Medicago PhosphoProtein Database: a repository for Medicago truncatula phosphoprotein data

Christopher M. Rose1, Muthusubramanian Venkateshwaran2, Paul A. Grimsrud3, Michael S. Westphall4, Michael R. Sussman3, Joshua J. Coon1,4 and Jean-Michel Ané2*

1 Department of Chemistry, University of Wisconsin, Madison, WI, USA
2 Department of Agronomy, University of Wisconsin, Madison, WI, USA
3 Department of Biochemistry, University of Wisconsin, Madison, WI, USA
4 Department of Biomedical Chemistry, University of Wisconsin, Madison, WI, USA

Edited by: Joshua J. Coon, Lawrence Berkeley National Laboratory, USA
Reviewed by: Dong Xu, University of Missouri, USA
Stefanie Wienkoop, University of Vienna, Austria

*Correspondence:
Jean-Michel Ané, Department of Agronomy, University of Wisconsin, 348 Moore Hall, 1575 Linden Drive, Madison, WI 53706, USA. e-mail: jane@wisc.edu

INTRODUCTION

Legumes, also known as Fabaceae, are a very large and economically important group of plants (Graham and Vance, 2003). They are used as food crops, forages, and green manure throughout the world. Most legumes can develop a nitrogen-fixing association with soil bacteria known as rhizobia, which results in the formation of root nodules (Jones et al., 2007). Rhizobia thrive inside these nodules and fix atmospheric nitrogen in exchange for a carbon source (Venkateshwaran and Ané, 2011). Many legumes also establish symbiotic associations with arbuscular mycorrhizal fungi which facilitate the acquisition of nutrients (phosphorous, nitrogen, etc.) and provide some protection against environmental stresses (Ruiz-Lozano et al., 1995; Schutzendubel and Polle, 2002; Bonfante and Genre, 2010).

Medicago truncatula (Medicago) is a well-established model for studying legume biology and, in particular, the molecular mechanisms mediating symbiotic associations (Cook, 1999). The Medicago research community has developed many genetic, proteomic, and genomic tools, including the recent release of its genome sequence (Huber, 2007).

Medicago truncatula (Medicago) is the model system for studying legume biology, making the study of its phosphoproteome essential. Here, we describe the Medicago Phosphoprotein Database (MPPD, http://phospho.medicago.wisc.edu), a repository built to house phosphoprotein, phosphopeptide, and phosphosite data specific to Medicago. Currently, the MPPD holds 3,457 unique phosphopeptides that contain 3,404 non-redundant sites of phosphorylation on 829 proteins. Through the web-based interface, users are allowed to browse identified proteins or search for proteins of interest. Furthermore, we allow users to conduct BLAST searches of the database using both peptide sequences and phosphorylation motifs as queries. The data contained within the database are available for download to be investigated at the user’s discretion. The MPPD will be updated continually with novel phosphoprotein and phosphopeptide identifications, with the intent of constructing an unparalleled compendium of large-scale Medicago phosphorylation data.

The ability of legume crops to fix atmospheric nitrogen via a symbiotic association with soil rhizobia makes them an essential component of many agricultural systems. Initiation of this symbiosis requires protein phosphorylation-mediated signaling in response to rhizobial signals named Nod factors. Medicago truncatula (Medicago) is the model system for studying legume biology, making the study of its phosphoproteome essential. Here, we describe the Medicago Phosphoprotein Database (MPPD, http://phospho.medicago.wisc.edu), a repository built to house phosphoprotein, phosphopeptide, and phosphosite data specific to Medicago. Currently, the MPPD holds 3,457 unique phosphopeptides that contain 3,404 non-redundant sites of phosphorylation on 829 proteins. Through the web-based interface, users are allowed to browse identified proteins or search for proteins of interest. Furthermore, we allow users to conduct BLAST searches of the database using both peptide sequences and phosphorylation motifs as queries. The data contained within the database are available for download to be investigated at the user’s discretion. The MPPD will be updated continually with novel phosphoprotein and phosphopeptide identifications, with the intent of constructing an unparalleled compendium of large-scale Medicago phosphorylation data.

Keywords: Medicago truncatula, phosphoproteome, ETD, CAD, proteomics

www.frontiersin.org
June 2012 | Volume 3 | Article 122 | 1
has not been studied extensively in vivo until we published a large-scale phosphoproteomic study in 2010 (Grimsrud et al., 2010). This work identified phosphorylation sites on proteins isolated from Medicago roots (from whole cell lysate and membrane-enriched fractions) using immobilized metal affinity chromatography (IMAC) and tandem mass spectrometry. The data collected from this large-scale study were used to create an online Medicago PhosphoProtein Database (MPPD; http://www.phospho.medicago.wisc.edu). The salient features of MPPD are discussed below.

**MASS SPECTROMETRY ANALYSIS OF THE MEDICAGO PHOSPHOPROTEOME**

The MPPD was populated using the phosphoproteomic workflow presented in Figure 1. Proteins were isolated from Medicago plant tissue and digested with trypsin, Lys-C, Glu-C, Arg-C, or Asp-N. To reduce sample complexity, resulting peptides were then fractionated by strong cation exchange chromatography. Phosphopeptides were then enriched by IMAC and analyzed using an electron transfer dissociation (ETD)-enabled LTQ Orbitrap mass spectrometer (Thermo-Fisher). To increase proteome coverage, both collisionally activated dissociation (CAD) and ETD (Syka et al., 2004) were used for peptide fragmentation. Spectra were searched against a Medicago protein database using the Open Mass Spectrometry Search Algorithm (OMSSA; Geer et al., 2004). Identifications were then filtered to 1% FDR at both the peptide and protein level. This analysis produced 3,457 unique phosphopeptides, 829 unique proteins, and 3,404 non-redundant sites of phosphorylation (Grimsrud et al., 2010). The entirety of this data is contained within the MPPD and is freely available for download.

**OVERVIEW OF THE MEDICAGO PHOSPHOPROTEIN DATABASE**

The MPPD (http://phospho.medicago.wisc.edu) is a web-based resource that allows users to search for a particular protein of interest, BLAST (Basic Local Alignment Search Tool; Altschul et al., 1990) a protein sequence, browse the entire phosphoproteomic

---

**FIGURE 1** | Proteomic workflow used to populate the Medicago Phosphoproteome Database. Proteins were isolated from Medicago root tissue and digested with trypsin, Lys-C, Glu-C, Arg-C, or Asp-N to create peptides. These peptides were then fractionated by strong cation exchange (SCX), enriched for phosphopeptides by immobilized metal affinity chromatography (IMAC), and sampled using nHPLC-MSMS. Peptides were fragmented via electron transfer dissociation (ETD) or collisionally activated dissociation (CAD). The results of this workflow are displayed in the table labeled phosphopeptide results. Modified from Grimsrud et al. (2010).
The homepage also provides detailed instructions for searching the database as well as an explanation of the various options for handling the data (i.e., search, browse, download). Here, we discuss each of these features and provide a guide to navigating the MPPD.

**SEARCHING THE MEDICAGO PHOSPHOPROTEIN DATABASE**

The MPPD allows users to search the database using three different criteria: (1) protein description, (2) BLAST protein sequence, and (3) peptide sequence/phosphorylation motif. For all search types multiple queries can be executed at one time. The search for protein description allows users to query the database using gene accession number, protein name, or any word contained in the protein description. Figure 2B displays the search results resulting from a search for Interacting Protein of DMI3 (IPD3). IPD3 interacts with and is phosphorylated by DMI3 and is a key regulator of legume–rhizobia symbiotic signaling (Messinese et al., 2007; Horvath et al., 2011). MPPD serves as a repository reporting of the number of protein identifications, proteins with shared peptides are grouped together in protein groups. The light blue bar in Figure 2A contains information regarding the protein group for IPD3. In this case, there was one protein group that contains two separate protein entries from the Medicago database. This group contained six peptides and received a p-score of 2.94e-33. Briefly, the p-score is calculated by multiplying the p-values of all peptides contained within the protein group and is used to conduct false discovery rate analysis at the protein level. In addition, the complete protein sequence is displayed with identified sites of phosphorylation appearing as lower case letters colored green or red with the colors representing localized and unlocalized sites of phosphorylation, respectively. Amino acids highlighted in green (localized sites of phosphorylation) indicate that the mass spectrometric data provided clear evidence for phosphorylation of the given residue. Amino acids highlighted in red indicate that mass spectrometric analysis had determined the presence of phosphorylation, but the site of phosphorylation could not be assigned to a specific residue with high confidence. In the case of unlocalized phosphorylation sites, all possible sites are highlighted in red.

Selecting an amino acid, such as s161, will display the characteristics of the spectra used to identify the given phosphorylation site (Figure 2C). This information includes the e-value and p-value provided by OMSSA, the precursor charge state, precursor theoretical mass, precursor experimental mass, precursor mass error, number of matching fragments, number of possible fragments, whether all phosphosites were localized, the dissociation method, the phosphopeptide sequence, and the reference in which the phosphosite was described.
rough characterization of the spectral quality as a lower e-value indicates the given MS/MS spectrum is likely associated with the phosphopeptide in question. The precursor mass error represents the deviation of the experimental mass from the theoretical mass. High accuracy mass analyzers (e.g., Orbitrap, FT-ICR, and TOF) enable greater specificity when searching protein databases and generally a precursor mass error less than 10 ppm will increase the number of peptide identifications (McAlister et al., 2010). The dissociation method is important to note, as certain peptides are more amenable to either collisional or electron-based fragmentation techniques (Jouany et al., 2008). To repeat the identification of phosphopeptides contained within the MPPD, it is important that the same fragmentation conditions be used. All possible phosphopeptide sequences are also displayed on this page. If the phosphosite was localized, only one sequence will appear per line with the modified residue appearing in lowercase. In the case of unlocalized phosphosites, all possible phosphopeptides will be displayed with each possible phosphopeptide appearing in lowercase.

Assessing sequence homology across species enables researchers to infer function of unknown proteins and make connections to research published on similar proteins in different systems. The MPPD allows users to conduct a BLAST search to query homology of proteins within the Medicago protein database with a user-provided protein sequence. This search returns all of the same protein attributes discussed above, but it adds a link to a flat file containing all alignments above the user-provided threshold. This flat file contains a list of the proteins matched, e-value score of each match, and a visual representation of the alignment. Phosphorylation motifs can help to elucidate potential kinases which alter the phosphorylation state of a particular amino acid (Schwartz and Oggi, 2005). The MPPD allows users to query the database by entering a short amino acid sequence or phosphorylation motif (e.g., Ser/Thr). Here, “x” is used as a single wild card amino acid while “*” is used for multiple wild card amino acids. This search is also case-sensitive, as lowercase letters signify a phosphorylated residue. The result of this search is the same for a protein description search (Figure 2B), but the results contain proteins that have the specified amino acid sequence or phosphorylation motif. This feature enables users who are interested in a particular kinase to determine if this protein kinase contains the motif of interest and if a phosphosite at this location was identified in our mass spectrometry analysis.

**BROWSING AND DOWNLOADING DATA**

The MPPD allows users to browse all protein groups contained within the database by selecting the "Browse" option in the tools menu. As described above, protein grouping occurs when multiple entries within the protein database share an identified peptide. To explain peptide identifications with the fewest number of proteins, these protein identifications are placed into one group. When the user selects the browse task, an online table appears, listing each identified protein group, including the associated protein number, number of proteins in the group, number of peptide identifications in the protein groups, the p-score for each protein, and the protein description of the longest protein in the protein group. As discussed above, the p-score is calculated by multiplying the OMSSA p-value for each peptide in the group and is used to calculate the false discovery rate for the given protein. Users can browse protein entries by changing pages, but to access a protein entry, users must copy the protein description and use the search tool.

To download text files containing all protein group, protein, and peptide identification data, users can select the download option on the tools menu. To download data right click on "Group Information," "Protein Information," or "Peptide Information" and select "Save As." Headers associated with each file are listed on the download page, as the text files do not contain headers.

**CONCLUSION**

As large-scale phosphoproteomic analysis of plant tissues continues to become more prevalent, tools are needed to enable facile access to data (Jayswan et al., 2012). Numerous databases for proteomic information in plants exist, including databases focused on sub-cellular fractions (AMPP, Kruft et al., 2001; AraFeron, Reumann et al., 2004; AtNoPDB, Brown et al., 2005; SUBA, Headwood et al., 2007; PIProT, Kieflmann et al., 2006; AT-Chloro, Ferro et al., 2010), single species (SpruceDB, Lipper et al., 2009; Soybean Proteome Database, Sakata et al., 2009; PhosPhAT, Dureck et al., 2010), multiple species (P3DB, Gao et al., 2009; PPDB, Sun et al., 2009), 2-D gel mapping (GelMap, Rode et al., 2011), and spectral data from mass spectrometry experiments (ProMex, Himmel et al., 2007). Here we highlight a publicly available web portal for Medicago phosphoproteins and phosphopeptides. This online database enables researchers to search for proteins of interest, BLAST for homologous proteins, search for phosphorylation motifs, browse, and download the data. This central repository contains all of the information for researchers to connect phosphosites identified in Medicago to other legumes, providing an invaluable resource for future studies pertaining to legume biology and, particularly, to the legume–rhizobia symbiosis. In addition to the data presented here, the Wisconsin Medicago Group (University of Wisconsin, Madison) has been very active in continuing phosphoproteomic characterization of Medicago in response to symbiotic signals. In particular, we are pursuing quantitative measurements of phosphorylation dynamics within Medicago in response to symbiotic stimuli. As we report our results, we will continue to build database tools and enable researchers to connect with our results. These tools will include the ability to query quantitative information of phosphorylation state alterations, allowing researchers to determine if proteins of interest are involved in a symbiotic cascade. In addition, future database applications will offer the opportunity for other researchers to upload their own data, creating a large, centralized source for all phosphoproteomic data relating to legumes.

**ACKNOWLEDGMENTS**

This work was supported by a grant from the National Science Foundation (NSF#0701846) to Michael R. Sussman, Joshua J. Coon, and Jean-Michel Ané. Christopher M. Rose was funded by an NSF Graduate Research Fellowship and NIH Traineeship (T32GM008305).
The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.