The problem P1 is defined and investigated in Motif-All, with the following input and output.

- **Input**: a set of phosphorylated peptides \( F \) (foreground data), a set of unphosphorylated peptides \( B \) (background data), the significance threshold \( \theta_{\text{sig}} \) and the support threshold \( \theta_{\text{sup}} \). Suppose we use \( \text{sig}(m) \) to denote the statistical significance evaluation function and use \( \text{sup}(m) \) to denote the support calculation function for each motif \( m \).

- **Output**: a set of motifs \( G \), where each \( m \in G \) satisfies: (1) \( \text{sig}(m) \leq \theta_{\text{sig}} \); (2) \( \text{sup}(m) \geq \theta_{\text{sup}} \).

In the input of problem P1, it is assumed that \( \text{sig}(m) \) returns a probability value like \( p \)-value for each motif \( m \); the smaller \( \text{sig}(m) \) is, the more significant the motif \( m \) is. Intuitively, \( \text{sig}(m) \) measures the over-expressiveness of \( m \) in \( F \) against \( B \). For instance, the binomial probability model is used in MMFPh for this purpose. The support for a motif \( m \) is defined as the percentage of phosphorylated peptides that can match this motif, i.e., \( \text{sup}(m) = \frac{\text{occ}(m)|F|}{|F|} \), where \( \text{occ}(m) \) is the number of peptides from \( F \) that match \( m \) and \( |F| \) denotes the size of the set \( F \).

Now we prove the following lemma:

**Lemma 1**: Motif-All is complete for the P1 problem formulation. In other words, this algorithm finds all motifs that have support values larger than or equal to \( \theta_{\text{sup}} \) and significance values smaller than or equal to \( \theta_{\text{sig}} \).

**Proof**: the completeness of Motif-All can be shown by the following two facts. The first is that the Apriori algorithm (Agrawal and Srikant, 1994) is complete, that is, all frequent motifs are enumerated in the 1st step of Motif-All. The second fact is that the 2nd step of Motif-all only prunes motifs whose significance values are larger than \( \theta_{\text{sig}} \).

In contrast, the MMFPh algorithm is unable to achieve such a completeness property for problem P1. Suppose \( m \) is a motif that has \( k \) fixed positions. We use \( m_{(-1)} \), \( m_{(-2)} \), ..., and \( m_{(-k)} \) to denote the motifs that are subsumed by \( m \), and each of them has exactly \( k-1 \) fixed positions. That is, the only difference between \( m \) and \( m_{(-1)} \) has the \( j \)-th fixed position in \( m \) is non-fixed in \( m_{(-1)} \). If all \( m_{(-1)}^{(h)} \)s are not significant, MMFPh’s motif growing algorithm (Algorithm 1 in the supplementary document of their paper) will not check \( m \) since each \( m_{(-1)}^{(h)} \) has no chance to enter the queue for possible extension. However, it is possible that \( m \) is statistically significant even though all \( m_{(-1)}^{(h)} \)s are not significant at all. As a result, MMFPh may miss some motifs that are both frequent and significant, although all its constituent motifs are not significant.

Here, we use a simple example to show this fact.

**Example 1**: Suppose we have the foreground and background data set as shown in Figure 1. We set the significance threshold \( \theta_{\text{sig}} = 0.1 \) and the support threshold \( \theta_{\text{sup}} = 0.5 \) (i.e. the occurrence threshold is 5 in MMFPh). Under such a parameter setting, three frequent motifs are listed in Table 1 together with their support values, significance values and overall scores. The significance value and overall score are calculated according to Equations (1) and (2) in Wang et al. (2013), respectively.

According to MMFPh’s motif growing algorithm, it is obvious that motif (I....S....) and (K....S.....) will be pruned because their
The statistical significance, support and overall score of three
output.

To Wang

animo acids, we use

Z.He and H.Gong

Fig. 1. Foreground and background data

Table 1. The statistical significance, support and overall score of three
frequent motifs

<table>
<thead>
<tr>
<th>Motifs</th>
<th>Significance</th>
<th>Support</th>
<th>Overall Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>L..S........</td>
<td>0.623</td>
<td>0.5</td>
<td>0</td>
</tr>
<tr>
<td>K..S.........</td>
<td>0.623</td>
<td>0.5</td>
<td>0</td>
</tr>
<tr>
<td>IK...S......</td>
<td>0</td>
<td>0.5</td>
<td>$+\infty$</td>
</tr>
</tbody>
</table>

To claim the completeness of MMFPh with respect to P2, one
has to show that it will not miss any motif that satisfies all three
requirements in the output. Unfortunately, this is also not true. As
shown in Example 1, (IK...S......) is such a maximal, significant
and frequent motif that cannot be reported by MMFPh. This fact
illustrates the reason that MMFPh is faster than Motif-All since it
reduces the search space at the cost of missing some significant and
frequent motifs.

From the viewpoint of solving P2, the current implementation
of MMFPh should be regarded as a fast algorithm that can
approximately identify most of target motifs. To identify all
maximal, statistically significant and sufficiently frequent motifs
with MMFPh, one has to use a new significance evaluation function
that has the so-called ‘downward-closure property’ (Agrawal and
Srikant, 1994). In other words, it can be proved that MMFPh is
able to guarantee the completeness if and only if its significance
evaluation function $\text{sig}(m)$ has the following property: if $\text{sig}(m) > \theta_{\text{sig}}$, then $\text{sig}(n) < \theta_{\text{sig}}$ for any $m \subseteq n$. Unfortunately, most statistical
significance evaluation functions do not have such a property.

Alternatively, one general solution is to post-process the output of
Motif-All. More precisely, we first find the set of all frequent motifs
$R$ using the Apriori-like algorithm. Then, the significance value of
each frequent motif in $R$ is calculated to derive a new set of frequent
and significant motifs $R_S$. Finally, each motif in $R_S$ is returned if it
is not subsumed by any other motif from the same set.

2.3 Summary

Although the MMFPh method is useful, it may miss some (maximal)
statistically significant and frequent motifs. Users should be aware
of such risk in selecting the appropriate method for their tasks.

Meanwhile, the complete search of Motif-All may generate
many false positive motifs. To distinguish truly meaningful
phosphorylation motifs from false ones, one needs to perform more
rigorous statistical validation such as permutation test (Gong and He,
2013). However, the problem of accurately and efficiently assessing
the statistical significance of phosphorylation motifs still remains
unsolved. More research efforts should be devoted to this direction
in future research.

Funding: The Natural Science Foundation of China under Grant No.
61003176

Conflict of Interest: None declared.

REFERENCES


Gong, H. and He, Z. (2011) Permutation methods for testing the significance

He, Z. et al. (2011) Motif-All: discovering all phosphorylation motifs. BMC
Bioinformatics, 12, 522.

Schwartz,D. and Gygi,S.P . (2005) An iterative statistical approach to the identification
of protein phosphorylation motifs from large-scale data sets. Nat. Biotechnol., 23,
1391–1398.

Bioinformatics, 28, 1562–1570.