Abstract: Biomarkers are of great importance in many fields, such as cancer research, toxicology, diagnosis and treatment of diseases, and to better understand biological response mechanisms to internal or external intervention. High-throughput gene expression profiling technologies, such as DNA microarrays and RNA sequencing, provide large gene expression data sets which enable data-driven biomarker discovery. Traditional statistical tests have been the mainstream for identifying differentially expressed genes as biomarkers. In recent years, machine learning techniques such as feature selection have gained more popularity. Given many options, picking the most appropriate method for a particular data becomes essential. Different evaluation metrics have therefore been proposed. Being evaluated on different aspects, a method’s varied performance across different datasets leads to the idea of integrating multiple methods. Many integration strategies are proposed and have shown great potential. This chapter gives an overview of the current research advances and existing issues in biomarker discovery using machine learning approaches on gene expression data.
Keywords: biomarker discovery; feature selection; gene expression; machine learning; statistical tests

INTRODUCTION

A biomarker is an indicator of a biological state, often in response to an intervention or the stage of a disease. Although biomarkers mostly refer to physiological or physical phenotypes, at the molecular level, a biomarker can indicate disease-associated molecular changes and may be useful in disease diagnosis (1, 2), various infections (3), neurological diseases (4), and for defining therapeutic targets (3). In toxicological studies, biomarkers are often used to define a set of differentially expressed genes or proteins in a toxic exposure or chemical risk assessment study (5–11). Data from various omics techniques, including transcriptomics, proteomics, and metabolomics, as well as epigenomics, are useful starting points for a biomarker discovery study (10, 12–15). In this chapter, we focus on the informative genes that can generally be used to distinguish samples from different groups, which can be normal or tumor tissues from human patients or tissues of animals that are exposed to toxic chemicals and their solvent controls, using gene expression data. Among the technologies for whole transcriptome gene expression profiling, DNA microarray and RNA sequencing (RNA-Seq) are the most popular (16).

On the methodology aspect, differential gene expression analysis has been the mainstream for its simplicity and interpretability. By comparing the mean expression values of different groups, we can measure the magnitude of difference between the groups, expressed as a fold change (FC), but it is important not to ignore the variance within each group. The genes of highly reproducible but comparably low difference in expression values are missed by looking solely at the FC (17). A statistical hypothesis test is usually applied, such as Student’s t-test, which considers both the difference between two groups’ mean values and the variability within each group. A p-value, which is the probability of obtaining an experimental result at least as extreme as the one observed under the null hypothesis, can be obtained from this kind of statistical tests. But such statistical tests usually require specific distributional assumptions; for example, the Student’s t-test is applicable if the values are normally distributed, which is rarely the case for gene expression data (17). In recent years, more and more concerns and debates about misuse of p-value have arisen (18–23). The choice of thresholds for FC and p-value can also significantly alter the interpretation of results (24).

In recent years, machine learning has been widely applied in biomarker discovery (3, 25–28). Machine learning applies mathematical approaches to train a model to learn from data for a particular task (29). The relevant machine learning techniques for biomarker discovery are classification and feature selection. Classification is a form of supervised learning where the algorithm is fed with labeled samples each represented by a set of features. The task is to learn a function that can predict the label of a sample from its features. In our case, the labels correspond to the different groups, and the features are the gene expression profiles. As in the case of gene expression data, the number of genes can be tens of
thousands (5). Feature selection is usually applied prior to classification or during classification, to remove noise or non-informative features to train a more precise and robust classifier (30, 31). Feature selection methods can generally be divided into three groups: (i) filter methods that select the features based on their correlation with the sample labels and are therefore independent of the classification procedure; (ii) wrapper methods which use an objective function (usually classification accuracy) to assess the importance of features, and (iii) embedded methods which are incorporated in the classifiers (32, 33). Since the selected features are informative in distinguishing samples from different groups, they can therefore also be regarded as biomarkers.

EVALUATION OF A BIOMARKER DISCOVERY METHOD

Several biomarker discovery methods have been proposed in the fast-developing machine learning field. A reasonable evaluation metric is necessary to choose the most appropriate biomarker discovery method. Two aspects have been addressed when talking about the performance of a biomarker discovery method: its stability and its ability to improve a classifier’s prediction accuracy (33–35). Another more direct way to assess performance is to look at the selected gene list given *a priori* knowledge of well-known biomarker sets which can be regarded as “gold standard” (36).

“Gold standard” gene sets

If *a priori* knowledge is available, such as the common gene mutations for breast cancer (37) or the common gene fusions for prostate cancer (38), at least conceptually, the relevant genes can be regarded as the true biomarker genes. In this case, evaluation of a biomarker discovery method becomes quite straightforward by simply comparing the selected gene set to the established “gold standard”. But establishing a high-quality “gold standard” becomes crucial to obtain both high precision (as many genes as possible are true biomarkers in the selected gene list) and sensitivity (as many true biomarkers as possible are selected from the whole gene list). To evaluate multiple RNA-Seq analysis workflows (including differential expression analysis), Williams *et al.* prepared a reference gene set based on results from four previous independent microarray and BeadChip studies (39). To reduce bias from one single statistical method, they employed both significance analysis of microarrays (SAM) (40) and limma (41, 42) and used the genes at the intersection of the two methods as the final reference. The resulting reference set was later used as “gold standard” in other studies to assess the performance of RNA-Seq analysis workflows or differential expression analysis methods (43, 44).

Stability

Ideally, the biomarkers should reflect the characteristics of the disease or exposure and be applicable to any sample in the data set. Thus, the biomarker discovery method should select a consistent set of genes disregarding minor changes in
the samples. However, in reality, due to differences between the samples, a biomarker discovery method will select different genes. The robustness of selecting similar gene sets even when the input data varies is called the stability of a method. The similarity of the selected gene lists can be used to define an evaluation metric reflecting the stability of the method.

Starting with two gene sets, Kalousis et al. (45) proposed to use the ratio between the number of genes contained in both sets (intersection) and the number of the set of genes contained in either (union) as the similarity index. Kuncheva et al. (46) pointed out that this index has a tendency to increase when there are more genes included in the list, which can encourage false positive results. They proposed to take into account the expected number of genes to be shared between the two sets as a modified index to solve that problem.

When it comes to a collection of gene sets, the similarity between them can be calculated by averaging all pairwise similarity indices (46). However, those similarity indices require that gene numbers in all gene sets are the same. Davis et al. (47) proposed a more flexible way to calculate similarity which allows various gene set sizes and can also directly calculate the similarity among more than two sets instead of in a pairwise fashion.

**Prediction accuracy**

A biomarker is an indicator of a biological state in response to an intervention, meaning that it can represent the characteristics of the samples in the intervened group compared with the control group. Compared with using the whole gene list to train a classifier that can distinguish the samples from different groups, training a classifier using biomarkers that already include the most distinctive information should give a comparative prediction performance or even a better one, since non-related and noisy genes can reduce the predictive ability of a classifier. Using several selected gene sets (potential biomarkers) to train classifiers, the prediction accuracy can reflect the quality of the corresponding gene set. A confusion matrix (48) is often used to evaluate the prediction performance of a classifier. Based on that, some evaluation measures such as Recall, Precision, area under a receiver operating characteristics curve, and so on, have been proposed to measure different performance aspects of a classifier (49).

**COMPARISON OF BIOMARKER DISCOVERY METHODS**

In the case where a well-established “gold standard” gene set is available, a simple comparison of the selected gene list to the reference list can assess the biomarker discovery method in question. But in most cases, such a true biomarker list is not available.

Before looking at the stability and prediction accuracy, which requires greater effort, a simple look at the gene list can still give some hints on the performance of the methods. Comparing the selected gene sets from multiple methods can shed some light on the exploration of the candidate methods, when the absolute performance is not of the highest concern. Blanco et al. (50) compared the genes identified as most relevant for discriminating sick and healthy patients as
produced by two different machine learning methods, random forest (51) and generalized linear models (52), and one classical gene expression analysis approach, edgeR (53). They found that random forest and edgeR tend to select similar gene sets compared with generalized linear models.

When the “gold standard” biomarker list is not available, and one still wants to assess the performance of a biomarker discovery method or compare multiple to select the best one for their study, stability and prediction accuracy can be used as evaluation metrics.

For a long time, improving prediction accuracy has been the focus of biomarker discovery methods. Lyons-Weiler et al. combined statistical tests with classification (17). They chose the threshold for FC and p-value which could help to achieve the highest classification accuracy. Comparing the F-score algorithm (from Support Vector Machines (SVM) (54)) with three popular differential expression analysis methods (limma, edgeR, DESeq (55)), Liang et al. (56) found that F-score algorithm obtained the best predictive performance when training an SVM classifier to predict stages of human embryonic development using single-cell RNA-Seq data. Schirra et al. evaluated the feature selection/classifier combinations that lead to an improved classification performance, and preferred filter methods when comparable prediction accuracy can be obtained for their higher interpretability (57).

Stability of biomarker discovery has gained more and more attention in recent years (32, 58, 59). A more complete evaluation of a biomarker discovery method should address both prediction accuracy and stability (33–35). In a previous study (33), on those two aspects, we compared the performance of both traditional statistical tests and machine learning methods: SAM, minimum redundancy maximum relevance (mRMR) (60), and characteristic direction (GeoDE) (36) on multiple datasets. We found that no single method outperforms the others on these two aspects across all tested datasets.

**ENSEMBLE OF MULTIPLE METHODS**

Since it is hard to tell which is the best one, another solution is to combine the potential methods. There are already studies showing that an ensemble of multiple feature selection methods can obtain a very satisfactory performance regarding both stability and prediction accuracy. The ensemble gene set can therefore be regarded as the final biomarker gene set (Figure 1).

Van IJzendoorn et al. combined statistical tests with machine learning techniques (61). On top of the significantly differentially expressed genes (adjusted p-value < 0.05), they applied random forest to select the most informative genes. By employing the ensemble feature selection concept, multiple biomarker discovery methods can be combined to take advantage of the strengths and overcome the weaknesses of the individual methods (62, 63). This approach is called function perturbation (32, 62). Similar to this logic, data perturbation refers to approaches applying one method on several data subsets generated from the original data set (for example using bootstrap (64)), and combining the results (58, 63, 65), an approach that has been shown to be able to improve the stability of the biomarker discovery method.
To take advantage of both data perturbation and function perturbation, we proposed to combine both of them (66) (Figure 2). In the phase of data perturbation, the stability of each method is calculated, and in the phase of function perturbation, when combining the results from multiple methods, their stabilities are used as their weights, so as to achieve the most robust final result. Testing on six microarray data sets from cancer studies, we found that the proposed framework achieved both high stability and prediction accuracy compared with the individual methods and the pure function perturbation.

**CONCLUSION**

In this chapter, we discussed biomarker discovery using gene expression data of the samples from different groups, usually a control group under normal biological status and a treated group with intervention or disease. The biomarker genes are therefore the responders to the intervention. Traditional statistical tests have been widely used to identify the differentially expressed genes as biomarkers for their simplicity and high interpretability. Such statistical tests are based on a hypothesis that the genes are independent of each other. This is not the case in a normal biological setting, since genes usually work together composing pathways...
Machine learning techniques, such as feature selection, have been applied with increasing frequency in biomarker discovery. Feature selection usually has fewer required assumptions compared with statistical tests. Many of them can take the interaction between genes and their joint power into consideration. The genes that are weak biomarkers by themselves but have a strong joint power can therefore be identified.

Another machine learning technique, classification, is also useful in biomarker discovery. Classification is not directly used to identify biomarkers but can be used to assess potential biomarkers selected by feature selection methods or statistical tests, since true biomarkers carry the characteristics of samples from the treated group compared with control group or vice versa and should therefore be informative in classifying the samples from different groups. The ability to improve a classifier’s prediction accuracy of a biomarker discovery method is widely used as an evaluation metric of candidate methods. We have seen that the choice of classification algorithm can highly affect the evaluation conclusion of the biomarker discovery methods (33), and using SVM to assess the performance of a feature selection method implemented in its own package together with other methods is unfair (56).
Besides prediction accuracy, a biomarker discovery method's stability has gained more attention in recent years. A good biomarker discovery method should provide a consistent biomarker list with some variance in the training samples, since the true biomarkers are intervention dependent (such as a disease or a toxicant exposure) and should be independent of the samples. There are many ways to calculate stability, but some of them tend to give a higher stability when more genes are included in the lists (46) and that is unfair for the methods that are stricter with redundant genes. Instead of looking only at the original gene list, Dessì et al. proposed to compare the lists in functional terms based on the molecular function Gene Ontology annotations, which has greater biological significance (35).

Many alternative approaches for improving a method's performance based on the aforementioned aspects have been proposed. One of them is feature selection ensemble, which combines the results of multiple biomarker discovery methods to take advantage of their strengths. It also solves the problem of having to choose the most appropriate method for a particular dataset since the performance of a method usually varies a lot across different datasets.

Besides assessing a biomarker discovery method on prediction accuracy and stability, one can also simply compare the candidate marker genes to a reference biomarker list, if such a “gold standard” exists. It is however difficult to be sure that the a priori knowledge is adequate, and that the list is complete and clear of false positives. Establishing such a reference list becomes extremely critical. Williams et al. applied two well-recognized methods (SAM and limma) on four independent datasets and used the intersected genes as reference (39). Biological a priori knowledge can also help in constructing such a reference list. Clark et al. made use of the relationship between differential STAT3 binding and differential gene expression in two subtypes of diffuse large B-cell lymphoma (DLBCL): germinal center B-cell-like (GCB) and activated B-cell-like (ABC) (36).

Biomarker discovery is a fast-growing field with many new ideas continuously being proposed. So far none are perfect, considering that the method is data dependent and no universal agreement on the evaluation of a method's performance has been established. However, devoted efforts are obviously enhancing progress in this field, which has a huge potential for providing a better understanding of disease diagnosis, prevention, and therapy, and for risk assessment of chemical toxicity.

Acknowledgement: This work was supported by the Research Council of Norway to the Digital Life Norway project dCod 1.0: decoding the systems toxicology of Atlantic cod (project no. 248840).

Conflict of interest: The authors declare no potential conflict of interest with respect to research, authorship and/or publication of this chapter.

Copyright and permission statement: The authors confirm that the materials included in this chapter do not violate copyright laws. Where relevant, appropriate permissions have been obtained from the original copyright holder(s), and all original sources have been appropriately acknowledged or referenced.
REFERENCES


