Phosphorylation motif discovery

3

3.1 Background and problem description

Posttranslational modification (PTM) is the chemical modification of a protein after its translation. Protein phosphorylation is one special kind of PTM that plays important roles in many cellular processes. Generally, phosphorylation can only happen on three docking sites: serine (S), threonine (T), and tyrosine (Y). A biological function usually involves a series of phosphorylation processes, in which protein kinases recognize specific protein substrates.

The advances in high-throughput techniques such as the tandem mass spectrometry make it possible to rapidly and directly recognize large-scale phosphorylation sites in a single experiment. Such sets of identified phosphopeptides provide valuable information for determining the specific kinase–substrate recognition.

Phosphorylation motifs are consensus amino acids that are aligned upstream and downstream of the phosphorylation sites. The discovery of phosphorylation motifs is to detect a set of motifs that occur more frequently in the set of phosphorylated peptides P (i.e., the foreground data) than that in the set of unphosphorylated peptides N (i.e., the background data). In other words, the identified phosphorylation motifs are all "over-expressed" in P. The phosphorylation motifs not only provide information about the specificities of the kinases involved, but also reveal the underlying regulation mechanism and facilitate the prediction of unknown phosphorylation sites.

One phosphorylation motif can be represented as a string with a single phosphorylated residue that is denoted with an underlined character, for example, S, T, or Y. The string representation of motifs consists of either conserved positions or wild positions (denoted by "."). For instance, "PS.D" is a phosphorylation motif that has two conserved positions. These two fixed residues on the conserved positions are "P" (one position on the left) and "D" (two positions on the right), respectively.

One motif A is said to be a k-motif if it has k conserved positions. If another motif B contains only a subset of these k amino acids at the corresponding positions in A, then B is called a submotif of A. For example, "PS.D" is a 2-motif and "S.D" is a 1-motif. In addition, "S.D" is a submotif of "PS.D."

Table 3.1 is a sample data set used for phosphorylation motif discovery. In this example, there are 10 phosphorylated peptides in the foreground data P and 10 unphosphorylated peptides in the background data N, respectively. Each peptide has 13 amino acids and its central position is the phosphorylation site. In this example data set, we can observe that "KMS" is an interesting phosphorylation motif because it appears five times in P but never occurs in N. In contrast, "K.S" is not a meaningful phosphorylation motif because its appearance frequencies in P and N are equal.

Data Mining for Bioinformatics Applications. http://dx.doi.org/10.1016/B978-0-08-100100-4.00003-X

Table 3.1 A sample data set used for phosphorylation motif discovery

Foreground data *P*	Background data *N*
GLKLKMSMQYPEG	EACPKHSWHTAHY
VLAYKMSWDPEVR	RGASKGSMVRFKG
QDPAKMSMQLATE	RQAVKPSWVARKY
LPYMKMSWLCSLA	LGVQKRSKHDRAH
HPGEKMSGALQDA	HVGAKRSADCANS
SWDTQASKTRDAL	DKHSQMSVRSQND
AKLQGWSTVTRGS	VCQADMSAVQRYS
VKQLAWSPVKMTS	SPWDTMSDYSDLQ
NLYWQTSEVLWRV	QKLQPMSPALLQG
HASDWPSEQPKMP	DPQHLMSRQEALG

Both the foreground data set and background data set consist of 10 peptide sequences, where the length of each peptide sequence is 13. The underlined central residue is the phosphorylation site.

3.2 The nature of the problem

If we consider the phosphorylation state as the class feature, the data sets used in phosphorylation motif discovery can be divided into two classes. Essentially, the problem of phosphorylation motif discovery can be modeled as a special discriminative pattern mining problem in the literature of data mining. More precisely, three mapping relationships between these two tasks can be observed.

First, by taking the position-specific amino acid set as the collection of all possible feature values, the peptide sequences in phosphorylation motif discovery correspond to the samples in discriminative pattern mining. Second, the phosphorylation motifs can be considered as the target patterns when mining discriminative patterns from class-labeled data sets. Last, the over-expression level of phosphorylation motifs can be calculated by statistical measures used for assessing the power of discriminative patterns.

Therefore, the task of phosphorylation motif discovery can be referred to as an example of discriminative pattern mining.

3.3 Data collection

To conduct the phosphorylation motif discovery, we first need to collect phosphorylation sites that have been experimentally verified. There are already more than 20 phosphorylation-related databases that store known phosphorylation sites.[1] Typically, these databases focus on phosphorylation sites of certain organisms. For instance, the PhosPhAt database is designed specifically to store and query Arabidopsis phosphorylation sites.

[1]In Ref. [1], some of the most popular phosphorylation-related databases are listed with sufficient details.

These verified phosphorylation sites in the databases are used for constructing the foreground data P. We also need to collect data that can be used to generate the background data N, which are peptides that cannot be phosphorylated. Hence, the entire protein sequence database of the target organisms or species should be collected as well in this phase.

3.4 Data preprocessing

Before conducting the discovery procedure for finding phosphorylation motifs, the raw data from the phosphorylation site database and the protein sequence database should be preprocessed to generate the tabular formed data. The typical preprocessing process works as follows:

- The peptides in the phosphorylation site database are first mapped back to their prospective proteins. Then, a fixed number of residues upstream and downstream of the phosphorylation site are reextracted from the protein sequences. Here the number of residues r is usually not large, for example, $r = 10$. This procedure guarantees that we can obtain the foreground data P with a fixed number of features.
- To generate the background data set N, the peptides of length $2r + 1$ with a phosphorylated residue in the center position are extracted from the corresponding protein database. Then, all peptides already in P are removed. Finally, a desired fraction of peptides are sampled from the remaining peptides to form the background data.

3.5 Modeling: A discriminative pattern mining perspective

The problem of phosphorylation motif discovery has been widely studied, and several effective algorithms have been proposed based on discriminative pattern mining techniques. On the one hand, some methods employ the exhaustive strategy and some approaches adopt the heuristic strategy to discover as many significant motifs as possible. On the other hand, some methods apply the frequency test to mine a set of frequent motifs as candidates first, and then employ the statistical significance evaluation to obtain significant motifs by filtering out insignificant ones, whereas other methods conduct these two assessments in a single stage to directly generate the significant motifs. The approaches proposed in Refs. [2,3] are two representative ones, which tackle the problem of mining phosphorylation motifs from the viewpoint of mining discriminative patterns. Here we will use the Motif-All algorithm [2] and the C-Motif algorithm [3] as examples to show how the discriminative pattern mining methods can be used to solve the problem of phosphorylation motif discovery.

3.5.1 The Motif-All algorithm

The Motif-All algorithm takes the support threshold and the significance threshold as input to find phosphorylation motifs from the set of phosphorylated peptides P and the set of unphosphorylated peptides N. As shown in Figure 3.1, this algorithm has the following steps:

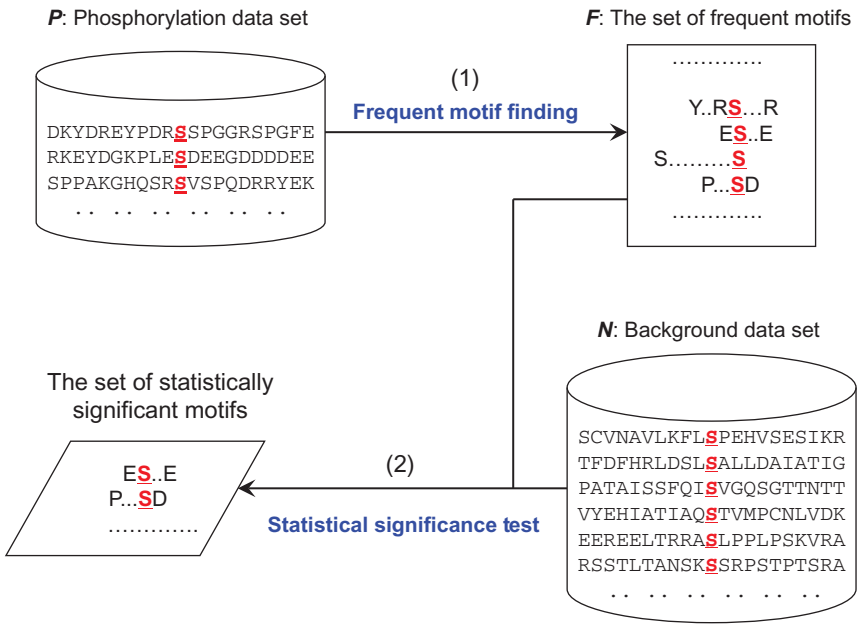

Figure 3.1 Overview of the Motif-All algorithm. In the first phase, it finds frequent motifs from P to reduce the number of candidate motifs. In the second phase, it performs the significance testing procedure to report all statistically significant motifs to the user.

1. Finding the set of all frequent motifs F from P using frequent pattern mining algorithms such as the Apriori algorithm [4]. A motif is said to be frequent if its support value (frequency) in P is no less than a user-specified support threshold.
2. Scanning the background set N to calculate the statistical significance for each motif in F. Those motifs that can pass the significance threshold are reported to the user. Popular measures such as the odds ratio and the risk ratio can be used for evaluating the statistical significance of motifs. The use of alternative significance measures will not change the nature of the problem and affect the property of the algorithm.

The Motif-All algorithm adopts the popular two-stage strategy used in discriminative pattern mining for phosphorylation motif discovery. This method is simple and effective. However, it evaluates the statistical significance of each motif without considering the effect of its subsets, making it possible to generate many redundant motifs whose over-expressiveness mainly comes from their submotifs. To alleviate this issue, the C-Motif algorithm [3] not only evaluates the statistical significance of each motif but also presents another measure called conditional significance to remove the effect of submotifs.

3.5.2 The C-Motif algorithm

C-Motif is implemented in a single stage where the frequency and the statistical significance values are tested at the same time, as shown in Figure 3.2. That is, this algorithm visits the candidate motifs in a breath-first manner. For each motif, the support,

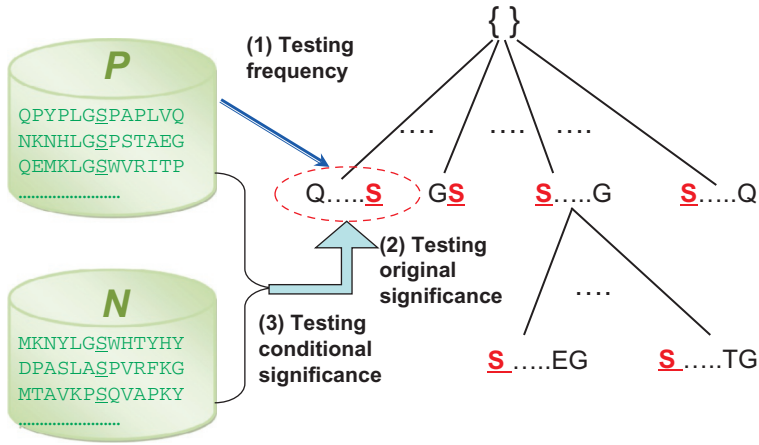

Figure 3.2 Overview of the C-Motif algorithm. The algorithm generates and tests candidate phosphorylation motifs in a breath-first manner, where the support and the statistical significance values are evaluated simultaneously.

the original statistical significance, and the conditional statistical significance are evaluated simultaneously. Here the original statistical significance is calculated over the data sets P and N. The calculation of conditional statistical significance is illustrated in Figure 3.3.

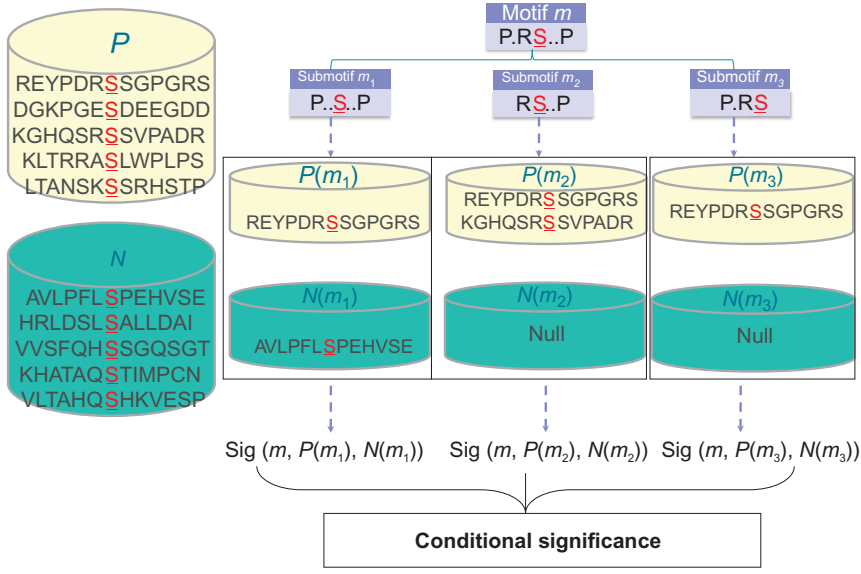

Figure 3.3 The calculation of conditional significance in C-Motif. In the figure, Sig(m, $P(m_i)$, $N(m_i)$) denotes the new significance value of m on its ith submotif induced data sets.

Because the peptide that contains one target motif must also contain its submotifs, the set of peptides that contain this motif must be a subset of the collection of peptides that contain its submotif. In particular, there are exactly k submotifs of size $k-1$ for one k-motif. For each submotif of size $k-1$, we can generate a set of peptides in which every peptide contains this submotif.

As shown in Figure 3.3, the motif "P.RS..P" is a 3-motif and has three submotifs of size two. Each submotif will induce a new foreground data set and a new background data set by selecting peptides that contain the corresponding submotif. On each submotif induced data set, the statistical significance of the motif "P.RS..P" is recalculated. The conditional significance of "P.RS..P" is obtained by aggregating the three new significance values.

The use of conditional significance is able to remove the effect of submotifs. As a result, more redundant motifs are filtered out by C-Motif.

3.6 Validation: Permutation *p*-value calculation

Although methods such as C-Motif can reduce the number of reported phosphorylation motifs significantly, the validation of discovered motifs is still a nontrivial issue. This is because no multiple testing corrections are carried out in most phosphorylation motif discovery algorithms, leading to an inaccurate motif significance assessment. As a result, there are still some false positives in the returned motifs.

In order to assess the significance of phosphorylation motifs more accurately, the permutation test is generally used to calculate the empirical p-value of each motif. The standard permutation procedure works as follows:

1. Formulate a permutation null hypothesis and choose a test statistic that will have different values under the null hypothesis and the alternative hypothesis. Here the null hypothesis is that the frequencies of motifs in foreground data and background data are the same. And the alternative hypothesis is that the frequency distributions of motifs in the foreground data and the background data are different. Furthermore, the choice of test statistic is flexible because one can use any statistical significance measures that are appropriate for evaluating the over-expressiveness of phosphorylation motifs.
2. Permute class labels (phosphorylated vs. unphosphorylated) in a way that is consistent with the null hypothesis of the test and the study designed to produce permuted data sets, then calculate the test statistic values of all tested motifs in these permuted data sets to generate the null distribution.
3. Calculate the original test statistic values of all tested motifs in the original data set. The p-values of all phosphorylation motifs can be calculated by locating the original test statistic values in the null distribution.

To illustrate the permutation procedure, one example is shown in Table 3.2. In this sample data set, there are two phosphorylated peptides and four unphosphorylated peptides whose class labels are denoted by "T" and "F," respectively. Table 3.3 shows one possible permuted data set after random permutation. That is, the new foreground

Table 3.2 The original foreground data *P* and the background data *N* before the permutation

Number	Peptide sequence	Class
1	DGYDRRYGDRYSPGGRSPGFE	T
2	DGNEVVEPVDYGKSKADDEFE	T
3	AEKKKTKKPSYPSSSMKSKVY	F
4	MTKDELTEEEYLSGKDYLDPP	F
5	RHKDSLAAAEYPDGMKVSNSH	F
6	GGTAVGKDLLYDGDSVKSSTD	F

In this example, there are two peptides in *P* and four peptides in *N*.

Table 3.3 One example randomized data set after permutation

Number	Peptide sequence	Class
1	DGYDRRYGDRYSPGGRSPGFE	F
2	DGNEVVEPVDYGKSKADDEFE	F
3	AEKKKTKKPSYPSSSMKSKVY	F
4	MTKDELTEEEYLSGKDYLDPP	T
5	RHKDSLAAAEYPDGMKVSNSH	T
6	GGTAVGKDLLYDGDSVKSSTD	F

data in the randomized data set are composed of the fourth peptide and fifth peptide instead of the first two peptides.

The standard permutation method is simple to implement and provides empirical *p*-values for motifs in which we are interested. However, the *p*-values for the motifs of different sizes are tested together in the standard permutation. That is, all motifs use the same permutation null.

Because the number of possible motifs increases rapidly with their sizes and motifs of the larger sizes are prone to have small *p*-values, the motifs of smaller size may be overwhelmed if we use the same permutation null. This problem will become increasingly serious as the size of motifs under investigation increases.

To overcome this limitation, Ref. [5] adopts a new permutation method that tests motifs of different sizes separately with different permutation nulls. The basic idea is to first test 1-motifs in the same manner as that in the standard permutation. In testing 2-motifs, the effects detected in the 1-motifs are incorporated into the construction of null distribution. Similarly, the effects detected in 1-motifs and 2-motifs are considered when testing the 3-motifs. This procedure continues until there are no frequent motifs with larger size that can be found for the significance testing.

3.7 Discussion and future perspective

Since the problem of phosphorylation motif discovery was introduced in 2005 by Schwartz and Gygi [6], many algorithms have been proposed from different angles [2,3,7–9]. Although some of these methods are designed regardless of the fact that the phosphorylation motif discovery problem is a discriminative pattern mining problem in essence, their key ideas have similar counterparts in the literature of discriminative pattern mining [10]. Hence, the advances in discriminative pattern mining will promote the development of more effective phosphorylation motif mining methods.

Despite the algorithmic advances in phosphorylation motif discovery, several challenging problems are still unsolved. To follow is a list of interesting and challenging problems that should be further investigated in the future.

First, although methods such as C-Motif can reduce the number of phosphorylation motifs, there are still many redundant or meaningless phosphorylation motifs that are reported to the users. Hence, it is necessary to develop more effective algorithms for further improving the precision of returned motifs. In particular, it is desirable to have a statistically sound significance measure that is able to remove the effect of submotifs in a natural manner.

Second, existing algorithms merely use the sequence data around the phosphorylation sites to conduct the analysis. This may prevent us from finding really biologically relevant patterns to derive useful scientific discoveries. To overcome this limitation, it is plausible to conduct motif search on expanded data sets that include additional features such as the three-dimensional protein structures.

Third, the construction of foreground data and background data in the literature is based on the direct extraction of a certain number of amino acids around a phosphorylation site. These amino acids from different peptides are aligned in a very rude manner. Indeed, it is highly necessary to perform sequence alignment with multiple sequence alignment algorithms before the motif extraction procedure. The challenge here is how to align thousands of short sequences rapidly and accurately.

Fourth, the permutation test procedure for testing the statistical significance of phosphorylation motifs has demonstrated its effectiveness in practice. However, there are several disadvantages in the direct permutation method. One is the p-values of a same motif may be inconsistent in different runs due to the effect of random sampling. Additionally, the computational cost of permutation-based method is very high. To obtain more accurate and stable results, more permutations have to be generated, rendering a more time-consuming task. These disadvantages limit the usability of the permutation-based method. Clearly, these disadvantages are caused by the inexact null distributions, that is, the null distribution generated in each run is only an approximate one. As a result, the permutation p-values of phosphorylation motifs calculated from it are also not exact. Hence, the algorithms that can generate an exact null distribution should be developed so as to obtain exact empirical p-values for accurately assessing the statistical significance of phosphorylation motifs.

Finally, the performance comparison of different algorithms is still not fully solved due to the lack of public benchmark data sets. In this regard, an alternative choice is to

generate simulated data with known ground truth. However, there is still no widely accepted simulation procedure for producing such synthetic data for performance evaluation. Thus, it is highly necessary to design a good simulator for this purpose.

References

[1] Y. Xue, et al., A summary of computational resources for protein phosphorylation, Curr. Protein Pept. Sci. 11 (2010) 485–496.
[2] Z. He, C. Yang, G. Guo, et al., Motif-all: discovering all phosphorylation motifs, BMC Bioinf. 12 (Suppl. 1) (2011) S22.
[3] X. Liu, J. Wu, H. Gong, et al., Mining conditional phosphorylation motifs, IEEE/ACM Trans. Comput. Biol. Bioinform. 11 (5) (2014) 915–927.
[4] R. Agrawal, R. Srikant, Fast algorithms for mining association rules, in: Proceedings of the 20th International Conference on Very Large Data Bases, 1994, pp. 487–499.
[5] H. Gong, Z. He, Permutation methods for testing the significance of phosphorylation motifs, Stat. Interface 5 (1) (2012) 61–73.
[6] D. Schwartz, S.P. Gygi, An iterative statistical approach to the identification of protein phosphorylation motifs from large scale data sets, Nat. Biotechnol. 23 (11) (2005) 1391–1398.
[7] A. Ritz, G. Shakhnarovich, A.R. Salomon, B.J. Raphael, Discovery of phosphorylation motif mixtures in phosphoproteomics data, Bioinformatics 25 (1) (2009) 14–21.
[8] Y.-C. Chen, et al., Discovery of protein phosphorylation motifs through exploratory data analysis, PLoS One 6 (5) (2011) e20025.
[9] T. Wang, et al., MMFPh: a maximal motif finder for phosphoproteomics datasets, Bioinformatics 28 (12) (2012) 1562–1570.
[10] X. Liu, J. Wu, F. Gu, et al., Discriminative pattern mining and its applications in bioinformatics. Brief. Bioinform (2015). http://dx.doi.org/10.1093/bib/bbu042.