New PhosphoPep help page and tutorial for users

Buttons used in PhosphoPep

 View available KEGG pathways for this protein (http://www.genome.jp/kegg/)[1]

 Start cytoscape network with this protein (http://www.cytoscape.org/)[2]

 View orthologs/homolog information (www.orthomcl.org/)[3]

 Search for protein interaction information in String (http://string.embl.de/)[4]

 Look up protein information in PeptideAtlas (http://www.peptideatlas.org/)[5]

 Search protein sequence at Scansite (http://scansite.mit.edu/)[6]

Importantly, as the current knowledge about cellular pathways is far from complete, only a portion of the phosphoproteins can be placed into pathways. This partial knowledge also applies to the orthologous protein information as well as to the prediction of a kinase for a given phosphorylation site.

Scores and numbers used in PhosphoPep

PeptideProphet	When interpreting tandem mass spectrometry data, it is crucial to determine if an identification is correct. The PeptideProphet computes a probability of a given fragment ion spectrum to be correctly assigned to a peptide sequence by a given database search algorithm and assigns a score accordingly[7, 8]. The range of the score is from 0 (worst) to 1 (best). Depending on the dataset or database the probabilities can slightly vary at a given threshold/score.
Tryptic Ends	As we analyze peptides in our tandem mass spectrometry experiments we have to digest the proteins using a protease. This is often done by using trypsin. Trypsin cleaves after arginine and lysine but exhibits also some unspecific cleavage[9]. Two tryptic ends means that both ends were specifically cut by trypsin.
Peptide Mass	Molecular mass of the (phospho)peptide
deltaCn	The deltaCn score (dCn) is a score computed by the Sequest[10] algorithm which we use to interpret tandem mass spectra. The dCn is the difference between the (normalized) cross-correlation parameter of the first- and second-ranked amino acid sequence assigned to a tandem mass spectrum. Simplified, the dCn tells you how much better

	the first (best) database search hit fits to a tandem mass spectrum than the second hit. In the case of phosphopeptides the dCn also correlates to the correctness of the phosphorylation site assignment within the phosphopeptide sequence[11].
# Obs	Number of times the phosphopeptide was identified in our experiments
# Mappings	Maps to # of gene models / maps to # of transcripts

How to assess the quality of a phosphopeptide identified using tandem mass spectrometry

In order to understand the basic methods of peptide identification using tandem mass spectrometry we strongly recommend studying the presentation which you can find under the link http://www.proteomesoftware.com/Proteome_software_pro_interpreting.html *

The presentation is easy to understand and represents a nice introduction to proteomics.

Of note: As the following text was written for users without any experience in mass spectrometry we attempted to describe each topic in a simplified manner, sometimes at the expense of accuracy. For users who wish to learn more about each topic we propose to read the literature given at the end of this tutorial.

When phosphopeptides are analyzed using liquid chromatography – tandem mass spectrometry and phosphopeptide sequences are assigned to the resulting spectra using database search algorithms, primarily two types of error can occur. The first type of error is the mis-assignment of the fragment ion spectrum to a peptide sequence[7, 8]. The second type of error is the mis-assignment of the site of phosphorylation in an otherwise correctly identified phosphopeptide[11].

Here we explain how each of the errors was assed and how the users of PhosphoPep can use the computed scores and some rules to judge if a phosphopeptide was correctly identified and the site correctly assigned.

* Proteome Software Inc.

Is the phosphopeptide correctly identified?

As mentioned above, one type of error in the automatic interpretation of tandem mass spectra is the mis-assignment of the fragment ion spectrum to a peptide sequence. This type of error can be estimated by applying statistical models such as the PeptideProphet[8] and/or by using decoy sequence databases[12].

All data loaded into PhosphoPep were assessed using both methods and we already applied a stringent cut off on all data. Therefore the false positive content in the case of the fly data is about 2.6 % (for yeast, worm and human this number is similar). This means that if you don't apply any further filter criteria about 1 out of 38 phosphopeptide entries is wrong. For bioinformatic large scale analyses this false positive rate is in most cases very acceptable, but for a biologist who wants to perform follow up experiments this can already be too high and therefore it is desirable to choose your own false positive rate. So how do you choose your own false positive rate?

One of the statistical tools to compute the false positive rate, the PeptideProphet[8], computes a score (ranging from 0 (worst) to 1(best)). This score is displayed for every peptide in PhosphoPep[13]. As mentioned above, we have already prefiltered the data, therefore the lowest PeptideProphet score you will find is 0.8. The closer the score is to 1.0 the lower is the chance that you pick a wrongly identified phosphopeptide. For example, at a Peptide Prophet cut off of 0.99 approximately 0.2 % of all entries (equal or above this score) are estimated to be false positive assignments (1 out of 500 phosphopeptide entries) for the fly dataset.

⊖ **Observed Phosphopeptides** view transitions

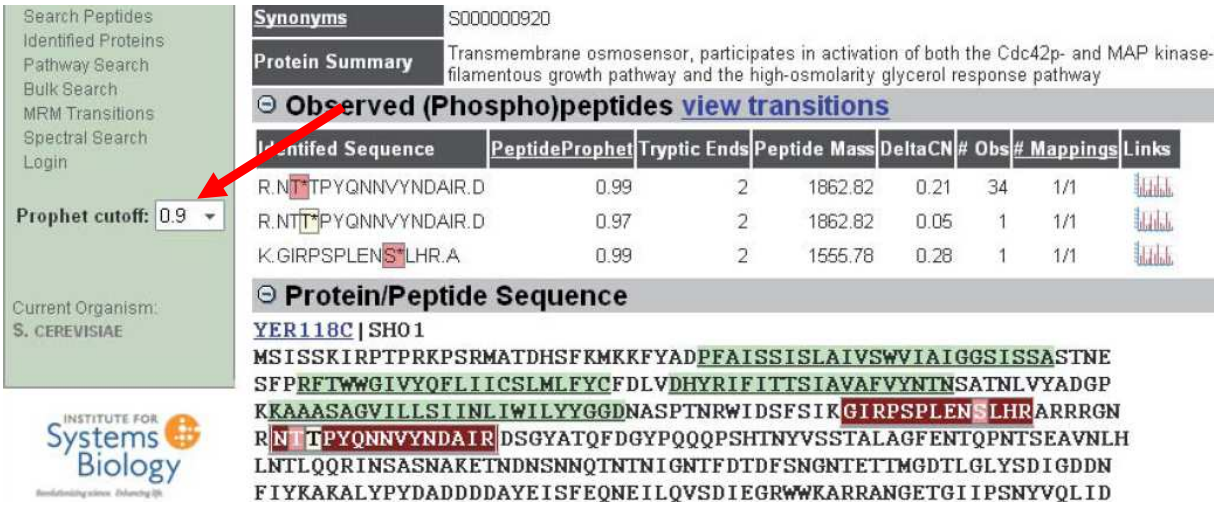
Identifed Sequence	PeptideProphet	Tryptic Ends	Peptide Mass	DeltaCN	# Obs	# Mappings	Links
R.TTS*SSFESEIK.S	0.92	2	1295.52	0.23	4	1/1	
K.SNGANRDS*SDLAPTLR.S	0.96	2	1753.78	0.16	19	1/1	
K.SNGANRDSS*DLAPTLR.S	0.84	2	1753.78	0.02	1	1/1	
K.RVS*DVLPK.R	0.95	2	993.52	0.21	6	1/1	

⊖ **Protein/Peptide Sequence**

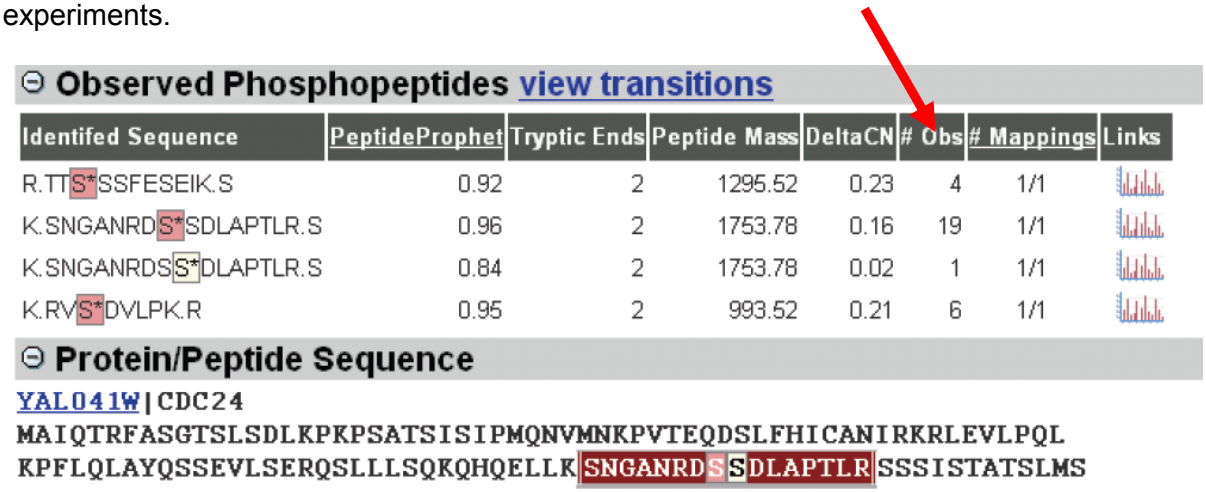
YAL041W|CDC24

MAIQTRFASGTSLSDLKPKPSATSISIPMQNVMNKPVTEQDSLFHICANIRKRLEVLPQL
KPFLQLAYQSSEVLSERQSLLLSQKQHQELLK**SNGANRDSSDLAPTLR**SSSISTATSLMS

With the button to the left you can choose the Prophet Score cut off on your own.



One further criterion which increases the certainty that a phosphopeptide was correctly identified is the "# Obs" which tells you how often a phosphopeptide was identified in our experiments.



The chance that a phosphopeptide which was identified multiple times is wrong is lower than that of a phosphopeptides that was just identified once (but keep in mind that this is only a rule of thumb and exceptions exist)[14].

So taken together, if you choose a phosphopeptide for follow up experiments make sure that it has a high PeptideProphet score and was observed multiple times.

Is the site of phosphorylation correctly assigned?

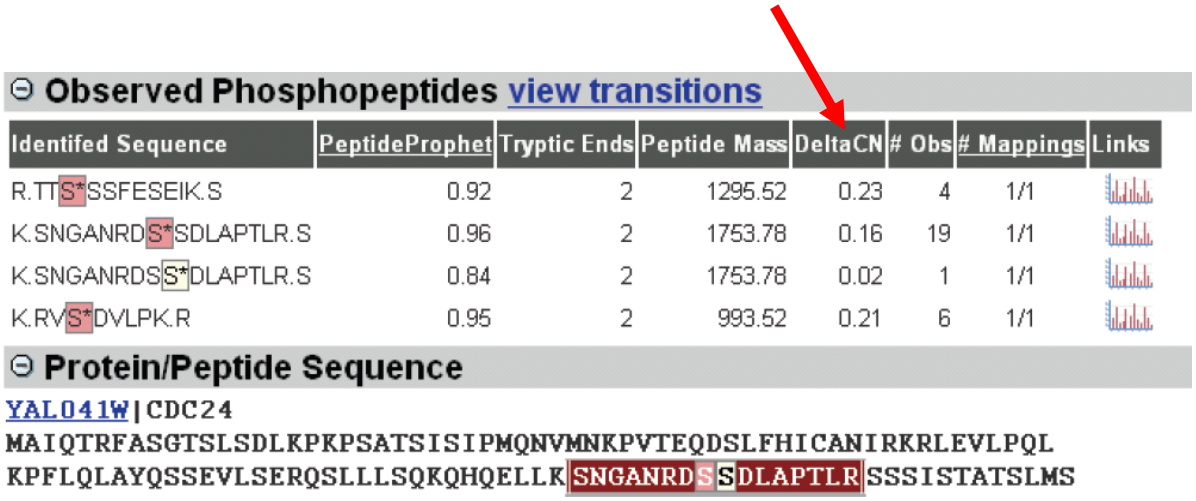
Often phosphopeptides are rich in serine and threonine residues which can sometimes puzzle the algorithm for the automatic interpretation of tandem mass spectra in regards to which serine/threonine(/tyrosine) was phosphorylated[11]. Therefore another type of error connected to phosphopeptides identified using tandem mass spectrometry is the mis-assignment of the site of phosphorylation in an otherwise correctly identified phosphopeptide[11].

This error was estimated by comparing the search engine output scores for the potential phosphorylated forms of a peptide, assuming that any hydroxy-amino acid in a phosphopeptide could be phosphorylated. Based on this estimation we highlighted the phosphopeptides either red (high probability of correct assignment) or yellow (low probability of correct assignment)[10, 11, 13].

As one typical approach to study protein phosphorylation is to mutate the site of phosphorylation to another amino acid residue it is advisable to ascertain that you choose the correct amino acid. There are several steps you can take in order to ensure that the site of phosphorylation was correctly assigned.

1) *Take a look at the dCn value.*

The first step to determine the certainty in the phosphorylation site assignment is to look at the dCn score. Simplified, the dCn tells you how much better the first (best) database search hit fits to a tandem mass spectrum than the second hit. Now if the first and second hits are the same phosphopeptide but the Sequest algorithm has problems to unequivocally assign the phosphorylation site, the score will be very low, often close to zero.



Observed Phosphopeptides view transitions

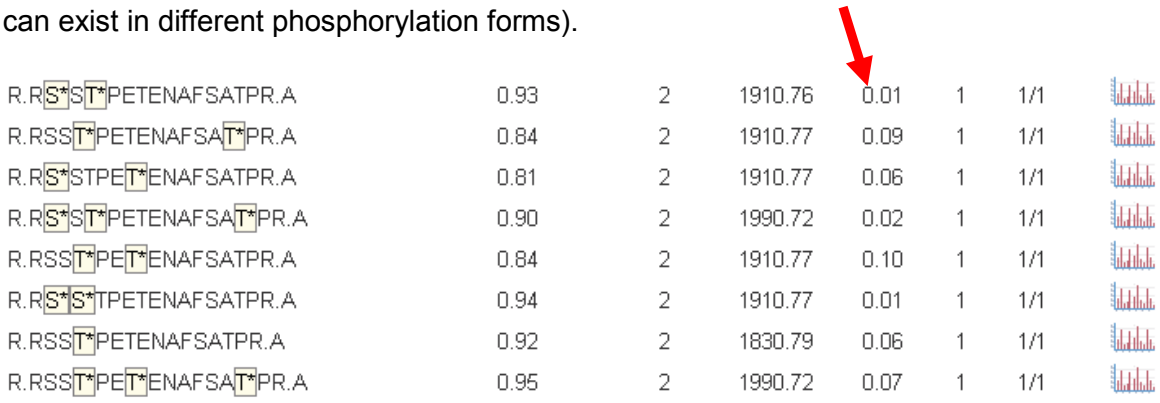
Identifed Sequence	PeptideProphet	Tryptic Ends	Peptide Mass	DeltaCN	# Obs	# Mappings	Links
R.TTS*SSFESEIK.S	0.92	2	1295.52	0.23	4	1/1	
K.SNGANRDS*SDLAPTLR.S	0.96	2	1753.78	0.16	19	1/1	
K.SNGANRDSS*DLAPTLR.S	0.84	2	1753.78	0.02	1	1/1	
K.RVS*DVLPK.R	0.95	2	993.52	0.21	6	1/1	

Protein/Peptide Sequence

YAL041W|CDC24
MAIQTRFASGTSLSDLKPKPSATSISIPMQNVMNKPVTEQDSLFHICANIRKRLEVLPQL
KPFLQLAYQSSEVLSERQSLLLSQKQHQELLK SNGANRDSSDLAPTLR SSSISTATSLMS

Again as a rule of thumb: The higher the dCn score the more certain is the phosphorylation site assignment. Normally, a score of dCn > 0.125 corresponds to a high certainty that the site is correctly assigned[11].

Below a phosphopeptide is shown which was identified several times but the site of phosphorylation could never be assigned with high certainty. As a result the same phosphopeptide exists in several versions in PhosphoPep. Such agglomerations of the same peptide with many different phosphorylation sites are a hint that the site is not well assigned (but keep in mind, some proteins are heavily phosphorylated and therefore the same peptide can exist in different phosphorylation forms).



R.RS*ST*PETENAFSATPR.A	0.93	2	1910.76	0.01	1	1/1	
R.RSST*PETENAFSAT*PR.A	0.84	2	1910.77	0.09	1	1/1	
R.RS*STPET*ENAFSATPR.A	0.81	2	1910.77	0.06	1	1/1	
R.RS*ST*PETENAFSAT*PR.A	0.90	2	1990.72	0.02	1	1/1	
R.RSST*PET*ENAFSATPR.A	0.84	2	1910.77	0.10	1	1/1	
R.RS*S*TPETENAFSATPR.A	0.94	2	1910.77	0.01	1	1/1	
R.RSST*PETENAFSATPR.A	0.92	2	1830.79	0.06	1	1/1	
R.RSST*PET*ENAFSAT*PR.A	0.95	2	1990.72	0.07	1	1/1	

2) Take a look at the kinase phosphorylation motif

An additional step to take in order to confirm a site of phosphorylation is to look at the possible kinase motif surrounding the phosphorylation site[15]. In the example below phosphorylation sites on the protein FUS3, a MAPK, are shown. Here it is not clear whether

R.IIDESAADNSEPTGQQ**S***GMTE**Y***VATR.W
or
R.IIDESAADNSEPTGQQSGM**T***E**Y***VATR.W

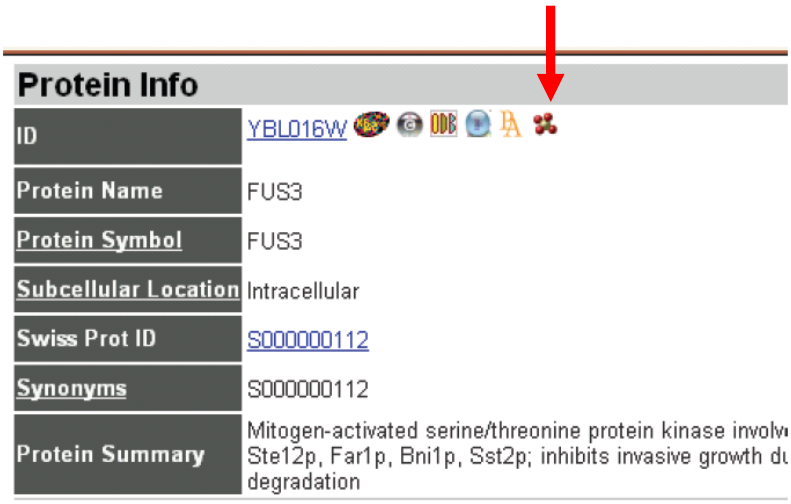
is correct. Knowing that the MAP kinases are activated by the phosphorylation in the **T**X**Y** motif we can assume that the R.IIDESAADNSEPTGQQSGM**T***E**Y***VATR.W is correct.

Observed Phosphopeptides view transitions

Identifed Sequence	PeptideProphet	Tryptic Ends	Peptide Mass	DeltaCN	# Obs	# Mappings	Links
R.IIDESAADNSEPTGQQS*GMTEY*VATR.W	0.93	2	2930.17	0.06	1	1/1	
R.IIDESAADNSEPTGQQSGMT*EY*VATR.W	0.85	2	2930.17	0.06	1	1/1	

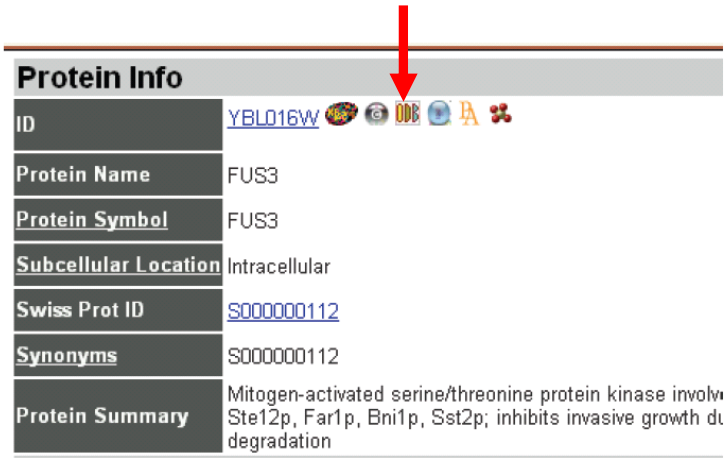
3) Predict the motif using Scansite

In case you do not have all kinase motifs memorized you can use the Scansite algorithm[6] to search the protein sequence for possible kinase motifs. For this simply click on the button "Search protein sequence at Scansite" in the "Protein Info" section.

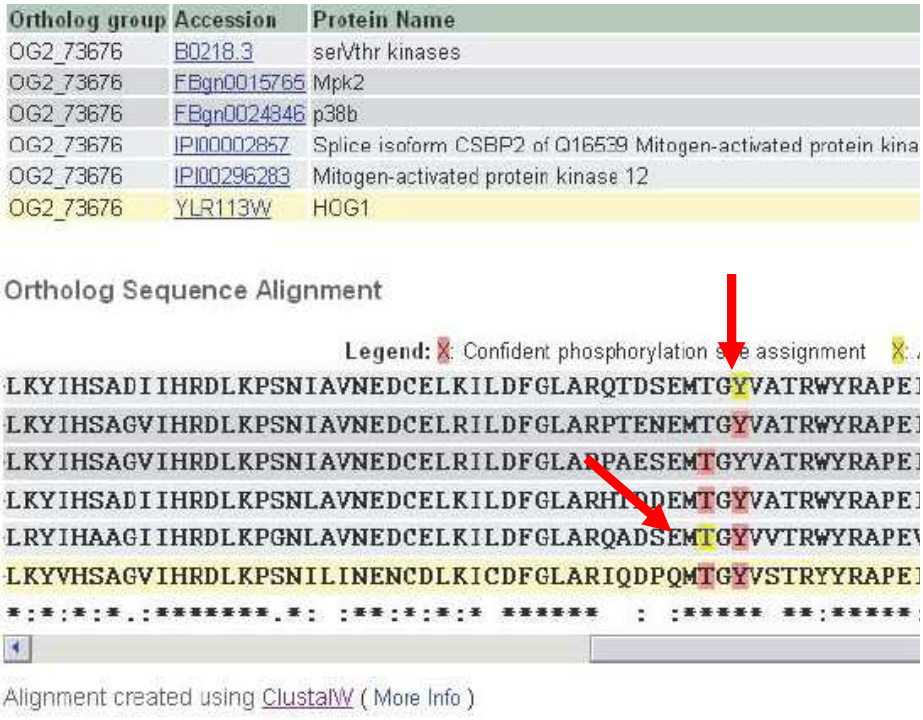


4) Check the evolutionary conservation of the site

You can also check whether your phosphorylation site of interest is evolutionary conserved which can be an additional indication for the correct assignment of a phosphorylation site. For this click on the button "View orthologs/homolog information"



and a new window will be opened, showing the alignment of the amino acid sequences with the identified phosphorylation sites between yeast, worm fly and human. In the example below, we can conclude that the unassigned phosphothreonine (highlighted in yellow) is correctly assigned and that in the top amino acid sequence either the tyrosine or threonine in the **T**X**Y** motif should be phosphorylated.



5) Take a look at the tandem mass spectrum

In order to assess whether the phosphorylation site was correctly assigned, it is always advisable to take a look at the tandem mass spectrum of the phosphopeptide. You can open it by clicking on the symbol

(The manual interpretation of tandem mass spectra can be, especially in the case of phosphopeptides, difficult. Therefore we recommend to study the following slides which are a nice introduction to this topic. You can find them under the URL

http://www.proteomesoftware.com/Proteome_software_pro_protein_id.html *

Observed Phosphopeptides view transitions

Identifed Sequence	PeptideProphet	Tryptic Ends	Peptide Mass	DeltaCN	# Obs	# Mappings	Links
R.TTS*SSFESEIK.S	0.92	2	1295.52	0.23	4	1/1	
K.SNGANRDS*SDLAPTLR.S	0.96	2	1753.78	0.16	19	1/1	
K.SNGANRDSS*DLAPTLR.S	0.84	2	1753.78	0.02	1	1/1	
K.RVS*DVLPK.R	0.95	2	993.52	0.21	6	1/1	

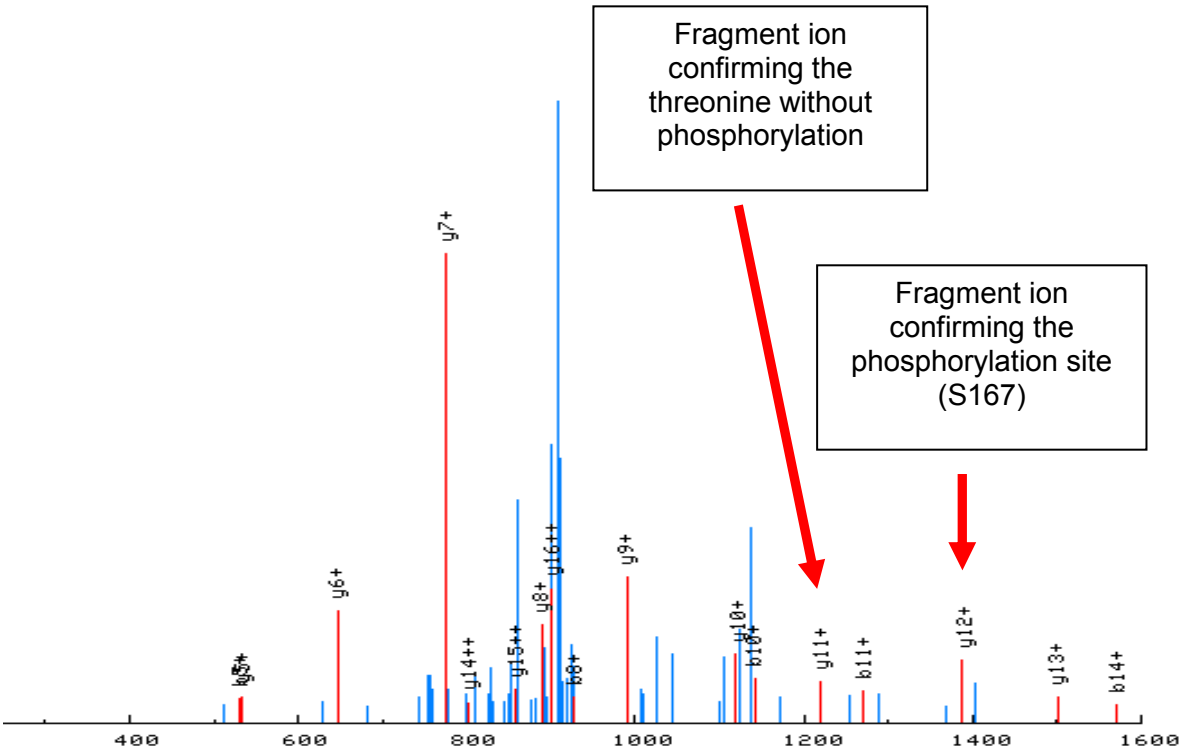
Protein/Peptide Sequence

YAL041W|CDC24
MAIQTRFASGTSLSDLKPKPSATSISIPMQNVMNKPVTEQDSLFHICANIRKRLEVLPQL
KPFLQLAYQSSEVLSERQSLLLSQKQHQELLK SNGANRDSSDLAPTLR SSSISTATSLMS

This will open a new window in which the tandem mass spectrum is displayed (See next page). In the upper window you see the tandem mass spectrum in which the fragment ion peaks are assigned with y-ion or b-ion together with a number (ion assignment nomenclature) as well as below the spectrum the amino acid sequence of the phosphopeptide is shown (a phosphoserine is indicated as "`S[167]`", a phosphothreonine as "`T[181]`" and a phosphotyrosine as "`Y[243]`"). Here you have to look for the following: left and right of the amino acid sequence the fragment ion signals which were found and could be assigned in the tandem mass spectrum are highlighted in red. In our example the question is, if the serine (at position 6) is phosphorylated LSLTDS$_{167}$TETIENNATVK or the adjacent threonine LSLTDST$_{167}$ETIENNATVK at position 7.

* Proteome Software Inc.

Of note, most spectra loaded into the PhosphoPep database are consensus spectra[16], which means that only repeatedly observed peptide fragment ions are shown. Noise signals were removed.



b^{1+}	b^{2+}	#	AA	#	y^{1+}	y^{2+}
114.1668	57.5871	1	L	17		
201.2450	101.1262	2	S	16	1803.8088	**902.4081**
314.4044	157.7059	3	L	15	1716.7306	**858.8690**
415.5095	208.2584	4	T	14	1603.5712	**802.2893**
530.5961	265.8017	5	D	13	**1502.4661**	751.7368
697.6542	349.3308	6	S[167]	12	**1387.3795**	694.1935
798.7593	399.8833	7	T	11	**1220.3214**	610.6644
927.8748	464.4411	8	E	10	**1119.2163**	560.1119
1028.9799	514.9936	9	T	9	**990.1008**	495.5541
1142.1393	571.5733	10	I	8	**888.9957**	445.0016
1271.2548	636.1311	11	E	7	**775.8363**	388.4219
1385.3586	693.1830	12	N	6	**646.7208**	323.8641
1499.4624	750.2349	13	N	5	**532.6170**	266.8122
1570.5412	785.7743	14	A	4	418.5132	209.7603
1671.6463	836.3268	15	T	3	347.4344	174.2209
1770.7789	885.8931	16	V	2	246.3293	123.6684
		17	K	1	147.1967	74.1021

Phosphorylation site

The inspection of the highlighted ions shows that indeed all peptide fragment ions, including the one corresponding to the phosphorylated serine as well as the non-phosphorylated

threonine, were identified and assigned, strengthening that the assigned serine phosphorylation is correct. In addition, take a look at the tandem mass spectrum. Here you can see that both assigned fragment ions are rather intense. In addition, the y^{11+} fragment ion at m/z 1,300.3 (1220.3 + 80) or the y^{11++} fragment ion at m/z 650.7 which would indicate that threonine 7 phosphorylated, are missing. The same is true for the b^{6+} ion at m/z 617.7 which would indicate that serine 6 is not phosphorylated. Both findings suggest the correct assignment of the phosphorylation site.

On the next two pages an example is shown in which the assignment of the correct site of phosphorylation is difficult. It is not clear if the highlighted serine

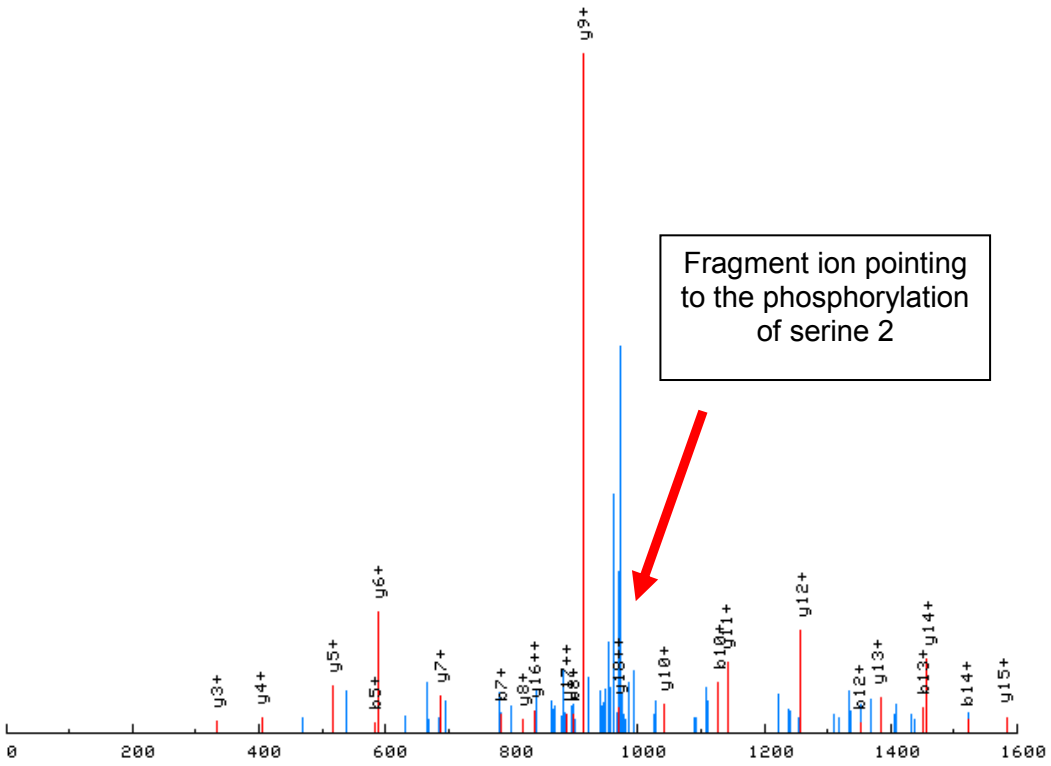
T**S***VSEAQNTQPQVANADAK

or the highlighted threonine

T*SVSEAQNTQPQVANADAK

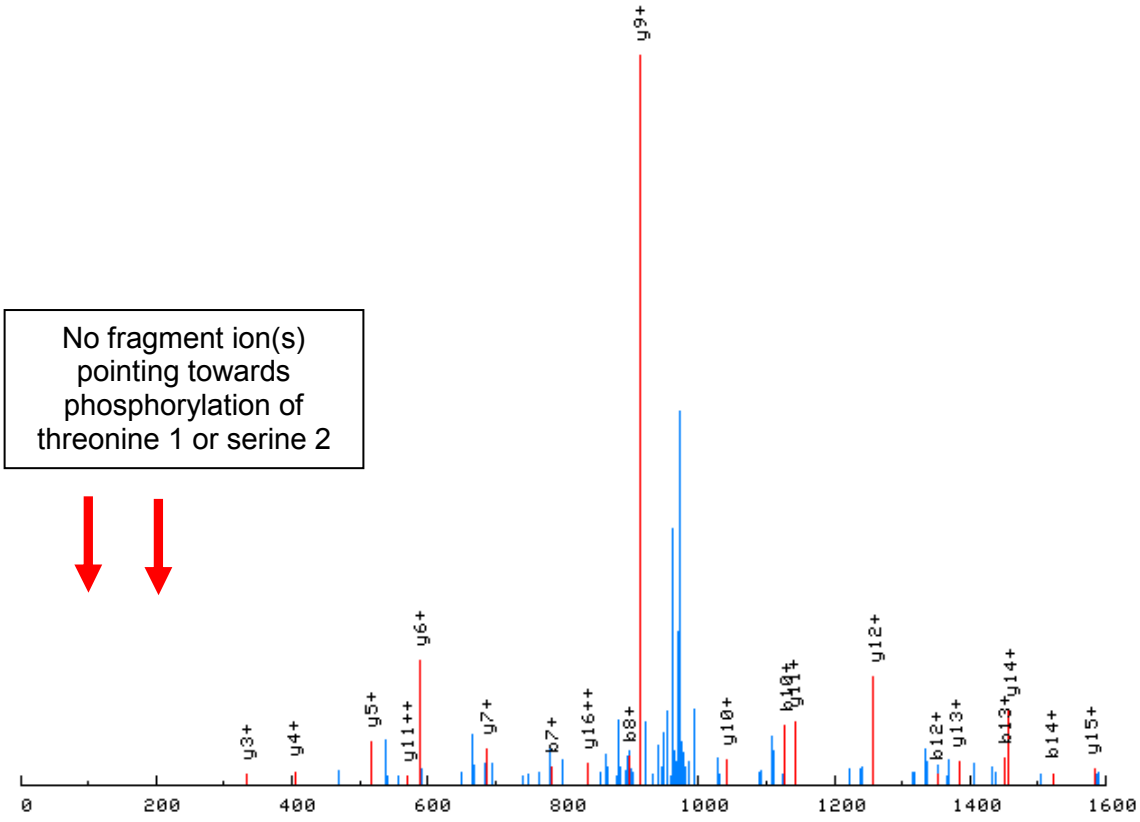
is phosphorylated.

First, most peptide fragment ions which could unequivocally distinguish the two phosphorylation sites are outside the recorded m/z range. Second, the y^{18++} fragment ion at m/z 969.97 which could indicate that the serine 2 is phosphorylated, but not threonine 1, is present at low relative intensity in a m/z region crowded with signals, therefore it is not sure whether this is a real fragment ion or noise.



b^{1+}	b^{2+}	#	AA	#	y^{1+}	y^{2+}
102.1125	51.5599	1	T	19		
269.1706	135.0890	2	S[167]	18	1938.9370	**969.9722**
368.3032	184.6553	3	V	17	1771.8789	**886.4432**
455.3814	228.1944	4	S	16	1672.7463	**836.8769**
584.4969	292.7521	5	E	15	**1585.6681**	793.3378
655.5757	328.2915	6	A	14	**1456.5526**	728.7800
783.7064	392.3569	7	Q	13	**1385.4738**	693.2406
897.8102	449.4088	8	N	12	**1257.3431**	629.1753
998.9153	499.9613	9	T	11	**1143.2393**	572.1234
1127.0460	564.0267	10	Q	10	**1042.1342**	521.5708
1224.1627	612.5850	11	P	9	**914.0035**	457.5055
1352.2934	676.6504	12	Q	8	**816.8868**	408.9471
1451.4260	726.2167	13	V	7	**688.7561**	344.8818
1522.5048	761.7561	14	A	6	**589.6235**	295.3155
1636.6086	818.8080	15	N	5	**518.5447**	259.7761
1707.6874	854.3474	16	A	4	**404.4409**	202.7242
1822.7740	911.8907	17	D	3	**333.3621**	167.1848
1893.8528	947.4301	18	A	2	218.2755	109.6415
		19	K	1	147.1967	74.1021

No fragment ion(s) pointing towards phosphorylation of threonine 1 or serine 2



b^{1+}	b^{2+}	#	AA	#	y^{1+}	y^{2+}
182.0924	91.5499	1	T[181]	19		
269.1706	135.0890	2	S	18	1858.9571	929.9823
368.3032	184.6553	3	V	17	1771.8789	886.4432
455.3814	228.1944	4	S	16	1672.7463	**836.8769**
584.4969	292.7521	5	E	15	**1585.6681**	793.3378
655.5757	328.2915	6	A	14	**1456.5526**	728.7800
783.7064	392.3569	7	Q	13	**1385.4738**	693.2406
897.8102	449.4088	8	N	12	**1257.3431**	629.1753
998.9153	499.9613	9	T	11	**1143.2393**	**572.1234**
1127.0460	564.0267	10	Q	10	**1042.1342**	521.5708
1224.1627	612.5850	11	P	9	**914.0035**	457.5055
1352.2934	676.6504	12	Q	8	816.8868	408.9471
1451.4260	726.2167	13	V	7	**688.7561**	344.8818
1522.5048	761.7561	14	A	6	**589.6235**	295.3155
1636.6086	818.8080	15	N	5	**518.5447**	259.7761
1707.6874	854.3474	16	A	4	**404.4409**	202.7242
1822.7740	911.8907	17	D	3	**333.3621**	167.1848
1893.8528	947.4301	18	A	2	218.2755	109.6415
		19	K	1	147.1967	74.1021

References

1. Kanehisa, M. et al. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* **34**, D354-357 (2006).
2. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498-2504 (2003).
3. Chen, F., Mackey, A.J., Stoeckert, C.J., Jr. & Roos, D.S. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.* **34**, D363-368 (2006).
4. von Mering, C. et al. STRING 7--recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res.* **35**, D358-362 (2007).
5. Desiere, F. et al. Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biol.* **6**, R9 (2005).
6. Obenauer, J.C., Cantley, L.C. & Yaffe, M.B. Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res.* **31**, 3635-3641 (2003).
7. Keller, A., Eng, J., Zhang, N., Li, X.J. & Aebersold, R. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol. Syst. Biol.* **1**, 2005 0017 (2005).
8. Keller, A., Nesvizhskii, A.I., Kolker, E. & Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74**, 5383-5392 (2002).
9. Picotti, P., Aebersold, R. & Domon, B. The Implications of Proteolytic Background for Shotgun Proteomics. *Mol. Cell. Proteomics* (2007).
10. Eng, J.K., McCormack, A.L. & Yates, J.R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976-989 (1994).
11. Beausoleil, S.A., Villen, J., Gerber, S.A., Rush, J. & Gygi, S.P. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nat. Biotechnol.* **24**, 1285-1292 (2006).
12. Elias, J.E. & Gygi, S.P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4**, 207-214 (2007).
13. Bodenmiller, B. et al. PhosphoPep--a phosphoproteome resource for systems biology research in Drosophila Kc167 cells. *Mol. Syst. Biol.* **3**, 139 (2007).
14. Reiter, L. et al. Protein identification false discovery rates for very large proteomics datasets generated by tandem mass spectrometry. *Submitted* (2008).
15. Olsen, J.V. et al. Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* **127**, 635-648 (2006).
16. Lam, H. et al. Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics* **7**, 655-667 (2007).