

Biological sequence motif discovery using *motif-x*.

Michael F. Chou¹ and Daniel Schwartz²

¹ Department of Genetics
Harvard Medical School
Boston, MA
Email: mchou(at)genetics.med.harvard.edu

² Department of Physiology and Neurobiology
University of Connecticut
Storrs, CT
Email: daniel.schwartz(at)uconn.edu

Abstract: The web-based *motif-x* program provides a simple interface to extract statistically significant motifs from large data sets such as MS/MS post-translational modification data and groups of proteins that share a common biological function. Users upload data files and download results using common web browsers on essentially any web-compatible computer. Once submitted, data analyses are performed rapidly on an associated high-speed computer cluster and they produce both syntactic and image-based motif results and statistics. The protocols presented demonstrate the use of *motif-x* in three common user scenarios.

Key terms: protein motif, phosphorylation, post-translational modification (PTM), motif discovery, *motif-x*, mass spectrometry, proteomics.

INTRODUCTION

Using the tools of mass spectrometry and spectral identification, large-scale proteomic experiments are now able to identify thousands of protein post-translational modifications (PTMs) in a single experimental run, and the technology to enrich for different modifications such as phosphorylation, acetylation, glycosylation, and others has been steadily improving. Because of the importance of post-translational modifications in normal and pathologic cellular physiology, the ultimate goal of measurement is to understand the underlying biological processes that lead to these modifications and the consequences thereof. Knowing, e.g., the preference of an enzyme for its natural substrates can help elucidate biological pathways in which they are involved.

Because part of the biochemical preference of an enzyme for a given substrate may be determined by residues immediately surrounding the site of action, biochemists have focused on identifying the critical neighboring residues that give rise to specific enzyme-substrate interactions. This pattern of residues along the short span of a protein or polypeptide is known as a short linear motif (or simply a *motif* for short). Motifs have historically been studied using mutagenesis experiments such as alanine scanning (Koyasu et al., 1994) and more recently, combinatorial library approaches (Hutti et al., 2004).

The *motif-x* algorithm for computationally determining short linear motifs was first described by Schwartz and Gygi in 2005 (Schwartz and Gygi, 2005), and an online implementation was made available by Chou and Schwartz at that time (<http://motif-x.med.harvard.edu/>).

Supplemental Data Files

Files providing supplemental data for all of the protocols presented in this unit may be downloaded at <http://www.currentprotocols.com/protocol/bi1315>.

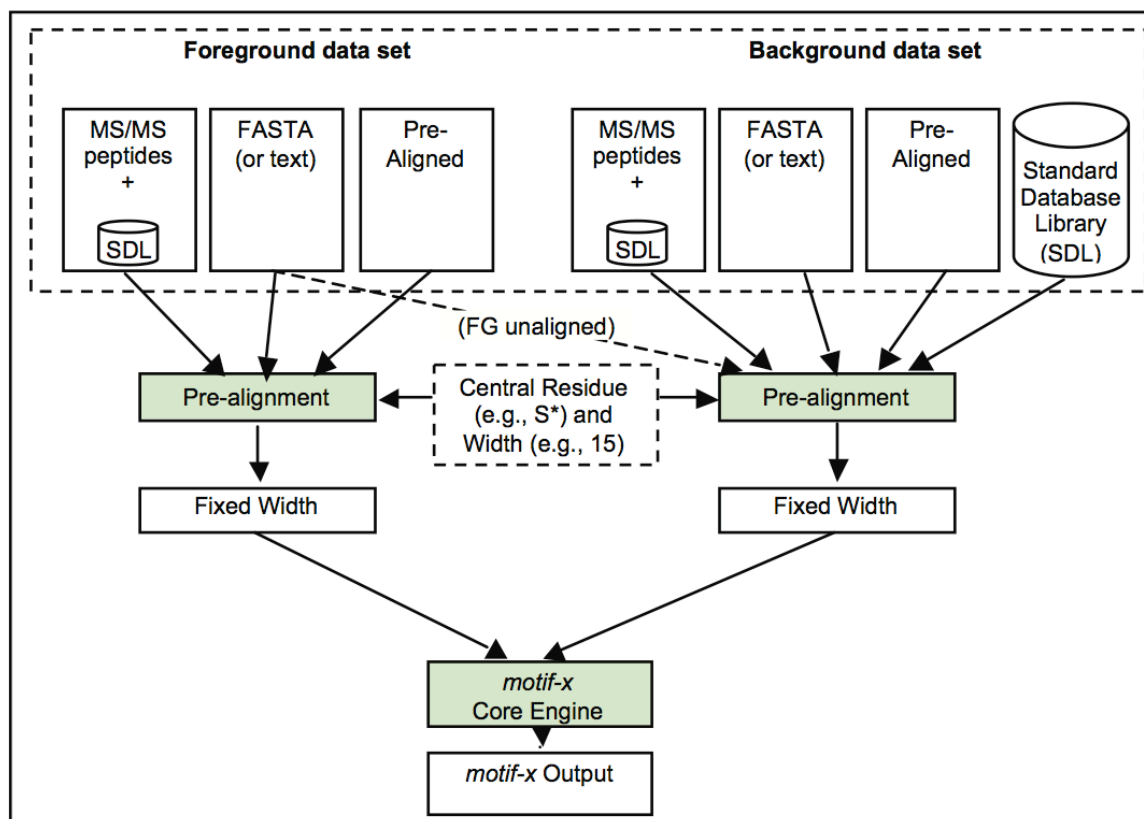
STRATEGIC PLANNING

As described in the examples that follow, *motif-x* was initially designed to extract statistically significant biological motifs surrounding sites of post-translational modification that are identified by mass spectrometry. However, as will be demonstrated below, *motif-x* can also be used to perform analyses using different data sources or biological conditions.

Flowchart 1 shows an overview of the variety of inputs available for use with *motif-x*. There are three possible types of sequence input data that are referred to as **foreground data sets** (the left side of the dashed box at the top). When a data set is submitted on the web server, the user specifies the type of input data, and, if necessary, *motif-x* translates that data into an internal format that it uses to find motifs. The three types of data that are used as input are:

(a) Tandem Mass Spectral (MS/MS) data that is often in the form of lists of tryptic peptides containing flags indicating which residues contain a PTM (this method is described in the Basic Protocol below, and is the most common application);

(b) Pre-aligned data sets which are increasingly being published by web sites containing post-translational modifications (this method is described in the Alternate Protocol 1); and



Flowchart 1

Overview of motif-x. The user provides foreground and background data sets for analysis (top dashed box) and other analysis parameters (middle dashed box). The use of MS/MS data sets always requires the specification of the organism it comes from (see text), and Standard Database Libraries (SDLs) for common model organisms are provided on the server for this and other purposes as shown. Light green boxes show processing steps while non-colored boxes indicate data inputs, data outputs or intermediate results.

(c) FASTA-formatted data sets which may be a set of proteins with known modifications or binding sites that the user believes to share a common motif (this method is described in the Alternate Protocol 2).

Table 1 summarizes the various researcher objectives that can be addressed using *motif-x*, as well as the format of the data used for each respective analysis.

Because residues that are significant for binding or substrate recognition can be found both upstream and downstream of a modification site, *motif-x* is designed to build motifs around a 'central residue'. So, for example substrates of the human protein kinase A (PKA) will often share the motif RRxS in common. That is, a serine on a substrate that is a phospho-acceptor will often have arginines at the second and third residues upstream from the serine which accounts for much of the specificity for that substrate modification site. Similarly for the class of proline-directed kinases (such as cyclin dependent kinases (CDKs) or mitogen-activated protein kinases (MAPKs)), substrates usually have the motif SP where a proline exists one residue downstream of the serine phospho-acceptor site. *motif-x* can take as input up to thousands of substrates and determine one or more statistically overrepresented motifs in such data sets.

By centering the analysis on a key residue (e.g., a serine known to be phosphorylated), *motif-x* is able to generate a list of overrepresented motifs that may contain residues both upstream and downstream from the key central residue.

Table 1. Researcher objectives met by using *motif-x* and sample formatting for various input data types.

| Researcher objective | Foreground format parameter that should be selected | Example input data |
|--|--|--|
| To find statistically overrepresented motifs from raw MS/MS PTM sequencing data. | MS/MS | DFDM#SFDSKLS*FDKDFFDK DTC@KDLFDS*FADGDR SDSASDAT*FFDSSAK |
| To find statistically overrepresented motifs from previously aligned sequence data (e.g., substrates of a kinase, where the central residue represents the modified site). | pre-aligned | CRPWPYSYAAKKF GTVTPDSRRGQAN LPQSYYSTKPVLK KDQASISEKRVTK |
| To find statistically overrepresented motifs directly from whole protein sequences. | FASTA | >Protein 1 MKAAADEKSLISAIDSYLPRQG ESLADRVSPKAKSFKGITGRDV HQFSTPNRKPVGYSIHKVDSIV SMCVKQTVI >Protein 2 MTGKASFEFVRLTKGGGIRSSR DKLGEICKPLFKGHPAVDPMTV TSVEPPKSLGKSEDLYFYIHAD CIEHQLLITD |
| To find statistically overrepresented motifs from a linguistic text file. | Text | Moby Dick Chapter 1 Call me Ishmael. Some years ago never mind how long precisely having little or no money in my purse, and nothing particular to interest me on shore, I thought I would sail about a little and see the watery part of the world. |

As described in the Basic Protocol, tandem mass spectrometry data is a natural form of input for such analyses because other programs such as SEQUEST (Eng et al., 1994) or Mascot (Perkins et al., 1999), can easily identify the actual modified residues in a sample. With regard to motif extraction, a potential drawback of MS/MS data sets is that they are fragmented (usually into tryptic peptides), completely unaligned, and each peptide may contain one or more modified residues. However, for this form of input data, *motif-x* has a special built-in preprocessing function which converts these data sets into aligned data sets whereby each modification is converted into a fixed-width peptide aligned on the modified residue (the native format used by the *motif-x* engine). In order to perform this conversion *motif-x* requires the user's specification of an organism.

The example in Alternate Protocol 1 has a goal similar to that of the basic protocol, and analysis proceeds the same way, except for the fact that the input data in this case is already in the native form with fixed-width peptides surrounding a key central residue.

Alternate Protocol 2 demonstrates the fact that data with a known modification site is not the only data format that can be handled by *motif-x*. Because each of the 20 normal amino acids has a unique biochemical functionality, it is possible to use any residue as a central residue even if a clear modification is not known. This works because *motif-x* is an unbiased, statistically based algorithm which does not make a *priori* assumptions regarding the function of the patterns it extracts.

An important aspect of the *motif-x* algorithm is the utilization of a "background data set" for the determination of significant residues. Unlike many other motif finding

algorithms that either use no background frequencies or simply use the frequencies of the 20 amino acids in the proteome as a background, *motif-x* (most typically) uses the entire proteome of the organism under study as a background to empirically determine the conditional probabilities required for significance as the algorithm proceeds.

Amongst other benefits, this allows *motif-x* to identify significant motifs containing correlated residues and motifs that are directly relevant to a particular organism.

Thus, the user must specify a background data set which, like the foreground data set, may be in a number of different formats, but is most often simply selected from the set of built-in FASTA-formatted Standard Database Libraries from a number of different model organisms (“SDLs” on the right side of the dashed box at the top of Flowchart 1).

Once launched, *motif-x* searches for any significant motifs in the foreground data set using the context of the background data set, and reports any motifs found to be overrepresented in the foreground data set with respect to the background data set.

The output is in both standard syntactic form, which indicates the most significant residues (e.g., RRxS), and in the form of a *motif-x* logo image. The logo indicates not only those residues that were significant enough to exceed the user defined threshold (such as both R residues in RRxS), but also the frequency of other residues which failed to exceed the significance threshold, yet nevertheless may be statistically overrepresented.

BASIC PROTOCOL

Extracting sequence motifs from MS/MS post-translational modification data.

Although the *motif-x* algorithm is general, and can be used to extract sequence motifs from data sets of varying formats, the most common *motif-x* analysis involves the extraction of motifs from large-scale post-translational modification data sets generated by tandem mass spectrometry experiments. The following protocol will use sample phosphorylation data obtained from the supplemental data of a study that investigated phosphorylation changes during the early differentiation of human embryonic stem cells (Van Hoof et al., 2009) (see Supplemental Data Files, File 1).

Necessary resources

Hardware

Any computer with Internet access.


Software

Any web browser (e.g., Internet Explorer, Firefox, Safari, Chrome, etc.).


Files

Performing the basic *motif-x* MS/MS protocol requires a list of peptides in SEQUEST (or similar) format (see Table 1). These peptides need not be contained within a file, as it is also possible to paste peptides directly into the *motif-x* input text box. Users may optionally decide to upload their own background data set, rather than use the provided *motif-x* background options. If users choose to upload foreground and/or background files, it is important to note that they should be .txt files, as other file types (e.g., .xls, .doc, .docx, .pdf, etc.) are not supported by *motif-x*. In general, if a file can be viewed using Notepad (on Windows) or TextEdit (in plain text mode on a Macintosh) without odd characters, the file will be suitable for uploading.

A sample MS/MS formatted foreground file corresponding to the same analysis illustrated in the protocol below is provided as a supplemental data file (see Supplemental Data Files, File 1).



v1.2 10.05.06



? job name: (optional)

? foreground data set (data set from which you wish to extract your motifs)

either
upload a file:

or
paste your data set:

```
YSVLNDDYFADVS#PLR
SGAVQAGSLGPGS#PVR
S#FEVEEVETPNSTPPR
KLANDFPLDLS#PVK
WLNSGRGDEASEEQNGS#SPK
WLNSGRGDEASEEQNGS#PK
ASPCS#SPTR
VGPAT#PSAQVGK
T#CILKLSLEK
LLSFS#PEEPPTLK
HIISATSLST#SPTELGR
```

? foreground format: FASTA pre-aligned text MS/MS

Extend From:

basic options

? central character:

? width:

? occurrences:

? significance:

? background:

advanced options

? upload background:

? background format:

? if background is MS/MS generated, extend from?:

? background central character:

help
faq

Please note:
 Periodic maintenance of the HMS computer cluster can cause problems with this website beyond our control. Please notify us (by emailing mchou@genetics.med.harvard.edu) if you do experience problems as we are not always aware of them.

By clicking the 'get motifs' button, you certify that you meet the requirements specified by the following disclaimer: The software provided on this website may be used freely by users from academic and non-profit organizations. Users from the commercial sector should contact Daniel Schwartz (daniel.schwartz@uconn.edu).

website created by Michael Chou and Daniel Schwartz ([Church Lab](#))
 © 2005-2008 The President and Fellows of Harvard College.

Figure 1

motif-x main data and parameter input page for the MS/MS example. This screenshot of the *motif-x* main data and parameter input page shows the appropriate parameters for the MS/MS example provided in the Basic Protocol. The data used for this example can be found in the supplemental data files (see Supplemental Data Files, File 1). An in-depth description of the various parameters can be found in the main text.

1. From your computer's web browser navigate to the *motif-x* web site at <http://motif-x.med.harvard.edu> and click on the *motif-x* logo at the top of the page to enter the *motif-x* data/parameters input page (see Figure 1).

2. In the blank text box labeled "job name" enter an optional title for the job.

3. Click the "Browse" button to upload a foreground data set or paste the data set directly into the large text box in the center of the page.

The foreground data is the data set from which you wish to extract motifs. Given the nature of sequencing by tandem mass spectrometry, these data should be input as peptide sequences, each separated by a carriage return. Special characters "#", "@", "", should be used to denote sites of protein modification (see Table 1 and Supplemental Data Files, File 1) by placing the special character immediately following the modified amino acid. Although this peptide modification syntax is standard for searches performed using SEQUEST, other common post-translational modification syntaxes include the addition of lowercase letters prior to the modification site (e.g., pS, pT, acK, (ox)M, etc.), a numerical mass change following the modification site (e.g., S(80), K(114), etc.), or signifying modification sites as lowercase letters themselves (e.g., s, t, y, k, etc.). Data using these alternative syntaxes will produce incorrect results, but can typically be converted into the SEQUEST syntax in a straightforward manner using the "find and replace" command in Microsoft Word or Excel (e.g., replacing all "S(80)" instances with "S#" will accomplish the task; however the final file must be plain text, not a Word or Excel*

file). It should be noted that with the exception of the special character denoting the modification site, motif-x removes all other non-alphanumeric characters from the data set prior to performing a motif search. Thus, “hyphens” denoting protein termini and “periods” denoting peptide fragment termini can be left in the data (e.g., the peptide sequence “-MEPSSWSGSES#PAENME.R” is acceptable). Sample MS/MS data is provided as a Supplemental Data File (see Supplemental Data Files, File 1) for readers wishing to follow along with the present protocol.

4. In the “foreground format” field select the “MS/MS” foreground format option and select an appropriate database in the drop-down menu from which to extend the peptides.

Since amino acid modification sites from MS/MS data sets often lack sufficient upstream and downstream residues necessary for the extraction of motifs, motif-x has a built-in “peptide extension” function. This built-in function maps user-inputted peptides back onto their appropriate proteomic database and gathers additional sequence information surrounding the modification sites as necessary. Users should therefore select the appropriate proteomic sequence database corresponding to the organism from which their foreground data sample was obtained. Users following the example presented here should select “IPI Human Proteome” as the extension database (see Figure 1).

5. Under the “basic options” heading enter an appropriate central character into the “central character” field.

Typically motif-x only takes a single character as input for this parameter; however, in the case of MS/MS formatted data motif-x will accept two characters as input. These two characters should reflect the appropriate syntax for modification sites in the foreground data set (discussed in step 3). In the example provided in Supplemental Data Files (File 1), phosphorylation sites are denoted by “#”, thus the characters “S#” should be used in the central character input field (see Figure 1). It is very important to note that this field is case sensitive, therefore inputting “s#” would yield no results for this example.

6. Enter the desired motif width into the “width” field.

Protein sequence motifs, especially those involved in protein post-translational modification tend to be between two and eight amino acids in length. Thus, choosing a length of 13 or 15 (i.e., six or seven amino acids on each side of the central modified residue) is expected to capture the vast majority of short-range sequence dependencies. Motif widths between 3 and 35 (odd lengths only) can be used for this parameter. However, choosing a motif width that is too narrow can result in the exclusion of motifs with critical longer-range dependencies and choosing a motif width that is too wide (without adjusting the significance threshold accordingly) can yield spurious motif results. In the example provided, a motif width of 15 has been chosen (see Figure 1).

7. Enter into the “occurrences” field the minimal number of times that you require your extracted motifs to occur within the foreground data set.

This parameter can be used to tune the specificity of motifs since motifs with greater specificity (i.e., more “fixed” positions) are expected to occur less often than those motifs with lower specificity. Users that wish to extract a maximal number of motifs should set this parameter to a low value (for example, “5”) and rely solely on the significance parameter (see step 8) to extract motifs. On occasion it may be useful to set this parameter as a fractional percent of the total number of modification sites in order to compare motifs with similar specificities across data sets that vary in size (e.g., to compare motifs from data sets of 300 and 3000 sites, one may opt to set the occurrences parameter to 5 and 50, respectively). In the example provided, an occurrence threshold of 5 has been chosen (see Figure 1).

8. Enter the significance threshold for the analysis into the “significance” field.

*The significance parameter corresponds to the binomial probability threshold necessary to “fix” each motif position during the motif-building phase of the algorithm. It is critical to note that this value does not take into account a correction for multiple hypotheses (such as the Bonferroni correction). On any given motif-x search step there are (number of possible characters at each position) * (number of nonfixed positions) hypotheses being tested. For example, in an S-centered analysis of width 15, there would be $(20) * (14) = 280$ hypotheses tested. To ensure an alpha-value of at least 0.05 by the Bonferroni method, one would need to divide the desired alpha-value by the total number of hypotheses tested (i.e., $0.05/280 = 0.00018$). Thus, for the previous example inputting a significance value of 0.00018 into motif-x would in fact correspond to a p-value of 0.05. The use of a motif-x significance threshold*

greater than 0.0005 is not suggested as it may result in the extraction of motifs that are not statistically significant. The default value of this parameter is 0.000001, which corresponds to an actual alpha-value of approximately 0.0003 for a protein motif analysis of width 15 after Bonferroni correction. A significance threshold of 0.000001 was chosen for the present example.

9. In the “background” field select the appropriate background data set for the analysis.

motif-x evaluates the statistical significance of positions within motifs in the foreground data set through a comparison to a background data set. For MS/MS analyses, the background data set should most-often correspond to the organismal proteomic database from which the sequencing data was obtained. Therefore, the peptide extension database (see step 4) and the background database should typically be the same. In the present example, since the foreground data set originated from human embryonic stem cells, the “IPI Human Proteome” was chosen as a background database (see Figure 1). The parameters under “advanced options” defining the background data set are not required for a basic motif-x run. These additional options are described in the “advanced parameters” section at the end of the protocols.

10. Press the “get motifs” button.

Pressing “get motifs” will launch an analysis on a node of the computer cluster and spawn a new web page as shown in Figure 2. This page will periodically reload until the results page is loaded. On occasion it may be necessary to manually reload the

page using your web browser. The motif-x algorithm is relatively fast, and jobs take under 5 minutes to perform. In an effort to minimize the burden on the computer cluster, motif-x jobs are limited to running within 15 minutes before a “timeout” message is generated. If users get this message they should try to simplify their analyses by either increasing the occurrence threshold or decreasing the significance threshold for their analysis. Please contact the authors if you believe you have an intractably large or complex data set to analyze.

11. View the *motif-x* results page.

The motif-x results page for an MS/MS format analysis is organized into four major regions: i) heading and parameters used (Figure 3), ii) data preprocessing statistics (Figure 4), iii) motif results with sequence logos (Figure 5), and iv) raw sequence data for each motif extracted (Figure 6).

Heading and parameters – This section provides information on the version of the motif-x software that was used; as well as a recapitulation of the parameters selected for the analysis.

Peptide pre-processing statistics – This section provides an accounting of the steps used to pre-process the MS/MS peptide data set into an aligned data set suitable for the motif-x algorithm. It is noticeable from the example provided that although 1091 phosphorylated peptides were initially inputted into motif-x, following pre-processing, only 853 serine-phosphorylated 15-mers remained.

XXXXmotif-XXXXX
v1.2 10.05.06

results

Parameters for this run:
 Job Name = Van Hoof et al regulated phospho hESC
 fgtype = 'ms'
 fgxtenddb = 'ipi.HUMAN.fasta'
 fgcentralres = 'S#'
 width = '15'
 occurrences = '5'
 significance = '0.000001'
 bgdb = 'ipi.HUMAN.fasta'
 bgtype = 'fasta'
 bgcentralres = 'S'

Job started Tue May 31 21:43:46 2011
 (Total elapsed time: 22 seconds)

Extracting motifs on the compute cluster...this can take up to several minutes depending on server load.
 You can bookmark this page if you want to check your results for up to a week.

Job Status:
 cluster-520601-shared_15m (motif-x_20110531-28119-82856005_Sx_vs_S_res.txt) **RUNNING**

Trying to predict PTMs in your proteins?
 Try our newly released [scan-x](#) tool for proteome-wide phosphorylation and acetylation prediction.

(Automatically re-checks in 10 seconds)

Figure 2

motif-x job submission page for the MS/MS example. This screenshot shows the job submission page users are brought to upon pressing the “get motifs” button. Although this page typically will auto-refresh, on occasion it may be necessary to manually refresh this page using the “Check Again” button or the web browser “reload” button to view the results.

XXXXmotif-XXXXX
v1.2 10.05.06

results

Parameters for this run:
 Job Name = Van Hoof et al regulated phospho hESC
 fgtype = 'ms'
 fgxtenddb = 'ipi.HUMAN.fasta'
 fgcentralres = 'S#'
 width = '15'
 occurrences = '5'
 significance = '0.000001'
 bgdb = 'ipi.HUMAN.fasta'
 bgtype = 'fasta'
 bgcentralres = 'S'

Job started Tue May 31 21:43:46 2011

Figure 3

motif-x version and parameters for the MS/MS example. This screenshot shows the uppermost portion of the *motif-x* results page for the MS/MS example described in the Basic Protocol.

Reasons for the decrease (outlined in the statistics found in this section) include: i) the redundancy of phosphorylation sites within the initial data set, ii) the inability to map certain peptides onto the database for proper extension, iii) the inability to get complete sequence information for some phosphorylation sites that were too close to protein termini, and iv) the inability to uniquely map all peptides back onto the database for unambiguous peptide extension. At the end of this section of the output, a hyperlink provides users with the ability to download their data set as a processed pre-aligned file to obviate the need for pre-processing the data in future motif-x analyses or to simply examine the result of the MS/MS pre-processing step.

Motif results and sequence logos – In addition to listing the extracted motifs, this section provides a table that outlines motif statistics. The columns of this table are titled: #, Motif, Motif Score, Foreground Matches, Foreground Size, Background Matches, Background Size, and Fold Increase. Motifs are listed in the order in which they are extracted by the algorithm, not with regard to statistical significance. Thus it should not be assumed that a motif found at a higher position in the list is more statistically significant than a motif found at a lower position in the list. All motifs extracted will, be significant at the alpha-value used for the significance parameter before the appropriate multiple hypothesis correction was calculated (see step 8). Therefore all motifs shown in Figure 5 are statistically significant at the $p < 0.0003$ level corresponding to the motif-x 0.000001 significance threshold that was specified for this particular run.

```
Results for (Central Foreground Residue: S# ; Background Residue: S)
Number of Peptides in Original Dataset: 1091
Number of Peptides in Original Dataset that are Unique: 921
Number of Peptides found in Database (ipi.HUMAN.fasta): 913
Number of Peptides NOT found in Database (ipi.HUMAN.fasta): 8
Number of central residues (residue = 'S#') mapped to the database : 949
Number of peptides without unique database mappings: 28
Number of peptides too close to protein termini: 36
Final Unique Target Peptides: 853
It took 51 seconds to preprocess foreground dataset
The input file has been converted to a pre-aligned file that may be used for subsequent runs of motif-x.
Right-click here to save it as a 'pre-aligned' dataset for possibly faster analysis in the future.
```

```
It took 8 seconds to preprocess background dataset
```

Figure 4

motif-x data pre-processing statistics for the MS/MS example. This screenshot shows the extensive pre-processing statistics provided for an MS/MS analysis, resulting from peptide mapping and extension. See text for greater detail.

Motifs Found

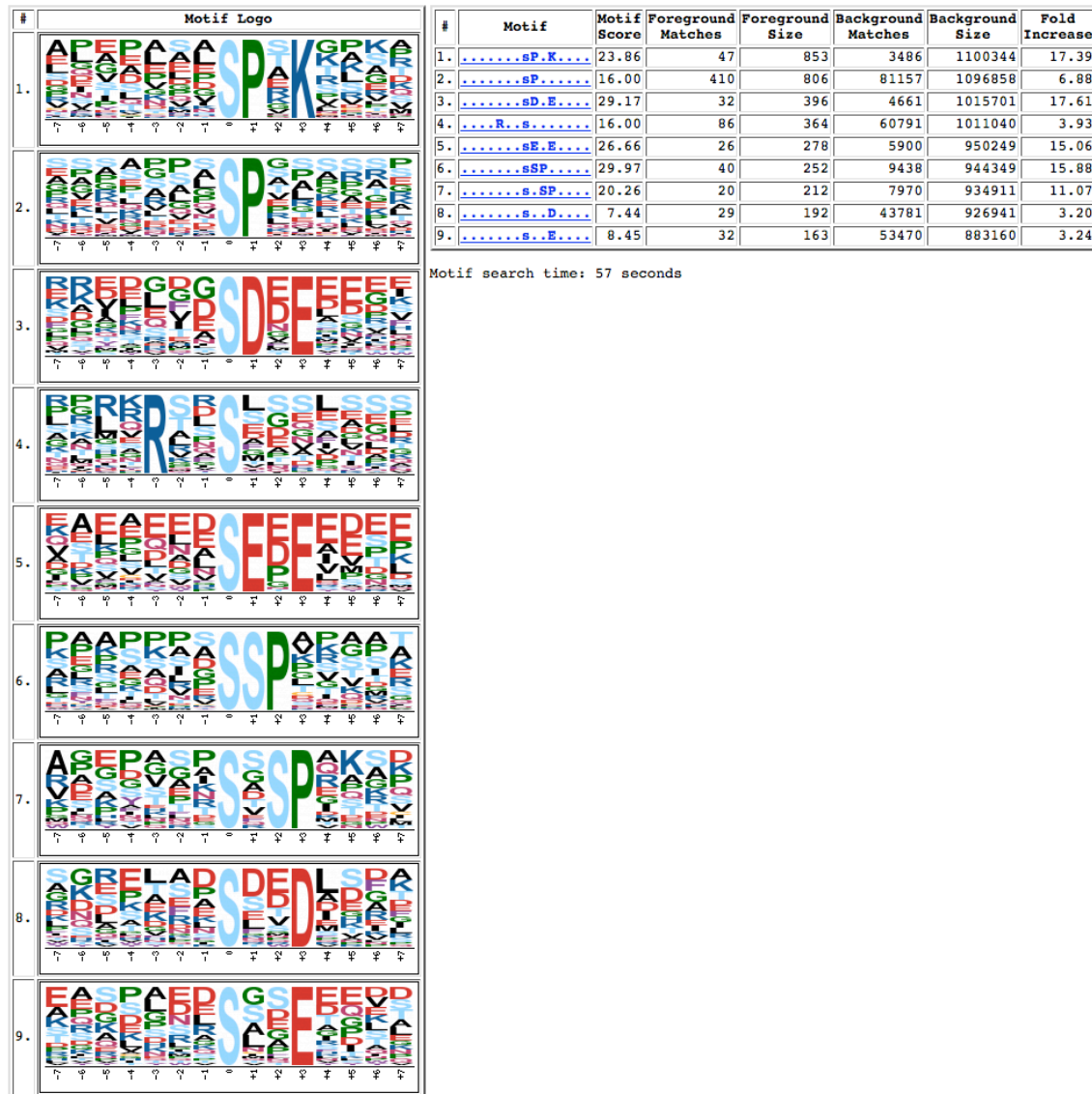


Figure 5

motif-x motif extraction results for the MS/MS example. This screenshot shows the actual *motif-x* results for the MS/MS example described in the Basic Protocol. Users attempting to reproduce the protocol using data from File 1 from the supplemental data files should obtain the same results. An in-depth description of the logos and table is provided in the main text.

Raw sequence data (motifs followed by their matching peptides):

.....sP.K....

ALAPAKESPRKGAAP
 PLEDTVLSPTKKRKR
 EEEDEALSPAKGQKP
 AGGSAALSPSKRKN
 AQLEPVASPAKKPKV
 ASAPAKESPRKGAAP
 AEEAAGASPAKANGQ
 VIEIEDASPTKCPIT
 IPEPKEPSPEKNSKK
 ADALPEHSPAKTSAV
 DDVPLSLSPSKRACA
 VQEEELLASPKLLED
 ALAKDMESPTKLDVT
 TPAPVEKSPAKKAT
 GSGQNEESPRKKA
 ENEDEQNSPPKGRK
 GIPLPAESPKKPKK
 EGGEASVPEKTSTT
 CLTLCLASPSKSTEM
 APPADFPSPRKSSGD
 NDFPLDLSPVKKRTR
 LPTVLPGPSKTRGQ
 RSGPRAPSPTKPLRR
 GVVSESDSPVKRPRG
 SEETPAISPKRRAP
 SVVSGLDSPAKTSSM
 RLGAGGSPKSPSA
 SQYVSGSPTKSKVT
 FMSETQSSPTKGVLM
 EVYELLDSPGVLLQ
 NGVAAEVSPAEEENP
 APSVSHVSPRKNPSV
 LKATVTPSPVKGKPK
 VFIDQNLSPGKGVVS
 TVVQKANSPEKPEEA
 EKISLSKSPKTKDPK
 CINAAPDSPSKQLPD
 DNIPEMPSPKMHGQ
 PFTLNSGSPKTKCSQ
 DPQQLQLSPLKGLSL
 SAKKVVVSPKVVAV
 RTGPPPISPSKRKF
 PLDCGSASPNKVASS
 KAPEKRASPKPASA
 LGTSDSDSPQKSSRD
 LNSSFETSPKVKWS
 LPSTNCQSPGKAIEN

.....sP.....

ASVSGPNPSETRRE
 RVQAGPCSPRRARG
 SPDKPGGSPSASRRK
 EVKVPASVVAQPKE
 RGVAPADSPAPRRS
 PVRAGGASPAASSTA
 CEAAEPCSPAAECE
 TPATGEQSPGARSRH
 EETECQSSPKHSSRE
 ASTASELSPKSKARP
 SATNKALSPVTSRTP
 QQAAGMLSPKTCGKE
 TGSTPPVSPTPSERS
 QEQECPPSPEPTRKE
 CTGAPGQSPGAGRAC
 NAVALSASPQLKEAQ
 HKMADCGSPFLGRRD
 SLASSNSPISQRRP
 NKGTPSQSPVVGSRQ
 EKGNVFSSPTAAGTP

Figure 6

motif-x raw sequence data for the MS/MS example. This partial screenshot shows the bottommost portion of the *motif-x* results page, corresponding to the raw sequences used to generate motifs in the motif-building phase of the algorithm. Users can jump to this portion of the results page by either clicking on a sequence logo or on the hyperlinked motif sequence in the results table.

The determination of whether a motif is known or novel can be accomplished using a variety of databases/tools (Edwards et al., 2008; Keshava Prasad et al., 2009) which store motif information. Additionally, directed Google searches using “x” to represent wildcard positions can often yield relevant primary literature sources of motif function. For example, searching for information on the first extracted motif in Figure 5 (.....sP.K....) by performing the Google search “SPxK phosphorylation motif” quickly reveals that the SPxK motif is the recognition sequence for cyclin dependent kinases (CDKs).

The “motif score” is calculated by taking the sum of the negative log probabilities used to fix each position of the motif. As such, higher motif scores typically correspond to motifs that are more statistically significant as well as more specific (i.e., greater number of fixed positions).

The “foreground matches” and “background matches” statistics indicate the number of peptides containing a given motif in those respective data sets following the removal of all peptides containing previously extracted motifs. Because of this iterative “set reduction” strategy, the “foreground matches” and “background matches” statistics may be less than or equal to the total number of instances of a given motif in the whole data set.

The “foreground size” and “background size” statistics indicate the total number of peptides contained in these data sets. The size of these data sets decreases

as motifs are extracted (i.e., down a column) due to the fact that peptides are removed from both the foreground and background data sets following motif extraction. The total number of foreground peptides not falling into any extracted motif class can therefore be calculated as the difference between the “foreground size” and the “foreground matches” of the final motif class (e.g., $163 - 32 = 131$ unclassified peptides in Figure 5).

The “fold increase” statistic is an indicator of the enrichment level of the extracted motifs. Specifically, it is calculated as $(\text{foreground matches}/\text{foreground size})/(\text{background matches}/\text{background size})$.

Frequency-based motif logos are provided to the left of the results table. In addition to indicating the fixed positions of the motif at full character height, these logos illustrate the amino acid composition of the “wildcard” positions within the motifs. Amino acids are sorted by their frequency at each position within the motif with the most frequent amino acids appearing closest to the top of the motif logo. Motif positions are labeled below the x-axis and residues are colored according to their chemical and physical properties.

Raw data – Clicking on either the links in the motif column of the results table, or on the sequence logos themselves, brings users to the raw peptide data used to extract the given motifs at the bottom of the motif-x results page. As such, the number of

peptides found for each motif in this section corresponds exactly to the number of “foreground matches” for the motif in column 2 of the motif-x results table.

ALTERNATE PROTOCOL 1

Extracting sequence motifs from pre-aligned data.

For algorithmic reasons, aligned peptides centered on a particular amino acid (typically a post-translational modification site) form the ideal data input format for *motif-x* and are thus included among the foreground format choices on the *motif-x* web site (see Table 1). This data format is also becoming increasingly adopted by protein post-translational modification databases (Hornbeck et al., 2004; Durek et al., 2010; Dinkel et al., 2011), as well as in the supplemental data of large-scale PTM studies (Matic et al., 2010; Rigbolt et al., 2011). The following protocol will use pre-aligned 15-mers centered on sumoylated lysine residues from human proteins to illustrate the use of *motif-x* on this type of foreground data format (see Supplemental Data Files, File 2).

Necessary resources

Hardware

Any computer with Internet access.

Software

Any web browser (e.g., Internet Explorer, Firefox, Safari, Chrome, etc.).

Files

Performing the *motif-x* pre-aligned protocol requires a specific list of peptides in pre-aligned format (see Table 1). These peptides need not be contained within a file, as it is also possible to paste peptides directly into the *motif-x* input text box. Users may optionally decide to upload their own background data set, rather than use the provided *motif-x* background options. If users choose to upload foreground and/or background files, it is important to note that they should be .txt files, as other file types (e.g., .xls, .doc, .docx, .pdf, etc.) are not supported by *motif-x*. In general, if a file can be viewed

using Notepad (on Windows) or TextEdit (in plain text mode on a Macintosh) without odd characters, the file will be suitable for uploading.

A sample pre-aligned formatted foreground file corresponding to the same analysis illustrated in the protocol below is provided as a supplemental data file (see Supplemental Data Files, File 2).

1. From your computer's web browser navigate to the *motif-x* web site at <http://motif-x.med.harvard.edu> and click on the *motif-x* logo at the top of the page to enter the *motif-x* data/parameters input page (see Figure 7).

2. In the blank text box labeled "job name" enter an optional title for the job.

3. Click the "Browse" button to upload a foreground data set or paste the data set directly into the large text box in the center of the page.

The foreground data is the data set from which you wish to extract motifs. Data should be uploaded or pasted in the same format as shown in the pre-aligned example in Table 1 or the raw sequence data provided in as a supplemental data file (see Supplemental Data Files, File 2). Peptides of equal width and centered on an amino acid of interest should be separated by a carriage return. In contrast to the previously described MS/MS format, which allows for special characters ("#", "@", "") to designate modification sites, the pre-aligned data format does not allow for the inclusion of special characters.*

4. In the "foreground format" field select the "pre-aligned" option.

This option indicates that the inputted data is already in the ideal pre-aligned format for motif-x, and that motif-x does not need to preprocess the data.

5. Under the “basic options” heading enter an appropriate central character into the “central character” field.

A single (case sensitive) character should be entered. As mentioned in Step 3 above, combinations of residues and special characters may not be used in pre-aligned analyses. In the sumoylation example presented here, since the central character of the pre-aligned data corresponds to the site of lysine modification, “K” is entered as the central character (see Figure 7). It should be noted however that all the peptides in the input data need not contain the same central character. In the case of pre-aligned input data containing multiple different central characters, motif-x will carry out the motif extraction procedure on only that subset of peptides bearing the central character specified by this field.

6. Enter the desired motif width into the “width” field.

In a pre-aligned analysis, the width parameter should correspond exactly to the width of the pre-aligned peptides. A width of 15 was entered for the present example.

7. Enter into the “occurrences” field the minimal number of times that you require your extracted motifs to occur within the foreground data set.

An explanation of this parameter is provided in the Basic Protocol, Step 7. An occurrences parameter of 5 was entered for the present example.

XXXXmotif-XXXX

v1.2 10.05.06



? job name: (optional)

? foreground data set (data set from which you wish to extract your motifs)

either
upload a file:

or

paste your data set:

```

TTEQPVAKQELVDFW
VGTQSSCGKSSVLESL
VNRPLTMKKEGIQTR
VQGNSSIKLELDASK
VSLTESMKMESGSPL
VSQPKSIKEEGEDLQ
VTTHPLAKDKMMNGG
WKLTDNIKYEDCEDR
WKLVENVKYEDIYED
YKKYKIKIVERVERE
YPSPTCVKSEMCPWM

```

? foreground format: FASTA pre-aligned text MS/MS Extend From:

basic options

? central character:

? width:

? occurrences:

? significance:

? background:

advanced options

? upload background:

? background format:

? if background is MS/MS generated, extend from?:

? background central character:

help faq

Please note:
 Periodic maintenance of the HMS computer cluster can cause problems with this website beyond our control. Please notify us (by emailing mchou@genetics.med.harvard.edu) if you do experience problems as we are not always aware of them.

By clicking the 'get motifs' button, you certify that you meet the requirements specified by the following disclaimer: The software provided on this website may be used freely by users from academic and non-profit organizations. Users from the commercial sector should contact Daniel Schwartz (daniel.schwartz@uconn.edu).

website created by Michael Chou and Daniel Schwartz ([Church Lab](#))
 © 2005–2008 The President and Fellows of Harvard College.

Figure 7

motif-x main data and parameter input page for the pre-aligned example. This screenshot of the *motif-x* main data and parameter input page shows the appropriate parameters for the pre-aligned example provided in Alternate Protocol 1. The data used for this example can be found in File 2 from the supplemental data files. An in-depth description of the various parameters can be found in the main text.

8. Enter the significance threshold for the analysis into the “significance” field.

An explanation of this parameter is provided in Step 8 of the previously described Basic Protocol. A significance of 0.000001 was entered for the present example.

9. In the “background” field select the appropriate background data set for the analysis.

motif-x evaluates the statistical significance of positions within motifs through a comparison to a background data set. It is highly preferable to select the background proteomic database corresponding to the organism from which the foreground data was derived. In the present human sumoylation site example the “IPI Human Proteome” is selected as a background database (see Figure 7) because the pre-aligned foreground data originated from human proteins.

10. Press the “get motifs” button.

An explanation of the result of pressing “get motifs” is provided in Step 10 of the previously described Basic Protocol.

11. View the *motif-x* results page.

The motif-x results page for pre-aligned analyses is nearly the same as the previously described results page for MS/MS analyses (see Step 11 of the Basic Protocol); however, due to the lack of peptide preprocessing in a pre-aligned analysis, the preprocessing statistics section is omitted. A screenshot of the motif-x results page for the sumoylation example is shown in Figure 8 so that readers following along may verify their results. These results indicate the extraction of 5 sumoylation motifs

XXXXmotif-XXXXX

v1.2 10.05.06

results

```
Parameters for this run:
Job Name = Pre-aligned Human sumoylation sites
fgtype = 'prealigned'
fgcentralres = 'K'
width = '15'
occurrences = '5'
significance = '0.000001'
bgdb = 'ipi.HUMAN.fasta'
bgtype = 'fasta'
bgcentralres = 'K'
```

Job started Thu May 12 19:10:23 2011

Results for (Central Foreground Residue: K ; Background Residue: K)
It took 257 seconds to preprocess background dataset

Motifs Found

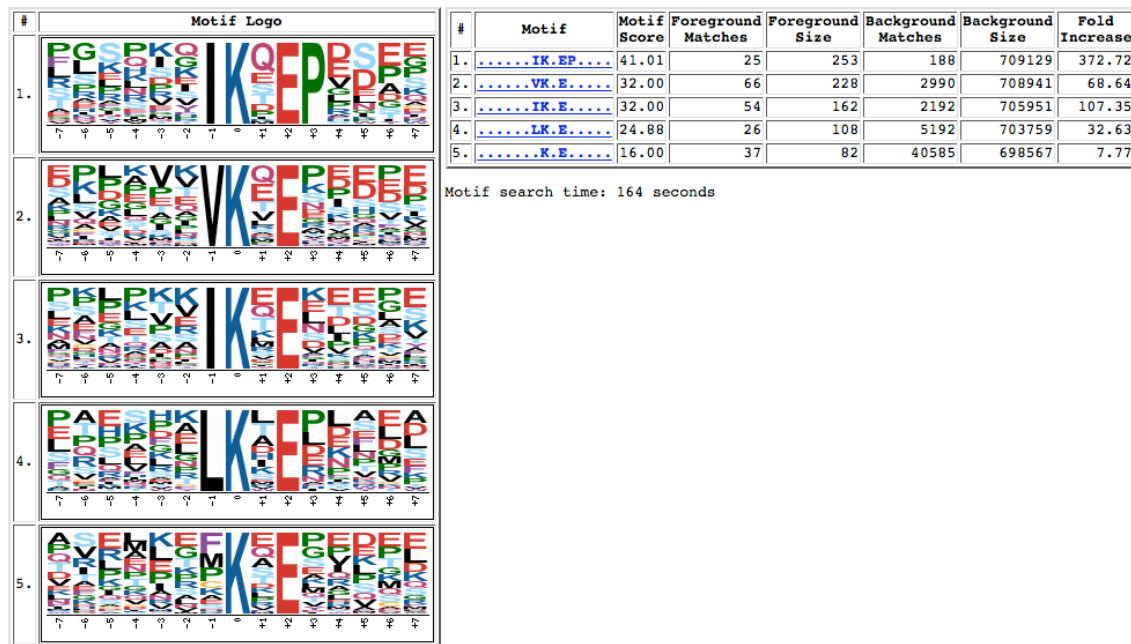


Figure 8

motif-x results page for the pre-aligned example. This screenshot shows the actual *motif-x* results for the pre-aligned example described in Alternate Protocol

1. Users attempting to reproduce the protocol using data from File 2 from the supplemental data files should obtain the same results. An in-depth description of the logos and table is provided in the main text.

(IKxEP, VKxE, IKxE, LKxE, and KxE), consistent with the experimentally verified sumoylation sequence specificity (ψ KxE, where ψ represents a hydrophobic residue).

ALTERNATE PROTOCOL 2

Extracting sequence motifs from whole protein sequence data in FASTA format.

Following experiments, researchers are often left with lists of proteins that share a common attribute (e.g., binding to the same partner, localizing to the same cellular compartment, etc.). Although the *motif-x* algorithm was initially designed with the intent of extracting protein post-translational modification motifs from aligned data, the algorithm is capable of extracting overrepresented sequence patterns from any sequence-based data set including whole protein data in FASTA format (see Table 1). It should be noted that patterns extracted from whole protein data need not be involved in protein post-translational modifications as *motif-x* extracts patterns statistically, without regard to cellular function. The following protocol will use a group of ten proteins selected from the ELM (Eukaryotic Linear Motif) database (Gould et al., 2010) known to be bound by Grb2-like Src Homology 2 (SH2) domains. Src Homology 2 domains are involved in a wide variety of protein signaling pathways, and are unified in their ability to bind phosphorylated tyrosine residues. The Grb2-like SH2 domain used in this analysis is known to specifically bind the YxN motif (Kessels et al., 2002). The data set used for this example is provided in Supplemental Data Files (see File 3).

Necessary resources

Hardware

Any computer with Internet access.

Software

Any web browser (e.g., Internet Explorer, Firefox, Safari, Chrome, etc.).

Files

Performing the *motif-x* FASTA protocol requires a specific list of proteins in FASTA format (see Table 1). These proteins need not be contained within a file, as it is also possible to paste proteins directly into the *motif-x* input text box. Users may optionally decide to upload their own background data set, rather than use the provided *motif-x* background options. If users choose to upload foreground and/or background files, it is important to note that they should be .txt files, as other file types (e.g., .xls, .doc, .docx, .pdf, etc.) are not supported by *motif-x*. In general, if a file can be viewed using Notepad (on Windows) or TextEdit (in plain text mode on a Macintosh) without odd characters, the file will be suitable for uploading.

A sample FASTA formatted foreground file corresponding to the same analysis illustrated in the protocol below is provided as a supplemental data file (see Supplemental Data Files, File 3).

1. From your computer's web browser navigate to the *motif-x* web site at <http://motif-x.med.harvard.edu> and click on the *motif-x* logo at the top of the page to enter the *motif-x* data/parameters input page (see Figure 9).
2. In the blank text box labeled "job name" enter an optional title for the job.
3. Click the "Browse" button to upload a foreground data set or paste the data set directly into the large text box in the center of the page.

The foreground data is the data set from which you wish to extract motifs.


**Harvard
Medical
School**

motif-x
v1.2 10.05.06

? job name: (optional)

? foreground data set (data set from which you wish to extract your motifs)
either
upload a file:

or
paste your data set:

```
>sp|P15391|CD19_HUMAN B-lymphocyte antigen CD19 OS=Homo sapiens
GN=CD19 PE=1 SV=6
MPPPRLLFFLFLTPMEVRPEEPLVVKVEEGDNAVLQCLKGTS DGPTQOLTWSRESPLKP
FLKLSLGLPGLGIHMRPLAIWLFI FNVSQMGGFYLCQPGPSEKAWQPGWTVNVEGSGE
LFRWVSDLGGLCCCLKNRSSEGPS SPSCKLMSPKLVVWAKDRDEINWCEPCLPDRSL
NQSLSQDLTMAPGSTLWLS CGVPPDVSVRGPLSWTHVHPKPKSLLSLELKDDRPARDMW
VMETGLLPRATAQDACKYYCHRCNLTMSF HLEITARPVLWHWLLRTGGWKVSAVTLAYL
IFCLCSLVGILHLQALVLRKRKRMTDPTRRFFKVTTPPGSGPQNOYGNVLSLPTPTSG
LGRAQRWAAGLGCTAPSYGNPSSDVQADGALGSRSPPGVGE EEEEGEGYEEPDSEEDSEF
YENDSNLQDQLSQDGSYENPEDEPLGFEDEDSFSNAESYENEDEELTOPVARTMDPLS
PHGSAWDPSREATSLGSQSYEDMRGILYAAPQLRSIRGQPCPNHEEDADS YENMDNPDGP
```

? foreground format: FASTA pre-aligned text MS/MS

basic options

? central character:

? width:

? occurrences:

? significance:

? background:

advanced options

? upload background:

? background format:

? if background is MS/MS generated, extend from?:

? background central character:

help **faq**

Please note:
Periodic maintenance of the HMS computer cluster can cause problems with this website beyond our control. Please notify us (by emailing mchou@genetics.med.harvard.edu) if you do experience problems as we are not always aware of them.

By clicking the 'get motifs' button, you certify that you meet the requirements specified by the following disclaimer: The software provided on this website may be used freely by users from academic and non-profit organizations. Users from the commercial sector should contact Daniel Schwartz (daniel.schwartz@uconn.edu).

website created by Michael Chou and Daniel Schwartz ([Church Lab](#))
© 2005-2008 The President and Fellows of Harvard College.

Figure 9

motif-x main data and parameter input page for the FASTA example. This screenshot of the *motif-x* main data and parameter input page shows the appropriate parameters for the FASTA example provided in Alternate Protocol 2. The data used for this example can be found in File 3 from the supplemental data files. An in-depth description of the various parameters can be found in the main text.

Data should be uploaded or pasted in the same format as shown in the FASTA example in Table 1 or the raw sequence data provided as a supplemental data file (see Supplemental Data Files, File 3). The FASTA data format allows one or more protein sequences in a text file each of which can be preceded by a description line that begins with a “greater than” (“>”) character, followed by biological sequence starting on the line below. In contrast to the previously described MS/MS format, which allows for special characters to designate modification sites (“#”, “@”, “”), the FASTA data format does not allow for the inclusion of special characters except in the description lines, which are ignored by motif-x. It is important not to simply merge all protein sequences together without description lines because sequences at the junctions would form non-meaningful peptide sequences when extracted by motif-x.*

4. In the “foreground format” field select the “FASTA” option.

This option indicates that the inputted data is in the FASTA data format and instructs the motif-x web tool on the proper preprocessing of the data.

5. Under the “basic options” heading enter an appropriate central character into the “central character” field.

A single (case sensitive) character should be entered. As mentioned in Step 3 above, combinations of residues and special characters may not be used in FASTA analyses. In the present example, since it is known a priori that the SH2 domain interacts with proteins bearing YxN motifs, “Y” is entered as the central character

(see Figure 9). However, performing a complete and unbiased motif search in a set of proteins requires 20 independent motif-x analyses centered at each of the 20 amino acids.

6. Enter the desired motif width into the “width” field.

Any odd value between 3 and 35 can be chosen as a motif width; however, most short linear protein motifs are under 15 amino acids long. In the present SH2 domain example, a width of 15 has been used (see Figure 9).

7. Enter into the “occurrences” field the minimal number of times that you require your extracted motifs to occur within the foreground data set.

This parameter can be used to tune the specificity of motifs since motifs with greater specificity (i.e., more “fixed” positions) are expected to occur less often than those motifs with lower specificity. Users that wish to extract a maximal number of motifs should set this parameter to a low value (e.g., 5) and rely solely on the significance parameter (see step 8) to extract motifs. It should be noted that raising the occurrences threshold to a high value (e.g., 50) on a relatively small FASTA data set (e.g., 5 proteins) is likely to result in no extracted motifs since it is unlikely that any sequence pattern would occur such a large number of times in a small data set. An occurrences threshold of 5 was used for the present example.

8. Enter the significance threshold for the analysis into the “significance” field.

An explanation of this parameter is provided in Step 8 of the previously described Basic Protocol. A significance threshold of 0.000001 was used for the present example.

9. In the “background” field select the appropriate background data set for the analysis.

motif-x evaluates the statistical significance of positions within motifs through a comparison to a background data set. Although it is highly preferable to select the background proteomic database corresponding to the organism from which the foreground data was derived, FASTA-formatted data sets provide a unique opportunity to use statistical characteristics from the foreground data set to derive a relevant background data set. On the motif-x web site this option is labeled “Use foreground, unaligned” in the drop-down menu. Because our SH2 domain example lacks an obvious proteomic background (due to the fact that the set of 10 proteins are derived from a variety of species), we have selected the “Use foreground, unaligned” option in this example to illustrate its usefulness (see Figure 9). Alternatively, users could select the background organism database that is most closely related to the foreground data set (in the present example, this would correspond to the IPI Human database).

XXXXmotif-XXXX

v1.2 10.05.06

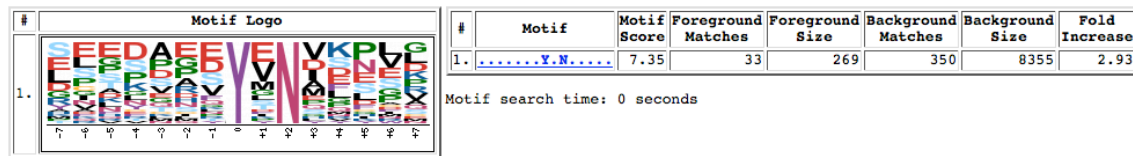
results

Parameters for this run:
 Job Name = Interactors of Grb2-like SH2 domain
 fgtype = 'fasta'
 fgcentralres = 'Y'
 width = '15'
 occurrences = '5'
 significance = '0.000001'
 bgdb = 'unaligned'
 bgtype = 'fasta'

Job started Mon May 16 14:10:36 2011

Results for (Central Foreground Residue: Y ; Background Residue: All (unaligned))
 It took 0 seconds to preprocess foreground dataset
 The input file has been converted to a pre-aligned file that may be used for subsequent runs of motif-x.
[Right-click here](#) to save it as a 'pre-aligned' dataset.
 It took 0 seconds to preprocess background dataset

Motifs Found



Raw sequence data (motifs followed by their matching peptides):

.....Y.N.....

HEEDADSYENMDNPD
 SEEDSEFYENDSNLG
 SLDCSREYVNVSPQ
 ELFDPPSYVNVQNLQ
 SVESCEDYVNVPESE
 SPSNAESYENDEEL
 EEPDPHQYVNDPFGK
 REDSARVYENVGLMQ
 LLPCTGDYMNMSPVG
 DGEAPDYENLQELN
 EETGSEYMNMDLGP
 KSLDPMVYMNDSPL
 EPKSPGEYVNIIEFGS
 SKRKGHEYTNIKYSL
 LSQDCGYENPEPEF
 RPPSAEYVSNALPVG
 YISKAQYVFNKRSRS
 TCSPPQYVNVQDPVR
 LQCLPREYVNAHCL
 DANVAVSYENQEPAC
 YVHVNATYVNVKVA
 GEVYEGVYTNHKGK
 ADEDEDDYPNGYLVV
 GPARLEYEENEKWR
 GSGPQNYGNVLSLP
 RILKVCYVNSNFPG
 SBNYMAPYDNYVPSA
 LGGTAPSYGNPSSDV
 ENKKNRYKNLPPD

Figure 10

motif-x results page for the FASTA example. This screenshot shows the actual *motif-x* results for the FASTA example described in Alternate Protocol 2. Users attempting to reproduce the protocol using data from File 3 from the supplemental data files should obtain the same results. An in-depth description of the logos and table is provided in the main text.

10. Press the “get motifs” button.

An explanation of the result of pressing “get motifs” is provided in the Basic Protocol, Step 10.

11. View the *motif-x* results page.

The motif-x results page for FASTA format analyses is nearly the same as the previously described results page for MS/MS analyses (see Step 11 of the Basic Protocol); however, due to the lack of significant peptide preprocessing in a FASTA analysis, the preprocessing statistics are minimal. A screenshot of the motif-x results page for the worked out example is shown in Figure 10. In the example, a single motif, YxN, was extracted from the data set corresponding to the known specificity of the Grb2 SH2 domain.

GUIDELINES FOR UNDERSTANDING RESULTS

The normal output of *motif-x* is a list of motifs for the given input data. Whereas previous studies sometimes indicate the existence of some flexibility at motif residues, *motif-x* deconvolutes input data into distinct and different motifs thereby avoiding over-generalizing its findings. For instance, a hypothetical motif shown in the literature as Tx[K/R][K/R] may be shown by *motif-x* as separate motifs for TxRR, TxKK and TxKR, but not TxRK, which means that the data set supplied did not achieve significance for that motif. It is possible that with a larger foreground data set, such a motif would

emerge as an additional motif in the series; however, given today's large data sets, it is more likely that such a motif is actually disfavored in comparison with the others.

The interpretation of the results will depend on the nature of the input data. If the user supplies substrate data that results from just one enzyme, then one or more motifs may emerge as preferred for that enzyme. However, shotgun proteomic data is likely to yield a collection of motifs that are the result of multiple enzymes. Such data is useful if one is predicting motifs that direct modification as opposed to understanding single enzyme-substrate specificity.

Note that the results page is a unique URL, and may be bookmarked and revisited for up to a week before it is archived. If the URL is no longer available, the user will get an error indicating that it is expired.

COMMENTARY

Background Information

Understanding cellular functions in normal and abnormal cells ultimately depends on understanding the molecular interactions within the cell, which includes enzymatic and binding reactions between proteins. Computational analyses of protein sequences, such as the determination of secondary structure, local hydrophobicity, signaling peptide sequences and domain architecture, have all increased our understanding of protein functions and mechanisms.

Within the enzyme-substrate subset of protein-protein interactions, it is often the chemical reactivity of a particular amino acid side chain within the substrate or enzyme that is a prerequisite for catalysis. Such residues thereby firmly fix at least one amino

acid in an enzyme or substrate that participates in the reaction. However, a repertoire of 20 amino acids still provides little specificity, and nature has evolved secondary sites within enzymes and substrates to further direct specificity.

Such short sequences within substrates or interaction partners have been shown in many cases to be necessary and/or sufficient to direct highly specific interactions including proteolysis and enzyme-catalyzed post-translational modifications, such as phosphorylation, sumoylation, glycosylation and others (Howard et al., 1991; Songyang and Cantley, 1995; Shakin-Eshleman et al., 1996; Sternsdorf et al., 1999).

Understanding the crucial residues and positions that are the primary determinants of substrate specificity or protein-protein interactions can be used to predict biological activities and mechanisms. This can be performed, for instance using a companion tool known as *scan-x* (<http://scan-x.med.harvard.edu>) described in an upcoming supplement of Current Protocols in Bioinformatics.

Partly because of the importance of phosphorylation and proteolysis in biology and partly because of ease of measurement, the discovery of motifs for kinases and proteases using biochemical means has historically led the field.

Recent advances in mass spectrometry are helping to provide thousands of new measurements with each high-throughput paper that is published, and now data sets for many protein post-translational modifications are quite large. However, understanding of the enzymes responsible for these modifications is lagging behind. Knowing the identity of preferred residues proximal to an enzymatically modified residue on a substrate improves understanding of interactions and helps to generate hypotheses about other

possible interactions. Using *motif-x* to discover motifs can provide important information about known enzymes and enzymes yet to be discovered in an unsupervised manner.

Often laborious bench work is required to understand biological mechanisms. The implementation of *motif-x* to deconvolute sequence data into potential interaction motifs can be used to dramatically improve the generation and testing of biological hypotheses.

Critical Parameters

Central Character (default: blank)

Once input data sets are processed, it is this residue that all foreground sequences will be centered upon. This is typically a site with a known or putative post-translational modification (PTM). For example, in the case of eukaryotic phosphorylation, one might use serine, threonine or tyrosine as a central residue. For mass spectrometry input data, the *motif-x* user will usually want to distinguish between a modified or unmodified residue in input data because, e.g., not every serine will be modified.

Therefore, in the foreground data set, the residues may be specified with or without a trailing symbol within MS/MS foreground data. Thus, in a list of hypothetical input sequences, one may have the tryptic peptide, MACRTISPWESR, where only the first S is phosphorylated. In that case, the input sequence should be modified to be MACRTIS*PWESR. If the input has more than one modified residue per sequence (e.g., MACRTIS*PWES*R) then *motif-x* will attempt to create a fixed length sequence surrounding each one and will be properly handled as two separate phosphorylation

events by *motif-x*. This format is the natural output format of SEQUEST, and the output from other spectral identification pipelines will have to be processed into this format if it is not already compliant. Once a decision has been made by the user to focus on a particular modified residue, that same symbol can be used as the central residue. So, in this instance, the user would specify 'S*' as the foreground central residue.

Note that any modified residue to be centered upon is limited to the repertoire of symbols: "*", "#", or "@" which must always be placed *after* the modified residue. While it is acceptable for foreground data sets to have multiple different symbols (indicating different modification types for a given peptide e.g., M@ACRTIS*PWES*R or MACRT*IS*PWES*R), if a modified residue is to be used as a central residue, the user must use the same residue and modifier symbol (e.g., 'S*') in the "central character" field (all other symbols will be ignored for that particular analysis). Note that it is also acceptable to use a single letter as the central residue, which will match every occurrence of that residue regardless of modification state.

Because each analysis is specific for a residue, to find motifs centered on each of the eukaryotic phosphorylated residues in a single data set, one can simply change the central residue from 'S*' to 'T*' to 'Y*' for each of three analyses.

Width (default: 15)

The default value for the width parameter is 15 residues (i.e., 7 residues on either side of the central residue). The user may increase this value to as high as 35, but we have rarely found residues that are significantly correlated if they are very distant from the central residue.

Occurrences (default: 20)

Normally, *motif-x* stops and outputs the current motif when it cannot find additional residues that are significantly different than the background. However, an additional stopping criterion occurs when the size of the foreground data set is reduced to below this occurrence threshold. Typically, this is set at some low number based on how few peptides are expected in each motif class. In general it is recommended that the significance threshold be set in such a way that minimum foreground occurrences are not invoked by the algorithm.

Significance (default: 0.000001)

Each step of the algorithm tests the significance of all residue/position pairs with respect to the background frequency of those same residue/positions. The significance threshold used by *motif-x* for this test is a user configurable parameter. Typically a statistical significance (α) would be set at 0.05 or 0.01, but each step of the *motif-x* algorithm is testing multiple hypotheses. Therefore this value must be reduced accordingly. Version 1.2 of *motif-x* does not automatically perform Bonferroni correction, so it must be done by the user to protect against the generation of false positives (Bonferroni, 1935).

We recommend at least dividing your desired alpha value by: $(\text{width} - 1) \times 20$. So, for instance, if you want to use an alpha of 0.05, and a width of 15 for protein analysis (i.e., 20 amino acids), then $0.05 / (14 \times 20) = 0.05 / 280 = 0.00018$. Thus, using a number higher than this will likely yield motifs that are not statistically significant. The

default is currently set at 0.000001, which corresponds to an alpha value of 0.0003 for a width of 15. This threshold is very conservative and generally works quite well.

Background (default: IPI Human Proteome)

As mentioned previously, the background data set is the source of statistics for each residue/position, and using the right organism will provide the most relevant results. Multiple proteomic Standard Database Libraries are provided on the *motif-x* web site, and they continue to be added over time by request to the authors.

A background data set size is typically more than an order of magnitude larger than the foreground data set size. Users may also provide background data sets in a number of different formats (pre-aligned, MS/MS and FASTA/text), but Standard Database Libraries are in FASTA format by default, and using a Standard Database Library is the most typical way to run *motif-x*.

A special background data set known as “Foreground Unaligned” can be generated from the foreground data set if the foreground is in FASTA format. Using this option, instead of processing a background data set to be centered on just one particular residue, it is aligned on *every* single residue and then used as the background. An example of the utility of this approach is shown in Alternative Protocol 2. Essentially, this allows for the discovery of motifs for a particular central residue evaluated in comparison to all other residues. “Foreground unaligned” can also be used when other suitable backgrounds cannot be determined.

Background central character (default: blank)

In general, one should set this to the same thing as the foreground central residue without a modifier, if any. So, for instance, if the foreground central residue is “S*”, then use “S” as the background central residue, if it is “T*” then use “T” as the background central residue, and so forth. If this field is left blank, it will default to the foreground central residue without any modifier.

Troubleshooting

Two extreme cases occur with some regularity:

- 1) **No Motifs Found**. When no motifs are found with the given input parameters, the message that is reported in the output is “Motifs Found: None”. While it is certainly possible that no motifs can be found at a significant level some suggested actions are:
 - a) Double check that you have inputted a non-empty foreground data set and that it contains a sufficient amount of data. It is nearly impossible to find significant motifs with a single protein, or only a dozen MS/MS peptides. There must be enough data to obtain at least one significant motif.
 - b) If performing an MS/MS analysis, download the pre-aligned data set from the hyperlink at the end of the first section (shown as “[Right-click here](#)” in Figure 4) to see if the pre-processing seems reasonable. There may be an obvious syntactic

error in the input data or specification of the central residue.

c) Double check the significance and occurrences thresholds to make sure they are reasonable. Setting an occurrence threshold at a low number, while maintaining the significance parameter, will still only yield statistically significant motifs even in a small set.

d) Increase the amount of foreground input data that you have. Note that identical peptides are not contained in *motif-x*'s analysis, so merely duplicating existing data is not an effective approach. Always use real data sets to get relevant results.

e) Closely examine the foreground data to make sure the central residue is represented correctly (i.e., "S" versus "s" or "T" versus "T*", etc.).

f) Make certain that the foreground data is significantly different from the background. Asking *motif-x* to find motifs in a data set that closely approximates the background will not yield significant results.

g) Some data sets simply do not have any significant motifs. For example, current ubiquitination data sets suggest that there is little evidence that short linear motifs in the immediate vicinity of the ubiquitinated lysine residue play a primary role in substrate specificity.

Note that significantly correlated residues may exist at a distance outside the scope of *motif-x* that nonetheless contribute to specificity. These other contributions to specificity may be particularly important in cases where *motif-x* is not otherwise able to find a strong motif in a data set.

- 2) *Timeouts*. Occasionally, an error message will be generated indicating that a “Timeout” has occurred. While *motif-x* usually runs very quickly, there are some data sets that cause the algorithm to take longer than 15 minutes of CPU time to run. This is usually due to extremely large data sets, and/or a significance parameter that is too large. However, it can also be due to an overloaded computer cluster. There are a number of things that can be done to try to generate motifs after a timeout. Any or all of them may alleviate the timeout:
 - a) Run the job again at a different time of day.
 - b) Decrease the value of the significance parameter to a very small number ($< 1e-6$ or smaller). Because decreasing the value of the significance parameter makes it harder to find significant motifs, reducing the significance threshold can actually reduce the output and the run time. If the job completes, the number of CPU seconds will be shown in the output, and if it is significantly less than 900 seconds (i.e., 15 minutes), then gradually increase the value of significance towards a desired value.

c) Reduce the size of the foreground data set.

d) Unfortunately, the computer cluster does occasionally become overloaded.

However, the job may still have been run. Before re-running jobs that look like they failed for no explainable reason, it is a good idea to try to wait a few minutes and use the “reload” button in the web browser while on the results page. In some cases, when it refreshes, output may be visible.

A more complete list of error messages and suggested remedies can be found in the Supplemental Data Files (see File 4). If you continue to experience problems despite following the troubleshooting guidelines, please contact the authors using the email address on the web site.

Advanced Parameters (optional)

Advanced options in the current version of *motif-x* (version 1.2) are limited.

Ability to upload a user-selected background data set

Users may upload a background data set by selecting “uploaded” under the background parameter, and browsing for the proper file by clicking on the “Browse” button next to the “upload background” dialog box in the “advanced options” section (see Figures 1, 7, and 9). With regard to formatting, background data sets may exist in the same four formats used for foreground data sets (i.e., FASTA, pre-aligned, text, and MS/MS – see

Table 1 for examples of correct formatting). As in the case of foreground data sets, backgrounds in MS/MS format (i.e., proteolytically digested sequence fragments) require users to indicate the proper database from which to extend using the drop down menu in the “advanced options” column. Also, as discussed in the preceding protocol, special characters (“#”, “@”, “*”) may only be used in conjunction with MS/MS-formatted backgrounds to denote modified residues. If a residue is not inputted into the background central character parameter, *motif-x* will automatically use the same central character that was inputted for the foreground data set, but without the modifier [so it is important to be explicit if using an MS/MS background (e.g., S*)].

Use of “text” format

The use of ‘text’ as a foreground or background format is not useful for proteomic analysis, and may be eliminated in future versions of the software.

Additional organism databases

Users should feel free to contact the authors to have additional commonly used background databases added to the *motif-x* web site.

Literature Cited

- Bonferroni, C.E. 1935. Il calcolo delle assicurazioni su gruppi di teste. In Studi in Onore del Professore Salvatore Ortu Carboni.13-60.
- Dinkel, H., Chica, C., Via, A., Gould, C.M., Jensen, L.J., Gibson, T.J., and Diella, F. 2011. Phospho.ELM: a database of phosphorylation sites--update 2011. *Nucleic Acids Res* 39:D261-267.
- Durek, P., Schmidt, R., Heazlewood, J.L., Jones, A., MacLean, D., Nagel, A., Kersten, B., and Schulze, W.X. 2010. PhosPhAt: the Arabidopsis thaliana phosphorylation site database. An update. *Nucleic Acids Res* 38:D828-834.
- Edwards, R.J., Davey, N.E., and Shields, D.C. 2008. CompariMotif: quick and easy comparisons of sequence motifs. *Bioinformatics* 24:1307-1309.
- Eng, J.K., McCormack, A.L., and Yates, J.R. 1994. An Approach to Correlate Tandem Mass-Spectral Data of Peptides with Amino-Acid-Sequences in a Protein Database. *Journal of the American Society for Mass Spectrometry* 5:976-989.
- Gould, C.M., Diella, F., Via, A., Puntervoll, P., Gemund, C., Chabanis-Davidson, S., Michael, S., Sayadi, A., Bryne, J.C., Chica, C., Seiler, M., Davey, N.E., Haslam, N., Weatheritt, R.J., Budd, A., Hughes, T., Pas, J., Rychlewski, L., Trave, G., Aasland, R., Helmer-Citterich, M., Linding, R., and Gibson, T.J. 2010. ELM: the status of the 2010 eukaryotic linear motif resource. *Nucleic Acids Res* 38:D167-180.

- Hornbeck, P.V., Chabra, I., Kornhauser, J.M., Skrzypek, E., and Zhang, B. 2004. PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation. *Proteomics* 4:1551-1561.
- Howard, A.D., Kostura, M.J., Thornberry, N., Ding, G.J., Limjuco, G., Weidner, J., Salley, J.P., Hogquist, K.A., Chaplin, D.D., Mumford, R.A., and et al. 1991. IL-1-converting enzyme requires aspartic acid residues for processing of the IL-1 beta precursor at two distinct sites and does not cleave 31-kDa IL-1 alpha. *J Immunol* 147:2964-2969.
- Hutti, J.E., Jarrell, E.T., Chang, J.D., Abbott, D.W., Storz, P., Toker, A., Cantley, L.C., and Turk, B.E. 2004. A rapid method for determining protein kinase phosphorylation specificity. *Nat Methods* 1:27-29.
- Keshava Prasad, T.S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D.S., Sebastian, A., Rani, S., Ray, S., Harrys Kishore, C.J., Kanth, S., Ahmed, M., Kashyap, M.K., Mohmood, R., Ramachandra, Y.L., Krishna, V., Rahiman, B.A., Mohan, S., Ranganathan, P., Ramabadran, S., Chaerkady, R., and Pandey, A. 2009. Human Protein Reference Database--2009 update. *Nucleic Acids Res* 37:D767-772.
- Kessels, H.W., Ward, A.C., and Schumacher, T.N. 2002. Specificity and affinity motifs for Grb2 SH2-ligand interactions. *Proc Natl Acad Sci U S A* 99:8524-8529.
- Koyasu, S., Tse, A.G., Moingeon, P., Hussey, R.E., Mildonian, A., Hannisian, J., Clayton, L.K., and Reinherz, E.L. 1994. Delineation of a T-cell activation motif

- required for binding of protein tyrosine kinases containing tandem SH2 domains. *Proc Natl Acad Sci U S A* 91:6693-6697.
- Matic, I., Schimmel, J., Hendriks, I.A., van Santen, M.A., van de Rijke, F., van Dam, H., Gnad, F., Mann, M., and Vertegaal, A.C. 2010. Site-specific identification of SUMO-2 targets in cells reveals an inverted SUMOylation motif and a hydrophobic cluster SUMOylation motif. *Mol Cell* 39:641-652.
- Perkins, D.N., Pappin, D.J., Creasy, D.M., and Cottrell, J.S. 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 20:3551-3567.
- Rigbolt, K.T., Prokhorova, T.A., Akimov, V., Henningsen, J., Johansen, P.T., Kratchmarova, I., Kassem, M., Mann, M., Olsen, J.V., and Blagoev, B. 2011. System-wide temporal characterization of the proteome and phosphoproteome of human embryonic stem cell differentiation. *Sci Signal* 4:rs3.
- 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.
- Shakin-Eshleman, S.H., Spitalnik, S.L., and Kasturi, L. 1996. The amino acid at the X position of an Asn-X-Ser sequon is an important determinant of N-linked core-glycosylation efficiency. *J Biol Chem* 271:6363-6366.
- Songyang, Z. and Cantley, L.C. 1995. Recognition and specificity in protein tyrosine kinase-mediated signalling. *Trends Biochem Sci* 20:470-475.

- Sternsdorf, T., Jensen, K., Reich, B., and Will, H. 1999. The nuclear dot protein sp100, characterization of domains necessary for dimerization, subcellular localization, and modification by small ubiquitin-like modifiers. *J Biol Chem* 274:12555-12566.
- Van Hoof, D., Munoz, J., Braam, S.R., Pinkse, M.W., Linding, R., Heck, A.J., Mummery, C.L., and Krijgsveld, J. 2009. Phosphorylation dynamics during early differentiation of human embryonic stem cells. *Cell Stem Cell* 5:214-226.

Key Reference

Schwartz and Gygi, 2005. See above.

Original description of the motif-x algorithm.

Internet Resources

<http://motif-x.med.harvard.edu>

Home page for the motif-x web tool (note: the protocols in this manuscript pertain to motif-x version 1.2).