Phosphorylation site prediction

4

4.1 Background and problem description

In an eukaryotic cell, it is estimated that about 30–50% of the proteins can be phosphorylated. In recent years, high-throughput studies have been able to identify new phosphorylation sites rapidly. Unfortunately, biological methods for phosphorylation event recognition are still costly and time consuming. In particular, the mass spectrometry-based techniques are biased toward abundant proteins and are difficult to provide specific information regarding the protein kinase–substrate interactions. Hence, the computational prediction method is potentially a useful alternative strategy for annotating the phosphorylation sites on the whole proteome scale.

Most proposed computational methods formulate the problem of phosphorylation site prediction as a binary classification problem, in which the class feature is the phosphorylation status. Although amino acids such as histidine and aspartate can also be phosphorylated, only serine (S), threonine (T), and tyrosine (Y) are considered frequently in most computational models.

4.2 Data collection and data preprocessing

As discussed in Chapter 3, we need to collect both the database of known phosphorylation sites and the corresponding protein sequence database to construct the data sets used for data analysis.

Fundamental to any classification problem, it is critical to generate training data and testing data for model construction and performance evaluation. In the context of binary classification, both the training data and testing data have two classes: positive data and negative data. Here, positive data contain a set of phosphorylated peptides, and negative data consist of a set of nonphosphorylated peptides. However, it is nontrivial to generate such kinds of data due to the uncertainty of phosphorylation events and the existence of unrecognized phosphorylation sites. As a result, different strategies have been proposed to construct the training/testing data.

Reference [1] provides a summary on both training and testing data construction methods in the literature. The phosphorylation site prediction task can be further divided into two more specific prediction tasks: non-kinase-specific phosphorylation site prediction and kinase-specific phosphorylation site prediction. For these two subtasks, the data construction methods for training data and testing data are different as well.

Data Mining for Bioinformatics Applications. http://dx.doi.org/10.1016/B978-0-08-100100-4.00004-1

4.2.1 Training data construction

Typical training data construction methods for non-kinase-specific phosphorylation site prediction are listed as follows:

1. *GTrainP*: The positive training data are composed of phosphorylated peptides that have been experimentally recognized.
2. *GTrainN1*: The negative training data contain only unphosphorylated peptides from phosphorylated proteins. One protein is called a phosphorylated protein if it has at least one S/T/Y residue that is known to be phosphorylated.
3. *GTrainN2*: The negative training data are composed of unphosphorylated peptides from both phosphorylated proteins and unphosphorylated proteins.
4. *GTrainN3*: The negative training data contain only unphosphorylated peptides from unphosphorylated proteins.
5. *GTrainN4*: This method first calculates the frequency distribution of amino acids of the peptides from GTrainN1. Then it randomly generates a set of new peptides according to the calculated frequency distribution as the negative training data.
6. *GTrainN5*: This method randomly generates a set of peptides with equal frequency distribution of amino acids as the negative training data.

To provide a vivid illustration, Figure 4.1 describes the constituent parts for the training data set generated from GTrainP, GTrainN1, GTrainN2, and GTrainN3.

The training data construction methods for kinase-specific phosphorylation site prediction are more diverse, as described below (see also Figure 4.2).

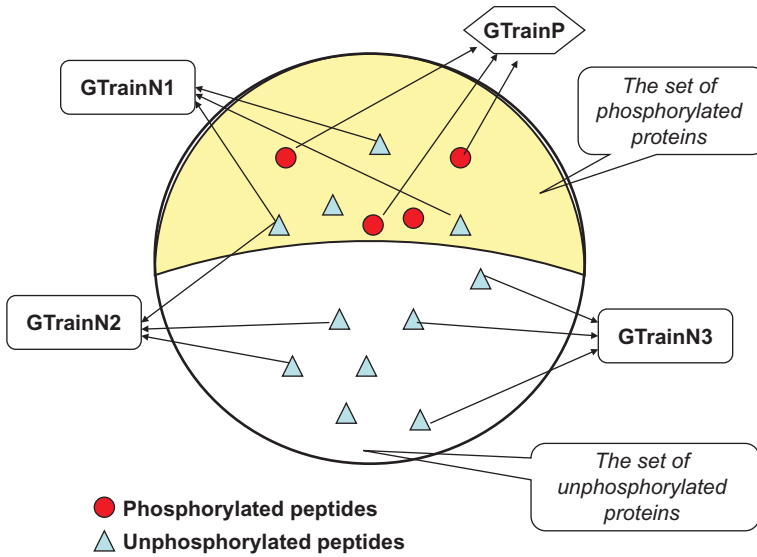

Figure 4.1 An illustration on the training data construction methods for non-kinase-specific phosphorylation site prediction. Here the shadowed part denotes the set of phosphorylated proteins and the unshadowed area represents the set of unphosphorylated proteins.

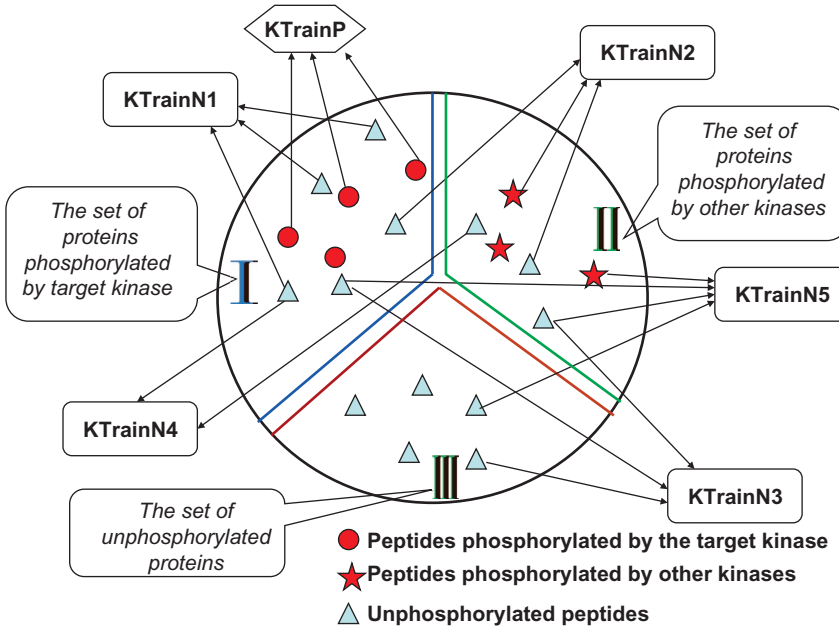

Figure 4.2 An illustration on the training data construction methods for kinase-specific phosphorylation site prediction. The proteins are divided into three parts: (**I**) the set of proteins that are phosphorylated by the target kinase, (**II**) the set of proteins that are phosphorylated by the other kinases, and (**III**) the set of unphosphorylated proteins.

1. *KTrainP*: The positive training data consist of verified phosphorylated peptides from phosphorylated proteins of the target kinase.
2. *KTrainN1*: The negative training data are composed of unphosphorylated peptides from phosphorylated proteins of the target kinase.
3. *KTrainN2*: The negative training data contain both unphosphorylated peptides from all phosphorylated proteins and those phosphorylated peptides from proteins that are not phosphorylated by the target kinase.
4. *KTrainN3*: The negative training data are constructed with unphosphorylated peptides from both phosphorylated proteins and unphosphorylated proteins.
5. *KTrainN4*: The negative training data are composed of unphosphorylated peptides from proteins that can be phosphorylated by any kinase.
6. *KTrainN5*: The negative training data include both unphosphorylated peptides from all proteins and phosphorylated peptides from proteins phosphorylated by other kinases.
7. *KTrainN6*: The negative training set are composed of unphosphorylated peptides whose phosphorylation residues are buried in the core of proteins phosphorylated by the target kinase. This method assumes that buried residues would not be physically accessible to any kinase, thus improving the quality of negative training data.

For both non-kinase-specific and kinase-specific predictions, the empirical comparison shows that different training data construction methods have different prediction performance and the difference is significant according to several statistical tests [1].

4.2.2 Feature extraction

To generate features for classifier training and testing, there are two widely adopted strategies in the literature.

On the one hand, one may directly view the data as a categorical data, where the amino acids at each position correspond to categorical feature values. For classifiers that cannot handle categorical data, each categorical feature can be transformed into 20 binary features, where each binary feature represents the presence or absence of a specific amino acid on that position with 1/0.

On the other hand, many methods extract some predefined features artificially. The most widely used features are the sequence compositions of amino acids surrounding phosphorylation sites. In addition, other more complex features are often adopted in different methods, such as the protein disorder features and features related with spatial amino acid compositions.

4.3 Modeling: Different learning schemes

In Ref. [2], existing phosphorylation site prediction tools are summarized and categorized from different viewpoints. These techniques differ in several ways: the machine learning or data mining techniques used; the features extracted from the set of peptides; whether predictions are kinase-specific or non-kinase-specific; and the construction of training and testing data.

Here we discuss the key modeling part from a different angle: the taxonomy of underlying machine learning principles.

4.3.1 Standard supervised learning

Almost all the existing prediction tools for phosphorylation sites fall into this category. That is, a classifier is first built from the given training data and then applied to predicting the class label of S/T/Y residues whose property is unknown. With respect to the classifier used, the most popular technique is support vector machine (SVM), which is exploited in many prediction tools (e.g., Musite [3]).

4.3.2 Active learning

Those tools based on supervised learning have been successfully applied to many organisms for phosphorylation site prediction. However, there is useful information inherent within the set of nonannotated S/T/Y sites that could be exploited for building more accurate classifiers. It is well-known that the number of nonannotated S/T/Y sites is much larger than the number of annotated ones.

To fully use the information in the S/T/Y sites whose phosphorylation status is still unknown, Ref. [4] presents an active learning strategy to train classifiers for phosphorylation site prediction.

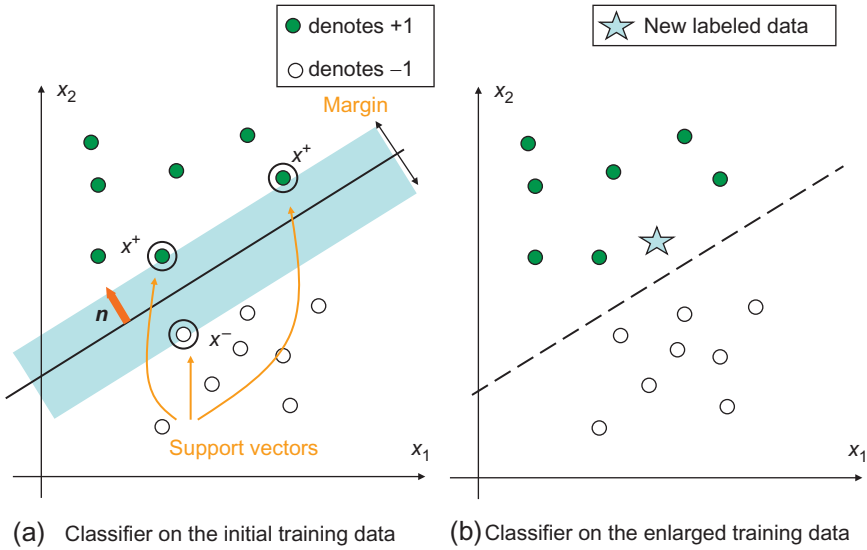

Figure 4.3 An illustration on the basic idea of the active learning procedure for phosphorylation site prediction. (a) The SVM classifier (solid line) generated from the original training data. (b) The new SVM classifier (dashed line) built from the enlarged training data. The enlarged training data are composed of the initial training data and a new labeled sample.

As shown in Figure 4.3a, an SVM classifier (solid line) is first built on the initial training data, where solid circles represents phosphorylated peptides (+1) and the empty circle denotes the unphosphorylated peptides (−1). The classifier is applied to evaluating a set of unlabeled peptides and the peptide with the highest classification confidence is marked with its new label. Then, this marked peptide, which is represented as a star in Figure 4.3b, is added into the training data set to generate a new enlarged training data set. On this new training data set, a new SVM classifier (dashed line) is learned. This update procedure is repeated until some stopping criteria are satisfied, for example, a sufficient number of new samples have been included in the augmented training data.

4.3.3 Transfer learning

In the context of standard supervised learning, it is assumed that the training data and future testing data come from the same feature space, that is, have the same distribution. However, in the application of phosphorylation site prediction, the lack of sufficient training data with respect to experimentally confirmed phosphorylation sites hampers the development of more accurate prediction models.

To alleviate this issue, a novel prediction method called PHOSFER for applying phosphorylation data from other organisms to enhance the accuracy of predictions in a target organism is presented in Ref. [5]. Essentially, this approach falls into the so-called transfer learning framework [6]. Transfer learning is helpful in case

we have a classification task in one domain of interest, but we only have sufficient training data in another domain of interest, where the latter may have a different data distribution [6]. In such cases, it would greatly improve the prediction performance if the knowledge in one domain can be transferred to the target domain successfully.

PHOSFER is a phosphorylation site prediction tool designed for organisms such as plants for which little phosphorylation site data are available. The soybean (Glycine max) is used as a test case to illustrate its feasibility and advantage. Basically, this method has the following steps, as described in Figure 4.4.

First of all, phosphorylation sites of nine different organisms that have been experimentally characterized are gathered from the corresponding online databases. In addition, the Basic Local Alignment Search Tool (BLAST) searches are used to determine the degree of conservation between phosphorylation sites in soybean and those in other organisms.

Then, the known phosphorylation sites from both soybean and other organisms are used as the training data. Each training peptide from other organisms is assigned a weight based on the degree of phosphorylation site conservation between soybean and the corresponding organism and the number of training peptides from that organism. The weights of training peptides from different organisms are determined according to the following principle: higher weights are given to peptides that are better conserved in soybean and from organisms with less known phosphorylation sites.

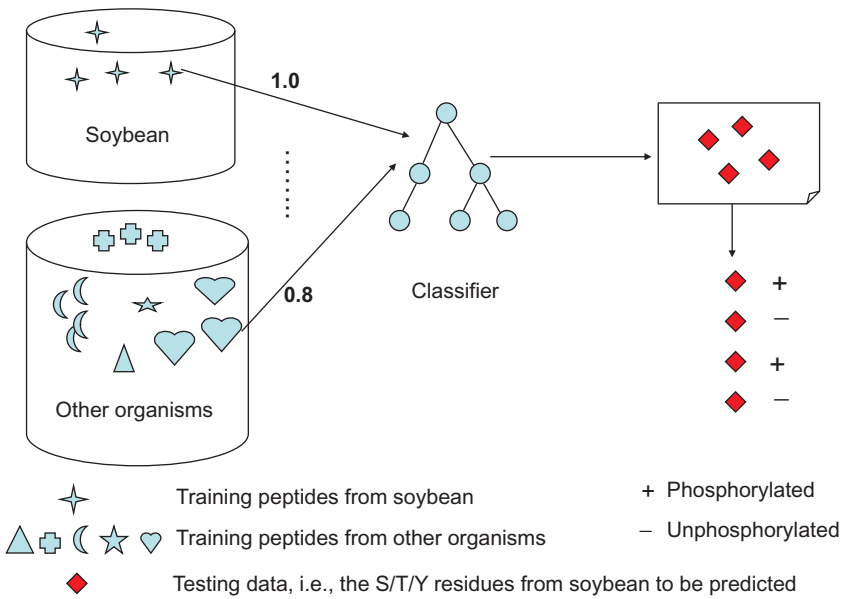

Figure 4.4 An overview of the PHOSFER method. The training data are constructed with peptides from both soybean and other organisms, in which different training peptides have different weights. The classifier (e.g., random forest) is built on the training data set to predict the phosphorylation status of remaining S/T/Y residues in the soybean organism.

Finally, the random forest classifier is generated on the weighted training data set collected from different organisms to predict the phosphorylation status of candidate S/T/Y residues in soybean. It has been demonstrated that the resultant predictor outperforms both the Arabidopsis-specific tools and a simpler machine-learning technique that uses only known phosphorylation sites from soybean.

4.4 Validation: Cross-validation and independent test

Cross-validation and independent test are widely used for evaluating the classification performance in the context of both non-kinase-specific and kinase-specific phosphorylation site prediction.

Cross-validation divides the training data into several disjointed parts of approximately equal size. Each part is selected in turn as the testing data, whereas the remaining parts are used as the training data. The prediction model built on the training data is then applied to predicting the class labels of testing data. This process is repeated until all parts have been masked once, and then the prediction accuracies across all blinded tests are combined to give an overall performance estimate.

Different from cross-validation, an independent test uses a third-party data set as the testing data. In the context of phosphorylation site prediction, there are three commonly used strategies:

1. *ID1*: The testing data are composed of a set of phosphorylated peptides and unphosphorylated peptides, which are extracted from proteins of the same species but have no overlap with the training data.
2. *ID2*: The testing data are constructed by selecting peptides from other species.
3. *ID3*: The testing data are sampled from the same set of peptides that is used for generating the training data.

4.5 Discussion and future perspective

The problem of phosphorylation site prediction is a typical sequence classification problem: to build a prediction model from a set of training sequences and classify unknown sequences into different categories. Different from the classification problem on feature vectors, sequences do not have explicit features. As a result, the sequence classification is a more challenging task than the classification on feature vectors.

Generally, the sequence classification methods can be divided into three categories [7]: feature-based methods, distance-based methods, and model-based methods. The feature-based method transforms a sequence into a feature vector and then applies conventional classification methods. The distance-based method uses a distance function to measure the similarity between sequences and then classifies the test sequence according to its nearest neighbors in the training data. The model-based method is built on generative models, which assume that the sequences in one class are generated from an underlying probabilistic model.

Currently, most phosphorylation site prediction methods are feature-based approaches, which extract predefined features from both the phosphorylated peptides and unphosphorylated peptides for constructing the predictive model. Because the classification performance is highly dependent on the chosen set of features, existing prediction tools have adopted different kinds of features in their methods. Even with sophisticated feature extraction approaches, the true relevant features are still difficult to capture because the domain-specific insight is either incomplete or hard to translate into effective features. Therefore, an alternative strategy is to automate the process of constructing effective features for the sequence classification (e.g., Ref. [8]). To date, there are still no algorithms for phosphorylation site prediction that are constructed with such automatic feature extraction strategies.

The distance-based methods and the model-based methods in sequence classification have not been widely used in phosphorylation site prediction. In the future, more research efforts should be devoted to these two categories with the goal that more accurate phosphorylation site prediction tools can be constructed.

On the other hand, the phosphorylation site prediction is largely modeled as a standard supervised learning problem. As presented in previous sections in this chapter, other learning schemes such as the active learning and the transfer learning have already been exploited to address this issue. Indeed, other learning schemes can be used to model the problem of phosphorylation site prediction as well. For instance, the transductive learning [9] is another choice, in which the samples that need to be predicted are already known in training the classifier. This learning scheme is particularly suitable to solve the phosphorylation site prediction problem because the samples that need to be predicted are those unknown S/T/Y sites in all the proteins across the whole proteome. Unfortunately, prediction tools based on this idea are still not available.

The cross-validation error estimation has been widely used in validation and performance comparison for phosphorylation site prediction. This procedure is "almost unbiased" when random sampling is used in fold generation. However, this is not true with separate sampling, where the positive data and negative data are independently sampled [10]. It has been shown that the classical cross-validation can have strong bias under the separate sampling in Ref. [10]. Therefore, to use cross-validation with separate sampling in phosphorylation site prediction in the future, one should use the separate-sampling version of cross-validation in Ref. [10] to avoid estimation bias in performance evaluation.

References

[1] H. Gong, X. Liu, J. Wu, Z. He, Data construction for phosphorylation site prediction, Brief. Bioinform. 15 (5) (2014) 839–855.

[2] B. Trost, A. Kusalik, Computational prediction of eukaryotic phosphorylation sites, Bioinformatics 27 (2011) 2927–2935.

[3] J. Gao, J. Thelen, A. Dunker, et al., Musite, a tool for global prediction of general and kinase-specific phosphorylation sites, Mol. Cell. Proteomics 9 (12) (2010) 2586–2600.

[4] J. Jiang, H. Ip, Active learning for the prediction of phosphorylation sites, in: D. Liu (Ed.), International Joint Conference on Neural Networks, 2008, pp. 3158–3165.

[5] B. Trost, A. Kusalik, Computational phosphorylation site prediction in plants using random forests and organism-specific instance weights, Bioinformatics 29 (6) (2013) 686–694.
[6] S. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2010) 1345–1359.
[7] Z. Xing, J. Pei, E.J. Keogh, A brief survey on sequence classification, SIGKDD Explor. 12 (1) (2010) 40–48.
[8] U. Kamath, K.D. Jong, A. Shehu, Effective automated feature construction and selection for classification of biological sequences, PLoS One 9 (7) (2014) e99982.
[9] V. Vapnik, Statistical Learning Theory, Wiley, New York, 1988.
[10] U.M. Braga-Neto, A. Zollanvari, E.R. Dougherty, Cross-validation under separate sampling: strong bias and how to correct it, Bioinformatics 30 (23) (2014) 3349–3355.