A Minimum Description Length Approach to the Multiple Motif Problem
Research Comps Defense

Anna Ritz, Brown University

April 11, 2008

Advisor: Ben Raphael

Committee Members:

Ben Raphael

Sorin Istrail

Chad Jenkins

Art Salomon

Contents

Posters and Publications:

Lulu Cao, Kebing Yu, Cindy Banh, Vinh Nguyen, **Anna Ritz**, Benjamin J. Raphael, Yuko Kawakami, Toshiaki Kawakami, Arthur R. Salomon. Quantitative time-resolved phosphoproteomic analysis of mast cell signaling. *Journal of Immunology* 179:5864-5876, 2007.

Anna Ritz, Benjamin J. Raphael, Arthur R. Salomon, Lulu Cao, Kebing Yu. Temporal Clustering of Protein Phosphorylation Networks, Poster Presentation, Center for Computational Molecular Biology (CCMB), Brown University, May 2007

Anna Ritz, Gregory Shakhnarovich, Benjamin J. Raphael. A Minimum Description Length Approach to Multiple Motif Finding, Poster Presentation, Workshop for Women in Machine Learning (WiML): October 2007

Anna Ritz, Gregory Shakhnarovich, Arthur R. Salomon, Benjamin J. Raphael. Identification of Protein Phosphorylation Motifs from Phosphopeptides. In preparation.

Abstract

Cells process information by passing signals between interacting proteins in a signaling network. These interactions are determined in part through patterns, or motifs, in the protein sequence. Recent technological advances make it possible to simultaneously measure many interactions in the cell, producing datasets that are mixtures of several motifs thus obscuring the identity of each motif.

We describe algorithms to discover multiple sequence motifs in such mixtures and to identify proteins that recognize the motifs. Our motif-finding algorithms derive a minimal set of motifs that distinguish a collection of measured sequences from a collection of background sequences using the principle of minimum description length (MDL) from information theory. For each identified motif, we define a motif specificity score that quantifies whether or not the sequences with a motif have a significant number of known interactions. Application of our algorithms to several recently published protein phosphorylation studies reveals several novel motifs that accurately identify important proteins in signaling networks.

1 Motivation

An organism's survival depends on the ability of its cells to perform specific tasks. Cells process information by passing signals between proteins in a signaling network. One processing mechanism for signal passing is phosphorylation, a chemical modification of a protein that acts as a functionality "on-off" switch by altering the protein's structure. *Phosphorylation* occurs when a protein kinase attaches a phosphate to a protein substrate. Conversely, dephosphorylation occurs when a protein phosphatase removes the phosphate from the protein substrate (Figure 1). Most knowledge about signaling networks has come from laborious and low-throughput experiments, where each experiment measures the interaction between a single pair of proteins. Recent technological advances make it possible to measure many protein phosphorylation states in the cell in a single experiment. Mass spectromtery, for example, measures the mass (and thus the phosphorylation state) of thousands of different proteins in a cell simultaneously.

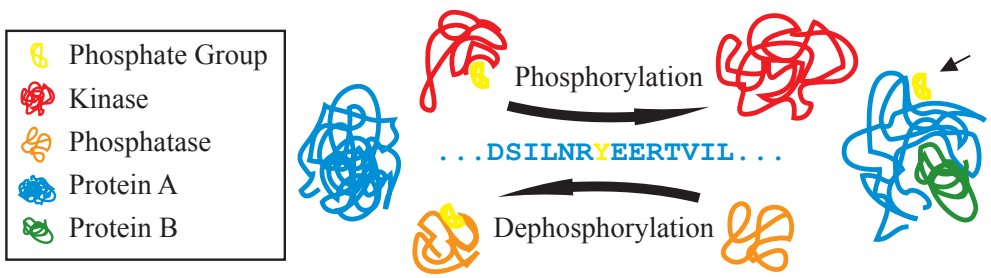

Figure 1: Phosphorylation and dephosphorylation affect the functionality of a protein. A kinase adds a phosphate at a specific location in Protein A's amino acid sequence (the yellow 'Y'). The chemical modification of Protein A (denoted by the small arrow) changes the protein's structure, and Protein A now interacts with Protein B to form a protein complex. A phosphatase removes the phosphate, returning Protein A to its native state. Note that phosphorylation might inhibit protein function and dephosphorylation might promote protein function, depending on the substrate.

Protein substrates are phosphorylated at specific positions in the protein sequence. Datasets produced by mass spectrometry experiments provide evidence of those locations, isolating short sequences (called *peptides*) of a fixed length L that surround the exact site of phosphorylation. Phosphorylation sites contain common sequence patterns, or motifs, that a kinase or phosphatase might recognize. Mass spectrometry datasets measure many kinase and phosphatase interactions,

producing a mixture of motifs present in the phosphorylated peptides. The goal is to identify phosphorylation motifs that distinguish phosphorylated peptides from unphosphorylated peptides. Identifying motifs in DNA and protein sequences is well-studied in computational biology, but the simulatenous measurement of many phosphorylated proteins introduces a new problem, the **Multiple Motif Problem**, that we state as follows.

Given: A collection of phosphorylated peptides of a fixed length L that are aligned such that the center amino acid is the same for all peptides.
Objective: A set of motifs that concisely describes the redundancy in the phosphorylated peptides.

Peptides are aligned on the same center letter because we look for kinases and phosphatases that target the same amino acid across all the peptides. A solution of the Multiple Motif Problem requires the definition of a motif and a criterion for comparing different sets of motifs. We adopt a motif model that consists of wildcard positions denoted by '`.`' that match any letter and conserved sequence positions denoted by brackets '`[]`' that match any of a list of letters. If there is more than one letter in a conserved position, we call it an *inexact* position. For example, `[DE]..pY..[IL]` is part of a documented motif [11] that has a phosphorylated tyrosine (denoted by the preceding lowercase `p`) and two conserved positions, and the peptide `E`**`D`**`AL`**`Y`**`PR`**`I`**`D` contains an *instance* of the motif. The Multiple Motif Problem is difficult because many motifs must be discovered simultaneously, and the similar chemical properties of amino acids means that motifs that often have several inexact positions. Further, the same peptide might contain an instance of more than one motif.

We frame the Multiple Motif Problem as a data compression problem and use the principle of minimum description length (MDL) [9] from information theory to solve it. The set of motifs that most parsimoniously describes the collection of peptides also reduces the amount of information required to transmit the data. This method allows motifs with arbitrary combinations of letters at conserved positions while restricting the complexity of the patterns. We developed two heuristics to find a local minimum of DL. These compare favorably to existing algorithms for the Multiple Motif Problem.

An important question is whether motifs identified via computational techniques are biologically relevant, which is hard to determine without biological experiments. We derive a Motif Specificity Score (MSS) that quantifies the extent to which the presence of a specific motif in a peptide indicates a known interaction with a kinase or phosphatase. This analysis is important for kinase/phosphatase interaction prediction.

2 Previous Work

Previous motif-finding approaches have been modified for multiple motif finding, but tend to produce overly-complicated or overly-simplified motifs. The well-known Expectation-Maximization-based algorithm MEME [2] finds multiple motifs in unaligned data by using a probabilistic erasing approach. In MEME, motifs of a fixed width w are represented as a position-weight matrix, where each position contains a letter distribution over an alphabet Σ; consequently, $w * |\Sigma|$ parameters are required to specify a motif. Finding many motifs with this representation requires a significant amount of data to correctly estimate the parameters, and MEME has not been rigorously tested on proteomic data. Tyrosine-centered mass spectromtery datasets usually have less than 1,000 short peptides, which is likely too little data for MEME to find multiple motifs.

A recently published method called Motif-X [20] finds multiple motifs by greedily extracting statistically significant motifs. However, the method only identifies motifs with a single letter at each conserved position, producing motifs with restricted expressive power. The algorithm finds the single most surprising letter at a position using the binomial distribution and the frequency of the amino acid in the background set. Motif-X then uses the letter/position pair to prune the dataset to find the next most significant letter/position pair and so on until there are no more significant pairs. This procedure produces a single motif. Once a single motif has been found, the phosphorylated peptides are pruned to remove all sequences with an instance of the motif. The motif discovery stage is repeated until there are no more statistically significant motifs in the dataset. There are two main limitations of this greedy algorithm. First, a single letter/position pair might not be significant on its own, but might be significant if considered with another letter in the same position. Since Motif-X only considers single letter/position pairs, it will not find a motif with inexact positions. Second, when peptides with a motif instance are removed from the dataset, there might be other motifs in the removed peptides that will not be identified by the method. The MDL approach addresses both of these issues.

Another algorithm called MDL-Pratt [4] greedily finds multiple motifs, but the algorithm's objective is fundamentally different than the Multiple Motif Problem. Instead of finding a set of motifs that best describes the sequences, MDL-Pratt partitions the data X into disjoint subsets $B_1, \ldots, B_k$ and motifs $m_1, \ldots, m_k$ such that $X = B_1 \cup \ldots \cup B_k$ and each sequence in B_i contains an instance of the motif m_i. This formulation defines a partitioning problem that does not allow peptides with an instance of more than one motif. Further, MDL-Pratt has the same limitations as Motif-X; namely, after finding a motif m_i and a subset of sequences B_i that contain instances of m_i, these sequences are removed from further consideration as the algorithm proceeds.

Lastly, the NetworKIN algorithm [14] is a recent attempt to map substrates with motif instances back to the kinase or phosphatase that targeted them using known protein-protein interaction networks. NetworKIN uses a database of known motifs (namely Scansite [16]) and a database of known protein interactions (namely STRING [22]) to link kinases and substrates. The NetworKIN algorithm considers each phosphorylation site separately, finding the shortest path in the STRING network between the phosphorylated and a candidate kinase. In our approach, we aim to find kinases and phosphatases that interact with many protein substrates in a single dataset.

3 Methods

We formulate the Multiple Motif Problem as the **MDL Multiple Motif Problem**, whose objective is to find a set of motifs that minimize the description length, an information-theoretic quantity that measures the amount of information (bits) required to represent (or *encode*) a collection of phosphorylated peptides. We consider the collection of N aligned peptides of fixed length L as a matrix $\mathbf{X}$. The alphabet Σ is the 20 amino acids, and the peptides are sequences from this alphabet. The goal is to describe $\mathbf{X}$ as mixture of sequences, some similar to the background and some with instances of an unknown number of motifs. We solve this problem by encoding $\mathbf{X}$ in terms of a set $\mathcal{M}$ of motifs and a background frequency distribution obtained from amino acid frequencies at each position in a larger set of unphosphorylated peptides. Since the background distribution is independent of the motif sets, we do not explicitly encode it. Thus, the number of bits $\Lambda(\mathbf{X}, \mathcal{M})$ required to encode $\mathbf{X}$ and $\mathcal{M}$ is the sum of bits required to encode the motif set and the data described by the motif set,

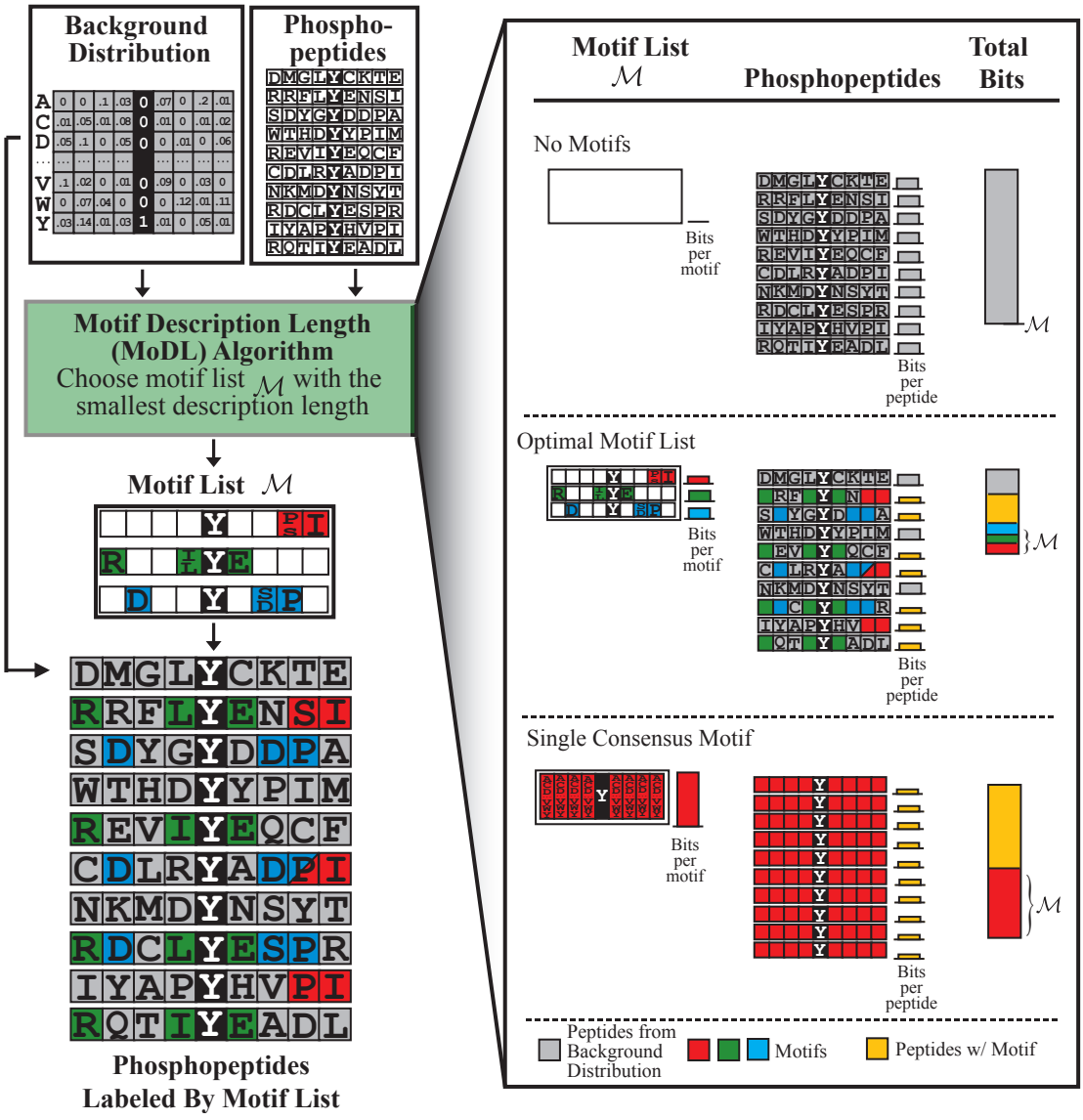

Figure 2: **Overview of the Motif Description Length (MoDL) Algorithms.** The input is a set of phosphorylated peptides and a background distribution representing the amino acid frequencies in a large set of unphosphorylated peptides. The MoDL algorithms use the description length, a measure of the amount of information (bits) required to represent the input phosphorylated peptides using a motif set $\mathcal{M}$ and the background distribution. The MoDL algorithms attempt to find the optimal motif set with minimum description length. For example, with an empty motif set (i.e. no motifs), each sequence must be described explicitly from the background distribution, yielding high description length (top right). On the opposite extreme, a single consensus motif succinctly describes all phosphorylated peptides in the input, but the consensus motif is itself complicated to describe because each amino acid at each position in the motif must be specified (bottom right). The optimal motif set includes only motifs that match several phosphorylated peptides, and minimizes the total description length required to represent both the motifs and the phosphorylated peptide sequences.

$$\Lambda(\mathbf{X}, \mathcal{M}) = \Lambda(\mathcal{M}) + \Lambda(\mathbf{X}|\mathcal{M}). \tag{1}$$

We now formulate the MDL Multiple Motif Problem as one of minimizing the description length $\Lambda(\mathbf{X}, \mathcal{M})$.

Given: An $N \times L$ matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$ of aligned sequences, where x_{ij} denotes a letter from an alphabet Σ at position j in sequence $\mathbf{x}_i$. A $|\Sigma| \times L$ matrix $\mathbf{P}$ of background distributions, where p_{st} is the frequency of letter $s \in \Sigma$ at position t.
Objective: Find a set of motifs $\mathcal{M}^* = \{m_1, \ldots, m_k\}$ that minimizes the description length:

$$\mathcal{M}^* = \operatorname*{argmin}_{\mathcal{M}} \Lambda(\mathbf{X}, \mathcal{M}).$$

For the purposes of this paper, $\mathbf{X}$ consists of sequences that have a fixed letter σ at the center position, typically a tyrosine (`Y`), serine (`S`), or threonine (`T`), letters of interest in phosphorylation studies. The sequences used to construct the background frequency matrix $\mathbf{P}$ have the same letter σ at the center position, and thus by $p_{\sigma t} = 1$ for the center position t. We modify $\mathbf{X}$ and $\mathbf{P}$ to be matrices of width $(L-1)$ because the center position does not need to be recorded. Below we describe how to compute $\Lambda(\mathbf{X}, \mathcal{M})$ for a given $\mathbf{X}$ and $\mathcal{M}$. We then describe two algorithms, **Mo**tif **D**escription **L**ength **Gr**eedy (MoDL-Gr) and **Mo**tif **D**escription **L**ength **En**umerative (MoDL-En), that find approximate solutions of the MDL Multiple Motif Problem (Figure 2).

3.1 Computing Description Length

We assume that the sequences $\mathbf{x}_1, \ldots, \mathbf{x}_N$ are independent, and so the description length of $\mathbf{X}$ is the sum of the description length of each sequence,

$$\Lambda(\mathbf{X}, \mathcal{M}) = \Lambda(\mathcal{M}) + \sum_{i=1}^{N} \Lambda(\mathbf{x}_i|\mathcal{M}). \tag{2}$$

Computing $\Lambda(\mathcal{M})$ requires encoding the motif set $\mathcal{M} = \{m_1, \ldots, m_k\}$, which we do by concatenating the encodings of the individual motifs preceded by an encoding of the value of k. Thus, if $K \geq k$ is an upper bound on the number of motifs,

$$\Lambda(\mathcal{M}) = \lceil \log_2 K \rceil + \sum_{i=1}^{k} \Lambda(m_i). \tag{3}$$

Each motif m_i is encoded in three parts:

1. $\mathbf{z}_i$ is an $(L-1)$-length binary row vector where $z_{ij} = 1$ indicates that j is a conserved position in m_i and $z_{ij} = 0$ indicates that j is a wildcard position in m_i. $\mathbf{z}_i$ requires $(L-1)$ bits to encode. Let $\gamma_i = \sum_j z_{ij}$ be the number of conserved positions in m_i.

2. Each conserved position t has a corresponding list of letters l_{it} in m_i. For each conserved position, there are two ways to encode l_{it}: as a $|\Sigma|$-length binary vector (called the *vector method*), or as a list of indices into Σ (called the *list method*). Encoding a conserved position with the vector method requires $|\Sigma|$ bits, one bit for each letter in the alphabet. Encoding a conserved position t with the list method requires $\lceil \log_2 |\Sigma| \rceil$ bits to encode the number of

letters in the list and $\lceil \log_2 |\Sigma| \rceil$ bits for each of the letters in the list. $\mathbf{c}_i$ is a row vector of length γ_i that denotes the number of bits required to encode each conserved position t using the list method, $c_{it} = (|l_{it}| + 1)\lceil \log_2 |\Sigma| \rceil$. The list method is more efficient at conserved position t when $c_{it} \leq |\Sigma|$. When the alphabet is the 20 amino acids, the vector method is more efficient for a conserved position t when there are more than three letters in the list ($c_{it} \geq 4$).

3. $\mathbf{s}_i$ is a γ_i-length binary column vector that specifies the encoding method for each conserved position, where $s_{it} = 1$ if the vector method is more efficient and $s_{it} = 0$ if the list method is more efficient for conserved position t in m_i. $\mathbf{s}_i$ requires γ bits to encode.

Thus, the description length $\Lambda(m_i)$ of motif m_i is

$$\Lambda(m_i) = |\mathbf{z}_i| + (|\Sigma| \times \vec{1})\mathbf{s}_i^T + \mathbf{c}_i(\vec{1} - \mathbf{s}_i)^T + |\mathbf{s}_i| \quad (4)$$

$$= (|\Sigma| \times \vec{1})\mathbf{s}_i + \mathbf{c}_i(\vec{1}^T - \mathbf{s}_i) + L + \gamma_i - 1, \quad (5)$$

where $\vec{1}$ is a row vector of ones and T denotes a transpose. For example, Table 1 shows the encoding of the motif `[DKIAS].Y.E`.

Variable	Value	Representation	Description Length
$\mathbf{z}_i$	[conserved,wildcard,wildcard,conserved]	`1001`	4 bits
$\mathbf{l}_{i1}$	{`D,K,I,A,S`}	`10100001100000010000` (vector for `D,K,I,A,S`)	20 bits
$\mathbf{l}_{i2}$	{`E`}	`00001` `00100` (# letters, index of `E`)	10 bits
$\mathbf{s}_i$	[vector method, list method]	`10`	2 bits
		Total:	36 bits

Table 1: Computing Λ(`[DKIAS].Y.E`).

We now turn to the task of computing $\Lambda(\mathbf{x}_i|\mathcal{M})$. We first describe how to use the background frequency matrix $\mathbf{P}$ to encode letters that are not part of a motif instance. For letter x_{ij} at position j in sequence $\mathbf{x}_i$, $p_{x_{ij}j}$ is the background frequency of that letter at position j. It has been shown that for a probability distribution over a set of characters (in our case, x_{ij}s), there exists a prefix code such that the description length of x_{ij} is $(-\log_2 p_{x_{ij}j})$ bits [8].[1] We construct an $N \times (L-1)$ matrix $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_N]^T$, where $b_{ij} = -\log_2 p_{x_{ij}j}$. Quantities that are common to all sets of motifs, such as $\mathbf{B}$ and $\mathbf{P}$, are not encoded. Let $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_k]^T$ be a $k \times (L-1)$ binary matrix that denotes the conserved positions for the k motifs. Sequence $\mathbf{x}_i$ is encoded in three parts:

1. $\mathbf{q}_i$ is a k-length binary row vector where $q_{ij} = 1$ if $\mathbf{x}_i$ contains an instance of motif m_j and 0 otherwise. $\mathbf{q}_i$ is encoded naively using k bits. However, if $\mathbf{x}_i$ has no motif instances, encoding $\mathbf{q}_i = \vec{0}$, the zero row vector, is redundant. Instead, an extra bit indicates whether $\mathbf{x}_i$ contains a motif instance or not; if it does, a k-bit vector encodes the motifs. Thus, encoding $\mathbf{q}_i$ requires $(1+k)^{1_{\mathbf{q}_i \neq \vec{0}}}$ bits, where $1_{\mathbf{q}_i \neq \vec{0}}$ is 1 if $\mathbf{x}_i$ contains a motif instance and 0 otherwise.

[1] This value might be a non-integer.

2. If $\mathbf{x}_i$ contains one or more instances of a motif, the *background letters* are the letters in positions not specified by the motifs (the wildcard positions). If $\mathbf{x}_i$ contains no motif instances, all letters in the sequence are background letters. The $(L-1)$-length vector $(\vec{1} - \mathbf{q}_i)\mathbf{Z}$ contains entries that are greater than 0 where $\mathbf{x}_i$ is a background letter. Define an $(L-1)$ row vector $\mathbf{a}_i = [a_{i1}, \ldots, a_{i(L-1)}]$ to be

$$a_{ij} = \begin{cases} 1 \text{ if } \left((\vec{1} - \mathbf{q}_i)\mathbf{Z}\right)_j > 0 \\ 0 \text{ otherwise} \end{cases}. \tag{6}$$

   $\mathbf{a}_i$ converts $(\vec{1} - \mathbf{q}_i)\mathbf{Z}$ to a binary vector. $\mathbf{b}_i$ contains the number of bits required to encode each letter in $\mathbf{x}_i$ using the background frequency, so encoding the background letters of $\mathbf{x}_i$ requires $\mathbf{a}_i\mathbf{b}_i^T$ bits, where the T denotes the transpose of $\mathbf{b}_i$.

3. The remaining letters to encode in $\mathbf{x}_i$ are the conserved positions. If a sequence $\mathbf{x}_i$ contains more than one motif instance, multiple motifs can represent a conserved position. For example, the sequence `DLYEE` contains instances of motifs `[DKIAS].Y.E` and `D[IL]Y..`, and the first position can be encoded using either motif. We choose the motif that requires the least number of bits to represent each conserved position t (in the example, the second motif would be chosen to encode the `D` in the first position). For each conserved position t, we find the length of the shortest letter list c_{jt} for motifs m_j that have an instance in $\mathbf{x}_i$ and has a conserved position at t. Define an $(L-1)$ row vector $\mathbf{d}_i = [d_{i1}, \ldots, d_{i(L-1)}]$ to be the shortest length of the letter lists at each position,

$$d_{ij} = \min_{\substack{m_l : q_{il} z_{lj} = 1, \\ 1 \le l \le k}} \lceil \log_2 c_{lj} \rceil. \tag{7}$$

   The conserved positions for $\mathbf{x}_i$ are the positions where $a_j = 0$, so encoding all the conserved positions in $\mathbf{x}_i$ requires $(\vec{1} - \mathbf{a}_i)\mathbf{d}_i^T$ bits.

Thus, the description length to encode $\mathbf{x}_i$ given a motif set $\mathcal{M}$ is

$$\Lambda(\mathbf{x}_i|\mathcal{M}) = (1+k)^{1_{\mathbf{q}_i \neq \vec{0}}} + \mathbf{a}_i\mathbf{b}_i^T + (\vec{1} - \mathbf{a}_i)\mathbf{d}_i^T. \tag{8}$$

For example, consider a uniform background matrix $\mathbf{B}$ where $b_{ij} = \log_2 20$ for $1 \le i \le N, 1 \le j \le L$. Table 2 shows the encoding of several sequences using the motif set $\mathcal{M} = \{$`[DKIAS].Y.E`, `D[IL]Y..`$\}$.

$\mathbf{x}_i$	$\mathbf{q}_i$	Bits	Background Letters	Bits	Conserved Letters	Bits
`DLYEE`	`11`	3	{`E`}	$\log_2 20$	{`D`,`L`,`E`}	$\lceil \log_2 1 \rceil + \lceil \log_2 2 \rceil + \lceil \log_2 1 \rceil$
`AKYME`	`10`	3	{`K`,`M`}	$2\log_2 20$	{`D`,`E`}	$\lceil \log_2 5 \rceil + \lceil \log_2 1 \rceil$
`SPYAR`	`00`	1	{`S`,`P`,`A`,`R`}	$4\log_2 20$	{}	0
`LRYEM`	`00`	1	{`L`,`R`,`E`,`M`}	$4\log_2 20$	{}	0
					Total:	$12 + 11\log_2 20$

Table 2: Computing $\Lambda(\mathbf{X}, \mathcal{M})$ given $\mathcal{M} = \{$`[DKIAS].Y.E`, `D[IL]Y..`$\}$ and the sequences above.

3.2 Algorithms for Minimizing Description Length

The search space of all motif sets $\mathcal{M}$ is very large - the number of motifs is exponential in the alphabet size $|\Sigma|$ and the sequence width L. This search space is reduced by only considering motifs that appear in the data, but the number is still too large to directly compute. Therefore, we have developed two heuristics to find a local minimum description length.

Algorithm 1 MoDL-Gr($\mathbf{X}$,$\mathbf{P}$)

$\mathcal{C} = \{c_1, \ldots c_R\} \leftarrow$ set of single-, double-, and triple-letter exact matches found in $\mathbf{X}$;
$t \leftarrow 0$; {iteration counter}
$\mathcal{M}^{(t)} \leftarrow \{\}$;
while $\Lambda(\mathbf{X}, \mathcal{M}^{(t)})$ has decreased in past l iterations **do**
  $\mathbf{W} \leftarrow (\mathcal{M}^{(t)} \setminus m_j), 1 \leq j \leq |\mathcal{M}^{(t)}|$ {operation 1}
    $\cup(\mathcal{M}^{(t)} \cup c_i), 1 \leq i \leq R$ {operation 2}
    $\cup(\mathcal{M}^{(t)} \cup c_i \setminus m_j), 1 \leq i \leq R, 1 \leq j \leq |\mathcal{M}^{(t)}|$ {operation 3}
    $\cup(\mathcal{M}^{(t)} \cup$ Merge(c_i,m_j)$\setminus m_j), 1 \leq i \leq R, 1 \leq j \leq |\mathcal{M}^{(t)}|$ {operation 4}
    $\cup(\mathcal{M}^{(t)} \cup$ Merge(c_i,m_j)$\setminus m_j \setminus m_k), 1 \leq i \leq R, 1 \leq j, k \leq |\mathcal{M}^{(t)}|, j \neq k$; {operation 5}
  $\mathcal{M}^{(t+1)} \leftarrow \text{argmin}_{\mathcal{W} \in \mathbf{W}} \Lambda(\mathbf{X}, \mathcal{W})$;
  $t \leftarrow t + 1$;
end while
$\mathcal{M} \leftarrow \text{argmin}_{\mathcal{M}' \in \{\mathcal{M}^{(0)}, \ldots, \mathcal{M}^{(t)}\}} \Lambda(\mathbf{X}, \mathcal{M}')$;
return $\mathcal{M}$;

Algorithm 2 Merge(m_i,m_j)

$v \leftarrow 0$; {conserved position counter}
for $t \leftarrow 1$ to $L - 1$ **do**
  if $\mathbf{c}_{it} \cup \mathbf{c}_{jt} = \{\}$ **then**
    $z_t \leftarrow 0$; {t is a wildcard position}
  else
    $z_t \leftarrow 1$; {t is a conserved position}
    $\mathbf{c}_v \leftarrow \mathbf{c}_{it} \cup \mathbf{c}_{jt}$; {create the list of letters}
    if $|\mathbf{c}_v| > 3$ **then**
      $s_v \leftarrow 1$; {use the vector method to encode t}
    else
      $s_v \leftarrow 0$; {use the list method to encode t}
    end if
    $v \leftarrow v + 1$;
  end if
end for
return $m = \{\mathbf{z}, \mathbf{C} = \{\mathbf{c}_1, \ldots, \mathbf{c}_{v-1}\}, \mathbf{s}\}$;

The algorithm **Mo**tif **D**escription **L**ength **Gr**eedy (MoDL-Gr) iteratively builds a motif set $\mathcal{C}$ from a set of simple *candidate motifs*(Algorithm 1). Candidate motifs are motifs with one, two, or three conserved positions and one letter per position (no inexact positions). We construct $\mathcal{C} = \{c_1, \ldots, c_R\}$ from the data $\mathbf{X}$. We initialize the motif set $\mathcal{M}^{(\prime)} = \emptyset$. At iteration $(t + 1)$, we construct a set $\mathbf{W}$ of potential motif sets from the motif set $\mathcal{M}^{(t)}$ from the previous iteration by performing the following operations:

1. Removing a motif m from $\mathcal{M}^{(t)}$.
2. Adding a motif $c \in \mathcal{C}$ to $\mathcal{M}^{(t)}$.
3. Adding $c \in \mathcal{C}$ to $\mathcal{M}^{(t)}$ and removing m from $\mathcal{M}^{(t)}$.
4. Merging $c \in \mathcal{C}$ with $m \in \mathcal{M}^{(t)}$, replacing m with the merged motif.
5. Merging $c \in \mathcal{C}$ with $m \in \mathcal{M}^{(t)}$, replacing m with the merged motif and removing another motif from $\mathcal{M}^{(t)}$.

Merging two motifs m_i and m_j means taking the union of the list of letters $\mathbf{c}_{it}$ and $\mathbf{c}_{jt}$ for each position t and updating the vector of conserved positions $\mathbf{z}$ and the vector of encoding methods $\mathbf{s}$ (Algorithm 2). The motif set $\mathcal{W} \in \mathbf{W}$ that has the lowest description length is chosen, and the loop proceeds until the description length has not decreased for l iterations. The purpose of the different operations used to build $\mathbf{W}$ is to try to reduce the chance of getting stuck in a local minimum.

Algorithm 3 MoDL-En($\mathbf{X}$,$\mathbf{P}$)

$\mathcal{C} = \{c_1 \ldots, c_R\} \leftarrow$ motifs with at most 2 conserved positions, where each position is either a single letter or a list of 2 letters;
$\underline{\mathcal{C}} = [\underline{c}_1, \ldots, \underline{c}_R] \leftarrow \mathcal{C}$ ordered such that $\Lambda(\mathbf{X}, \{\underline{c}_i\}) \leq \Lambda(\mathbf{X}, \{\underline{c}_{i+1}\})$;
$\mathcal{M}_s \leftarrow \underline{c}_1$;
$\mathcal{M}_d \leftarrow \operatorname{argmin}_{\{\underline{c}_i, \underline{c}_j\}:1 \leq i < j \leq 100} \Lambda(\mathbf{X}, \{\underline{c}_i, \underline{c}_j\})$;
$\mathcal{M}_t \leftarrow \operatorname{argmin}_{\{\underline{c}_i, \underline{c}_j, \underline{c}_k\}:1 \leq i < j < k \leq 25} \Lambda(\mathbf{X}, \{\underline{c}_i, \underline{c}_j, \underline{c}_k\})$;
$\mathcal{M} \leftarrow \operatorname{argmin}\{\mathcal{M}_s, \mathcal{M}_d, \mathcal{M}_t\}$;
return $\mathcal{M}$;

The algorithm **Mo**tif **D**escription **L**ength **En**umerative (MoDL-En) computes the description length of all motifs that contain at most two conserved positions and contain at most two letters per conserved position (Algorithm 3). The motifs with the smallest description length are then used to compute the description length for different combinations of two and three motifs, and the motif set $\mathcal{M}$ with the smallest description length is output.

We compare the MoDL algorithms to Motif-X [20], an algorithm where motifs are iteratively discovered by considering the most significant letters in the data according to a binomial model. Motif-X is available only as a webserver, so we implemented the algorithm in Matlab to run our own experiments. The published Motif-X algorithm required two modifications to ensure a fair comparison with the MoDL algorithms. First, the published Motif-X score is calculated as the negative log of the p-value of the binomial distribution given the *pruned* dataset, i.e. the dataset that was used to discover the motif. This score is inaccurate because sequences with the motif might have been thrown out in prior iterations. Instead, we use the negative log of the p-value of the complete data as the score to compare Motif-X with the MoDL algorithms. Second, the Motif-X webserver removes non-unique sequences from the datasets (D. Schwartz, personal communication), which might yield inaccurate motifs because peptides from multiple proteins might be the same when trimmed to peptides of length L. We remove sequences from the foreground and background with the same protein name and phosphorylation site. For these experiments, the minimum number of occurrences for a motif was 5 and the significance threshold was $p = 10^{-6}$.

In order to evaluate the performance of MoDL-Gr and MoDL-En, we tested both algorithms on synthetic datasets with motifs planted at different frequencies. We randomly extracted 100 mouse peptides of length $L = 7$ with a `Y` in the center position. We planted an instance of the motif `[DE]..pY..[IL]` at a frequency between 0 and 75 percent, generating eight different

Planted Freq.	Motifs Identified MoDL-Gr	MoDL-En	Motif-X
0%	pY[AGHLPRSWY][FMPQY]	pY[HS]	
5%	[DE]..pY	pY[HS]	
10%	pY..I [DE]..pY	[DE]..pY pY..I	
15%	D..pY..[GIL]	D..pY..[IL]	D..pY
20%	[DE]..pY..[IL]	[DE]..pY..[IL]	D..pY
25%	[DE]..pY..[IL]	[DE]..pY..[IL]	D..pY
50%	[DE]..pY..[IL]	[DE]..pY..[IL]	E..pY D..pY
75%	[DE]..pY..[IL]	[DE]..pY..[IL]	E..pY D..pY

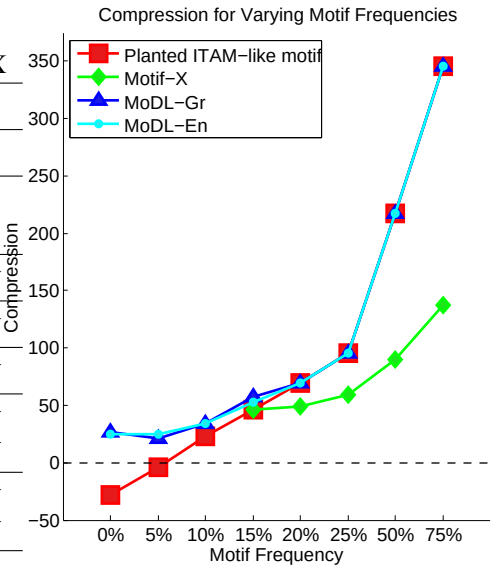

Figure 3: (Left) Motifs identified in synthetic data with the ITAM-like motif [DE]..pY..[IL] planted at frequencies from from 0% to 75%. Motif-X does not return any motifs if the frequency of the planted motif is below 15%. (Right) Compression, defined as the difference between the description length of the data with no motifs and the description length of the data with the indicated motif set, for each of the synthetic datasets. At low frequencies the planted motif does not minimize DL because there are too few motif instances compared to the background.

datasets with varying motif frequencies. This motif is part of the Immunoreceptor Tyrosine-based Activation Motif (ITAM) [11] and has two possible letters at each of two conserved positions. The planted motif frequency indicates the percent of of motif instances; for example, at 20% frequency there are approximately 5% of the sequences in the dataset matching D..pY..I, D..pY..E, E..pY..I,and E..pY..E. Motif-X cannot identify this motif because Motif-X only identifies motifs with a single letter at each conserved position. However, we expected Motif-X to recover all the conserved positions and the letters at each conserved position as multiple motifs.

The MoDL algorithms identify part of the planted motif at 10% frequency and perfectly recover the planted motif at 20% (Figure 3). By comparison, Motif-X fails to identify all conserved positions of the planted motif. One likely reason for this failure is that after the first letter of the motif is identified, all peptides that do not contain that letter are removed from the data before the next iteration. For example, suppose D..pY is the first letter found; nearly all of the sequences with the motif also have an I *or* an L in the last position. D..pY..I or D..pY..L will not be as significant as D..pY..[IL], so Motif-X will not add the I or the L to the motif.

The performance of the algorithms is quantified by computing the *compression*, defined as the difference between the description length of the data with no motifs and the description length with the motif set (Figure 3). We find that the compression of the MoDL algorithms increase as the frequency of the planted motif increases. At 20% frequency, the compression found by the MoDL algorithms is identical to the compression of the planted ITAM motif.

3.3 Motif Validation

The existence of a phosphorylation motif suggests that a kinase or a phosphatase has a preference for the phosphorylated peptides containing instances of the motif. We refer to such a set as a *motif*

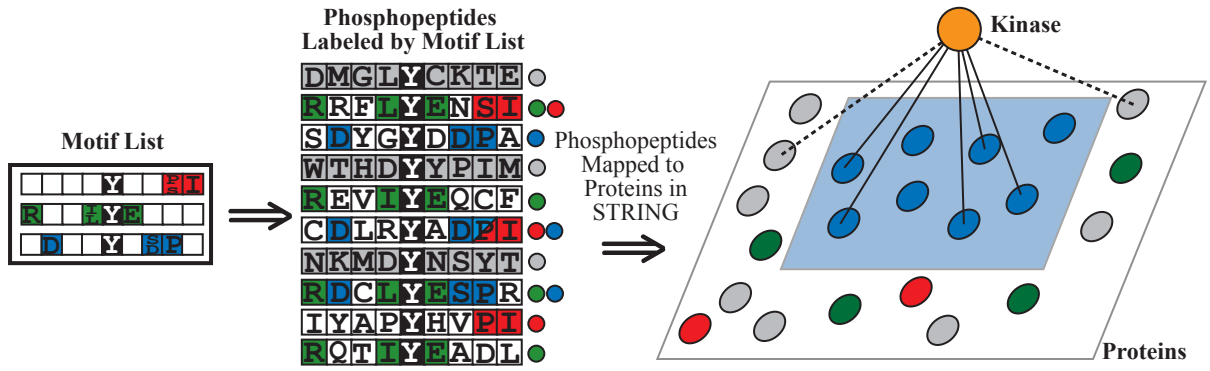

Figure 4: Computing the *Motif Specificity Score (MSS)* between a kinase and a *motif group*, the set of proteins that contain a motif instance. We map each phosphorylated peptide in the dataset to the corresponding protein in the STRING protein-protein interaction database [22]. The proteins are colored according to the motif instances they contain at the phosphorylation site. Proteins not containing motif instances are colored gray. To find the MSS for the blue motif `D..pY.[SD]P`, we consider all proteins in the motif group (blue plane). Solid lines denote interactions between the kinase and the blue motif group, and the dotted lines denote interactions between the kinase and proteins not in the blue motif group. A kinase will have a high MSS if the number of solid lines is significantly greater than the number of dotted lines.

group. For each motif group and each kinase or phosphatase, we compute the *Motif Specificity Score (MSS)* that quantifies whether the kinase or phosphatase has more interactions with the motif group than expected by chance (Figure 4). We compute the MSS for a motif m and a kinase k as follows. Let N be the number of total proteins in the dataset, M be the number of proteins that contain an instance of the motif m, and J be the number of interactions between the kinase k and the dataset; J is determined by an independent source and will be described later. The enrichment of interactions between k and the motif group is given by the hypergeometric p-value:

$$Pr[\geq l \text{ interactions}] = \sum_{i=l}^{J} \frac{\binom{J}{i}\binom{N-J}{M-i}}{\binom{N}{M}}. \tag{9}$$

We define the MSS to be

$$MSS(m,k) = -\log_{10}(Pr[\geq l \text{ interactions between kinase } k \text{ and a motif group defined by } m]). \tag{10}$$

Kinases or phosphatases are considered significant if $MSS \geq 1.3$, corresponding to a p-value of less than 0.05. A high MSS for a particular kinase/phosphatase indicates that a statistically significant number of the known interactions between the kinase/phosphatase and proteins in the dataset are interactions with proteins in the motif group. We use the STRING 7.1 database [22], a compilation of experimentally measured and predicted (from literature mining or cross-species comparisons) protein-protein interactions to compute the MSS. Note that STRING records interactions only between proteins, and contains no information about interaction sites on a protein. Thus we map each phosphorylated peptide in the input onto the corresponding protein. We use only "high-confidence" (> 0.7) edges between proteins with phosphorylated peptides and kinases/phosphatases in STRING, and some of these may not be interactions of phosphorylation or dephosphorylation since STRING fails to specify the type of interaction. This process will cause

some ambiguity for proteins with multiple phosphorylation sites. Moreover, STRING does not contain any loops in the network, and thus systematically ignores autophosphorylation, a common feature of signalling pathways. Despite these difficulties we find that a number of motifs give high MSS.

4 Results

We applied the MoDL algorithms to four phosphoproteomic datasets: mouse mast cell signaling [5], human HER2 signaling [23], human HeLa EGFR signaling [17], and signaling in various cancer cell lines [19]. In the HER2 signaling experiment [23], two cell lines (parental 'P' and a retrovirally transduced clone '24H') were stimulated with Epidermal Growth Factor (EGF) or Heregulin (HRG). These four datasets are considered separately and a fifth dataset combines all phosphoproteins measured from any experiment. The cancer cell lines studied in the last dataset include Karpas 299, Su-DHL-1, NIH/3T3, and Jurkat. We use all tyrosine-centered peptides of length L from the proteome of the species as the background set for each experiment. Table 3 presents the results of motif-finding using the MoDL algorithms and using Motif-X. In all cases, at least one of the MoDL algorithms outperform Motif-X by finding a motif set with lower description length. Moreover, in nearly every case, MoDL-Gr and MoDL-En produce at least one motif with a higher score according to Motif-X's scoring method (the negative logarithm of the binomial p-value [20]) than the best motif found by Motif-X.

The kinase LYN is known to be a crucial member of the mast cell signaling pathway[12], and strikingly the MoDL algorithms discover two motifs with high MSSs for LYN in the mast cell dataset (Table 4). The motif `[DE]..pY[ADESTY]`, identified by MoDL-Gr, appears in eight proteins that interact with LYN according to STRING, and indeed all eight of these proteins are known substrates of LYN including the FcIgE receptors [11], SYK [5], Bruton's Tyrosine Kinase (BTK) [15], DOK1 [13], and a complex of SKAP55 and FYB/SLAP130 [5]. The motif `[DE]..pY[ADESTY]` resembles the first half of the known ITAM motif `[DE]..pY..[IL]` [11] targeted by Src family kinases like LYN. While the discovered motif, `[DE]..pY[ADESTY]`, has a lower MSS for LYN than the ITAM motif (1.70 vs. 1.96), the MoDL-Gr motif appears in twice as many LYN substrates. MoDL-En produces a similar motif `D..pY` that has a high MSS for both FYN (1.4320) and LYN (1.3263), both of which are involved in FcIgE signaling [18].

Interestingly, both the known ITAM motif and the motif `[IL]pY[DE]` discovered by MoDL-En produce high MSSs for the Platelet-derived Growth Factor Receptors PDGFRa (2.0762) and PDGFRb (1.4721). This suggests potential crosstalk between the FcIgE signaling pathway and the PDGFR signaling pathway.

In the HER2 dataset (Table 5), we discover several motifs in the different experimental conditions. MoDL-Gr and Motif-X find the motif `pY..P` in all conditions, and this motif is part of the known ABL consensus motif `A.VIpYAAP` [21]. Notably, ABL has the highest MSS of all kinases and phosphatases, and MSS for ABL is highest (6.127) in the parental EGF-stimulated experiment. Many of the proteins in the `pY..P` motif group are known ABL targets, but not all. In particular, GRF1 is in the motif group but there is no interaction between ABL and GRF1 recorded in STRING. However, GRF1 is phosphorylated by the BCR-ABL fusion protein at the measured phosphorylation site Y1106 [7], whose peptide sequence matches the motif (`NEEENI`**`pY`**`SV`**`P`**`HDS`). Thus, the motif and MSS are useful for predicting new kinase substrates. The motif `pY..[PV]` found in the parental cell line datasets is similar to the ABL consensus and also has a high MSS for ABL. MoDL-Gr also identifies a motif `[DENS][DNPRS].pY` that has high MSS (2.638)

Dataset	MoDL-Gr Motifs	Score	MoDL-En† Motifs	Score	Motif-X‡ Motifs	Score
Cao et. al.	[DE]..pY[ADESTY]	64.75	[IL]pY[DE]	43.73	D..pYE	30.31
142 sequences	IpY	21.62	D..pY	31.60	IpY	21.62
					E..pY	13.64
Compression	**159.30 Bits**		**131.06 Bits**		**37.92 Bits**	
W-Y. et. al.	[DE]..pY	31.65	[DE]..pY	31.65	D..pY	19.45
P_EGF	pY..[PV]	20.74	pY..[PV]	20.74	pY..P	16.12
229 sequences						
Compression	**122.92 Bits**		**122.92 Bits**		**114.82 Bits**	
W-Y. et. al.	[DE]..pY	30.18	[DE]..pY	30.18	D..pY	16.27
P_HRG	pY..[PV]	18.78	[DP].pY	15.69	pY..P	15.84
191 sequences						
Compression	**100.80 Bits**		**105.90 Bits**		**92.22 Bits**	
W-Y. et. al.	[ADEN][ADLP].pY	47.75	[DE]..pY	35.15	D..pY	17.41
24H_EGF	pY..P	18.92	[DP].pY	18.96	E..pY	16.71
225 sequences					pY..P	18.92
Compression	**140.17 Bits**		**130.17 Bits**		**83.78 Bits**	
W-Y. et. al.	[DENS][DNPRS].pY	43.65	[DE]..pY	33.04	D..pY	17.98
24H_HRG	pY..P	16.08	[DP].pY	17.66	pY..P	16.08
209 sequences						
Compression	**119.09 Bits**		**120.26 Bits**		**103.52 Bits**	
W-Y. et. al.	[DE]..pY	42.26	pY..P	31.95	D..pY	25.20
All Exp.	pY.[DESV][LPV]	48.46	pY[AD]	20.74	pY..P	31.95
299 sequences					E..pY	16.37
					pY.VP	34.19
Compression	**211.50 Bits**		**196.21 Bits**		**33.36 Bits**	
NPM-ALK	H.G[EV][KN]PpY.C..[CR]G	22.06	pY..V	29.68	pY..V	29.68
248 sequences			[DE]..pY.....[GK]	29.70	E..pY	20.33
			[ST]...[IP]pY	28.50	IpY	18.34
Compression	**358.41 Bits**		**243.03 Bits**		**144.11 Bits**	
c-Src	pY[DS]	34.73	pY[DS]	34.73	pYS	18.85
185 sequences					pYD	14.52
Compression	**100.02 Bits**		**100.02 Bits**		**82.35 Bits**	
Jurkat	[DY]...pY	22.24	[DE]..pY	30.24	D...pY	18.05
184 sequences			[DY]...pY	22.24	pY..P	16.71
Compression	**72.94 Bits**		**109.60 Bits**		**95.51 Bits**	

†The min. # of occurrences for a motif is 5% of the dataset size.

‡The threshold parameter is 10^{-6} and the min. # of occurrences is 5% of the dataset size.

Table 3: Motifs identified by each algorithm in each dataset. The first four Wolf-Yadlin et. al. experiments measure two cell lines (parental 'P' and a clone '24H') stimulated under two conditions ('EGF' and 'HRG'). The last Wolf-Yadlin et. al. experiment ("All Exp.") consists of phosphorylated peptides measured in any of the the four experimental conditions. Score is the negative logarithm of the binomial p-value computed using the percentage of phosphorylated peptides with the motif and the percentage of background peptides with the motif. Motif-X aims to maximize this score. Compression is the difference between the the description length of the motif set and the description length of the null model (no motifs). The MoDL algorithms aim to maximize this value.

Kinase	MSS	# Interactions in Motif Group	Total # Interactions	Interacting Proteins in Motif Group
Motif: [DE]..pY[ADESTY] FROM: MoDL-Gr (found in 45 proteins)				
LYN	1.7007	8	11	BTK,DOK1 p62,FcIgER β,FcIgER γ,FYB,GAB2,SKAP55R,SYK
Motif: [IL]pY[DE] FROM: MoDL-En (found in 24 proteins)				
PDGFRb	1.4721	4	7	PI 3 p85 α,PI 3 p85 β,PLC γ 1,PLC γ 2
Motif: D..pY FROM: MoDL-En (found in 32 proteins)				
FYN	1.4320	5	8	FcIgE R γ,GAB2,PI 3 p85 α,SHC,SKAP55R
LYN	1.3263	6	11	BTK,DOK1; p62dok,FcIgE R β,FcIgE R γ,GAB2,SKAP55R,
ZAP70	1.3263	6	11	FcIgE R γ,Fyn,GAB2,PI 3 p85 α,PI 3 p85 β,SHC
Motif: D..pYE FROM: Motif-X (found in 12 proteins)				
TRKA	1.5463	2	3	PI 3 p85 α,PI 3 p85 β
ILK	1.5463	2	3	PI 3 p85 α,PI 3 p85 β
Motif: [DE]..pY..[IL] FROM: ITAM (found in 11 proteins)				
FYN	2.5667	4	8	FcIgE R γ,FYB,SHC,SKAP55R
PDGFRa	2.0762	2	2	PLC γ 1,SHC
LYN	1.9626	4	11	FcIgE R β,FcIgE R γ,FYB,SKAP55R
TRKB	1.6228	2	3	PLC γ 1,SHC
RET	1.3454	2	4	PLC γ 1,SHC

Table 4: **Motif Specificity Scores for the motifs discovered in the mast cell dataset of Cao et. al.** The ITAM motif [DE]..pY..[IL] is from the literature [11] and was not recovered by an algorithm. Motif Specificity scores were computed for the 115 proteins in this dataset appear in the STRING database, and only 'high-confidence" interactions in STRING (score >0.7) were considered. Kinases with MSS $\geq$ 1.3 are shown.

for the phosphatase PTPN11/SHP2, a known member of the EGFR/HER2 pathway, and a motif `[ADEN][ADLP].pY` with high MSS for ERBB4 (1.846).

Finally, the NPM-ALK condition from the Rush et. al. dataset produces the specific motif `H.G[EV][KN]PpY.C..[CR]G`. This motif matches multiple phosphorylated peptides in five zinc finger proteins (ZFPs): hypothetical protein MGC12466 (also known as zinc finger protein 670 [10]), similar to ZFP 91, TIP20, ZNF24, and ZNF264. Additionally, this motif matches 11 peptides from these five proteins that were not measured as phosphorylated. Each of the 21 motif hits in the peptide sequences occur exactly after the zinc finger domain indicated by the consensus sequence `C.{2-4}C...F.....L..H...H` [6], where the first `H` in the motif is the last histidine in the zinc finger domain consensus. In the Rush et. al. dataset (Table 6), the common motifs `pY..P` and `[DE]..pY` from the Jurkat cell line condition again have a high MSS for ABL; here, however, `[DE]..pY` has the higher score. There is a significant overlap of measured phosphorylated peptides between this experiment and the Wolf-Yadlin et. al. data (specifically DOK1, PXN, ZAP-70, PI34K p85, and PLG-γ) that contribute the high ABL MSS score for both datasets.

5 Discussion

We have described an MDL-based formulation of the Multiple Motif Problem and two MoDL-based algorithms to discover protein phosphorylation motifs. We also defined a motif specificity score (MSS) to identify a kinase or phosphatase that interacts with a given motif. The MoDL motif-finding algorithms outperform Motif-X on phosphotyrosine datasets according to several criteria: reduction in description length, Motif-X's statistical score, and the motif specificity score (MSS). Another advantage of the MoDL method is that it does not require a choice of parameters such as the significance thresholds required by Motif-X. The motifs discovered with the MoDL algorithms are quite short, which is consistent with earlier studies [20] and various databases of phosphorylation motifs [1, 3, 16]. Since the sequence specificity of such short motifs will be very low, *de novo* prediction of phosphorylation sites using sequence motifs alone will likely yield many false positives. Nevertheless, we have derived the Motif Specificity Score, which combines motifs with prior knowledge of protein interactions.

Computing the MSS of each motif we discovered, we identified several kinase/phosphatase-substrate interactions. Many of these relationships are consistent with known interactions in the studied pathways, but we also obtain several novel predictions of interactions that were not recorded in the STRING database (e.g. the ABL-GRF1 interaction in the Wolf-Yadlin et. al. dataset).

We have demonstrated that the combination of phosphoproteomic data, motif-finding and prior knowledge of protein interactions, as recorded in protein-protein interaction databases, is a powerful paradigm for linking kinases/phosphatases to their substrates in an experimentally stimulated signaling pathway.

6 Future Work

Further improvements to the MoDL algorithms are possible. In particular, in all biological datasets, we observed that the MoDL algorithms return at most three motifs. Thus, the metric of minimum description length might be too restrictive for identifying the *biologically* important qualities of phosphorylation motifs. Notably, the motif `[DE]..pY[ADESTY]` identified by MoDL-Gr in

Experiment	Motifs									
	D..pY$^{(x)}$	E..pY$^{(x)}$	[DE]..pY$^{(g,e)}$	pY..P$^{(g,x)}$	pY.VP$^{(x)}$	pY..[PV]$^{(g,e)}$	[DP].pY$^{(e)}$	[ADEN][ADLP].pY$^{(g)}$	[DENS][DNPRS].pY$^{(g)}$	pY[AD]$^{(e)}$
P_EGF	PYK2(1.684)		PTPN12(1.505)	ABL(6.127)		ABL(3.763)				
74 Proteins	ZAP70(1.457)			HCK(2.131)		TYK2(1.737)				
				KIT(1.676)		LYN(1.444)				
P_HRG	$\emptyset$		PTPN12(1.541)	ABL(3.340)		ABL(1.871)	HCK(2.144)			
51 Proteins			MET(1.331)	HCK(2.571)		PDGFRa(1.463))	LYN(1.733)			
				ILK(1.510)		HCK(1.463)	ErbB4(1.715)			
24H_EGF	FMS(1.683)	$\emptyset$	HCK(1.490)	ABL(2.339)			MET(1.953)	FLT3(1.846)		
33 Proteins	ZAP70(1.348)		FMS(1.490)				FMS(1.757)	ErbB4(1.846)		
	JAK1(1.348)						ErbB4(1.318)	TEC(1.846)		
24H_HRG	AXL(1.882)		ZAP70(1.873)	ABL(3.467)			MET(2.648)		PTPN11(2.638)	
59 Proteins	ZAP70(1.838)		PTPN12(1.619)	HCK(1.715)			HCK(2.570)		SYK(2.513)	
	JAK1(1.583)		FLT1(1.324)	CTK(1.410)			PTPN11(2.234)		MET(2.504)	
All	AXL(2.175)	FYN(1.828)	SYK(2.145)	ABL(5.142)	PTPRH(1.788)					INSR(2.404)
Experiments	ITK(1.649)	EphB1(1.425)	FYN(1.552)	HCK(2.647)	CSK(1.523)					FYN(1.828)
148 Proteins	SYK(1.443)		AXL(1.541)	MET(2.307)	TGFbR1(1.267)					PTPRM(1.806)

pY.[DESV][LPV]$^{(g)}$ did not produce any kinases or phosphatases with significant MSSs.

^{x}Motif found with Motif-X, ^{g}Motif found with MoDL-Gr, ^{e}Motif found with MoDL-En

Table 5: **Kinases and phosphatases with the highest Motif Specificity Scores (MSSs) in the HER2 (Wolf-Yadlin et. al.) dataset [23].** The first four rows give the results for datasets from two cell lines (parental 'P' and a clone '24H') stimulated under two conditions ('EGF' and 'HRG'). The last row ("All Experiments") consists of phosphorylated peptides measured in any of the the four experimental conditions. For each experiment and each motif, the three kinases or phosphatases with the greatest MSS (indicated in parentheses) above 1.3 are reported. The empty set $\emptyset$ indicates that the motif was found, but there were no kinases or phosphatases with significant MSSs for the given dataset.

Kinase/ Phosphatase	MSS	# Interactions in Motif Group	Total # Interactions	Interacting Proteins in Motif Group
Dataset: c-Src, Motif: pYD FROM: Motif-X (found in 14 proteins)				
PTPRH	1.6908	2	2	P130Cas,RA70
Dataset: Jurkat, Motif: pY..P FROM: Motif-X (found in 20 proteins)				
LCK	2.0201	5	8	CD28,Dok2,Dok1,PXN,ZAP70
ABL	1.4715	4	7	Dok2,Dok1,PXN,ZAP70
Dataset: Jurkat, Motif: [DE]..pY FROM: MoDL-En (found in 36 proteins)				
ABL	3.1218	7	7	cortactin,Dok2,Dok1,PXN,PLC-γ1,PI3K p85-α,ZAP70
RET	2.1867	5	5	Dok2,5 Golgin-84,Dok1,PXN,PLC-γ1
MET	1.7328	4	4	cortactin,PLC-γ1,PI3K p85-α,Ets-1
PTPN11	1.7328	4	4	Lck,PXN.PLC-γ1,PI3K p85-α
SRC	1.4995	8	12	cortactin,Dok2,GIT1,Dok1,PXN,PLC-γ1,PI3K p85-α,Ets-1
LCK	1.4987	6	8	Dok2,Dok1,PXN,PLC-γ1,PI3K p85-α,ZAP70

Table 6: **Motif Specificity Scores for the motifs discovered in the Rush et. al. datasets.** Motif Specificity scores were computed for the proteins in these datasets that appear in the STRING database, and only 'high-confidence" interactions in STRING (score $>$0.7) were considered. Kinases and phosphatases with MSS $\geq$ 1.3 are shown.

the Cao et al. dataset includes several motifs for c-Src kinase that were identified by Motif-X in the Rush et. al. dataset [20]; MoDL-Gr either found a more specific motif or combined several biologically distinct motifs into one. The MoDL algorithms can be modified to return a user-specified number of motifs, allowing the user to incorporate prior knowledge about the number of interactions expected in a dataset. Additionally, the same letters tend to appear at multiple positions in many motifs (for example D,E, and P). While the current motif representation used in MoDL does not allow gaps with varying length, MDL-Pratt [4] incorporates variable-length gaps in their description length computation.

Another possibility is to reduce the alphabet of amino acids; the chemical properties of some amino acids are similar, and an 11-letter alphabet might reduce the motif search space without loss of motif specificity. Incorporating the MSS score in the motif discovery stage will explicitly identify motifs with high MSSs in the protein-protein interaction network. Finally, the biological utility of this method will be improved if it is freely available to researchers; thus, a webserver is in development that will allow aligned sequences as input and motifs and significant kinases and phosphatases as output.

References

[1] Ramars Amanchy, Balamurugan Periaswamy, Suresh Mathivanan, Raghunath Reddy, Sudhir Gopal Tattikota, and Akhilesh Pandey. A curated compendium of phosphorylation motifs. *Nat Biotechnol*, 25(3):285–286, 2007.

[2] T L Bailey and C Elkan. The value of prior knowledge in discovering motifs with MEME. *Proc Int Conf Intell Syst Mol Biol*, 3:21–29, 1995.

[3] Sudha Balla, Vishal Thapar, Snigdha Verma, Thaibinh Luong, Tanaz Faghri, Chun-Hsi Huang, Sanguthevar Rajasekaran, Jacob J del Campo, Jessica H Shinn, William A Mohler,

Mark W Maciejewski, Michael R Gryk, Bryan Piccirillo, Stanley R Schiller, and Martin R Schiller. Minimotif Miner: a tool for investigating protein function. *Nat Methods*, 3(3):175–177, 2006.

[4] A. Brazma, I. Jonassen, E. Ukkonen, and J. Vilo. Discovering patterns and subfamilies in biosequences. *Proc Int Conf Intell Syst Mol Biol*, 4:34–43, 1996.

[5] L. Cao, K. Yu, C. Banh, V. Nguyen, A. Ritz, B.J. Raphael, Y. Kawakami, T. Kawakami, and A.R. Salomon. Quantitative time-resolved phosphoproteomic analysis of mast cell signaling. *J. Immunol.*, 179:5864–5876, 2007.

[6] R.D. Finn, J. Mistry, B. Schuster-Bckler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S.R. Eddy, E.L. Sonnhammer, and A. Bateman. Pfam: clans, web tools and services. *Nucleic Acids Res.*, 34:D247–251, Jan 2006.

[7] V.L. Goss, K.A. Lee, A. Moritz, J. Nardone, E.J. Spek, J. MacNeill, J. Rush, M.J. Comb, and R.D. Polakiewicz. A common phosphotyrosine signature for the Bcr-Abl kinase. *Blood*, 107:4888–4897, 2006.

[8] P.D. Grunwald, editor. *The Minimum Description Length Principle*. MIT Press, 2007.

[9] P.D. Grunwald, I.J. Myung, and M. Pitt, editors. *Advances in Minimum Description Length: Theory and Applications*. MIT Press, 2004.

[10] M. Hirakawa. HOWDY: an integrated database system for human genome research. *Nucleic Acids Res.*, 30:152–157, 2002.

[11] S.A. Johnson, C.M. Pleiman, L. Pao, J. Schneringer, K. Hippen, and J.C. Cambier. Phosphorylated immunoreceptor signaling motifs (ITAMs) exhibit unique abilities to bind and activate Lyn and Syk tyrosine kinases. *J. Immunol.*, 155:4596–4603, 1995.

[12] M.H. Jouvin, M. Adamczewski, R. Numerof, O. Letourneur, A. Vall, and J.P. Kinet. Differential control of the tyrosine kinases Lyn and Syk by the two signaling chains of the high affinity immunoglobulin E receptor. *J. Biol. Chem.*, 269:5918–5925, 1994.

[13] X. Liang, D. Wisniewski, A. Strife, R. Shivakrupa, B. Clarkson, and M.D. Resh. Phosphatidylinositol 3-kinase and Src family kinases are required for phosphorylation and membrane recruitment of Dok-1 in c-Kit signaling. *J. Biol. Chem.*, 16:13732–13738, 2002.

[14] Rune Linding, Lars Juhl Jensen, Gerard J Ostheimer, Marcel A T M van Vugt, Claus Jorgensen, Ioana M Miron, Francesca Diella, Karen Colwill, Lorne Taylor, Kelly Elder, Pavel Metalnikov, Vivian Nguyen, Adrian Pasculescu, Jing Jin, Jin Gyoon Park, Leona D Samson, James R Woodgett, Robert B Russell, Peer Bork, Michael B Yaffe, and Tony Pawson. Systematic discovery of in vivo phosphorylation networks. *Cell*, 129(7):1415–1426, 2007.

[15] A.J. Mohamed, B.F. Nore, B. Christensson, and C.I. Smith. Signalling of Bruton's tyrosine kinase, Btk. *Scand. J. Immunol.*, 49:113–118, 1999.

[16] John C Obenauer, Lewis C Cantley, and Michael B Yaffe. Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res*, 31(13):3635–3641, 2003.

[17] Jesper V Olsen, Blagoy Blagoev, Florian Gnad, Boris Macek, Chanchal Kumar, Peter Mortensen, and Matthias Mann. Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell*, 127(3):635–648, 2006.

[18] V. Parravicini, M. Gadina, M. Kovarova, S. Odom, C. Gonzalez-Espinosa, Y. Furumoto, S. Saitoh, L.E. Samelson, J.J. O'Shea, and J. Rivera. Fyn kinase initiates complementary signals required for IgE-dependent mast cell degranulation. *Nat. Immunol.*, 3:741–748, 2002.

[19] J. Rush, A. Moritz, K.A. Lee, A. Guo, V.L. Goss, E.J. Spek, H. Zhang, X.M. Zha, R.D. Polakiewicz, and M.J. Comb. Immunoaffinity profiling of tyrosine phosphorylation in cancer cells. *Nat. Biotechnol.*, 23:94–101, 2005.

[20] Daniel Schwartz and Steven P Gygi. An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. *Nat Biotechnol*, 23(11):1391–1398, 2005.

[21] Z. Songyang and L.C. Cantley. Recognition and specificity in protein tyrosine kinase-mediated signalling. *Trends Biochem. Sci.*, 20:470–475, 1995.

[22] Christian von Mering, Lars J Jensen, Michael Kuhn, Samuel Chaffron, Tobias Doerks, Beate Kruger, Berend Snel, and Peer Bork. STRING 7–recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res*, 35(Database issue):358–362, 2007.

[23] Alejandro Wolf-Yadlin, Neil Kumar, Yi Zhang, Sampsa Hautaniemi, Muhammad Zaman, Hyung-Do Kim, Viara Grantcharova, Douglas A Lauffenburger, and Forest M White. Effects of HER2 overexpression on cell signaling networks governing proliferation and migration. *Mol Syst Biol*, 2:54, 2006.