

Using the *scan-x* Web site to predict protein post-translational modifications.

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 recent plethora of proteomic mass spectrometry data is providing evidence that almost every protein in the cell undergoes some form of post-translational modification. We describe a protocol to use the *scan-x* Web site to view predicted acetylation sites in the human proteome and predicted phosphorylation sites in the human, mouse, fly and yeast proteomes with high specificity. This tool is accessible from virtually any computer with a Web browser. The only requirement is a means of searching for a protein of interest in one of the represented organisms.

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

INTRODUCTION

Protein post-translational modifications (PTMs) direct a wide variety of cellular functions, processes and pathways. Recent advances in large-scale mass spectrometry experiments have identified thousands of PTMs including sites of phosphorylation and acetylation (Hornbeck et al., 2004; Bodenmiller et al., 2007). Despite these advances, direct identification of these sites is not comprehensive for most proteins due to a variety of experimental and technical reasons. Because PTMs are often the result of enzymes with specific preferences for residues surrounding a modification site, these preferences may be used to predict modification sites on an arbitrary protein sequence.

Here, we describe a method to find predicted modification sites on a protein of interest using the *scan-x* Web site. These potential PTM sites were generated using a computational approach that uses existing modifications to predict unknown modifications at other sites in an organism's proteome with high specificity. The *scan-x* program was designed to take the output of *motif-x* (Schwartz and Gygi, 2005) analysis runs (<http://motif-x.med.harvard.edu>, see unit 13.15 on *motif-x*) and independently scan protein sequence files for the occurrences of these motifs in other proteins or proteomes.

Each motif identified by *motif-x* contains certain positions with fixed residues, which are considered critical because of their extreme statistical significance. To predict potential modification sites, first *scan-x* uses these fixed residues as a pattern to search within protein sequence files, and only matches

sequences that contain these patterns. For example, the motif RRxS only matches proteins containing that particular pattern of fixed residues (where 'x' can be any residue). Second, after an exact match has been made, a score based on residues proximal to the fixed residues using the motif's Position Weight Matrix (PWM) is generated. The score is strongly positive for sequences that are well correlated with the PWM signature, and can be negative for those that do not correlate well with the PWM. A cross-validation methodology can be used to determine the sensitivity and specificity of scores above a particular threshold (Schwartz et al., 2009).

Table 1. Sources of Protein Data Used for PTM Prediction in Each Organism.

Organism	Proteomic data source	Web reference
Human and Mouse	IPI database	http://www.ebi.ac.uk/IPI/
Fly	FlyBase	http://flybase.org/
Budding yeast	<i>Saccharomyces</i> Genome Database (SGD)	http://www.yeastgenome.org/

Because of the cross-validation methodology, determination of score thresholds for a given sensitivity and specificity requires several runs through the *motif-x* and *scan-x* pipeline. Such analyses can produce very large output files and consume significant computational resources. Therefore, the *scan-x* Web site allows users to search a database of predicted phosphorylation and

acetylation sites that have been pre-computed using *scan-x* without having to perform cross-validation. These scans have been filtered such that only sites that which have been scored at the 95% or 99% specificity levels are presented. Currently, the *scan-x* Web site allows users to search for predictions of phosphorylation within the proteomes of human, mouse, fly and budding yeast as well as acetylation predictions within the human proteome.

In order to perform a search, the user must specify an organism, modification type, and a specificity level (either 95% or 99%). The 99% specificity level provides greater stringency, and thus greater certainty of predictions at the expense of sensitivity. The databases used in this search were obtained from the sources listed in Table 1.

Because the pre-computed *scan-x* prediction results for a given organism is prohibitively large for Web viewing, in addition to the other criteria, the *scan-x* Web server requires users to specify a particular protein of interest. In addition to the other criteria, the user must specify a protein of interest to search for within the proteome of the selected organism. However, matching a protein of interest by name is subject to the shifting changes in nomenclature that different databases adopt. Fortunately, short peptide subsequences greater than 7 residues in length are rarely repeated in an organism's proteome and can act as a signature to uniquely identify a given protein. Thus, the *scan-x* Web site provides two different methods to search for a protein of interest: (1) by using all or part of a protein name, and (2) by using a small peptide fragment from that protein.

BASIC PROTOCOL

Using the *scan-x* Web server to view predicted post-translational modification sites on a protein of interest.

The *scan-x* algorithm uses high-quality post-translational modification (PTM) motifs extracted with the *motif-x* approach to make PTM predictions on a protein of interest. The *scan-x* Web site has been designed to allow users to search precompiled *scan-x* phosphorylation and lysine acetylation (human proteome only) prediction results in the human, mouse, fly, and yeast proteomes (see Table 2).

The protocol below illustrates the use of the *scan-x* Web site to find predicted phosphorylation sites on the Retinoblastoma-associated protein (Rb1) in mouse.

Table 2. Available Organisms and Specificity Levels for each Type of Predicted PTM.

PTM	Organism(s) available for prediction	Specificity levels available for prediction
Serine phosphorylation	Human Mouse Fly Yeast	95% and 99%
Threonine phosphorylation	Human Mouse Fly Yeast	95% and 99%
Tyrosine phosphorylation	Human	95% and 99%
Lysine acetylation	Human	95% and 99%

Necessary resources

Hardware

Any computer with Internet access.

Software

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

Files

None (however a protein sequence fragment and/or the protein name is needed).

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

2. In the “select a data set to search within” options, select the appropriate organism, modification, and specificity level you wish to use for the search.

Whole proteomic scans at the 95% and 99% specificity levels were precalculated, and by definition, refer to 5% and 1% false positive rates, respectively. Simplified, the false positive rate indicates the number of truly unmodified sites within a protein that are expected to be incorrectly called as modified. For example, a 95% specificity (5% false positive rate) would suggest that in a protein containing 100 serine and threonine residues (none of which have been previously shown to be phosphorylated), on average, 5 will likely be incorrectly predicted as modified, while at the 99% specificity level, only a single incorrectly predicted site is expected. Because it is virtually impossible to characterize a position within a protein as “unmodified” under all possible conditions, the indicated specificity levels in fact serve as a lower

bound on the true specificity of the tool. One should be careful not to interpret the specificity level in terms of the prediction results (i.e., a 95% specificity level does not indicate that out of 20 predicted phosphorylation sites only 1 will be mischaracterized). In the Rb1 protein example provided we have selected “mouse ser/thr phosphorylation at 95% specificity” since we wish to find the maximal number of potential phosphorylation sites in the mouse Rb1 protein (see Figure 1).

3. In the “search by gene name” and/or “search by partial amino acid sequence” fields input the appropriate information.

When carrying out a search by gene name, scan-x searches the headers of the FASTA files within the given proteome. Since the human and mouse data sets are derived from the IPI protein database, the fly data set is derived from the FlyBase protein database, and the yeast data set is derived from the SGD database, only inputted text that exactly matches FASTA protein descriptors used in those databases will return anticipated results. A more effective method of searching involves the use of partial amino acid sequences which, when long enough (typically 7 amino acids or greater), can serve as unique protein identifiers. In the Rb1 protein example (Figure 1), we have inputted a 15 amino acid long fragment (EDDPAQDSGPEELPL) spanning positions 24 to 38 within the protein (although any segment of 7 amino acids or greater within the Rb1 protein could have been used for the present analysis). It is important to note that scan-x only uses this sequence fragment to search for proteins of

interest, not to actually scan the sequence fragment for modification sites. Thus, it is unnecessary (and not advisable) to paste an entire protein sequence into the “search by partial amino acid sequence” text box because a difference of a single amino acid (e.g., a polymorphism) can cause a protein to not be located. If both a gene name and a partial amino acid sequence are inputted, scan-x will return the union of the search results.

scan-x
v1.1 03.04.2009

Harvard Medical School

Welcome to the *scan-x* search page. *scan-x* is a software tool designed to find motifs (identified using *motif-x*) within any sequence data set. The first large scale scan was performed using all available human, mouse, fly and yeast phosphorylation and acetylation data to perform a scan for undiscovered modification sites.

For more information, please see:
Schwartz D, Chou MF, Church GM (2009). Predicting protein post-translational modifications using meta-analysis of proteome-scale data sets. *Mol Cell Proteomics* 8(2):365-79.
[Abstract](#) | [PDF](#) | [Free Full-text](#) | [Supplemental Data](#) | [PubMed](#)

Using a part of the gene name, or a short unique fragment from your protein of interest, you may search our current prediction results for phosphorylation on your favorite proteins from Human, Mouse, Drosophila and Yeast protein databases, and for acetylation on Human protein databases. As described in the paper, each data set contains predicted hits at a certain level of specificity. Generally speaking, the higher the specificity, the lower the sensitivity and vice versa. The total number of predictions for each data set as of February 2009 are described in the text of the paper.

select a data set to search within	<input type="radio"/> human ser/thr phosphorylation at 95% specificity
	<input type="radio"/> human ser/thr phosphorylation at 99% specificity
	<input type="radio"/> human tyr phosphorylation at 95% specificity
	<input type="radio"/> human tyr phosphorylation at 99% specificity
	<input type="radio"/> human lys acetylation at 95% specificity
	<input type="radio"/> human lys acetylation at 99% specificity
	<input checked="" type="radio"/> mouse ser/thr phosphorylation at 95% specificity
	<input type="radio"/> mouse ser/thr phosphorylation at 99% specificity
	<input type="radio"/> fly ser/thr phosphorylation at 95% specificity
	<input type="radio"/> fly ser/thr phosphorylation at 99% specificity
	<input type="radio"/> yeast ser/thr phosphorylation at 95% specificity
	<input type="radio"/> yeast ser/thr phosphorylation at 99% specificity

search by gene name* (minimum of 3 characters)

AND/OR**

search by partial amino acid sequence (e.g. MLPEDKE, minimum 7 letters)

*Searches are not case sensitive. When searching by gene name, human and mouse gene descriptions are derived from the [IPI database](#), drosophila gene descriptions are derived from [FlyBase](#), and yeast gene descriptions are derived from the [SGD](#).

**Providing genes AND sequence will return the union of the search results.

PLEASE NOTE: The use of automatic query software is prohibited.

By clicking the 'Search' 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
scan-x main search page. This screenshot of the *scan-x* main parameters/search page shows the appropriate parameters for the mouse Rb1 example provided in the basic protocol. An in-depth description of the various parameters can be found in the main text.



scan-x
v1.1 03.04.09

results

Parameters for this search:

```
genename =
sequence = EDDPAQDSGPEELPL
dataset = mouse ser/thr phosphorylation at 95% specificity
```

Job started Tue May 31 22:19:32 2011

(Total elapsed time: 4 seconds)

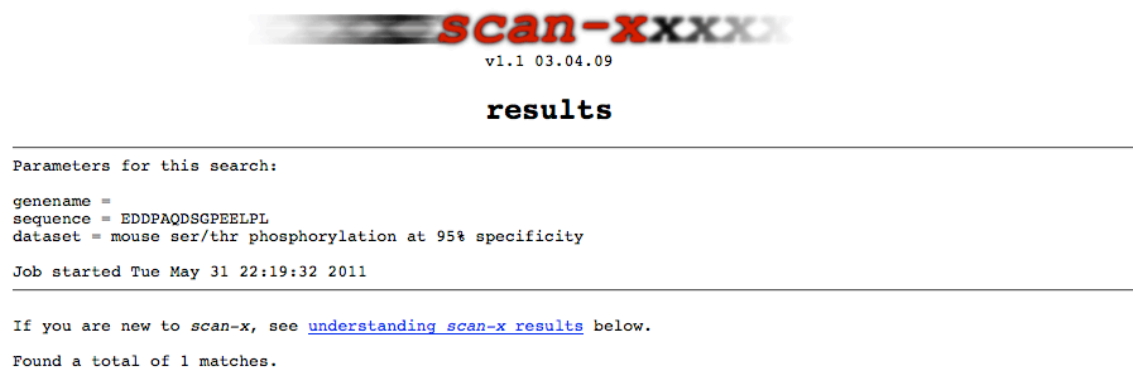
Running job on the compute cluster...this can take up to several minutes depending on server load.
You can bookmark this page if you want to check your results for up to a week.

Job Status:
cluster-521398-shared_15m (motif-x_20110531-24039-97454362_res.txt) **RUNNING**

(Automatically re-checks in 10 seconds)

Figure 2

scan-x job submission page for the Rb1 example. This screenshot shows the job submission page users are brought to upon pressing the “search” 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.



scan-x
v1.1 03.04.09

results

Parameters for this search:

```
genename =
sequence = EDDPAQDSGPEELPL
dataset = mouse ser/thr phosphorylation at 95% specificity
```

Job started Tue May 31 22:19:32 2011

If you are new to *scan-x*, see [understanding scan-x results](#) below.

Found a total of 1 matches.

Figure 3

scan-x Web search heading and parameters for the Rb1 example. This screenshot shows the uppermost portion of the *scan-x* results page for the Rb1 example described in the basic protocol.

4. Press the “Search” button.

Pressing the “Search” button will 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. scan-x searches should not take more than a couple of minutes to perform.

5. View the scan-x results page.

The scan-x results page can be subdivided into three major sections: (1) heading and parameters (Figure 3), (2) prediction results (Figure 4), and (3) the key to understanding scan-x results (Figure 5).

Heading and parameters: This section provides information on the version of the scan-x software that was used as well as a recapitulation of the parameters selected for the analysis (e.g., the gene name and/or the sequence fragment used in the search).

Prediction results: The prediction results are shown in a form similar to FASTA-formatted sequences (see Figure 4) with position numbers located at the start of each line. The rank of the protein prediction in the context of the entire proteome is provided above the FASTA-like descriptor, followed by the

total protein score (i.e., the sum of all scan-x scores) and the maximum scan-x score within the protein. Thus, in the mouse Rb1 example provided, the Rb1 protein ranked 9,197 out of a total of 31,368 mouse proteins and had total protein and maximum protein scan-x scores of 121.781 and 22.014, respectively. Predicted phosphorylation sites correspond to the central residues of the 15-amino acid peptides located beneath the protein sequence. The Rb1 example provided contains 9 predicted phosphorylation sites (Figure 4). To the right of each 15-amino acid peptide is the corresponding scan-x score and the position number of the modification site (in parentheses). The 15-amino acid peptides containing the predicted modification sites are color-coded according to the motif position weight matrix under which they were predicted. Black residues correspond to the fixed positions within the motif, green residues correspond to amino acids that are over-represented in the PWM (and thus contribute positively to the overall scan-x score), red residues correspond to amino acids that are under-represented in the PWM (and thus contribute negatively to the overall scan-x score), and gray residues correspond to amino acids for which there is no information (and are therefore neutral to the overall scan-x score). It is possible to infer the kinases involved in each predicted phosphorylation site on the basis of the motif used in the prediction. In the mouse Rb1 example provided, S243, T350, S601, S605, S800, and T814 were all predicted under either SP or TP motifs and therefore imply phosphorylation by a generalized proline directed kinase, while position S804 is part of a more specific SPxxSP motif and therefore may involve the

action of GSK3-beta kinase. It should be noted that the predicted modification sites all exceed the minimum specificity levels specified by the analysis regardless of its scan-x score; however, peptides with higher scores have a greater likelihood of being correct predictions. Additionally, scan-x scores should only be compared across peptides bearing the same number of “fixed” motif positions (i.e., black residues), with more “fixed” positions indicating a higher confidence prediction.

Key to understanding scan-x results: This self-explanatory box is found at the bottom of every scan-x results page and serves as guide to interpreting scan-x results.

```

9197/31368) score = 121.781 22.014
>IPI:IPI00121418.1 Mus musculus (Mouse) RETINOBLASTOMA-ASSOCIATED PROTEIN.
1 MPPKAPRRAAAAPPPPPPPPPREDDPAQDSGPEELPLARLEFEIEEBEFIALCQKLVDPHVRRERAWLTWEKVSSVDGILEGYIQKKELWGCIFIA
  EDDFAQDSGPEELPL 11.711 (31)
101 AVDLDEMPFTFTELOKSIETSVYKFFDILLKEIDTSTKVDNAMSRLKKNVNLCALYSKLERTCELIYLTQPSSALSTEINSMLVLRKISWITFLAKGEVL
201 QMEDDLVISFQMLCVVDYFIKFSPPALLREPYKTAAPINGSPTPRRQNRSARIAKQLENDTRIEIVLCKEHECNIDEVKNVYKNFIPFINSLGIV
  AAIPINGSPTPRRG 18.361 (243)
301 SSNGLPEVESLSKRYEEVYLKKNKDLARLFLDHDKTLQTDPIDSFETERTPRKNNPDEEANVTPHTPVRTVMNTIQQLMVILNSASDQPSLENLISYFNN
  DSFETERTPRKNNPD 9.844 (350)
401 CTVNPKENILKRVKDVGHIFKEKFAVAVGQCVDIGVQRYKLVRLYYRVMESMLKSEEBERLSIQNFSKLLNDNIFHMSSLACALEVVMATYSRSTLQHL
501 DSGTDLSPFWILNVLNKAQDFYKVIKVEANLTREMIKHLERCEHRIMESLAWLSDSPLFDLIKQSKDGECPDLEPACPLSLPLOGNHTAADMYL
  TAADMYSPLRSPKX 15.892 (601)
  MYLSPLRSPKRTST 10.685 (605)
601 SPLRSPKRTSTTRVNSAANTETAASAFHTQKPLKSTSLALFYKVVYRLAYLRLNLTLCARLLSDHPELEHIWTLFQHTLQNEVELMRDRHLDQIMMCS
701 MYGICKVKNIIDLKFKIIVTAYKDLPHAAQETFKRVLIREEEFDSIIVFYNVSMQRLKTNILQVASTRPPPLSPIPHIPRSPYKFSSSPLRIPGCGNIYIS
  YASTRPPPLSPIPHI 10.230 (771) PCGNIYISPLKSPYK 22.014 (800)
  IYISPLKSPYKISEG 11.888 (804)
801 PLKSPYKISEGLPTPTKMTPTSRILVSGICESFOTSEKFKINQMVCSNDRVLKRSAGCGNPPKLNVRFDIEGAEADGSKHLPAESKFPQQLAEMTST
  KISEGLPTPTKMTPT 11.154 (814)
901 RTRMQKQRMNESKDVSNKEEK

```

Figure 4

scan-x prediction results for the Rb1 example. This screenshot shows the scan-x results for the Rb1 example described in the basic protocol. An in-depth explanation of the results view is provided in the text.

```

Key to understanding scan-x results:

Protein_Rank/Total_Proteins) score = Total_Protein_Score Maximum_Score
>Protein_Name
  1 MLPEDKEADSLRCNISVKAVKKEVEKKLRCLLADLPLPPELPGDDLSKSPEEKKTATQLHSKRRPKICGPYGETKEKDIDWGKRCVDFDI . . .
      ISVKAVKKEVEKKLR 16.512 (22)

scan-x predicted hits follow sequence data and are aligned with the sequence.

The results are color coded with:
GREEN meaning the specific residue is over-represented in the motif,
RED meaning the specific residue is under-represented in the motif,
BLACK meaning the residue is an essential (i.e. fixed) part of the motif, and
GRAY meaning that there is no information about that residue in the motif.

Following the amino-acid sequence is the scan-x score, and the position of the central residue is in parentheses.

NOTE: While some scan-x predictions will contain experimentally verified modification sites, due to sensitivity
thresholds you will typically not see all of these known sites for a protein. So, in addition to this scan-x
output, you may wish to check the sources of our training data including:
Swiss-Prot, Phospho.ELM, PhosphoSite and PhosphoPep.

```

Figure 5

Understanding scan-x results text box. This screenshot shows a key for understanding *scan-x* results, which can be found at the bottom of every *scan-x* search results page.

GUIDELINES FOR UNDERSTANDING RESULTS

The *scan-x* Web site search output contains a number of features that allow interpretation of the results. This includes the motif's syntactic match (indicated by black letters), statistically over- or under-represented characters (indicated by green or red letters respectively), the predicted modification position, and a score that is related to the significance of the correlation with the motif's PWM. Higher scores generally have better correlations to the found motif. All predicted sites shown in the output have score values above the requested specificity threshold (i.e., 95% or 99%).

scan-x search results reflect phosphorylation or acetylation site predictions across all proteins of the represented organisms at a chosen specificity level. In general, short simple syntactic motifs alone are easily found within proteomes, so purely syntactic matches provide only one level of specificity (e.g., in a random amino acid sequence, the SP motif would be expected to occur once in

approximately every 400 residues). *scan-x* provides an additional level of specificity based upon the significance score in combination with the syntactic match. This combined approach also factors in the contribution of residues that neighbor the exact match site.

The final output of a *scan-x* Web search includes entire protein sequences (subject to the protein input search criteria) that have at least one highlighted motif. If a protein of interest is correctly identified by name or peptide sequence, then the failure to find it in the output or the failure to find a prediction within a given protein means that a score did not exceed the selected specificity cutoff.

Finally, the results page has a unique URL that 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

As biologically related data sets become larger, they provide the means to determine many statistically significant features, such as the post-translational motifs described here. By using the *scan-x* Web site to search for proteins of interest, one can identify potential phosphorylation or acetylation modification sites that can lead to further biological hypotheses and experiments.

The *scan-x* Web site is the result of many cross-validated searches through proteomic databases using an internal version of *scan-x* to determine sensitivity and specificity as described in Schwartz, Chou, and Church (2009). At

this point, the nature of such large cross-validation analyses does not readily lend itself to routine analysis on a Web server; therefore, the current *scan-x* Web site is designed to allow searches through previously validated *motif-x/scan-x* runs.

Cross-validation has also indicated an improvement in sensitivity and specificity over existing global post-translational modification prediction approaches (Obenauer et al., 2003; Ingrell et al., 2007; Gnad et al., 2011), thus making *scan-x* a useful starting point to explore potential post-translational modification sites on proteins of interest.

Critical Parameters

Given the current lack of comprehensive post-translational modification data for all tissue types, in some cases, training data may not be entirely representative of actual modifications that occur in nature. Therefore, it is not unusual for one organism to contain a prediction that is lacking in a homologous protein due to lack of within-organism training data. To obtain a more comprehensive understanding of phosphorylation modifications, we recommend that users look for predictions at the level of both 95% and 99% stringency, as well as in homologous proteins in other organisms.

There are no default parameters.

Troubleshooting

The failure of a search to return a substrate protein with a prediction may be the result of (1) lack of evidence of phosphorylation or acetylation at that site; (2) a failure to find a predicted motif at the given specificity level, (3) failure to find a protein by name or peptide fragment; or (4) potential bias or deficiencies in the underlying training set used by the authors to predict motifs.

Table 3. Potential *scan-x* error messages and Explanations.

Error message	Suggested Action
You must select a gene name or a sequence to search on	It is necessary to restrict the search to at least a gene name and/or a peptide fragment.
Gene name must be at least 3 characters (if used)	Gene names are optional, but if used, they must have at least 3 characters.
Sequence must be at least 7 letters (if used)	Sequence fragment searching is optional, but if used, must have at least 7 residues.
You must specify a data set to search within	Use the radio buttons on the search page to choose a PTM, database and specificity level.
Cannot locate results for _____ they may have been purged	A previously submitted job may have expired and been purged after a week.
Cannot locate results for jobID _____. They may have been purged, or are not yet available.	Either the URL was not typed correctly, or more likely, a previously submitted job may have expired and been purged after a week.

If you believe that a protein of interest should have a predicted modification, and you have not yet searched by sequence fragment, then: Locate the protein sequence for your selected organism through some other means (e.g. NCBI, Swissprot, etc.), copy a small 7-10 residue sequence from the retrieved peptide,

and paste it into the “search by partial amino acid sequence” field. Clear all characters from the “search by gene name” field and search again.

Table 3 provides a more detailed explanation of current *scan-x* error messages.

Literature Cited

- Bodenmiller, B., Malmstrom, J., Gerrits, B., Campbell, D., Lam, H., Schmidt, A., Rinner, O., Mueller, L.N., Shannon, P.T., Pedrioli, P.G., Panse, C., Lee, H.K., Schlapbach, R., and Aebersold, R. 2007. PhosphoPep--a phosphoproteome resource for systems biology research in *Drosophila* Kc167 cells. *Mol Syst Biol* 3:139.
- Gnad, F., Gunawardena, J., and Mann, M. 2011. PHOSIDA 2011: the posttranslational modification database. *Nucleic Acids Res* 39:D253-260.
- 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.
- Ingrell, C.R., Miller, M.L., Jensen, O.N., and Blom, N. 2007. NetPhosYeast: prediction of protein phosphorylation sites in yeast. *Bioinformatics* 23:895-897.
- Obenauer, J.C., Cantley, L.C., and Yaffe, M.B. 2003. Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res* 31:3635-3641.
- Schwartz, D., Chou, M.F., and Church, G.M. 2009. Predicting protein post-translational modifications using meta-analysis of proteome scale data sets. *Mol Cell Proteomics* 8:365-379.
- 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.

Key Reference

Schwartz, Chou, and Church, 2009. See above.

Original description of the scan-x algorithm.

Internet Resources

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

Home page for the scan-x version 1.1 Web tool.

<http://phosphosite.org>

A large database of experimentally determined post-translational modification sites.

<http://phospho.elm.eu.org>

A large database of experimentally determined phosphorylation sites.

<http://scansite.mit.edu>

An alternative, widely used Web tool for kinase-specific phosphorylation prediction.