the strength of the separation.

A recent application of PCA involved visualizing the expression pattern of kidney microRNA in the presence and absence of ischemic reperfusion injury in mice. The study demonstrated a strong separation in the expression patterns of injured vs. non-injured mice, thus justifying further studies to identify miRNA biomarkers of ischemic reperfusion injury.

A caveat to PCA is that a complete data matrix is required to compute the components. In many cases missing values can be imputed using straightforward model-based approaches [53–55], which works well for microarray data. However, with newer omic technologies, such as mass-spectrometry based proteomics and metabolomics, the data is left-censored and thus imputation of the missing data can cause a severe bias in the components due to the misrepresentation of the variance structure of the data. Methods such as projection pursuit [56] can be used to overcome the missing data challenge. Figure 1 displays a principal component analysis (PCA) when performed using a limit-of-detection (LOD) imputation (Figure 1A) versus a PCA analysis using an alternate approach, such as Sequential Projection Pursuit (SPP), that does not require imputation (Figure 1B) [57] In this example proteomics dataset there are two factors as associated with eight C57BL/6 mouse lung tissue samples as previously described [58] The first is diet induced obesity described as regular weight (RW) and obese (OB) and the second includes sham controls (SC) and exposure to lipopolysaccharide (LPS). Figure 1 is color coded to highlight the clear distinction in the obesity factor. Both the imputation-based PCA and SPP approaches appear to separate the obesity factor well, although the SPP approach visually shows a clearer pattern. The plots are overlaid with red triangles that represent all mouse samples that had >28% of the observed peptides categorized as missing (not detected in that sample). The LOD imputation approach shows a severe bias based on missing data, the red triangles are all clustered to the top and left of the plot. Furthermore, the scales of the axes are extreme in comparison to the SPP approach showing that the missing data is driving this separation, not a biological mechanism. Thus even common tasks such as PCA can be more accurately interpreted when expert knowledge is applied to the underlying the data structure.

2.4.2. Network analysis—The results of many data-driven techniques for analyzing large scale datasets are often best represented in network form. Using similarity measures such as Pearson's correlation, Euclidean distance, or mutual information, computational approaches can reveal similarity between entities' expression profiles, thus implying some kind of relationship. These kinds of networks can provide a very intuitive way of visualizing complicated data that reflects the underlying structure of the dataset. Figure 2A shows a

coexpression network inferred from this gene expression analysis in macrophages responding to application of nanoparticles [59]. Since the system proceeds through a series of steps over time, the network arranges nodes (genes) in a temporal arrangement, finally ending at a state very similar to its starting state. This elucidation of temporal relationships contrasts with the results shown in Figure 2B. This example of a standard heatmap visualization of gene expression data, with a number of modules identified by color to the right (as defined using hierarchical clustering) fails to reveal the temporal relationship between the clusters.

As discussed above, structural analysis can include identification of clusters or communities in the network, which can suggest mechanisms of regulation. Analysis of network structure can also identify nodes with high centrality, as determined by metrics such as degree (number of edges connected to a given node), or betweenness (number of shortest paths between all nodes that pass through a given node) [60]. Nodes with high centrality, particularly betweenness, have been shown to be enriched for critical regulator function [59, 61, 62]; thus construction of a network that represents biological relationships of some kind can lead to discovery of new regulatory mechanisms.

Data-driven network analysis was used in one study designed to identify novel markers for chronic lymphocytic leukemia (CLL). Networks based on correlation-derived connectivity revealed a cluster containing several known CLL markers. A set of new candidate biomarkers from this cluster was isolated by identifying genes whose expression was predictive of IgVH mutation, which is a known diagnostic tool for CLL [38]. In another approach termed SVM-RCE (Recursive Cluster Elimination), clusters from gene expression networks are iteratively screened for classification ability using an SVM. Clusters with low predictive power are removed and the network is rebuilt with the remaining genes. This process is repeated until a target number of clusters remains in the rebuilt network. Yousef et al, applied this approach to pre-existing published datasets from cutaneous T-cell lymphoma and from airway epithelial cells in smokers with and without lung cancer, comparing the predictive power of SVM-RCE to the original signatures published for the chosen datasets. In both cases there was a significant improvement in accuracy [63], illustrating the power of iterative analyses.

3. Knowledge-driven approaches to biomarker identification

The rapid maturation of genomics and proteomics technologies has overwhelmed scientists with a prodigious amount of high-throughput experimental data in the last two decades. The bottleneck for life science studies has shifted from generating the data to interpreting results so as to derive insights into biological mechanisms. The increasing use of systems biology approaches has prompted researchers to integrate heterogeneous data into existing knowledge bases in order to facilitate the understanding of disease and biological process mechanisms at a systematic level. In this section, we provide descriptions for several common knowledge-based approaches used in cancer biomarker discovery and examples of their latest applications.

3.1 Protein-Protein Interactions (PPI)

Protein-protein interactions (PPI), the physical binding interactions between proteins, play a key role in many cellular processes [64]. To understand the mechanisms underlying biological processes such as disease progression at a molecular level, it is critical to identify, characterize and interpret PPIs. Most PPI studies have focused on two areas: experimental identification and characterization of protein interactions (including populating protein and domain interaction databases based on experimental data); and application of computational approaches to predict protein and domain interactions based on the experimental findings.

In this section, we focus on the applications that integrate the available knowledge of protein interaction networks with experimental data sets in order to facilitate biomarker discovery. Although the complete interaction map of proteins from any species does not yet exist, some studies have estimated that the total number of protein interaction types (classes of distinct interactions in terms of structure) is limited to a rather small number, around 10,000 [65]. Considering that protein interactions are highly specific [66] and that our knowledge of PPIs has been rapidly increasing, it will be possible in the near future to predict and/or interpret new interactions based on existing interaction networks, which will have important implications to understanding the cellular networks that give rise to biomarkers.

Increasingly, knowledge-based approaches, including the integration of PPI networks with experimental and clinical datasets, have been applied in biomarker discovery. For instance, Xiong et al, applied PPIs to biomarker identification for lung cancer by extracting synergistic gene pairs from a microarray dataset of 66 samples [67] Specifically, the logic status of a PPI was determined by the relative expression of the corresponding gene pair, which was used in the assessment of cancer phenotype via a SVM as a classifier. A total of 16 gene pairs were identified with strong association with the phenotype for human lung cancer, and three of them (Pafah1b1-Ndel1, Cav1-Src and Nos3-Cav1) displayed a skewed distribution in cancer samples. In addition, a novel potential PPI between Src and Cav-1 was identified, contributing new insights into potential mechanisms of lung cancer. Ideker, et al. have combined decision trees with PPI networks allowing identification of subnetworks and combinatorial logic that relates them to cancer [68]. Though its application to biomarkers is not explicit, this study in particular demonstrates the value of combining statistical approaches with existing knowledge to derive enhanced biomarkers.

3.2. Pathway Analysis

Pathway analysis, i.e., the analysis of expression data for functionally related genes, is another form of knowledge that can be integrated into biomarker identification studies. Pathway analysis focuses on the identification of differentially expressed functionally related genes (pathways), rather than single genes, from gene expression data. The ultimate goal of this approach is to develop a comprehensive understanding of disease-related mechanisms at a molecular level [69]. Despite the varieties of methods available for specific pathway-based analysis, all have adopted a fairly similar perspective. In general, pathway-based analysis strategies consist of the following three components: i) choosing sets of genes, generally using a data-driven process (for example, differential expression); ii) asking a biologically relevant question (formulating a hypothesis) about functions that may be involved; and iii) choosing an effective statistical test to answer the question [70].

At the stage of choosing gene sets, the choice is made whether or not to pre-select sets of genes. Pre-selected sets of genes can be used to test specific hypotheses about whether specific pathways are significantly differentially expressed between phenotypes. Many conventional statistical tests can be used to answer this question (e.g. Fisher’s exact test and Chi-square test). However, this simple and straightforward approach suffers from several shortcomings, such as the inflation of the probability of Type I error in multiple hypothesis testing, and has lost popularity in large-scale data analysis [71].

One of the major disadvantages of an analysis that focuses on pre-selected gene sets is that it ignores all genes not included in the pre-selected list. Therefore, the current trend is to use global strategies that investigate all expressed genes without pre-selection. Among the approaches that avoid the use of pre-selected gene sets, Gene Set Enrichment Analysis (GSEA) has been widely accepted since its appearance [72, 73]. It essentially compares the difference in expression of a set of genes against the remainder of the genes between two

phenotypes; this is also referred as a competitive approach. We will discuss it separately in the next section.

Many approaches available for pathway analysis can be viewed as self-contained tests, which try to answer the question; “Is one gene set differentially expressed between distinct phenotypes?” Self-contained tests can be categorized into either univariate or multivariate tests. In general, these tests are easy to interpret and are often favored over tests that compare one set of genes against a larger set, so-called competitive tests, like GSEA [74]. Pathway analysis approaches can also be used to compare differential expression of functional groups when individual genes may not be shared between two systems being compared. We have recently described one such approach in which significantly enriched functional groups in different systems were used to compare the responses of these systems to infection with influenza over time [75].

The ability of pathway analysis approaches to identify common functional elements in noisy data sets is illustrated by a recent analysis comparing the results of plasma protein-based and cell line-based proteomic analyses [76]. Mass spectrometry based analysis of plasma samples from 40 women diagnosed with breast cancer and 40 healthy controls identified 254 statistically differentially expressed proteins, of which 25 were further classified as “activated” plasma proteins based on the pathway analysis and literature curation to serve as pathway biomarker candidates. The top three enriched pathways included complement and coagulation cascades, regulation of actin cytoskeleton and focal adhesion. Cross-validation against two proteomics studies using breast cancer cell lines showed that there was a higher degree of similarity between cell lines and plasma at the level of pathways, compared to individual proteins. [62].

3.2.1. Gene-Set Enrichment Analysis (GSEA)—Gene Set Enrichment Analysis (GSEA) is a strategy for gene expression data analysis based on pathway knowledge that has had a significant influence on the general framework for analyzing high-throughput gene expression data. This computational method focuses on finding statistically significant differences between two biological states, e.g., phenotypes, using sets of functionally related genes (chosen from prior biological knowledge, i.e., knowledge-based), rather than individual genes [72, 73]. The three key elements involved in this method are: 1) ranking all genes in a dataset according to their expression differences between two biological states, and calculating an enrichment score (ES) for each gene set; 2) estimating the significance level of each ES using a permutation test procedure; 3) adjusting for multiple hypothesis testing.

Since its appearance, the gene-set-based strategy has been widely applied to many types of datasets and has demonstrated significant advantages, including robustness and biological relevance. Besides changing the focus from individual genes to groups of genes, another underlying advantage of this approach is its ability to extract common features from datasets derived from different platforms at the level of the functionally related gene-set rather than the single-gene. Using GSEA, two independent studies in lung cancer were compared and showed a strong correlation and a large overlap of the significantly enriched gene sets between the two studies [72]. This is significant because traditional analyses comparing differentially expressed genes found no significant similarities between the two datasets, highlighting the power of using knowledge-based approaches for noisy and multi-source data.

To extend the range of applicability of GSEA, several modifications have been made to its basic framework. For example, parametric analysis of gene set enrichment (PAGE) was developed for larger datasets and to address the need to decrease computation time [77];

Jiang et. al, have proposed a method that allows for adjustments based on other covariates [78]; and Woolf and colleagues have extended this approach to a generally applicable gene set enrichment (GAGE) for handling pathway analyses for datasets with different sample sizes and experimental designs [79]. However, some researchers have challenged the sensitivity of the GSEA approach and shown that in some cases two simple procedures based on the one-sided z-test and the χ^2 test outperform GSEA [80].

Here we describe two recent studies that used GSEA and GSEA-related approaches to discover biomarker candidates. In a lung cancer study testing the hypothesis that increased angiogenesis was related to decreased survival in non-small cell lung cancer, GSEA approaches identified a panel of regulatory microRNAs (including miR-155, miR-21, and miR-106a) that were significantly associated with both angiogenesis and decreased survival [81]. In another study, GSEA was used to identify signaling pathways associated with the response of cervical tumors to chemoradiation therapy, as monitored by glucose uptake. GSEA analysis identified over-expression of the P13K/Akt pathway in association with an incomplete metabolic response to therapy. Since an incomplete metabolic response is known to be associated with poor survival, these results suggested that targeted inhibition of the PI3K/AKT pathway may improve patient response to chemotherapy [82].

3.3. Text Mining

Text mining has a long and varied history outside of the bioscience field. This technique started to appear in biomedical literature in the late 1990s and has experienced a surge in popularity over the last decade [83, 84]. Biomedical text mining refers to the use of automated methods to explore the prodigious amount of knowledge available in the existing biomedical literature to benefit researchers. In contrast to other fields of application, the most popular text mining tools used in both bioscience and medical fields have been developed by bioscientists, rather than text mining specialists. Examples include the applications Chilibot, Textpresso, and PreBIND [85].

In general, there are three major steps involved in biomedical text mining: i) recognizing terms, ii) looking for relationships between these terms, and iii) discovering new relationships. The task of recognizing various terms, also referred to as named–entity recognition, is a process by which information is retrieved from selected and/or relevant documents/literature by computer scanning. A variety of entities, such as protein name, cell type, gene mutation, and disease, can be recognized during the process [86]. Some algorithms and machine-learning tools can be designed to consistently recognize an individual entity under different names, synonyms, homonyms, and acronyms [87, 88].

The second step is looking for relationships between terms, i.e., information extraction. The simplest and most intuitive way to identify relationships is by using co-occurrence-based methods. This type of method looks for concepts occurring in the same unit of text (for example, a sentence or an abstract) under the assumption that terms that show up in the same place are most likely related to each other, e.g., a gene mutation and a disease are often mentioned in the same abstract [89]. However, two other sophisticated methods, consisting of knowledge-based and statistical approaches, are more commonly employed. Knowledge-based methods are self-explanatory and involve integrating a general knowledge of linguistics and biology at certain levels in order to recognize a relationship between terms. In contrast, the statistical approaches apply classifiers at different levels of text. In practice, the three approaches can be fruitfully combined. For example, co-occurrence can be used as an initial baseline before the statistical process step, followed by knowledge-based post-processing.

At the third stage, text mining is used to unearth previously undiscovered relationships that are hidden in the literature. These relationships can be used to prioritize biomarkers by highlighting important biological links between proteins and known pathways of importance, for example. The explosion of scientific data has made it virtually impossible for life scientists to read everything related to their research projects from either a historical or contemporary perspective. The use of text mining techniques can significantly reduce the time required for searching relevant literature, and dramatically increase the chances for discovering new connections.

In biomarker discovery studies, text mining can be used as a powerful discovery and validation tool. For example, Deng et al, have developed a multi-platform strategy, link-test, to cross-link experimental datasets at both transcriptomic (microarray data) and proteomic (mass spectrometry data) levels in an effort to discover new prostate cancer biomarkers which could outperform the current biomarker, prostate specific antigen or PSA. Cross-validation results showed high prediction accuracy using the identified biomarker candidates, which were further validated by text mining of prostate-cancer-related genes from OMIM (Online Mendelian Inheritance in Man) [90]. Another example, PubMeth, is a freely accessible cancer methylation database that combines a text mining approach with manual reviewing and annotation. It can be used as an efficient way to discover novel methylation markers of cancers. The earlier version of this approach was first reported in 2008 [91]. More recently, the authors have been able to combine experimental data with literature results with a co-occurrence-based method [92].

4. Evaluation of biomarker identification approaches

Development of approaches to identification of robust biomarkers requires careful attention to methods to evaluate performance. A common problem with many of the approaches discussed above is that use of an inappropriate evaluation scheme can result in vastly overstated performance results [43]. In terms of biomarker identification, this has certainly been a source of many problems contributing to the lack of robustness of identified biomarkers [93].

Common approaches used to evaluate the quality of biomarker signatures include the determination of Receiver Operating Characteristics (ROC) and the measurement of the area under the curve (AUROC). This is accomplished by varying some stringency parameter (p-value cutoff, for example) and plotting the specificity versus 1-sensitivity. The AUROC is equal to 1.0 if the method classifies all positive and negative examples correctly, and 0.5 for random class assignment. This metric has the advantage of combining two important components, the type I and type II error rates, or i.e., the portion of false positive predictions and false negative predictions that are made by the method.

Proper evaluation of a biomarker identification method involves establishing a set of experimental data that will be used to parameterize the method and a separate, independent data set that can be used to test the predictions made for consistency. Evaluation of the performance of the method on the independent test set is referred to as cross-validation, and is commonly repeated with different partitions to arrive at an estimate of performance. Experimental validation of identified biomarkers in a completely independent data set representing an appropriate experimental system can provide very good evaluation, but is generally prohibitive in terms of time and money, and thus is rarely performed in a comprehensive fashion. In many cases approaches based on either cross-validation or bootstrapping are used to evaluate data sets [94].

Bootstrapping refers to the process of repeatedly dividing a single data set into training and testing sets that can be used to assess the overall performance of the classification method.

The bootstrapping paradigm, and other resampling techniques, can provide estimates of classification performance on 'independent' data, but can suffer from unknown or unaccounted for relationships in the data. That is, if different observations are not independent, bootstrapping methods can still provide overstated results [94]. However, the down-selection of biomarkers can be effectively performed in this manner if the bootstrapping algorithm is employed within the feature selection. This is computationally expensive, but can dramatically improve the identification of potential candidate biomarkers when independent test data is not available [94]. In addition, a particular problem for biomarker identification is that individual datasets often are themselves biased, for example, by selection of subjects. This means that biomarkers identified from one set of subjects may not work in another set of subjects selected using different criteria [43]. Meta-analysis of data from multiple independent studies is one way to address this issue, but requires that the measurements taken by each study be comparable, a very difficult proposition given the plethora of platforms for data generation and the wide variety of data processing pipelines. These issues can be partially solved by employing non-parametric versions of gene pair expression methods, such as the top scoring pairs algorithm (see above).

5. Expert Opinion

The amount of thought and effort devoted to the identification and characterization of biomarkers has exploded over the past decade, accompanied by a substantial increase in the number and sophistication of analytical tools available for selecting candidate biomarkers from complex datasets. Our review has focused on describing the fundamental tool sets currently available, and the development of increasingly sophisticated approaches. With over 3000 publications on 'biomarker discovery' listed in PubMed for the past 5 years, and 461 in 2011 alone, it would be impossible to comprehensively review each individual effort. Yet despite this well-documented effort, the FDA has been approving only 1 to 3 new biomarkers for clinical use each year [10]. Clearly, the current procedures for identifying biomarkers that can withstand a rigorous validation process and subsequently demonstrate true clinical utility are not working.

We propose a new synthesis of data-driven and expert knowledge-based approaches to concurrently mine the power of statistical algorithms for selection while guiding the process to include weighting factors derived from expert-knowledge. Bayesian approaches have always included weighting factors by assigning prior probabilities, but there are many different ways to accomplish this. This is both a problem of quantifying expert knowledge and of identifying the appropriate knowledge bases to utilize that capture expert knowledge pertinent to the question being posed. Expert knowledge exists in the form of existing experimental data, previous information from experts in the field, and community-assembled knowledge sources such as functional annotation. Integrating these kinds of knowledge with powerful statistical approaches has the promise of identifying more robust and clinically relevant biomarkers. One potential benefit of this combination is that expert knowledge can be used to filter out spurious or low-quality biomarkers to yield genes, proteins, or metabolites that are able to more accurately and confidently predict, diagnose, or quantify disease. A second benefit is that this kind of analysis can help identify higher levels of abstraction, for example groups of genes related in a pathway, that serve as improved biomarkers, by virtue of their ability to capture many variants of an adversely affected disease-related process. Many approaches using Bayesian models that employ standard priors without specifying particular variables fail to truly capture the diversity of biological information [95]. One relatively recent approach has developed an empirical Bayes method in which prior information about pathways and networks is selected and weighed in an automated objective process [96].

The development of a semi-automated, iterative process that can be optimized around a metric of biomarker quality offers a potential solution for testing the effects of different expert parameters. The metric used for evaluation should reflect the fidelity of the biomarker with standard measures of binary classification performance: accuracy, sensitivity, specificity, and especially positive predictive value. Ground truth for these studies is the phenotypic expression of each subject under study, for example survival times or other clinical indicators of disease. The metric should also incorporate the risk associated with the measurement. For mature diagnostic tests, this would account for the consequences of incorrectly predicting the onset (or non-onset) of a disease. It is also useful to incorporate an element of cost in the assessment. The cost factor must be appropriate to the stage of the test. While out-of-pocket cost to the patient is an appropriate cost element for a mature clinical test, it must be remembered that experimental verification and validation also includes costs for sample acquisition and sample analysis.

We are currently developing an approach that relies predominantly on the use of expert knowledge to determine the structure of the statistical analyses, and to combine the results of statistical analyses in a manner that incorporates the known complementarity of the underlying biological functions. An example of this process would involve the initial application of a data reduction approach, such as PCA or supervised clustering, on an unselected global dataset (such as a transcriptomic or proteomic analysis). The features identified as discriminatory between the desired phenotypes would then be subjected to a functional analysis, such as GSEA, to identify the functional components underlying the statistical separation. Once the functional groups responsible for discrimination have been identified, statistical approaches are used to identify those features within a functional group that were the most robust surrogate for the behavior of the group using a signature quality metric defined for the clinical application, as described above. This process can be iterated to optimize the signature quality metric until the desired performance had been obtained.

Iteration might proceed by first identifying groups of genes (e.g., from broad functional groups), that provide the best classification between control and disease samples. These groups could then be broken into successively more detailed and informative subgroups, and classification methods re-parameterized. In this way groups of genes could be identified that were optimally informative from the standpoint of disease classification as well as the underlying biological functions. For many disease processes, including cancer, this approach could provide biomarkers that are more robust than traditional approaches.

While the overall focus of this process is to improve the performance of biomarkers for the specific category of intended use (such as early detection, prognosis, or response to therapy), the melding of statistical and expert-driven approaches also ensures that there is a discernible biological rationale underlying the choice of biomarkers. This direct link to biological function has the potential to drive the development of improved therapeutic strategies by identifying the complementary and/or parallel biological functions that contribute most strongly to discrimination between cases and controls.

Acknowledgments

Preparation of this review was conducted under Signatures Discovery Initiative, a component of the Laboratory Directed Research and Development Program at Pacific Northwest National Laboratory, a multiprogram national laboratory operated by Battelle for the U.S. Depart-ment of Energy under Contract DE-AC05-76RL01830. Authors received additional support from the National Cancer Institute's Early Detection Research Network [IAA Y1-CN-0002-1 to KDR] and Clinical Proteomic Tumor Analysis Consortium [CA160019 to KDR and JEM].

References

1**. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther. 2001; 69(3):89–95. Commentary. Provides detailed definitions and concepts of different types of biomarkers. [PubMed: 11240971]

2**. Ludwig JA, Weinstein JN. Biomarkers in cancer staging, prognosis and treatment selection. Nat Rev Cancer. 2005; 5(11):845–856. Review. A comprehensive review on different types of cancer biomarkers. [PubMed: 16239904]

3. Executive summary: Standards of medical care in diabetes. Diabetes Care. 2010; 33(Suppl 1):S4–S10. [PubMed: 20042774]

4. Kim C, Tang G, Pogue-Geile KL, Costantino JP, Baehner FL, Baker J, Cronin MT, Watson D, Shak S, Bohn OL, et al. Estrogen receptor (ESR1) mRNA expression and benefit from tamoxifen in the treatment and prevention of estrogen receptor-positive breast cancer. J Clin Oncol. 2011; 29 (31): 4160–4167. [PubMed: 21947828]

5. Konecny G, Slamon DJ. HER2 testing and correlation with efficacy of trastuzumab therapy. Oncology (Williston Park). 2002; 16(11):1576, 1578. [PubMed: 12469932]

6. Friedman LS, Ostermeyer EA, Lynch ED, Szabo CI, Anderson LA, Dowd P, Lee MK, Rowell SE, Boyd J, King MC. The search for BRCA1. Cancer Res. 1994; 54 (24):6374–6382. [PubMed: 7987831]

7*. Welcsh PL, King MC. BRCA1 and BRCA2 and the genetics of breast and ovarian cancer. Hum Mol Genet. 2001; 10(7):705–713. Review. Describes biological functions of BRCA1/2 in DNA repair and transcriptional regulation. [PubMed: 11257103]

8**. Welcsh PL, Owens KN, King MC. Insights into the functions of BRCA1 and BRCA2. Trends Genet. 2000; 16(2):69–74. Review. A comprehensive review on biological functions of BRCA1 and BRCA2. [PubMed: 10652533]

9*. Bast RC Jr. CA 125 and the detection of recurrent ovarian cancer: a reasonably accurate biomarker for a difficult disease. Cancer. 2010; 116(12):2850–2853. Commentary. Describes several applications of CA125 serving as a biomarker for ovarian cancer. [PubMed: 20564390]

10*. Zhang Z, Chan DW. The road from discovery to clinical diagnostics: lessons learned from the first FDA-cleared in vitro diagnostic multivariate index assay of proteomic biomarkers. Cancer Epidemiol Biomarkers Prev. 2010; 19(12):2995–2999. Uses Ova1 as a demonstration of the lessons, challenges and strategies associated with identification of cancer biomarkers in clinical diagnostics. [PubMed: 20962299]

11. Zhang Z, Yu Y, Xu F, Berchuck A, van Haaften-Day C, Havrilesky LJ, de Bruijn HW, van der Zee AG, Woolas RP, Jacobs IJ, et al. Combining multiple serum tumor markers improves detection of stage I epithelial ovarian cancer. Gynecol Oncol. 2007; 107(3):526–531. [PubMed: 17920110]

12. Zhang Z, Bast RC Jr, Yu Y, Li J, Sokoll LJ, Rai AJ, Rosenzweig JM, Cameron B, Wang YY, Meng XY, et al. Three biomarkers identified from serum proteomic analysis for the detection of early stage ovarian cancer. Cancer Res. 2004; 64(16):5882–5890. [PubMed: 15313933]

13*. Chou R, Croswell JM, Dana T, Bougatsos C, Blazina I, Fu R, Gleitsmann K, Koenig HC, Lam C, Maltz A, et al. Screening for prostate cancer: a review of the evidence for the U.S. Preventive Services Task Force. Ann Intern Med. 155(11):762–771. Discussion of utility issues that must be met for cost-effective use of biomarkers, using PSA as the example. [PubMed: 21984740]

14. Daberkow D 2nd. Screening for asymptomatic cancers. J La State Med Soc. 1997; 149(8):285–290. [PubMed: 9260456]

15. Bertenshaw GP, Yip P, Seshaiah P, Zhao J, Chen TH, Wiggins WS, Mapes JP, Mansfield BC. Multianalyte profiling of serum antigens and autoimmune and infectious disease molecules to identify biomarkers dysregulated in epithelial ovarian cancer. Cancer Epidemiol Biomarkers Prev. 2008; 17 (10):2872–2881. [PubMed: 18843033]

16. Mor G, Visintin I, Lai Y, Zhao H, Schwartz P, Rutherford T, Yue L, Bray-Ward P, Ward DC. Serum protein markers for early detection of ovarian cancer. Proc Natl Acad Sci U S A. 2005; 102(21):7677–7682. [PubMed: 15890779]

17. Merritt MA, Parsons PG, Newton TR, Martyn AC, Webb PM, Green AC, Papadimos DJ, Boyle GM. Expression profiling identifies genes involved in neoplastic transformation of serous ovarian cancer. BMC Cancer. 2009; 9 :378. [PubMed: 19849863]

18. Amonkar SD, Bertenshaw GP, Chen TH, Bergstrom KJ, Zhao J, Seshaiah P, Yip P, Mansfield BC. Development and preliminary evaluation of a multivariate index assay for ovarian cancer. PLoS One. 2009; 4(2):e4599. [PubMed: 19240799]

19. Yurkovetsky Z, Skates S, Lomakin A, Nolen B, Pulsipher T, Modugno F, Marks J, Godwin A, Gorelik E, Jacobs I, et al. Development of a multimarker assay for early detection of ovarian cancer. J Clin Oncol. 2010; 28(13):2159–2166. [PubMed: 20368574]

20. Mok SC, Wong KK, Chan RK, Lau CC, Tsao SW, Knapp RC, Berkowitz RS. Molecular cloning of differentially expressed genes in human epithelial ovarian cancer. Gynecol Oncol. 1994; 52(2): 247–252. [PubMed: 8314147]

21. Palmer C, Duan X, Hawley S, Scholler N, Thorpe JD, Sahota RA, Wong MQ, Wray A, Bergan LA, Drescher CW, et al. Systematic evaluation of candidate blood markers for detecting ovarian cancer. PLoS One. 2008; 3(7):e2633. [PubMed: 18612378]

22. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, et al. A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med. 2002; 347(25):1999–2009. [PubMed: 12490681]

23. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, Baehner FL, Walker MG, Watson D, Park T, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med. 2004; 351(27):2817–2826. [PubMed: 15591335]

24. Fan C, Oh DS, Wessels L, Weigelt B, Nuyten DS, Nobel AB, van't Veer LJ, Perou CM. Concordance among gene-expression-based predictors for breast cancer. N Engl J Med. 2006; 355(6):560–569. [PubMed: 16899776]

25. Mosley JD, Keri RA. Cell cycle correlated genes dictate the prognostic power of breast cancer gene lists. BMC Med Genomics. 2008; 1:11. [PubMed: 18439262]

26. Desmedt C, Sotiriou C. Proliferation: the most prominent predictor of clinical outcome in breast cancer. Cell Cycle. 2006; 5(19):2198–2202. [PubMed: 16969100]

27. Venet D, Dumont JE, Detours V. Most random gene expression signatures are significantly associated with breast cancer outcome. PLoS Comput Biol. 2011; 7(10):e1002240. [PubMed: 22028643]

28. Guttery DS, Hancox RA, Mulligan KT, Hughes S, Lambe SM, Pringle JH, Walker RA, Jones JL, Shaw JA. Association of invasion-promoting tenascin-C additional domains with breast cancers in young women. Breast Cancer Res. 2010; 12(4):R57. [PubMed: 20678196]

29. Autier P, Boniol M, Gavin A, Vatten LJ. Breast cancer mortality in neighbouring European countries with different levels of screening but similar access to treatment: trend analysis of WHO mortality database. Bmj. 2011; 343:d4411. [PubMed: 21798968]

30*. Nugent R, Meila M. An overview of clustering applied to molecular biology. Methods Mol Biol. 2010; 620:369–404. Review. A broad overview of both attribute- and similarity-based clustering as applied to biological problems. A good place to start to get an idea of what clustering methods are commonly used and how they work. [PubMed: 20652512]

31*. Boutros PC, Okey AB. Unsupervised pattern recognition: an introduction to the whys and wherefores of clustering microarray data. Brief Bioinform. 2005; 6(4):331–343. Review. Discusses integrating unsupervised machine-learning techniques with biological information. Provides examples of how different approaches to clustering microarray data can be used. [PubMed: 16420732]

32. O'Dwyer D, Ralton LD, O'Shea A, Murray GI. The proteomics of colorectal cancer: identification of a protein signature associated with prognosis. PLoS One. 2011; 6(11):e27718. [PubMed: 22125622]

33. Conrads TP, Anderson GA, Veenstra TD, Pasa-Tolic L, Smith RD. Utility of accurate mass tags for proteome-wide protein identification. Anal Chem. 2000; 72(14):3349–3354. [PubMed: 10939410]

34*. Anil J. Data clustering: 50 years beyond K-means. Pattern Recognition Letters. 2010; 31(8):651–666. Overview of clustering, clustering methods and new algorithms in clustering. Discusses challenges in selection of clustering algorithms and comparison of clustering methods.

35*. Kerr G, Ruskin HJ, Crane M, Doolan P. Techniques for clustering gene expression data. Comput Biol Med. 2008; 38(3):283–293. Review. Addresses the limitations in clustering, and provides a framework of the evaluation of clustering gene expression analyses. [PubMed: 18061589]

36*. Madeira SC, Oliveira AL. Biclustering algorithms for biological data analysis: a survey. IEEE/ACM Trans Comput Biol Bioinform. 2004; 1(1):24–45. Survey. A intensive summary of methods and applications of biclustering in biology. [PubMed: 17048406]

37. Aranday-Cortes E, Hogarth PJ, Kaveh DA, Whelan AO, Villarreal-Ramos B, Lalvani A, Vordermeier HM. Transcriptional profiling of disease-induced host responses in bovine tuberculosis and the identification of potential diagnostic biomarkers. PLoS One. 2012; 7(2):e30626. [PubMed: 22359547]

38. Zhang J, Xiang Y, Ding L, Keen-Circle K, Borlawsky TB, Ozer HG, Jin R, Payne P, Huang K. Using gene co-expression network analysis to predict biomarkers for chronic lymphocytic leukemia. BMC Bioinformatics. 2010; 11 (Suppl 9):S5.

39. Zucknick M, Richardson S, Stronach EA. Comparing the characteristics of gene expression profiles derived by univariate and multivariate classification methods. Stat Appl Genet Mol Biol. 2008; 7(1):Article7. [PubMed: 18312212]

40. Wang H, Gottfries J, Barrenas F, Benson M. Identification of novel biomarkers in seasonal allergic rhinitis by combining proteomic, multivariate and pathway analysis. PLoS One. 2011; 6(8):e23563. [PubMed: 21887273]

41. Gentles AJ, Plevritis SK, Majeti R, Alizadeh AA. Association of a leukemic stem cell gene expression signature with clinical outcomes in acute myeloid leukemia. Jama. 2010; 304(24): 2706–2715. [PubMed: 21177505]

42*. Noble WS. What is a support vector machine? Nat Biotechnol. 2006; 24 (12):1565–1567. Primer. An excellent short primer on SVMs and their applications for classification in large-scale biological data. [PubMed: 17160063]

43. Barla A, Jurman G, Riccadonna S, Merler S, Chierici M, Furlanello C. Machine learning methods for predictive proteomics. Brief Bioinform. 2008; 9 (2):119–128. [PubMed: 18310105]

44. Schrauder MG, Strick R, Schulz-Wendtland R, Strissel PL, Kahmann L, Loehberg CR, Lux MP, Jud SM, Hartmann A, Hein A, et al. Circulating micro-RNAs as potential blood-based markers for early stage breast cancer detection. PLoS One. 2012; 7(1):e29770. [PubMed: 22242178]

45. Johansson H, Lindstedt M, Albrekt AS, Borrebaeck CA. A genomic biomarker signature can predict skin sensitizers using a cell-based in vitro alternative to animal tests. BMC Genomics. 2011; 12:399. [PubMed: 21824406]

46*. Kingsford C, Salzberg SL. What are decision trees? Nat Biotechnol. 2008; 26(9):1011–1013. Primer. A primer on how decision tree approaches work and how they are applied to classification problems in biology. [PubMed: 18779814]

47. Diaz-Uriarte R, Alvarez de Andres S. Gene selection and classification of microarray data using random forest. BMC Bioinformatics. 2006; 7:3. [PubMed: 16398926]

48. O'Bryant SE, Xiao G, Barber R, Huebinger R, Wilhelmsen K, Edwards M, Graff-Radford N, Doody R, Diaz-Arrastia R. A blood-based screening tool for Alzheimer's disease that spans serum and plasma: findings from TARC and ADNI. PLoS One. 2011; 6(12):e28092. [PubMed: 22163278]

49. van Dijk SJ, Feskens EJ, Heidema AG, Bos MB, van de Rest O, Geleijnse JM, de Groot LC, Muller M, Afman LA. Plasma protein profiling reveals protein clusters related to BMI and insulin levels in middle-aged overweight subjects. PLoS One. 2010; 5(12):e14422. [PubMed: 21203453]

50**. Eddy JA, Sung J, Geman D, Price ND. Relative expression analysis for molecular cancer diagnosis and prognosis. Technol Cancer Res Treat. 2010; 9(2):149–159. Review. In-depth review of relative expression approaches to identification of biomarkers in cancer. [PubMed: 20218737]

51. Geman D, d'Avignon C, Naiman DQ, Winslow RL. Classifying gene expression profiles from pairwise mRNA comparisons. Stat Appl Genet Mol Biol. 2004; 3:Article19. [PubMed: 16646797]

52. Raychaudhuri S, Stuart JM, Altman RB. Principal components analysis to summarize microarray experiments: application to sporulation time series. Pac Symp Biocomput. 2000:455–466. [PubMed: 10902193]

53. Celton M, Malpertuy A, Lelandais G, de Brevern AG. Comparative analysis of missing value imputation methods to improve clustering and interpretation of microarray experiments. BMC Genomics. 2010; 11:15. [PubMed: 20056002]

54. Liew AW, Law NF, Yan H. Missing value imputation for gene expression data: computational techniques to recover missing data from available information. Brief Bioinform. 2011; 12(5):498–513. [PubMed: 21156727]

55. Schafer JL. Multiple imputation: a primer. Stat Methods Med Res. 1999; 8 (1):3–15. [PubMed: 10347857]

56. Webb-Robertson BJM, Jarman KH, Harvey SD, Posse C, Wright BW. An improved optimization algorithm and Bayes factor termination criterion for sequential projection pursuit. Chemometrics and Intelligent Laboratory Systems. 2005; 77(1–2):149–160.

57. Webb-Robertson BJ, Jarman KH, Scott HD, Posse C, Wright BW. An improved optimization algorithm and a Bayes factor termination criterion for sequential projection pursuit. Chem Intell Lab Sys. 2005; 77:149–160.

58. Webb-Robertson BJ, Matzke MM, Jacobs JM, Pounds JG, Waters KM. A statistical selection strategy for normalization procedures in LC-MS proteomics experiments through dataset-dependent ranking of normalization scaling factors. Proteomics. 2011; 11(24):4736–4741. [PubMed: 22038874]

59. McDermott JE, Archuleta M, Thrall BD, Adkins JN, Waters KM. Controlling the response: predictive modeling of a highly central, pathogen-targeted core response module in macrophage activation. PLoS ONE. 2011; 6(2):e14673. [PubMed: 21339814]

60**. Mason O, Verwoerd M. Graph theory and networks in Biology. IET Syst Biol. 2007; 1(2):89–119. A comprehensive technical description of graph theory and network applications in the biology field. [PubMed: 17441552]

61. McDermott JE, Archuleta M, Stevens SL, Stenzel-Poore MP, Sanfilippo A. Defining the players in higher-order networks: predictive modeling for reverse engineering functional influence networks. Pac Symp Biocomput. 2011:314–325. [PubMed: 21121059]

62. Diamond DL, Syder AJ, Jacobs JM, Sorensen CM, Walters KA, Proll SC, McDermott JE, Gritsenko MA, Zhang Q, Zhao R, et al. Temporal proteome and lipidome profiles reveal hepatitis C virus-associated reprogramming of hepatocellular metabolism and bioenergetics. PLoS Pathog. 2010; 6(1):e1000719. [PubMed: 20062526]

63. Yousef M, Ketany M, Manevitz L, Showe LC, Showe MK. Classification and biomarker identification using gene network modules and support vector machines. BMC Bioinformatics. 2009; 10:337. [PubMed: 19832995]

64. Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dumpelfeld B, et al. Proteome survey reveals modularity of the yeast cell machinery. Nature. 2006; 440(7084):631–636. [PubMed: 16429126]

65. Aloy P, Pichaud M, Russell RB. Protein complexes: structure prediction challenges for the 21st century. Curr Opin Struct Biol. 2005; 15(1):15–22. [PubMed: 15718128]

66. Shoemaker BA, Panchenko AR. Deciphering protein-protein interactions. Part I. Experimental techniques and databases. PLoS Comput Biol. 2007; 3 (3):e42. [PubMed: 17397251]

67. Xiong J, Liu J, Rayner S, Li Y, Chen S. Protein-protein interaction reveals synergistic discrimination of cancer phenotype. Cancer Inform. 2010; 9 :61–66. [PubMed: 20458363]

68. Dutkowski J, Ideker T. Protein networks as logic functions in development and cancer. PLoS Comput Biol. 2011; 7(9):e1002180. [PubMed: 21980275]

69*. Schadt EE. Molecular networks as sensors and drivers of common human diseases. Nature. 2009; 461(7261):218–223. Review. Describes the linkages between different types of large-scale and high-dimensional molecular data and complex disease networks focusing on genetic information. [PubMed: 19741703]

70. Emmert-Streib F, Glazko GV. Pathway analysis of expression data: deciphering functional building blocks of complex diseases. PLoS Comput Biol. 2011; 7(5):e1002053. [PubMed: 21637797]

71. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009; 4(1):44–57. [PubMed: 19131956]

72*. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America. 2005; 102(43):15545–15550. Describes the algorithm employed in the gene set enrichment analysis (GSEA). [PubMed: 16199517]

73. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E, et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet. 2003; 34 (3):267–273. [PubMed: 12808457]

74. Goeman JJ, Buhlmann P. Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics. 2007; 23(8):980–987. [PubMed: 17303618]

75. McDermott J, Shankaran H, Eisfeld A, Belisle S, Neuman G, Li C, McWeeney S, Sabourin C, Kawaoka Y, Katze M, et al. Conserved host response to highly pathogenic avian influenza virus infection in human cell culture, mouse and macaque model systems. BMC Syst Biol. 2011 in press.

76. Zhang F, Chen JY. Discovery of pathway biomarkers from coupled proteomics and systems biology methods. BMC Genomics. 2010; 11(Suppl 2):S12.

77. Kim SY, Volsky DJ. PAGE: parametric analysis of gene set enrichment. BMC Bioinformatics. 2005; 6:144. [PubMed: 15941488]

78. Jiang Z, Gentleman R. Extensions to gene set enrichment. Bioinformatics. 2007; 23(3):306–313. [PubMed: 17127676]

79. Luo W, Friedman MS, Shedden K, Hankenson KD, Woolf PJ. GAGE: generally applicable gene set enrichment for pathway analysis. BMC Bioinformatics. 2009; 10:161. [PubMed: 19473525]

80. Irizarry RA, Wang C, Zhou Y, Speed TP. Gene set enrichment analysis made simple. Stat Methods Med Res. 2009; 18(6):565–575. [PubMed: 20048385]

81. Donnem T, Fenton CG, Lonvik K, Berg T, Eklo K, Andersen S, Stenvold H, Al-Shibli K, Al-Saad S, Bremnes RM, et al. MicroRNA signatures in tumor tissue related to angiogenesis in non-small cell lung cancer. PLoS One. 2012; 7 (1):e29671. [PubMed: 22295063]

82. Schwarz JK, Payton JE, Rashmi R, Xiang T, Jia Y, Huettner P, Rogers BE, Yang Q, Watson M, Rader JS, et al. Pathway-Specific Analysis of Gene Expression Data Identifies the PI3K/Akt Pathway as a Novel Therapeutic Target in Cervical Cancer. Clinical cancer research : an official journal of the American Association for Cancer Research. 2012; 18(5):1464–1471. [PubMed: 22235101]

83. Evans JA, Rzhetsky A. Advancing science through mining libraries, ontologies, and communities. The Journal of biological chemistry. 2011; 286(27):23659–23666. [PubMed: 21566119]

84. Tanabe L, Scherf U, Smith LH, Lee JK, Hunter L, Weinstein JN. MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling. Biotechniques. 1999; 27(6):1210–1214. 1216–1217. [PubMed: 10631500]

85. Cohen KB, Hunter L. Getting started in text mining. PLoS Comput Biol. 2008; 4 (1):e20. [PubMed: 18225946]

86. Krauthammer M, Nenadic G. Term identification in the biomedical literature. J Biomed Inform. 2004; 37(6):512–526. [PubMed: 15542023]

87. Rzhetsky A, Seringhaus M, Gerstein MB. Getting started in text mining: part two. PLoS Comput Biol. 2009; 5(7):e1000411. [PubMed: 19649304]

88. Rodriguez-Esteban R. Biomedical text mining and its applications. PLoS Comput Biol. 2009; 5(12):e1000597. [PubMed: 20041219]

89*. Jenssen TK, Laegreid A, Komorowski J, Hovig E. A literature network of human genes for high-throughput analysis of gene expression. Nat Genet. 2001; 28(1):21–28. An early example of

automated extraction of gene-to-gene co-citation networks using publicly available databases. [PubMed: 11326270]

90. Deng X, Geng H, Bastola DR, Ali HH. Link test--A statistical method for finding prostate cancer biomarkers. Comput Biol Chem. 2006; 30(6):425–433. [PubMed: 17126079]
91. Ongenaert M, Van Neste L, De Meyer T, Menschaert G, Bekaert S, Van Criekinge W. PubMeth: a cancer methylation database combining text-mining and expert annotation. Nucleic acids research. 2008; 36(Database issue):D842–846. [PubMed: 17932060]
92. Ongenaert M. Epigenetic databases and computational methodologies in the analysis of epigenetic datasets. Adv Genet. 2010; 71:259–295. [PubMed: 20933132]
93**. Feng Z, Prentice R, Srivastava S. Research issues and strategies for genomic and proteomic biomarker discovery and validation: a statistical perspective. Pharmacogenomics. 2004; 5(6): 709–719. A perspective piece that describes the statistical requirements, challenges, and strategies for the discovery and validation of disease-related biomarkers focusing on early detection of cancer. [PubMed: 15335291]
94. Jiang W, Simon R. A comparison of bootstrap methods and an adjusted bootstrap approach for estimating the prediction error in microarray classification. Statistics in medicine. 2007; 26(29): 5320–5334. [PubMed: 17624926]
95. Mukherjee S, Pelech S, Neve RM, Kuo WL, Ziyad S, Spellman PT, Gray JW, Speed TP. Sparse combinatorial inference with an application in cancer biology. Bioinformatics. 2009; 25(2):265–271. [PubMed: 19038985]
96. Hill SM, Neve RM, Bayani N, Kuo WL, Ziyad S, Spellman PT, Gray JW, Mukherjee S. Integrating biological knowledge into variable selection: an empirical Bayes approach with an application in cancer biology. BMC Bioinformatics. 2012; 13(1):94. [PubMed: 22578440]

HIGHLIGHTS

- Despite a decade of research, the translation of biomarker studies into FDA approved clinical tests has been low
- More effective integration of data-driven and expert-driven strategies for biomarker discovery and identification is seen as key to improved success
- Data-driven approaches to biomarker discovery focus on data reduction, data classification, and data visualization
- Expert knowledge-driven approaches feature curated knowledge of protein-protein interactions, pathways, gene function, and ontologies
- Objective metrics for evaluating biomarker performance are key to using a iterative approach for improving performance
- Integrating data-driven and knowledge-based approaches for biomarker identification has the potential to improve performance and leverage translational applications

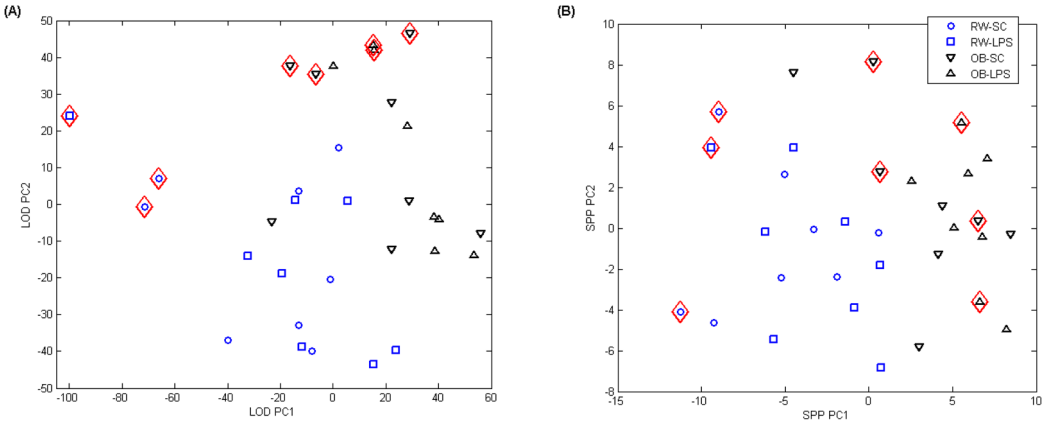

Figure 1.

PCA-based visualization of a proteomics data biomarker study. (A) PCA was performed on a proteomics dataset with a simple limit-of-detection based imputation of missing data and (B) no imputation using Sequential Projection Pursuit. The axes in both plots represent the first principal component (X axes) and the second principal component (Y axes).

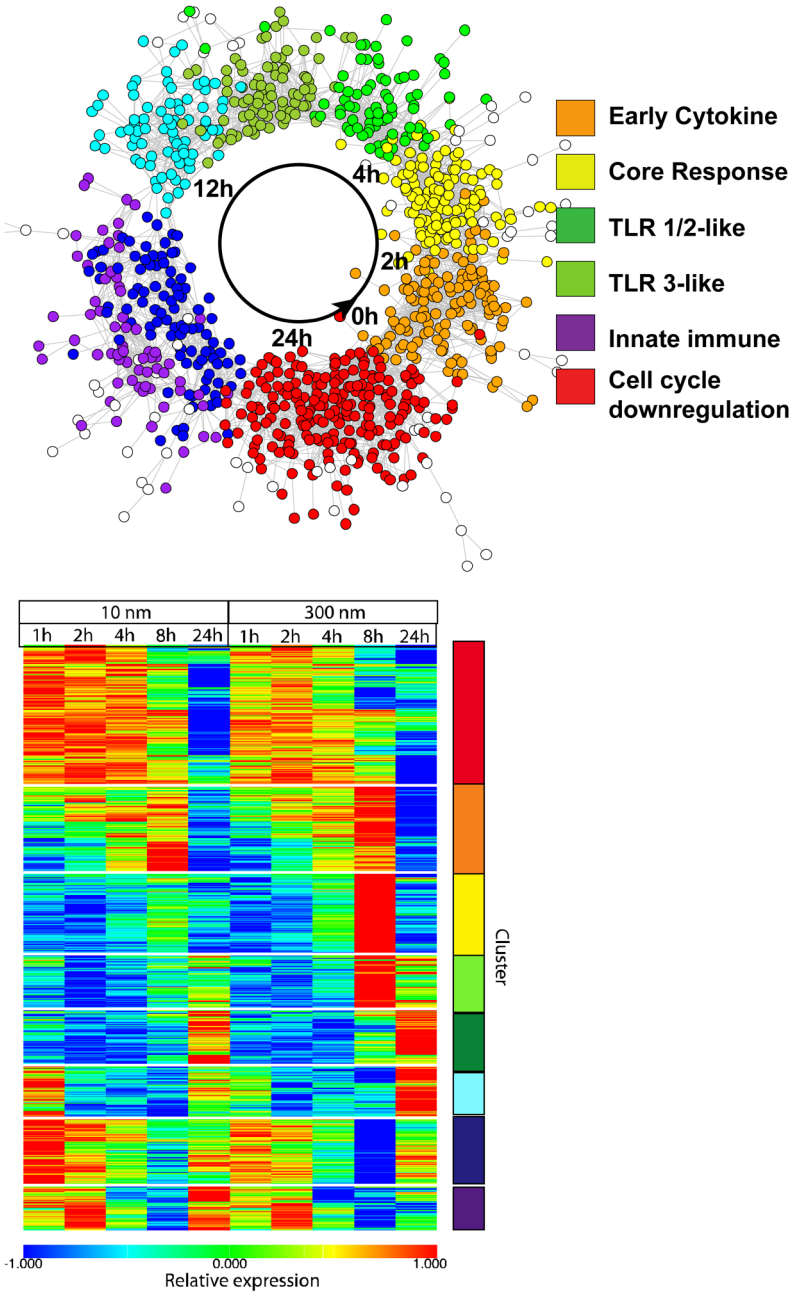

Figure 2.

A. Coexpression network of macrophage response to nanoparticle exposure showing the progression of functional modules through time. B. Standard heatmap of macrophage transcriptional response to nanoparticle exposure with modules defined using hierarchical clustering.

Table 1

Statistical methods applied to the data-driven analysis approaches.

Category[a]	Statistical Methods*	Applications	Concerns
Data Reduction (R)	Trend analysis (C)	Extracting underlying patterns	Unsupervised learning; good for the initial step of data analysis
	Clustering (V)	Subgrouping data; biomarker verification	
Classification (C)	Regression	Simple and straightforward	
	Support vector machine	High dimensional data; can be applied to nonlinear data	Work best for binary classification; requires good kernel function
	Decision trees and random forest	Recursive partitioning; effectively discriminate data	Overfitting training set
	Artificial neural networks	An adaptive approach, Model relationships in large complex datasets	Overfitting training set; non-intuitive models
	Gene relationship analysis	Comparison between patients, normalization between datasets	
Visualization (V)	Principal component analysis (R)	Visualize key patterns in a reduced dimension	Imputation required
	Network analysis (C)	Reveal similarity and relationships among clusters, recognize hubs in networks	

[a]Some methods are relevant to multiple categories, as indicated: R, data reduction; C, classification; V, visualization