

## PROTEIN SEARCH TUTORIAL

Users have three options to find a protein of interest: (a) search by protein identifier, (b) search by sequence (here ubiquitin), or (c) search by protein description. In addition, the search can be restricted to a single plant species by using the dropdown menu selection option below.

The image shows a web interface for protein search with several annotated components:

- Navigation Bar:** Contains links for HOME, PROTEIN SEARCH (highlighted with a dashed box), PTM SEARCH, PTM BLAST, EXPERIMENTS, PTM INFO, BROWSE/DOWNLOAD, SUBMIT, and RESOURCES.
- Section Header:** "Protein Search" with a blue question mark icon and a "Help pop-up" box.
- Search Options:** Three buttons labeled "SEARCH BY IDENTIFIER", "SEARCH BY SEQUENCE" (highlighted in blue), and "SEARCH BY DESCRIPTION".
- Annotation a) Search Protein ID:** Points to the "SEARCH BY IDENTIFIER" button. Example text: "e.g. 'AT4G05320.1' or 'Q3EAA5' (UniProtKB)".
- Annotation b) Search protein sequence:** Points to the "SEARCH BY SEQUENCE" button. Example text: "e.g. ubiquitin sequence".
- Annotation c) Search description:** Points to the "SEARCH BY DESCRIPTION" button. Example text: "e.g. 'polyubiquitin'".
- Form Fields:**
  - Sequence:** A text input field containing "MQIFVKLTGKITLEVSSDTIDNVKAKIQDKEGIPPDQQRLLF/". Below it, text reads: "String of amino acids (min 5 characters)" and "Exact matches only returned".
  - Species:** A dropdown menu with "Arabidopsis thaliana" selected.
- Optional Restriction:** A box labeled "Optional: restrict search to one species" with an arrow pointing to the Species dropdown.
- Search Button:** A red "SEARCH" button with a "Start query" annotation box pointing to it.

After pressing the 'search' button, any results will appear below the query box. All proteins fulfilling the criteria will be listed in the search results table. Note that the result table also includes protein splice forms! All columns can be sorted, including a description column, species abbreviation, cross-references protein identifiers of PLAZA or UniProtKB (requiring identical protein sequence) or the amount of PTM sites and types.

Search Results						
⬆⬆ : sort columns						
⬆⬆ ID	⬆⬆ Description	⬆⬆ Species	⬆⬆ #PTMs	⬆⬆ #PTM Types	⬆⬆ PLAZA Gene ID	⬆⬆ Uniprot ID(s)
AT4G02890.3	Ubiquitin family protein	ath	268	5	AT4G02890	A0A178V886 J7FN14 Q3E7T8 Q8H159-2
Click for protein overview						
AT4G02890.4	Ubiquitin family protein	ath 	268	5	AT4G02890	A0A178V886 J7FN14 Q3E7T8 Q8H159-2
Species protein ID (here Araport11) NOTE: splice forms of protein are included!	Protein description	Species, here: <i>Arabidopsis thaliana</i>	# PTM sites and types in protein	External IDs, cross-referenced if 100% protein identity		

By clicking a protein identifier the PTM protein sequence overview is launched, as example we show here the protein encoded by polyubiquitin 10 (AT4G05320.1). Below a general protein info header with description and cross-references, a PTM table (left, green border), PTM protein sequence overview (top-right, red border) and protein domain/site table is provided (bottom-right, blue border). These are interactively connected to each other. For instance, by default all PTM checkboxes are selected in the PTM table. Removing a specific checkbox will remove the highlighting in the protein sequence overview. Note that a color legend can be displayed and also by hovering over a modified amino acid, the modification(s) will appear in a pop-up box. Similarly, a protein domain can be selected, e.g. here all ubiquitin domains were selected, and the domain will be underlined in the PTM protein sequence overview. In the PTM table additional information is found such as the type of PTM with corresponding protein position, the originating (plain) peptide identified by MS, the respective publication and a confidence color-coding. By clicking the MS study, the experiment overview is launched. If localization probabilities or differential abundance estimates (log2 fold change and significance) are available, these are displayed as well. Log2 fold changes are displayed in a heatmap-like gradient (green is upregulated, red is downregulated). In case the significance estimate was below the threshold employed in the respective study, this is also highlighted in green (note this was not the case here). The PTM table can be exported by clicking the 'Export results' button.

PTM Table

Show confidence meta-data

CSV export

SHOW CONFIDENCE >

Sort

PTM Type	Mod AA	Pos	Peptide	Exp ID	Conf	Log2 FC	P/Q val	Loc Prob
<input checked="" type="checkbox"/>	mo	M	1	MQIFVK	5a	0.032		
<input checked="" type="checkbox"/>	ac	K	6	VKLTGK	97			
<input checked="" type="checkbox"/>	ac	K	6	MQIFVKLTGK	97			
<input checked="" type="checkbox"/>	ub	K	6	MQIFVKLTGK	100			
<input checked="" type="checkbox"/>	ub	K	11	TLTGKITLEVSSDTIDNVK	99			
<input checked="" type="checkbox"/>	ub	K	29	AKIQDKEGIPPDQQR	99			
<input checked="" type="checkbox"/>	ub	K	33	IQDKEGIPPDQQR	99			
<input checked="" type="checkbox"/>	ub	K	33	AKIQDKEGIPPDQQR	100			
<input checked="" type="checkbox"/>	ac	K	48	LIFAGKQLEDGR	94a	-0.206	0.240	1.000
<input checked="" type="checkbox"/>	ub	K	48	LIFAGKQLEDGR	3			1.000

Checked: display PTM on overview

Experiment ID #  
Click: experiment details  
Hover: title experiment

Peptide matches proteins encoded multiple gene loci  
On hover: display genes

Confidence estimate:  
High  
Medium  
Low  
Details: see further

EXPORT RESULTS

Sequence

Length: 381

Highlighted PTMs

SHOW PTM COLOR LEGEND

Checked: remove highlighting

Domains & Sites

Clear highlighted range

Interpro Domains

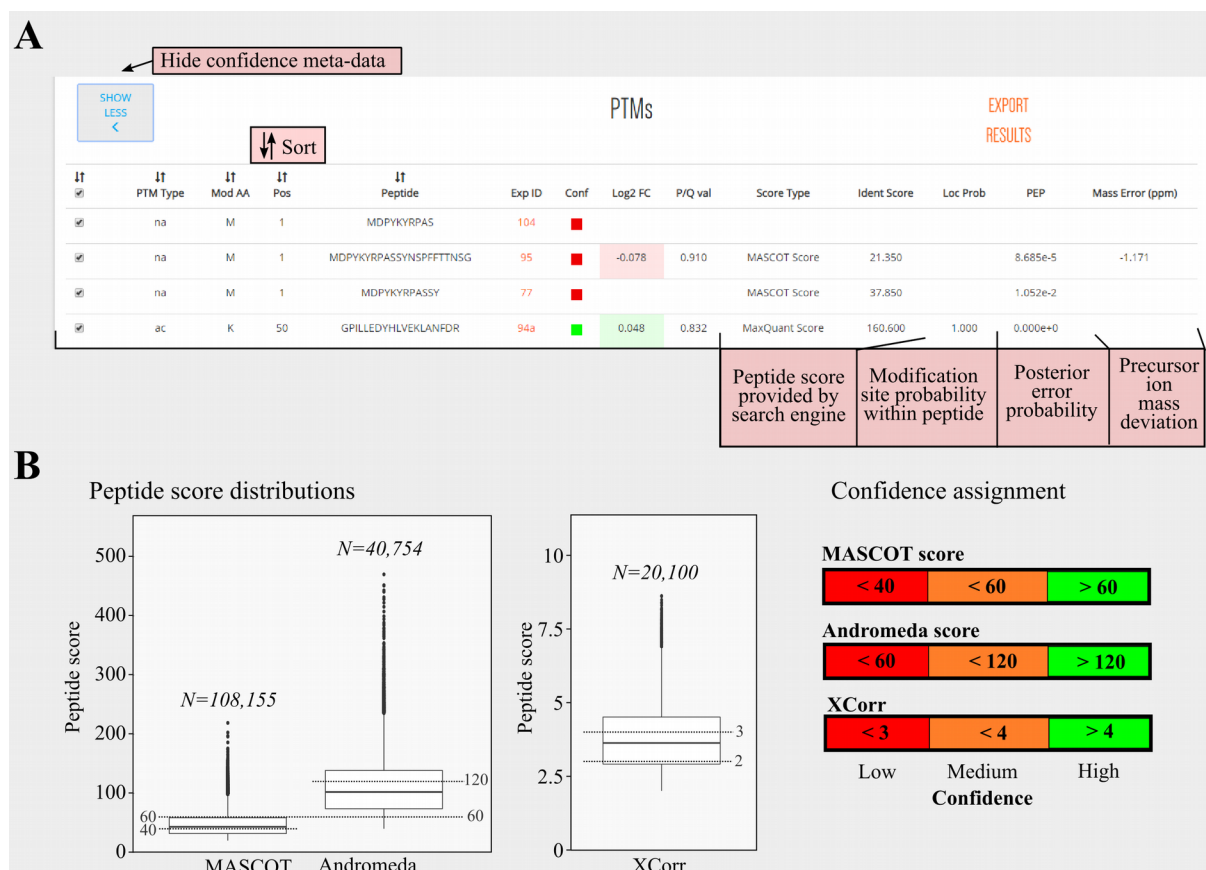
Show	IPR ID	Description	From	To
<input checked="" type="checkbox"/>	IPR000626	Ubiquitin domain	1	76
<input checked="" type="checkbox"/>			77	152
<input checked="" type="checkbox"/>			153	228
<input checked="" type="checkbox"/>			229	304
<input checked="" type="checkbox"/>			305	380

Protein domain or UniProtKB annotated sites

Protein overview: PTM color legend

ID	PTM Type	Color
mo	Methionine Oxidation	
ac	Acetylation	
ub	Ubiquitination	
nt	N-terminal	
ph	Phosphorylation	
	Multiple types	

Details of the confidence meta-data collected can be consulted by clicking 'SHOW CONFIDENCE'. Below, we can view the extended version (figure panel A) including these confidence estimates reported by experiments, including peptide scores, posterior error probability (PEP), modification site localization probability and/or precursor mass deviation. Peptide scores are measured by search engines and score how a tandem mass spectrum matches a peptide from the searched protein database. Most frequently reported scores (used search algorithms) are the MASCOT ion score (MASCOT, Perkins et al., 1999), the Andromeda score (built-in MaxQuant software suite, Cox et al. 2011) and the cross-correlation score (XCorr, originally for SEQUEST, Eng et al., 1994). Distributions of these scores can be consulted in the figure panel B below. For these three search engines minimal peptide score thresholds were used. MASCOT ion scores were required to be at least 20, Andromeda scores 40 and XCorr scores at least 2. Next to peptide scores, which are highly differing and dependent on the search algorithm used, the PEP provides a more unified confidence estimate and can be considered as a "local FDR" that expresses the chance that a given peptide-to-spectrum match was incorrect. Most PEP values reported here were measured by software such as MaxQuant (Cox and Mann 2008), Proteome Discoverer (Thermo Scientific) or post-processing algorithms such as Percolator (Käll et al., 2007). Lastly, beside peptide-level confidence measurements, modification localization probability within a peptide can be assessed by algorithms such as PhosphoRS (Taus et al., 2011) or the PTM Score implemented in MaxQuant (Olsen et al., 2006). Here, we required a modification site localization probability of at least 0.75, when reported. Based on the peptide scores provided, PTMs are categorized as being low, medium or high confident (figure panel B - right). Assessing reliability of PTMs is a crucial step as false positive identification may occur in mass spectrometry identification results. In this aspect, careful inspection of experimental details remains therefore advisable.



## REFERENCES

- Cox, J., Neuhauser N., Michalski A., Scheltema R.A., Olsen J.V. and Mann M.** (2011) Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794-1805.
- 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. *J. Am. Soc. Mass Spectrom.* **5**, 976-989.
- Käll, L., Canterbury J.D., Weston J., Noble W.S. and MacCoss M.J.** (2007). Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **4**, 923-925.
- Olsen, J.V., Blagoev, B., Gnad, F., Macek, B., Kumar, C., Mortensen, P. and Mann, M.** (2006) Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* **127**, 635-648.
- 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.
- Taus, T., Köcher, T., Pichler, P., Paschke, C., Schmidt, A., Henrich, C. and Mechtler, K.** (2011) Universal and confident phosphorylation site localization using phosphoRS. *J. Proteome Res.* **10**, 5354-5362.