Biomarker discovery

8

8.1 An introduction to biomarker discovery

As suggested by the Biomarkers Definitions Working Group, a *biomarker* is "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention." With the advent of molecular biology and modern medicine, scientists begin to search for the biomarkers associated with disease processes at the molecular level. The molecular targets for biomarkers include biological molecules such as genes, transcripts, proteins, metabolites, and regulatory RNAs.

Generally, two basic styles are used in the research for biomarker discovery: hypothesis driven and discovery driven. The hypothesis-driven methods typically have clear candidate markers, which are usually identified during the study of disease processes. In contrast, the discovery-driven methods have no clear targets, which identify candidate biomarkers by analyzing the biological data sets through computational methods.

The advent of high throughput technologies has provided an unprecedented opportunity for identifying biomarkers from large-scale "omics" data in the discovery-driven style. Computationally, the biomarker discovery problem can be modeled as the problem of selecting features that can effectively distinguish cases from controls in a classification model. The biomarker discovery algorithms have been investigated for a long time, and numerous computational methods have been proposed from different research communities. In this chapter, the identification of biomarkers is studied from the viewpoint of data mining.

The major computational challenge in biomarker discovery is the so-called high-dimensional small-sample problem. That is, there are a huge number of genes or proteins/peptides that represent potential biomarkers in the data set. In contrast, the number of cases and controls is usually less than one thousand. As a result, the number of samples is significantly less than the number of features.

8.2 Data preprocessing

The discovery of biomarkers from high-throughput "omics" data is a very challenging task. It requires the collection and analysis of biological data in a quite complex pipeline. Typically, the discovery task is divided into different stages to reduce the complexity. To clarify this point, the biomarker discovery from mass spectrometry (MS) data is used as an example here to illustrate the preprocessing step and the analysis procedure.

MS has proved to be a useful tool for generating protein profiles of tissue, serum, and urine samples for case–control studies [1]. Ideally, the MS device should only generate peaks that are produced from ionized proteins or peptides. However, the

Data Mining for Bioinformatics Applications. http://dx.doi.org/10.1016/B978-0-08-100100-4.00008-9

acquired mass spectra contain not only true peaks that correspond to proteins or peptides of scientific interest but also noisy signals. Therefore, the following preprocessing steps should be conducted before the execution of core biomarker selection procedure, as presented in Figure 8.1.

1. *Feature extraction*: The first step is the extraction of "signals" or "features" from each spectrum. This involves many low-level signal processing tasks such as smoothing, peak detection, and peak quantification. Feature extraction is the most critical step in biomarker identification since all subsequent analysis steps use extracted features as input.
2. *Feature alignment*: Feature alignment establishes the correspondence among biological features extracted from different spectra. In other words, the aim of feature alignment is to produce a two-dimensional table that can be used in biomarker selection.
3. *Feature transformation (optional)*: A set of new features is created, where each new feature is obtained from the transformation or combination of old features. For instance, some methods (e.g., Ref. [2]) find a group of correlated proteins from the protein–protein interaction network, and then transform each group into a new feature for subsequent feature selection and classification. As shown in the example of Figure 8.2, the new feature T_2 is obtained by combining features F_3, F_5, and F_6.

8.3 Modeling

After the preprocessing step, the transformed data in the form of a two-dimensional table is used to generate a subset of features as markers. From the angle of data mining and machine learning, such a task is often modeled as a feature selection problem.

Generally, feature selection methods can be categorized based on how they are coupled to the classification model [3]. As shown in Figures 8.3–8.5, there are

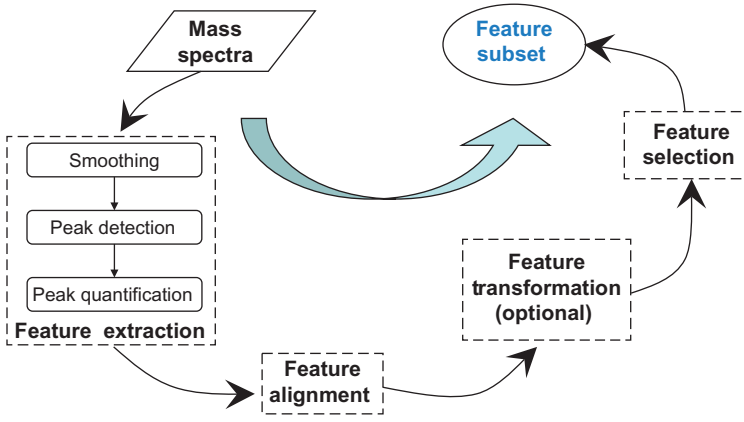

Figure 8.1 A typical data analysis pipeline for biomarker discovery from mass spectrometry data. In this workflow, there are three preprocessing steps: feature extraction, feature alignment, and feature transformation. After preprocessing the raw data, feature selection techniques are employed to identify a subset of features as the biomarker.

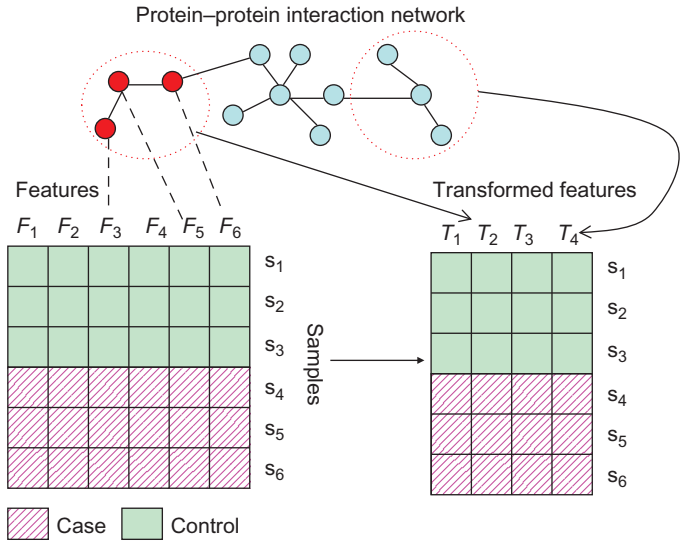

Figure 8.2 An illustration of feature transformation based on protein–protein interaction (PPI) information. The PPI information is used to find groups of correlated features in terms of proteins. These identified feature groups are transformed into a set of new features for biomarker identification.

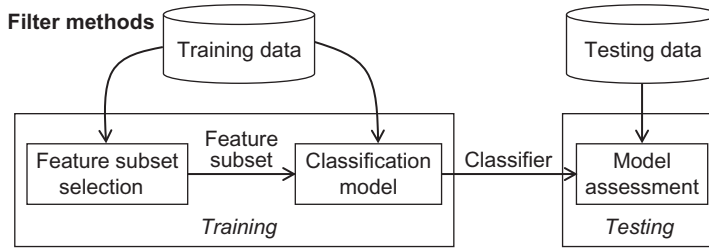

Figure 8.3 Filter methods for feature selection. In the filter method, the goodness of a feature subset is evaluated using only the intrinsic properties of the data.

generally three different types of feature selection methods: filter methods, wrapper methods, and embedded methods.

- *Filter methods*: The feature subset generation and evaluation is conducted as a preprocessing step, which is independent of the classification model used.
- *Wrapper methods*: The selection criterion of feature subset is determined by the performance of the classification model, which is strongly dependent on the specific classification method used.
- *Embedded methods*: The feature subset generation and evaluation is embedded as an integral part of the classification algorithm.

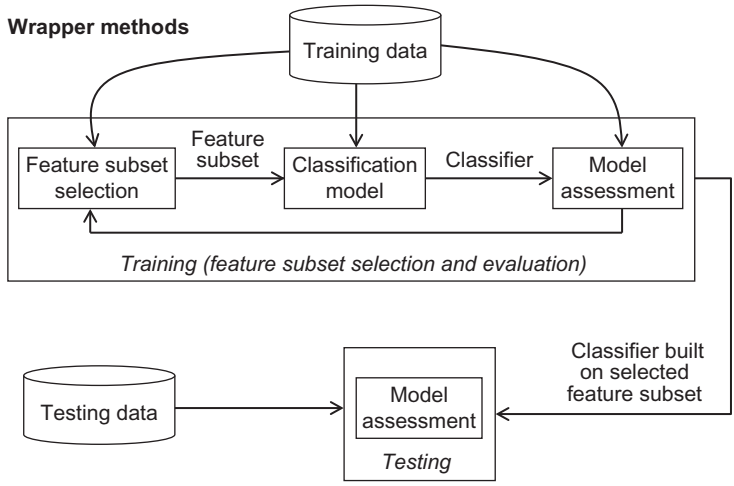

Figure 8.4 Wrapper methods for feature selection. In the wrapper method, the feature subset selection is based on the performance of a classification algorithm.

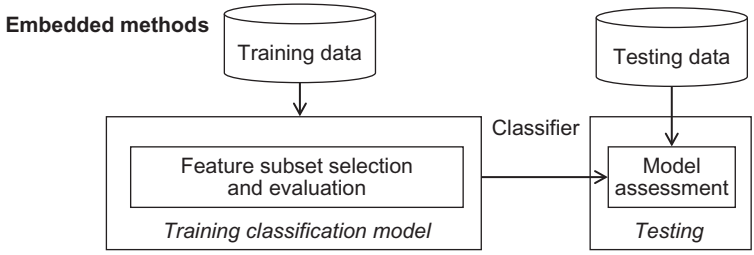

Figure 8.5 Embedded methods for feature selection. In the embedded method, the selection of feature subset is integrated with the construction of the classifier.

Among these methods, filter approaches are simple and flexible. Wrapper methods and embedded approaches directly optimize the classification performance during the feature selection process at the cost of incurring a high computational burden and probably over-fitting the classification model.

Here an unpublished feature selection method for biomarker discovery is introduced for the purpose of illustration. This method is named as *F*eature ranking using *B*inary *T*hreshold *C*lassifier (FBTC). It is a filter method, in which each feature is evaluated individually according to its classification capability estimated by a binary threshold classifier. The use of a binary threshold classifier in ranking features enables us to reduce the effect of variation in biological feature values, whereas evaluating each feature separately is computationally efficient.

Figure 8.6 illustrates the FBTC method, which outputs the ranking values of all m features. The basic idea is to evaluate each feature individually according to its

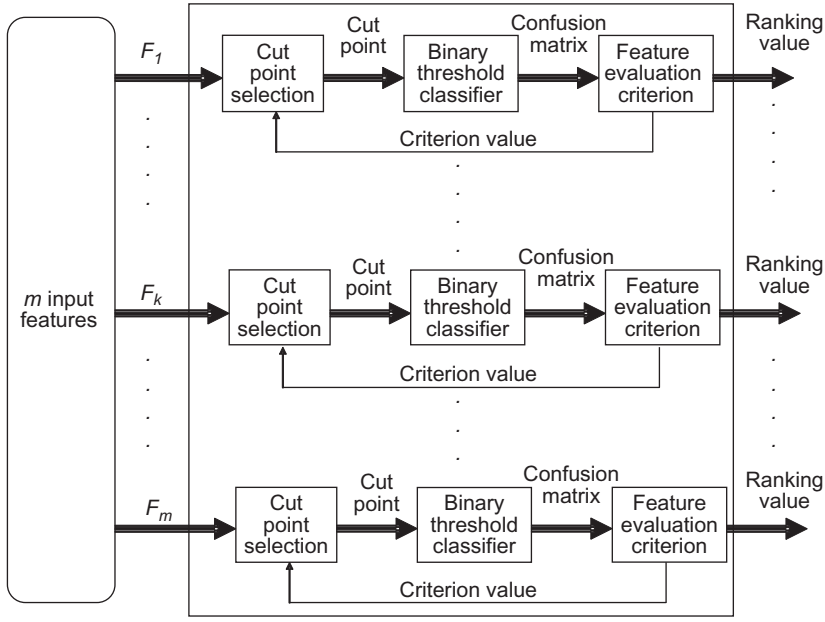

Figure 8.6 An illustration of the FBTC method.

classification power estimated by a binary threshold classifier. In the following, the details of this feature selection method are explained step by step.

8.3.1 Cut point selection

For a given feature F_k, we order all S sample values in the training set in a nondecreasing manner, that is, $F_k(X_1) \leq F_k(X_2) \leq \cdots \leq F_k(X_S)$. According to the procedure of the binary threshold classifier, it is sufficient to consider one cut point (e.g., $(F_k(X_h)+F_k(X_{h+1}))/2$) in each interval of $[F_k(X_h), F_k(X_{h+1})]$. In total, we only need to consider $S-1$ cut points for evaluation.

8.3.2 Binary threshold classifier

The basic idea of a binary threshold classifier on a numeric feature F_k is very simple. It consists of a cut point p_i and a target class c (1 corresponds to the positive class and 0 corresponds to the negative class). The decision function $BTC(X_h \mid p_i, c, F_k)$ is defined as:

$$BTC(X_h|p_i,c,F_k) = \begin{cases} c, & \text{if } F_k(X_h) > p_i \\ \bar{c}, & \text{if } F_k(X_h) \leq p_i \end{cases} \tag{8.1}$$

where $F_k(X_h)$ represents the feature value of a sample X_h and p_i denotes the threshold.

The use of a binary threshold classifier is based on the following considerations.

Why threshold classifier: A threshold classifier is robust against noise because it does not rely on precise feature values, which makes it suitable for handling the coefficient of variation in biological features.

Why binary classifier: The rationale can be discussed from two perspectives. First, it is natural to use a binary classifier because we are handling a binary classification problem (cancer vs. normal). Second, such a binary classifier is simple and easy to implement.

8.3.3 Feature evaluation criterion

Classification performances are typically evaluated using measures such as sensitivity and specificity. In the confusion matrix given in Table 8.1, cases (people have a certain disease) are considered as "positives" (class 1) and controls (people who have no that disease) are considered as "negatives" (class 0). In the horizontal direction of Table 8.1, $P = TP + FN$ and $N = FP + TN$, where P denotes the number of positive samples and N represents the number of negative samples. Vertically, $PP = TP + FP$ and $PN = FN + TN$, where PP denotes the number of predicted positives and PN denotes the number of predicted negatives. Correspondingly, some common evaluation metrics are defined in Table 8.2 as well.

Table 8.1 Confusion matrix defines four possible scenarios when classifying samples in the context of biomarker discovery

		Predicted class		
		Case	**Control**	**Row total**
Actual class	Case	True positives (TP)	False negatives (FN)	P
	Control	False positives (FP)	True negatives (TN)	N
Column total		PP	PN	S

Table 8.2 Performance metrics for evaluating classifiers

Names	**Abbreviations and definitions**
Specificity	$Spec = TN/N$
Sensitivity	$Sens = TP/P$
Positive predictive value	$PPV = TP/PP$
Negative predictive value	$NPV = TN/PN$
Accuracy	$ACC = (TP + TN)/S$
Balanced accuracy	$BACC = (Spec + Sens)/2$

We can use different measures in Table 8.2 to evaluate the importance of features. Here another more general measure is used, which is a function of measures in the confusion matrix:

$$U(p_i, c, F_k) = \max\left\{TP_i^c - \pi_N FP_i^c, TN_i^c - \pi_P FN_i^c\right\}, \tag{8.2}$$

where π_N denotes the *penalty* of false positives and π_P is the *penalty* of false negatives. The subscript and superscript indicate that outcomes are from the condition of cut point p_i and target class c. In general, any value in the interval $[0, +\infty)$ is meaningful as a proper assignment for π_N and π_P.

The feature evaluation criterion is defined as:

$$FEC(F_k) = \arg \max_{p_i, c \in \{0, 1\}} U(p_i, c, F_k), \tag{8.3}$$

where p_i goes through $S-1$ cut points to find the best classification result provided by this feature.

The setting of two penalty parameters is critical in feature ranking. Furthermore, several specific choices have clear practical implications. Suppose that the maximal value of $U(p_i, c, F_k)$ for feature F_k is achieved at the cut point p_i, at which the following properties can be observed.

- If $\pi_N = 1$ and $\pi_P = 1$, then the maximal ACC value is also achieved at the cut point p_i.
- If $\pi_N = P/N$ and $\pi_P = N/P$, then the maximal $BACC$ value is also achieved at the cut point p_i.
- If $\pi_N = \pi_P \to +\infty$, then the maximum value of PPV or NPV is also achieved at the cut point p_i under the constraint of $NPV = 1$ or $PPV = 1$.

Clearly, the FEC criterion is closely related with common ACC and $BACC$ measures by assigning certain values to π_N and π_P. In other words, FEC is a more general measure for feature ranking.

After ranking all features, we can select a subset of features with higher ranks and carry out classification using these features.

8.4 Validation

Besides the wet-lab validation through biological assays and clinical trials, several commonly used evaluation criteria are used for validating the reported biomarkers in the data-analytic phase of biomarker discovery.

1. *Classification performance*: One important goal of the biomarker discovery is to generate a subset of features that can accurately distinguish different classes such as the different stages of some diseases. Therefore, the prediction performance of the classification model built on the selected feature subset is an outstanding criterion. The most commonly used performance measure is the classification accuracy, which is defined as the percentage of correctly classified test samples. In some special cases, more concerns are given to class-specific performance measures such as sensitivity and specificity, which are defined as the proportion of correctly classified samples in the positive and the negative class, respectively.

2. *The size of feature subset*: To date, the number of candidate biomarkers that can be identified and validated for diagnostic purposes is rather limited. Therefore, most studies focus on biomarkers that are composed of a handful of features. In other words, the number of selected features should be as small as possible. Based on this fact, the number of selected features can serve as a criterion for the biomarker evaluation as well. In biomarker identification, minimizing the number of selected features might be even more important than improving the classification accuracy.
3. *Stability*: Traditionally, the classification accuracy is used as the major criterion for biomarker selection. In high-dimensional biological data, it can be observed that many different feature subsets have the same or similar predictive performance. If there is only one true biomarker, it is difficult to distinguish true biomarkers from false ones effectively only according to the classification performance. Therefore, the nonreproducibility of reported markers has become one major obstacle in biomarker discovery. Recently, the stability of a selected feature subset has become a new criterion for biomarker validation [4]. The stability of a feature subset is a good indicator of marker reproducibility. Although stability cannot override classification performance in the evaluation of candidate markers, it is a useful auxiliary criterion when the candidate feature subsets have similar predictive accuracy.

When the classification accuracy is used as the performance measure, the cross-validation approach is widely adopted for a rigorous evaluation. There are several variants of cross-validation in the literature. The k-fold cross-validation partitions the data into approximately k parts, in which one part is used for testing and another $k-1$ parts are used for feature selection and model training. The leave-one-out cross-validation takes only one sample as the testing sample and uses all remaining samples as the training samples.

To evaluate the stability of biomarkers (feature subset), one strategy is to randomly sample many subsets of the original training data. The feature selection method is performed on these subsamplings of the training data to check whether the target feature subset can be identified frequently. This procedure is illustrated in Figure 8.7.

8.5 Case study

In this section, the method in Ref. [5] is used as an example to illustrate the data analysis process in biomarker discovery. To identify biomarkers that can distinguish hepatocellular carcinoma (HCC) from cirrhosis, mass spectra from 84 HCC patients, 51 cirrhotic patients, and 80 healthy individuals are first generated.

In feature extraction, each mass spectrum was smoothed with the lowess smoothing method and was normalized by dividing its total ion current. Then, the slope of the peaks is used as the criterion for peak detection. The detected peaks correspond to target features in biomarker discovery.

In feature alignment, detected peaks from different samples are aligned together if their locations are no more than a given threshold.

In feature selection, a wrapper approach is exploited in which support vector machine (SVM) is used as the classifier. The proposed method reported a set of eight peaks as biomarkers. The SVM classifier built with these peaks achieved very

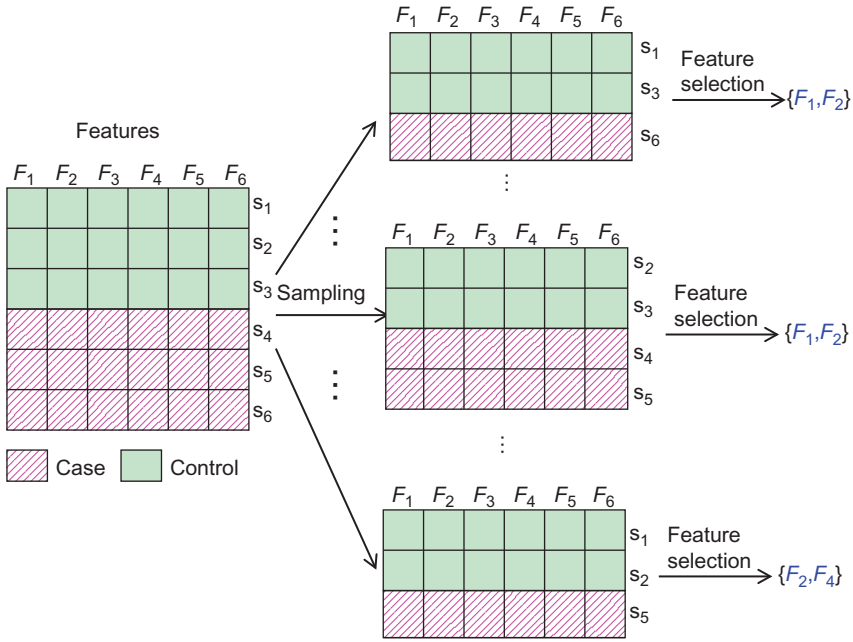

Figure 8.7 Stability evaluation by randomly sampling the original training data. Suppose multiple random subsets of the original training data set are generated. For each random subset of samples, the feature selection method is used to identify a feature subset. If the candidate biomarker (feature subset) identified from the original data has good stability, then it should occur frequently in the set of feature subsets obtained from the randomly selected data sets.

high classification accuracy in distinguishing HCC from cirrhosis in an independent testing data set.

8.6 Discussion and future perspective

During the past decades, feature selection methods have been used as the workhorse for identifying biomarkers in different applications. However, the biomarker discovery problem poses some new computational challenges that cannot be fully addressed by the traditional feature selection techniques. For instance, the stability of selected feature subset with respect to sampling variations has long been under-considered. As summarized in Ref. [4], there are already some stable feature selection methods for biomarker discovery in the literature. However, many data-analytical issues remain unsolved. For instance, how does one directly measure the stability of a feature subset? Is it possible to explicitly control the stability of a reported feature subset in the feature selection procedure?

The reason for the failure of identifying really useful clinical biomarkers from "omics" data sets of genes, transcripts, proteins, or other significant biological

molecules is very complicated. One important factor is that the true relevant biological molecules cannot be captured and recorded in the data set to be analyzed for biomarker selection. In this case, it is impossible to discover biomarkers successfully. Therefore, besides developing more sophisticated biomarker identification algorithms, it is more critical to improve the data coverage of wet-lab technologies for profiling biological molecules.

References

[1] Z. He, R.Z. Qi, W. Yu, Bioinformatic analysis of data generated from MALDI mass spectrometry for biomarker discovery, Top. Curr. Chem. 331 (2013) 193–210.
[2] H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, T. Ideker, Network-based classification of breast cancer metastasis, Mol. Syst. Biol. 3 (2007) 140.
[3] M. Hilario, A. Kalousis, Approaches to dimensionality reduction in proteomic biomarker studies, Brief. Bioinform. 9 (2) (2008) 102–118.
[4] Z. He, W. Yu, Stable feature selection for biomarker discovery, Comput. Biol. Chem. 34 (4) (2010) 215–225.
[5] H. Ressom, R. Varghese, S. Drake, et al., Peak selection from MALDI-TOF mass spectra using ant colony optimization, Bioinformatics 23 (5) (2007) 619–626.