DEVELOPMENT OF A PLANT PROTEIN PHOSPHORYLATION DATABASE AND A WEB-BASED PROTEIN PHOSPHORYLATION PREDICTION TOOL

A Thesis

Presented to

The Faculty of the Graduate School

At the University of Missouri

In Partial Fulfillment

Of the Requirements for the Degree

Master of Science

By

CURTIS BOLLINGER

Dr. Dong Xu, Advisor

MAY 2012

The undersigned, appointed by the dean of the Graduate School, have examined the thesis entitled

DEVELOPMENT OF A PLANT PROTEIN PHOSPHORYLATION DATABASE AND A WEB-BASED PROTEIN PHOSPHORYLATION PREDICTION TOOL

Presented by Curtis Bollinger

A candidate for the degree of

Master of Science

And hereby certify that, in their opinion, it is worthy of acceptance.

Dr. Dong Xu

Dr. Yi Shang

Dr. Scott Peck

ACKNOWLEDGEMENTS

First I would like to thank my advisor Dr. Dong Xu for his support and guidance throughout my undergraduate and graduate studies at the University of Missouri. I would like to thank the members of the Digital Biology Laboratory. In particular, I would like to thank Qiuming Yao, Jianjiong Gao, and Shiyuan “Cynthia” Chen who I worked with on my research. I would also like thank Dr. Yi Shang and Dr. Scott Peck for agreeing to review my master thesis. Dr. Yi Shang has also been an excellent professor and very supportive of me.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

In this project, two efficient and intuitive user-interfaces were designed and developed to be highly accessible and versatile for bioinformatics researchers. Two web based user interfaces were designed for two bioinformatics tools, Plant Protein Phosphorylation Database (P^3DB) and Musite.net. These tools were developed primarily for protein phosphorylation research.

P^3DB is a web interface to a database containing only information on plant protein phosphorylation. The web interface of P^3DB was redesigned from a previous version to provide a significantly better user experience and additional features. This was done using dynamic web coding techniques and a more modular template file structure.

Musite.net is a web-based version of a protein phosphorylation prediction tool and has been expanded to allow predictions of other posttranslational modifications. Musite.net implements the main features of its desktop counterpart as well as some unique features. Musite.net is extremely dynamic in order to function similarly to a desktop application. Musite.net was made very accessible by implementing uniform resource locator (URL) submissions and an application programming interface (API).

Dynamic web coding allowed several user tasks to be simplified. Additionally, some features were implemented that would have otherwise been impossible with a static webpage. A template file structure allows future maintenance and development to be simplified compared to a less modular system.

1. INTRODUCTION

Proteins perform several essential functions in living cells. Protein posttranslational modification (PTM) is often necessary for a protein to become fully functional. PTM is important for many cellular processes such as signaling, cellular differentiation, protein degradation, protein stability, gene expression regulation, protein function regulation, and protein interactions [1].

Two bioinformatics tools were developed for protein PTM research. P^3DB is a bioinformatics resource specifically for plant protein phosphorylation research data. Protein phosphorylation is the most common type of PTM that affects protein degradation, cell cycle, and growth. Musite.net is a bioinformatics tool for protein PTM site prediction with options for general and kinase specific phosphorylation site prediction.

P^3DB is a web interface to a database that consolidates all plant protein phosphorylation related information into one location. There are other protein phosphorylation databases, but none that specifically focus on plants (see table 1.1). Plant phosphorylation data can also be viewed from multiple levels including the entire protein sequence, a single phosphorylation site, or a peptide. All the data for P^3DB is stored locally on the server and the database is actively updated with new data from credible research studies. The web interface of P^3DB has been completely rewritten with improvements over the original version. The new user interface is more interactive and easier to use because of dynamic features.

Database	Website	Species	PTM Type
UniProt/Swiss-Prot [7]	http://uniprot.org/	All species	All
dbPTM [8]	http://dbptm.mbc.nctu.edu.tw/	All species	All
PHOSIDA [9, 10]	http://www.phosida.com/	Multiple species	Multiple
P3DB [11]	http://p3db.org/	Plants	Phosphorylation
PhsophoPep 12, 13]	http://www.phosphopep.org/	Yeast, Worm, Fly, Human	Phosphorylation
Phosphomouse [14]	https://gygi.med.harvard.edu/phosphomouse/	Mouse	Phosphorylation

Table 1.1: A sampling of protein PTM databases.

Musite.net is a web server version of the standalone Musite application. Musite is a protein phosphorylation prediction tool. Musite.net uses the same prediction algorithm as the standalone version, but on a webserver platform. Musite.net implements an intuitive user interface similar to the standalone version to provide continuity when users switch between the different versions. Musite.net can be accessed from any machine with a web browser and internet making it very accessible. An application programming interface (API) and uniform resource locator (URL) submission allow other bioinformatics web resources to easily interact with Musite.net. The interface of Musite.net has several advantages over similar prediction tools. Multiple predictions can be performed at once and results can be compared easily using the dynamic and interactive interface. A specificity slider allows the specificity threshold to be dynamically changed while many other tools have a fixed threshold. The results are also quickly generated instead of being emailed to the user.

Tool Name	Website	PTM Type
Musite [3, 4]	http://musite.sf.net and http://musite.net	Multiple
PROSITE [15, 16]	http://ca.expasy.org/prosite/	Multiple
ScanSite [17]	http://scansite.mit.edu/	Multiple
DISPHOS [18]	http://www.dabi.temple.edu/disphos/	Phosphorylation
NetPhos [19,20]	http://www.cbs.dtu.dk/services/NetPhos/	Phosphorylation
KinasePhos [21]	http://kinasephos2.mbc.nctu.edu.tw/	Phosphorylation

Table 1.2: A sampling of protein PTM prediction tools.

P^3DB and Musite.net complement each other in the type of data they provide. P^3DB contains high quality data from experimental studies. This data provides Musite.net with high quality training data for plants. Musite.net has limits on the accuracy of predictions, but still provides a more general view of where phosphorylation sites may be in a protein. P^3DB uses the URL submission feature of Musite.net. This allows researchers to submit a protein from P^3DB to Musite.net giving them an additional perspective on the data. PTM prediction gives a more general view of where PTM sites may be in a protein.

P^3DB and Musite.net both make use of a template file structure for organizing the back-end code. This allows for simplified development and updates. Both also implement several dynamic web coding features which simplify tasks for users. P^3DB and Musite.net have been tested on all major browsers and perform well.

2. PLANT PROTEIN PHOSPHORYLATION DATABASE

The goal of Plant Protein Phosphorylation Database (P^3DB) is to provide everything a researcher needs in one place on plant protein phosphorylation. P^3DB maintains a database of only the highest quality data. This is useful for many types of researchers including computational biologists, and bioinformaticians. Various levels of information on plant protein phosphorylation are provided throughout the pages of P^3DB. The main pages are Home, Search, Browse, Data Sets, Download, and Contact.

2.1 Home Page

The home page of P^3DB is designed to make the purpose of P^3DB clear and provide easy and intuitive access to content for new and returning users. Like every page of P^3DB, a menu bar at the top of the page allows access to the most important pages of the site at any time. A quick-search function just below the main menu bar allows users to quickly make a database query from the home page. The quick-search performs a simple keyword search on the description and organism of all proteins in the database and the results are displayed on the browse page. A more advanced search can be performed from the search page. The first content presented on the home page is all the major statistics related to the database. These statistics give a basic overview of what the database of P^3DB contains. More detailed information on the database can be found on the data sets page. A large news panel in the center of the home page shows the most

recent information and updates for P^{3}DB. Previous news and updates can be viewed by clicking the arrow on the edge of the panel.

The home page introduces users to the intuitive interface of P^{3}DB which is consistent throughout the entire site. Anything that is dynamic and interactive is colored in a brighter green with white text. This lets users know what they can interact with just by viewing the page. All web-links are colored in a reddish-brown color which contrasts the green background so they stand out.

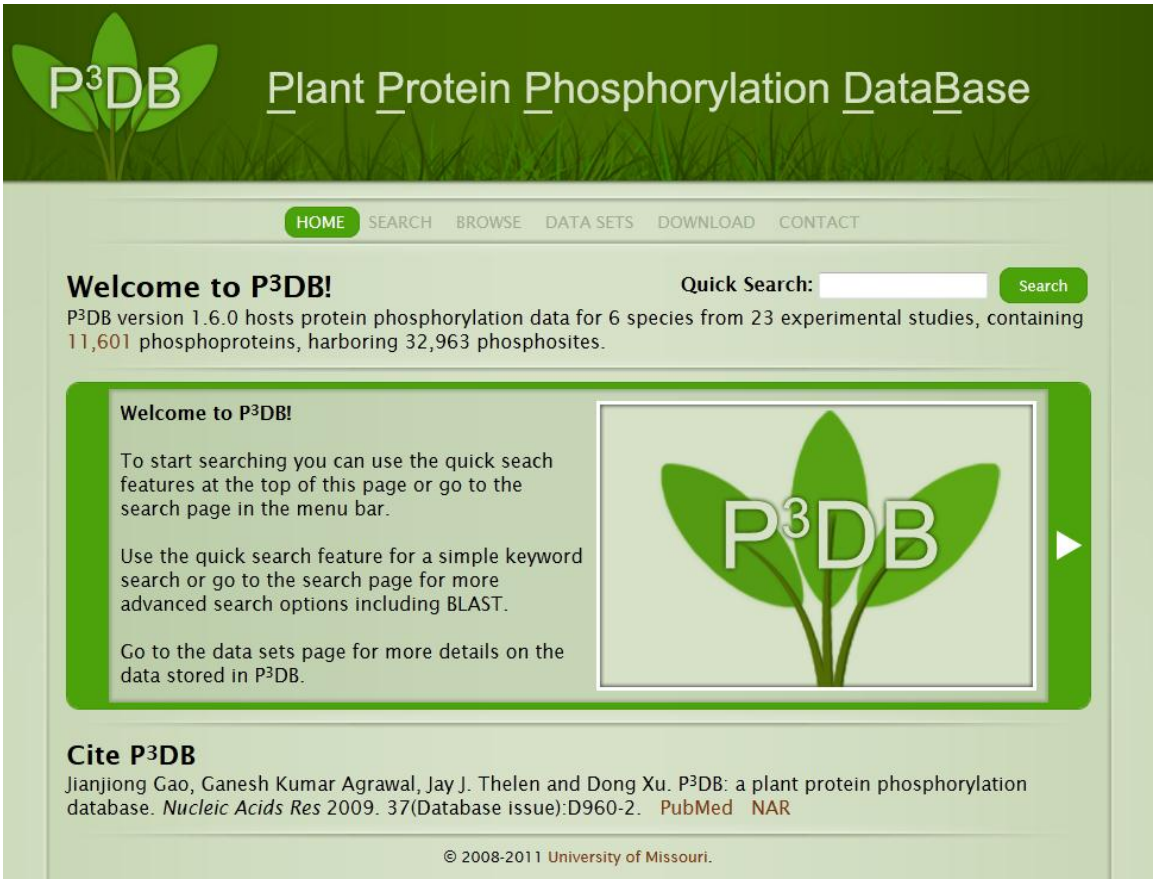

Figure 2.1: Appearance of P^{3}DB from the home page.

2.2 Search Page

The search page provides several methods for filtering the database of P^3DB. Users are able to search by exact match or use Basic Local Alignment Search Tool (BLAST). A BLAST finds sequences that are similar to a given sequence with a similarity above a specified threshold (e-value). Searches and BLASTs provide different levels of filtering the database for proteins or peptides based on what a user is interested in and the search criteria they have available.

A protein queries the database for exact matches with given search criteria. A protein search can be performed using search criteria for organism type, research group or reference, accession number, and description. An organism type can be selected from a list of all plant species stored in the database. A research group or reference can be chosen from a list of sources that provided data for P^3DB. An accession number may be given as either a partial or complete number. The database currently contains more than thirty-five different types of accession numbers for the convenience of researchers. Most accession numbers begin with a standardized set of characters signifying the source of the accession number. Searching for these standardized set of characters as a partial accession number could be useful for users interested in results that contain a specific type of accession number. The description search criteria can be given as either a partial or complete description text. Submitting a protein search redirects to the browse page where a table of results is displayed.

A peptide search is similar to a protein search with slightly different search criteria. Peptide searches use search criteria for organism type, reference, and an amino acid sequence. The amino acid sequence is required when performing a peptide search. The sequence should be an exact subsequence or a complete match of peptides stored in the database.

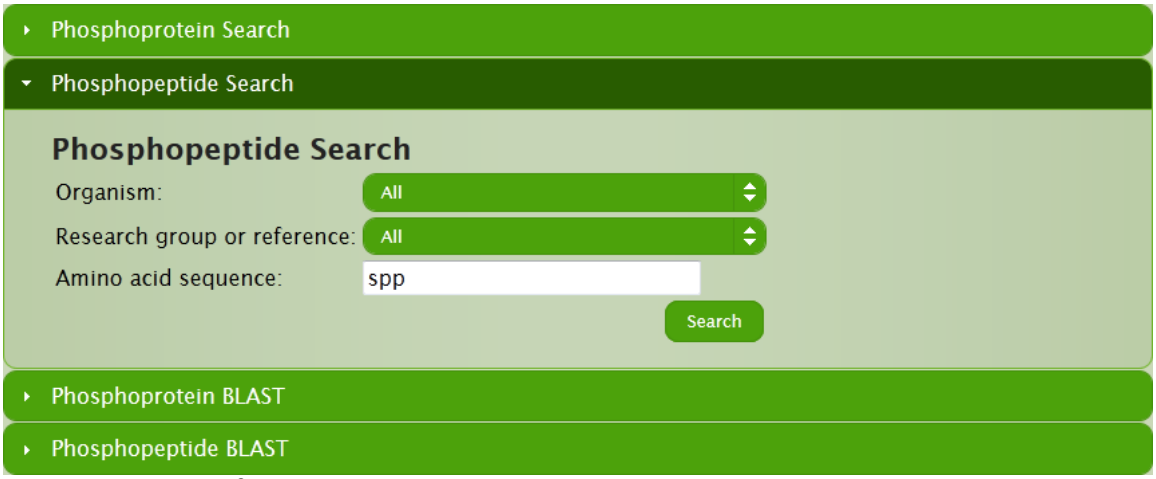

Figure 2.2: P^{3}DB peptide search with a queried amino acid sequence of "spp".

A protein or peptide BLAST uses search criteria for organism type, research group or reference, amino acid sequence, and e-value threshold. The amino acid sequence is required and must be in FASTA format. Sequences can be entered with or without a header line, line numbers, or spaces. The sequence can be a complete protein sequence, partial sequence, or peptide. By default, a phosphoprotein BLAST e-value threshold is set to 0.1 and a phosphopeptide BLAST e-value threshold is set to 100. Predefined options are given for the e-value above and below the default threshold. Results are based on sequence similarity and are initially sorted by best match. Results can also be organized by organism.

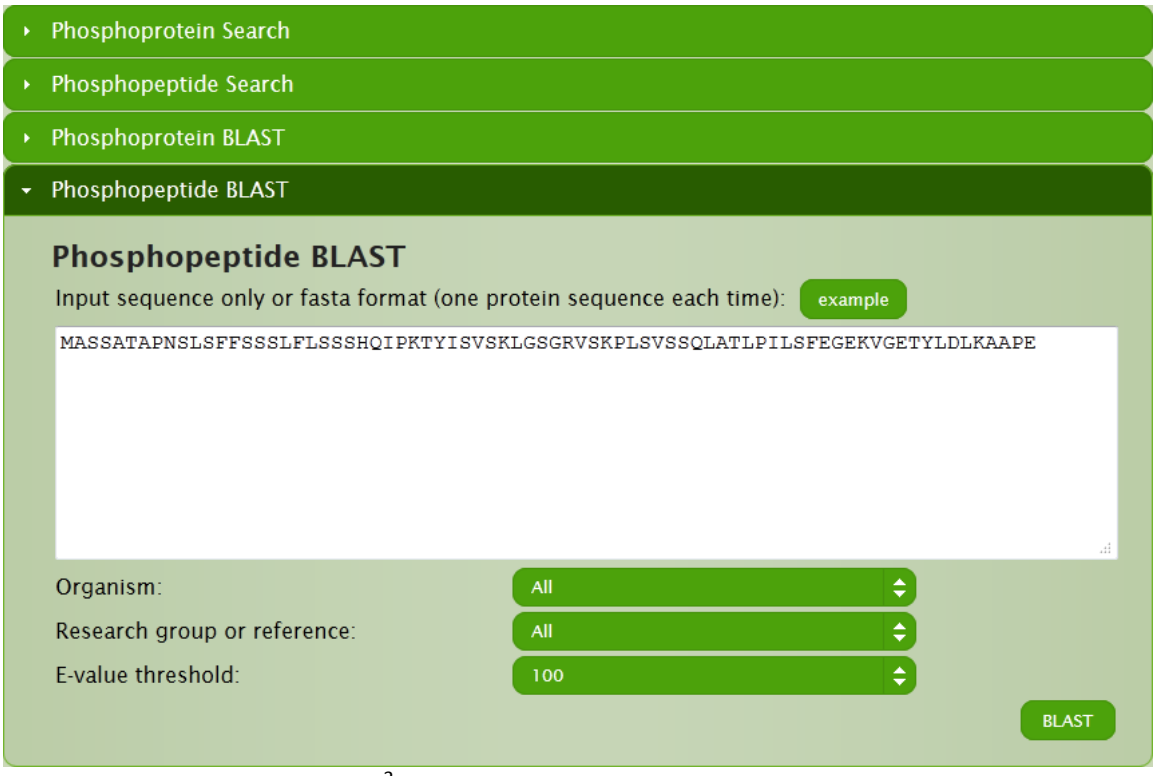

Figure 2.3: P[3]DB peptide BLAST with a query sequence.

The content of the page utilizes a jQuery-UI accordion widget. jQuery-UI is a JavaScript library with several useful user interface features. The accordion widget allows content to be contained in tabs that can be expanded or collapsed. Each search or BLAST function is contained in an accordion tab (see figure 2.2). The accordion widget utilizes the accordion tabs to display all the search or BLAST functions in a clear and compact way. This is better than a single static page because the content is not displayed all at once which would make the page very long and hard to navigate. This is also better than multiple static pages that need to load, and cause a delay, when changing pages. Changing functions in the accordion widget is dynamic so the page does not need to be reloaded. Future search functions can easily be implemented as an additional accordion tab.

All search and BLAST functions are processed on the server and the results are displayed on the browse page. Search functions simply query the database using the search criteria defined by the user and return the results to the browse page. BLAST functions appear to be the same as search functions to the user, but an extra step is involved. The server sends the search criteria to a BLAST application which returns the BLAST results. The results are then displayed on the browse page.

2.3 Browse Page

The browse page displays results from database queries. If the user goes directly to the browse page, all proteins in the database are displayed. This is equivalent to performing a protein search from the search page with no search criteria. The results table is intended to give users enough information so they are able to choose the individual proteins or peptides they are interested in. More detailed information can be found on a specific protein or peptide by clicking the associated details button. All accession links in the protein description column of the results table go to the associated protein page. Links to the source of the accession can be found on the associated protein, peptide, or phosphorylation site page.

All search and BLAST results contain a protein description and organism type column. Results can be sorted by any column. For peptide search results, a peptide column shows exact subsequence matches. The exact match of the queried sequence is underlined and any phosphorylation sites are highlighted with red text (see figure 2.4).

Show 10 entries Search:

Peptide	Peptide	Protein Description	Organism
Details	DLAGGGDSSPYSNQEDISSPQRELPETQSPPLK	DFCI GI:TC69789,	*Brassica napus*
Details	DLAGGGDSSPYSNQEDISSPQRELPETQSPPLK	DFCI GI:TC69789,	*Brassica napus*
Details	DLAGGGDSSPYSNQEDISSPQRELPETQSPPLK	DFCI GI:TC69789,	*Brassica napus*
Details	DLAGGGDSSPYSNQEDISSPQRELPETQSPPLK	DFCI GI:TC69789,	*Brassica napus*
Details	DLAGGGDSSPYSNQEDISSPQRELPETQSPPLK	DFCI GI:TC69789,	*Brassica napus*
Details	DLAGGGDSSPYSNQEDISSPQRELPETQSPPLK	DFCI GI:TC69789,	*Brassica napus*
Details	DLAGGGDSSPYSNQEDISSPQRELPETQSPPLK	DFCI GI:TC69789,	*Brassica napus*
Details	DLAGGGDSSPYSNQEDISSPQRELPETQSPPLK	DFCI GI:TC69789,	*Brassica napus*
Details	ELPETQSPPLK	DFCI GI:TC69789,	*Brassica napus*
Details	RNSPAYYSPPR	Glyma:Glyma02g41510.2,	*Glycine max*

Showing 1 to 10 of 1,445 entries (filtered from 245,542 total entries) First Previous 1 2 3 4 5 Next Last

Figure 2.4: P^{3}DB browse page result table after a peptide search with a queried amino acid sequence of “spp”.

All BLAST results display an e-value column. For peptide BLAST results, a sequence alignment column shows the similarity between a matched portion of the queried sequence and peptides in the database (see figure 2.5). The first line shows the matched portion of the queried sequence. The third line shows the matched peptide from the database. The middle line shows how the two sequences compare to each other. An amino acid denotes an exact match. A “+” denotes a close match, but not exact. A space denotes no match. All phosphorylation sites are highlighted with a red background and white text meaning they can be clicked on. Clicking on a phosphorylation site leads to the associated phosphorylation site page.

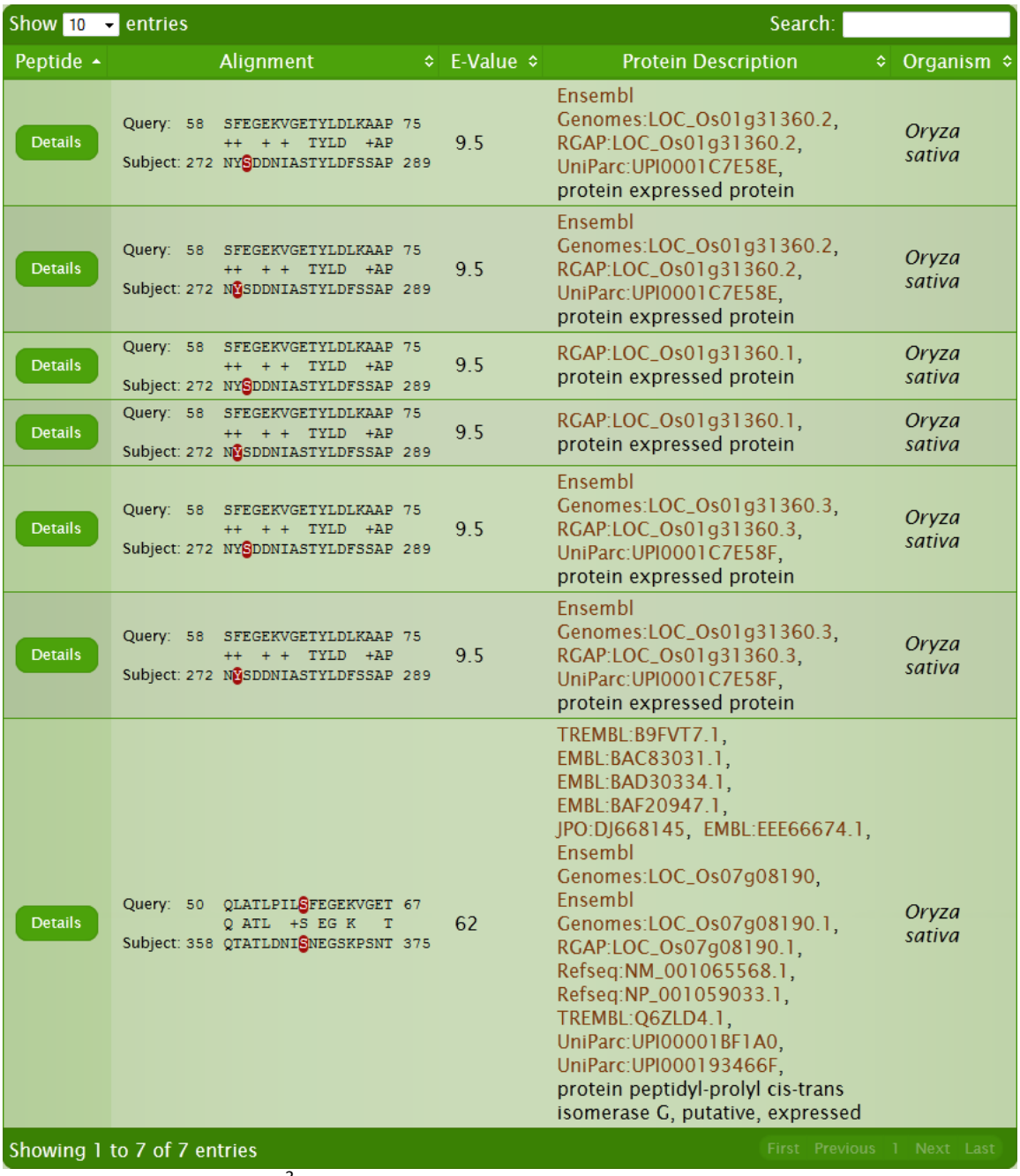

Figure 2.5: P[3]DB browse page result table after a peptide BLAST.

The database for P[3]DB contains over 10,000 proteins and over 30,000 phosphorylation sites, so loading every search result at once would be slow and impractical. Users are likely to never view many of the results from any particular query, so it is not necessary to retrieve them until explicitly requested. At the same time, it is

convenient for the user to dynamically search, sort, and change the number of displayed results. To compromise between these two conflicting conditions, the result table is populated using asynchronous JavaScript and XML (Ajax). Ajax allows content to be loaded from the server without the page reloading. An Ajax call to the server is made each time a dynamic function is used. The server returns the necessary data and the table is repopulated dynamically. This provides all the benefits of dynamic functions with short loading times.

The BLAST function does not use Ajax to load data. Instead, all BLAST results are returned at once and not stored. Time would only be wasted if each result page was loaded separately because all the results would be generated for each request. This makes the loading of BLAST results slower than search results because more data is loaded at one time. The loading time is typically less than ten seconds with the current hardware and at the current database size. Because all the result data is already loaded, pagination is very fast.

2.4 Protein, Peptide, and Phosphorylation Site Pages

All results on the browse page link to either a protein or peptide page. The protein and peptide pages both link to at least one phosphorylation site page. The protein, peptide, and phosphorylation site pages provide varying levels of information on phosphorylation site data. This provides more flexibility and in-depth information than similar resources. Researchers can view phosphorylation site data from multiple

perspectives to gain a better understanding of the data. Some researches may also be interested in a particular level of phosphorylation site information.

These three pages link to each other in a tree-like fashion. A protein may link to several phosphorylation sites and a phosphorylation site may link to several peptides. The peptide and phosphorylation site pages both link back to their corresponding protein page (see figure 2.6). This is useful if users want to view a different peptide or phosphorylation site page that belongs to the same protein. The way these three pages connect to each other reflects the database architecture.

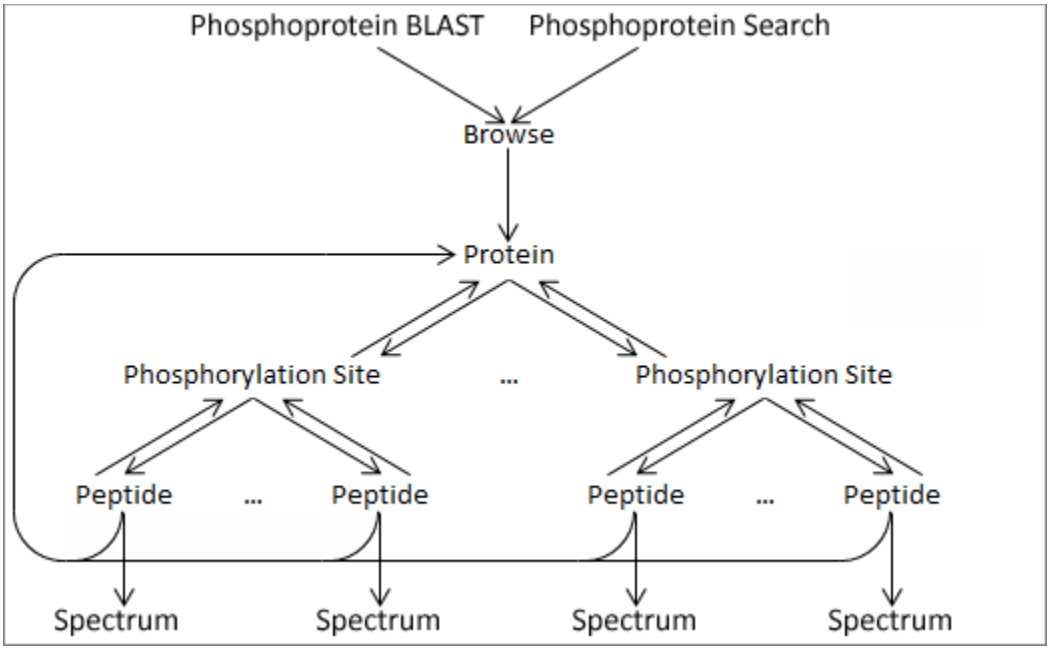

Figure 2.6: Flow diagram showing how the P^3DB protein page, phosphorylation site page, peptide page, and spectrum page are linked together.

The protein page displays specific information for a particular protein. The protein sequence is displayed with known phosphorylation sites highlighted with a red background and white text. All accession numbers that belong to the protein are listed at the bottom of the page and link to other databases. These accessions also appear on

the phosphorylation site and peptide pages. The displayed phosphorylation sites can be filtered by the experimental study that provided them using the reference select box. The page is dynamically updated when a reference option is chosen. The reference choice can be chosen on all three pages and the choice is remembered when switching between the protein, phosphorylation site, and peptide pages. It can be very useful for researchers to switch between references to compare data provided by each study. Some studies may get similar results. More information on a phosphorylation site can be found by clicking the site in the protein sequence or the phosphorylation site data table. This leads to an associated phosphorylation site page.

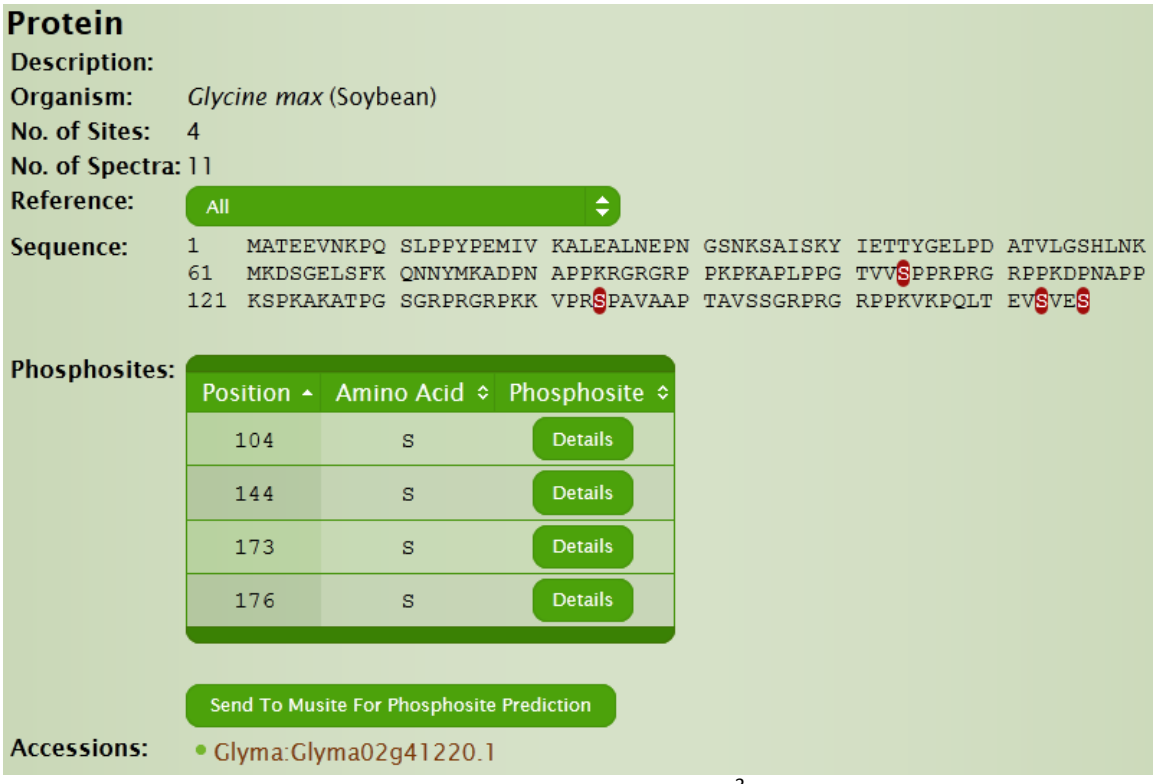

Figure 2.7: Protein page of P^{3}DB.

The phosphorylation site page is similar to the protein page, but with more targeted information. The position of the phosphorylation site within its associated

protein is given. The page shows the surrounding sequence of the phosphorylation site with only the targeted phosphorylation site highlighted in red. When a specific reference is chosen, not all, a view reference button is displayed next to the reference select box. This button links to an external page which provides detailed information on the reference. A list of peptides that the phosphorylation site belongs to is given. The targeted phosphorylation site is highlighted with a yellow background and red text. This is to distinguish it from other phosphorylation sites which are highlighted with a red background and white text. Clicking on a details button for a peptide will link to a peptide page.

Figure 2.8: Phosphorylation site page of P^{3}DB.

The peptide page is also similar to the protein page, but with more targeted information. The location of the peptide is given. This is the position of the first amino acid of the peptide in its associated protein. The page shows the peptide sequence with all phosphorylation sites highlighted in red. A list of all phosphorylation sites within the

peptide and their positions within the protein are given. Clicking on a phosphorylation site in the peptide sequence or the details button for a phosphorylation site will link to a phosphorylation site page. The bottom of the peptide page displays a table with mass spectrum analysis information. Some peptides link directly to a viewable and interactive mass spectrometry graph (see figure 2.9). The mass spectrometry graph provides very detailed mass spectrum information for researchers. The viewer is provided by European Bioinformatics Institute (EBI) [2].

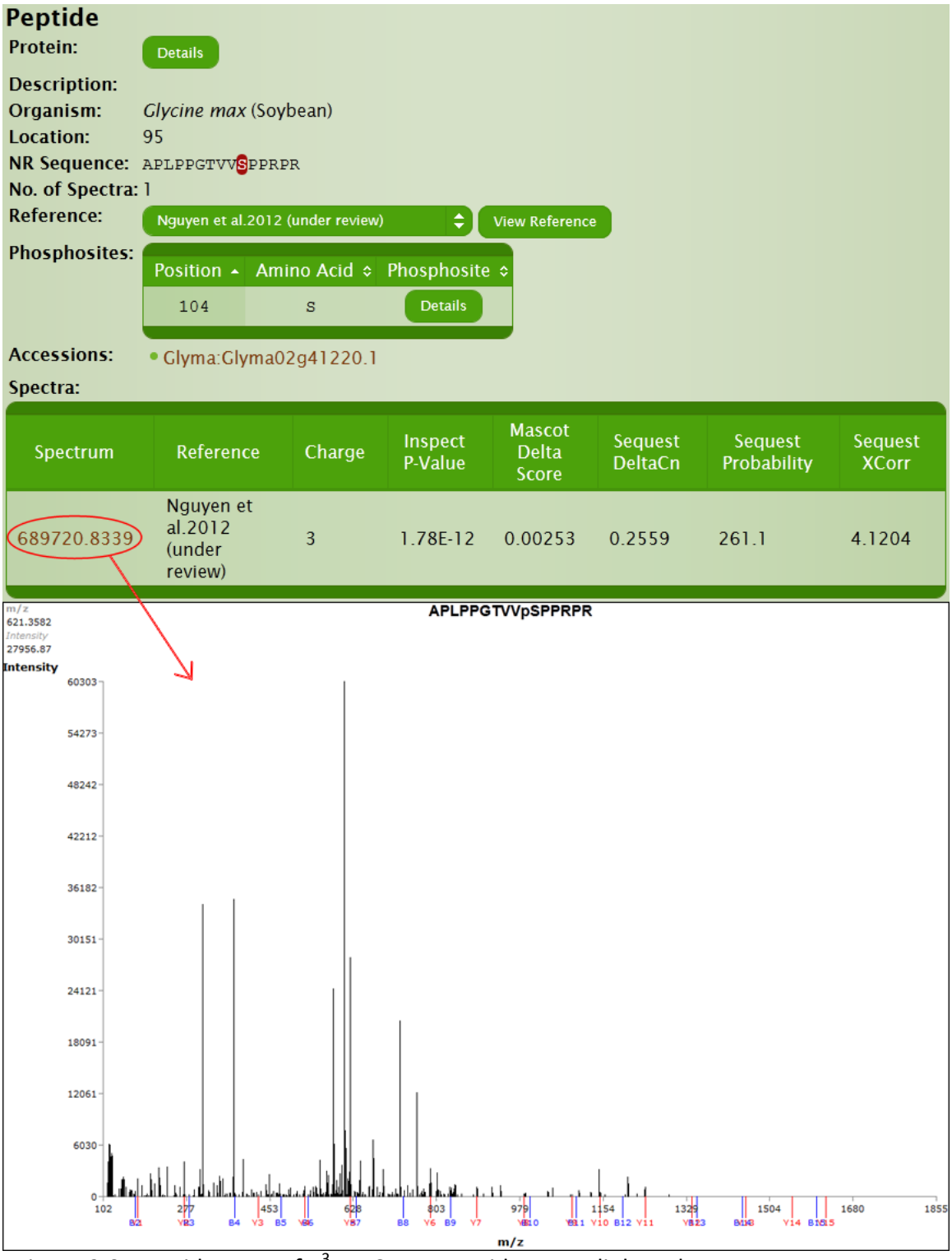

Figure 2.9: Peptide page of P[3]DB. Some peptide pages link to the mass spectrum page.

2.5 Data Sets Page

The data sets page provides statistics on where all the data in P^{3}DB originated. There are three tables on this page where each one has more detailed information than the last. The first and most general table gives the total number of experimental studies, organisms, proteins, and phosphorylation sites in the database. The second table gives the total number of proteins and phosphorylation sites for each organism (see table 2.1). The third table gives the number of proteins and phosphorylation sites separated by experimental study and organism including the totals for each experimental study.

The data sets page automatically updates as data is added to the database. The current data sets range from the year 2004 to 2012. Future enhancements are planned to allow users to upload their own datasets from the interface of P^{3}DB. Uploaded datasets will be reviewed for credibility before being added to the database.

The current version of P^{3}DB, version 1.6.0, contains protein phosphorylation data for six species from twenty-three experimental studies. There are currently 11,601 proteins, containing 32,963 phosphorylation sites in the database.

Organism	Phosphoproteins	Phosphosites
Arabidopsis thaliana	3,930	13,623
Brassica napus	325	818
Glycine max	1,451	2,739
Medicago truncatula	980	3,351
Oryza sativa	4,829	12,317
Zea mays	86	115
Total:	11,601	32,963

Table 2.1: Shows the number of proteins and phosphorylation sites stored in P^{3}DB by organism.

2.6 Download and Contact Pages

The download page allows users to download data from P^3DB. A subset of data can be downloaded by specifying an organism or experimental study. Data can be downloaded in several different formats. When a user submits a download request a usage-information pop-up is presented asking for their name, organization and email. The collected information is used by P^3DB administrators to better understand the user base. Future development may be targeted for the majority type of user based on gathered usage data. This pop-up is designed to be as unobtrusive as possible while still getting the user's attention. Several options for closing the pop-up are given. There is a skip button at the bottom and a close button at the top. Additionally, if the user clicks anywhere outside of the pop-up it will close. The screen around the pop-up darkens to make the pop-up more noticeable. A checkbox gives the option for users to subscribe to an announcement mailing list to receive updates on the development of P^3DB. Users can also subscribe to the mailing list on the contact page. The contact page contains information on who to contact for various issues.

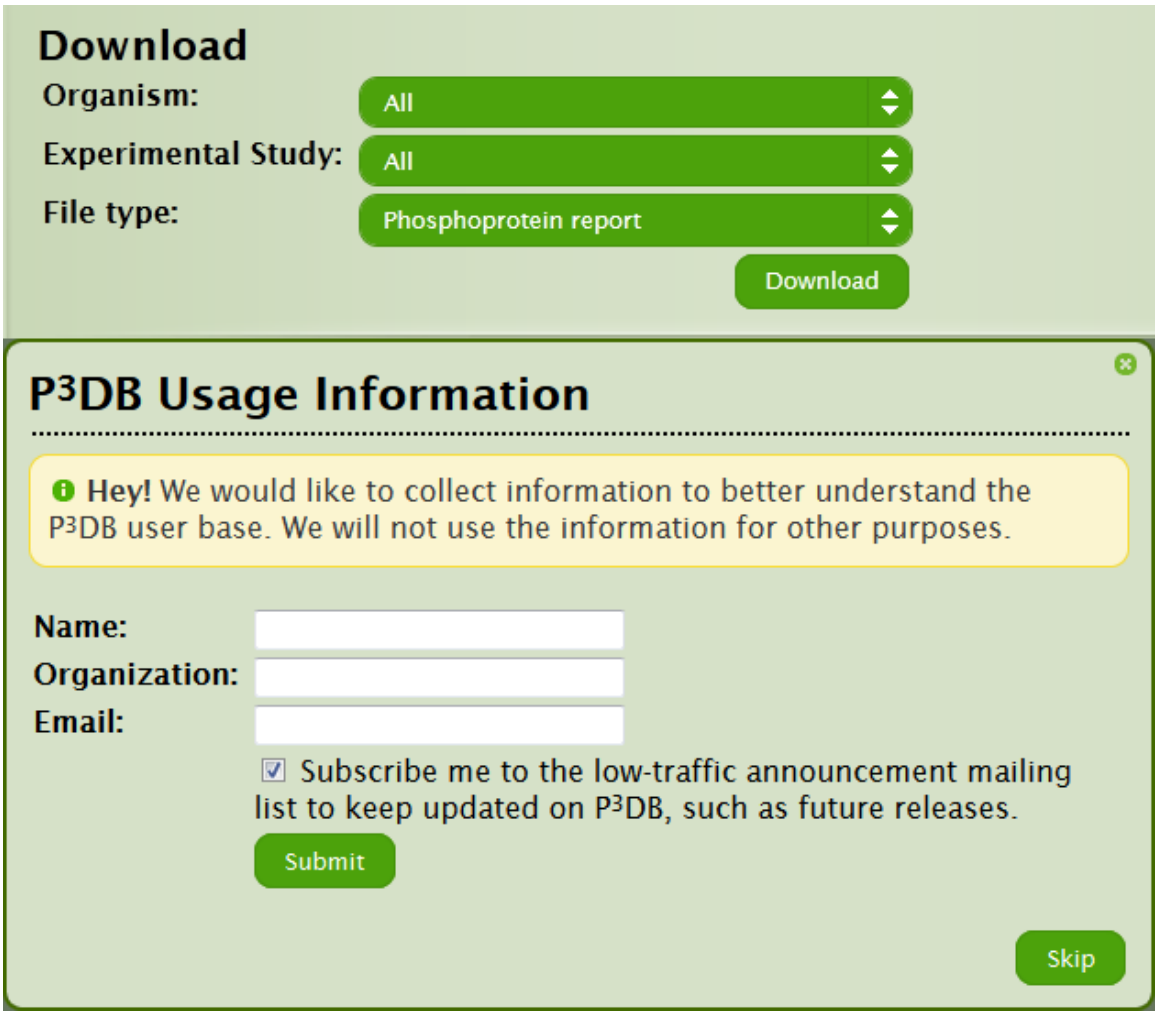

Figure 2.10: P[3]DB download page and usage information pop-up.

3. MUSITE.NET

Musite.net is a web server version of Musite. Musite is a standalone bioinformatics application specifically designed for large-scale prediction of protein phosphorylation sites. Musite.net uses many of the same prediction models as the standalone Musite, and has also been expanded to predict other post translational modifications (PTM) aside from phosphorylation.

Musite.net is more accessible than the standalone Musite because it does not need to be downloaded. Users do not need to worry about updating because updates are applied to the server. Musite.net also has benefits over similar prediction tools. Musite.net has several pre-made models for the types of predictions researchers are commonly interested in. Results are generated on the fly so users to not need to wait for an email or other method of receiving results. This allows users to make fast, high-quality predictions.

3.1Musite.net Design

Users are encouraged to use both the Musite application and Musite.net. Musite.net was designed to have an interface similar to the standalone desktop version Musite to provide continuity for users when they switch between the two tools. To accomplish this, nearly every feature of Musite.net is dynamic to make it similar to a desktop application. Unlike most websites, Musite.net presents very little static information and instead executes functions with dynamic results. Musite.net uses tabs similar to the

standalone desktop version instead of multiple pages of a website that can be navigated (see figure 3.1). The browser is never refreshed when switching between tabs and content is dynamically loaded when a new tab is created. Content always stays on the page, hidden at times, until the page or browser is closed. This eliminates the need to store the user's current session or cache results on the server which decreases server load.

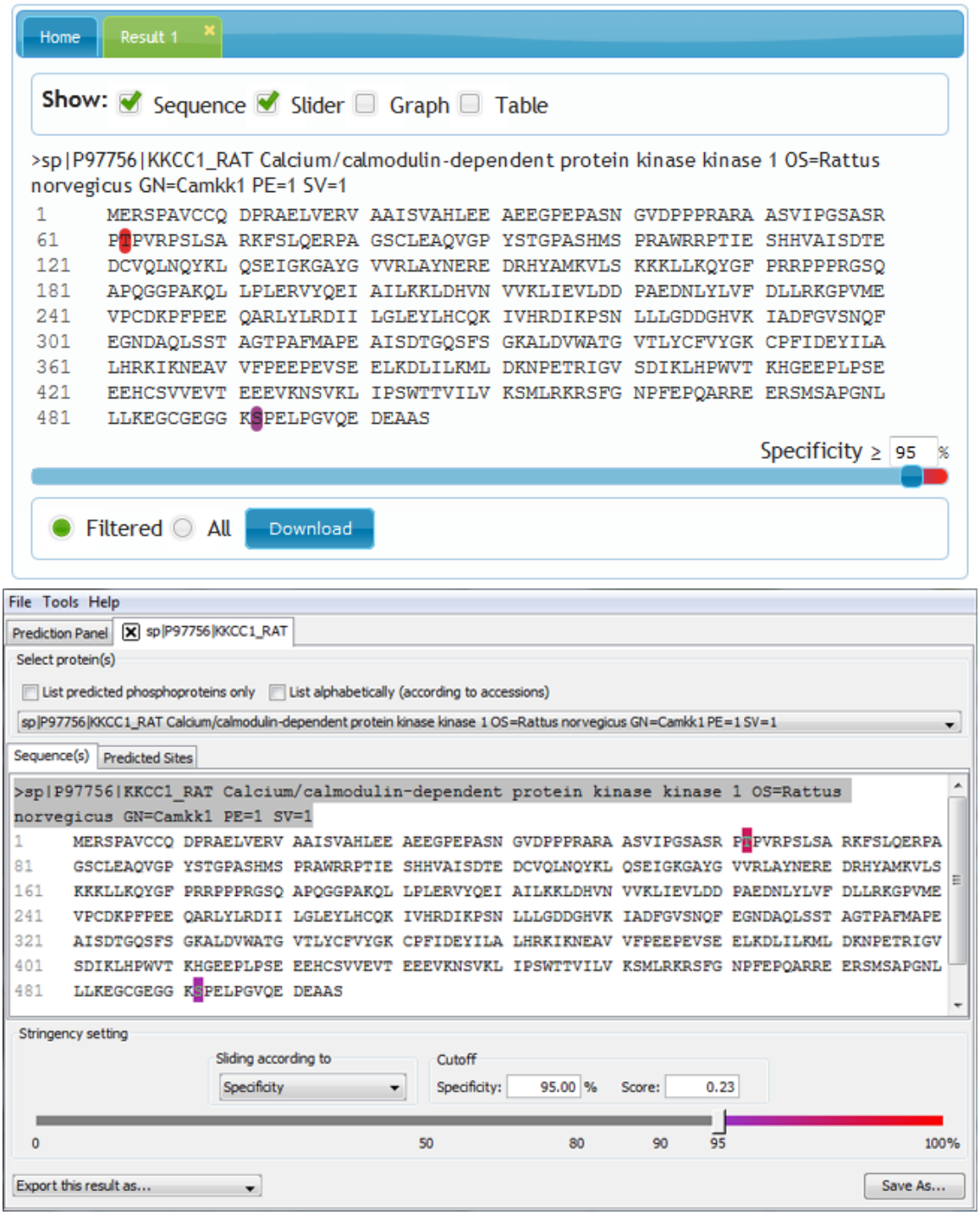

Figure 3.1: Similarity comparison of Musite.net (top) and Musite (bottom).

3.2 Prediction Method

Phosphorylation site prediction can be modeled as a binary classification problem. High quality data from several sources was used for the predictive model. Three types of features were extracted from the data to be used for classification including K-nearest neighbor (KNN) score, disorder value, and amino acid frequencies. The KNN score is the percentage of nearest neighbors with phosphorylation sites. Multiple KNN scores of 0.25%, 0.5%, 1%, 2%, and 4% of the bootstrap training dataset were used as KNN features. Disordered regions of a protein are regions without a well-defined structure. Phosphorylation sites are more frequently found in these regions. Average predicted disorder scores with window sizes of one, five, and thirteen were used as disorder features. Twenty amino acid frequencies (one for each amino acid) with a window size of thirteen were used as amino acid frequency features.

Bootstrap aggregation, or bagging, with support vector machines (SVM) was used for classification. There are many more non-phosphorylation sites (negative training samples) than there are phosphorylation sites (positive training samples) creating an unbalanced dataset. Bootstrap aggregation balances the number of positive and negative training samples to a default of 2000 each. A default of five SVMs was used for classification, each one classifying a different bootstrap. The results of all the trained SVMs were averaged together to get a final result [3].

3.3 Home Tab

The home tab is where users input their query sequence(s) or accession(s). This tab is always available and can never be closed. If Accessions and sequences are input at the same time then they are both processed and appear in the same result tab. An example link is provided next to each input. Clicking on an example link will dynamically load example sequences or accessions in their respective input fields. Currently UniProt accessions can be used. More types of accessions are planned for the future. There are currently several PTM prediction models to choose from including phosphorylation, acetylation, glycosylation, methylation, palmitoylation, sulfation, and SUMOylation (see table 3.1).

Sequences must be entered in FASTA format. Sequences can be entered with or without a header, line numbers, or spaces, for convenience. The sequence input box can be dragged vertically up or down so that more sequences are visible (see figure 3.2). Musite.net is not intended for large volume predictions like the standalone version because all users share the same server. Input is limited to one-hundred protein sequences per submission to avoid overloading the server. Users are encouraged to download the Musite application for large volume predictions.

The home page introduces users to an intuitive theme that is consistent throughout the site. Anything clickable is colored in a glossy blue. Links are similarly colored in a bright blue to stand out. Anything currently active is colored a bright green.

Alerts and tool-tips are colored yellow. This quickly allows users to visually understand how they can interact with the site.

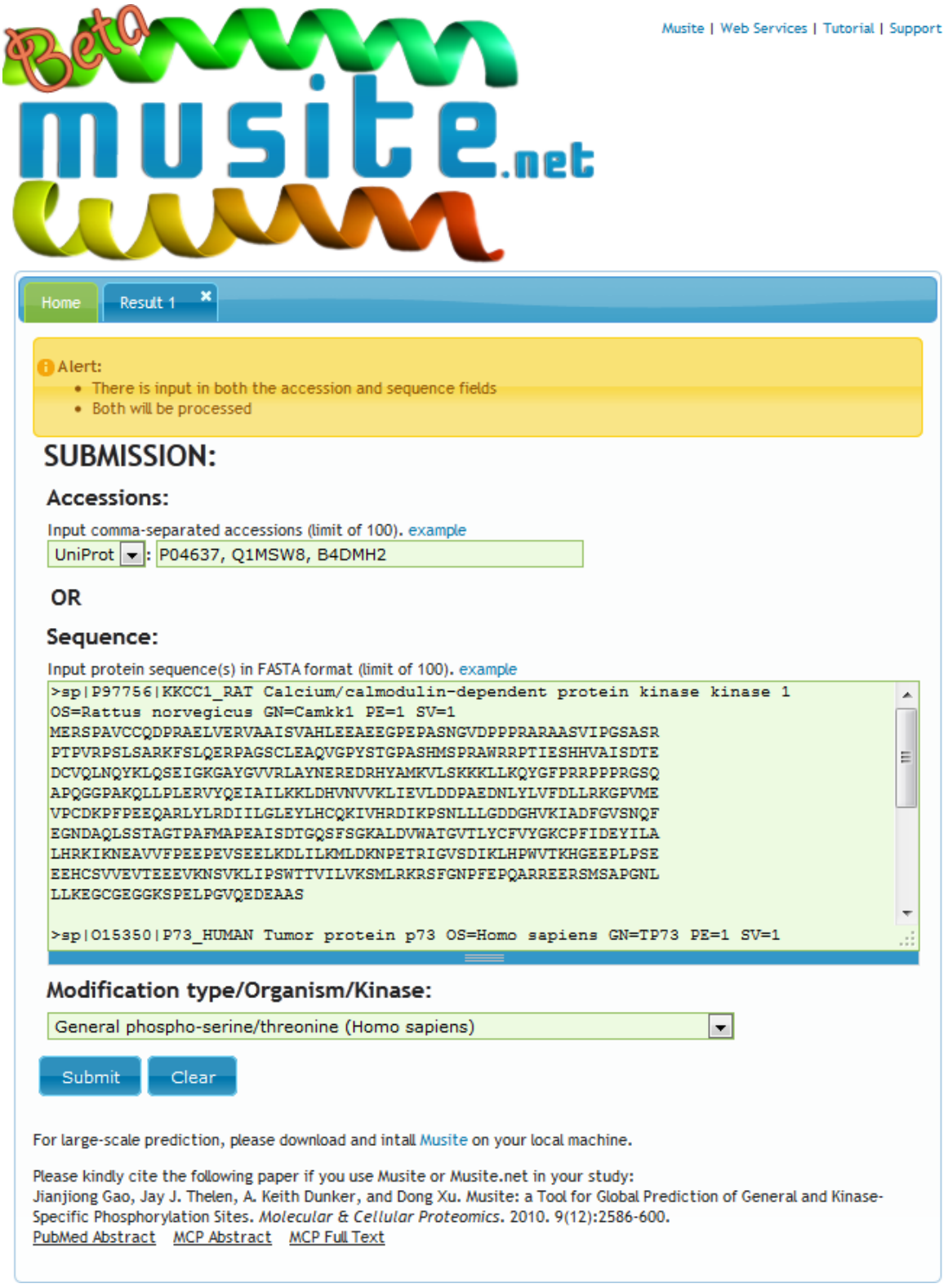

Figure 3.2: Musite.net home page.

Modification type/Organism/Kinase
General phospho-serine/threonine (Homo sapiens)
General phospho-tyrosine (Homo sapiens)
General phospho-serine/threonine (Mus musculus)
General phospho-tyrosine (Mus musculus)
General phospho-serine/threonine (Arabidopsis theliana)
General phospho-serine/threonine (Caenorhabditis elegans)
General phospho-serine/threonine (Drosophila melanogaster)
General phospho-serine/threonine (Saccharomyces cerevisiae)
General phospho-serine/threonine (Eukaryote)
General phospho-tyrosine (Eukaryote)
Kinase-specific phospho-serine/threonine (Ataxia telangiectasia mutated)
Kinase-specific phospho-serine/threonine (Cyclin dependent kinase 1)
Kinase-specific phospho-serine/threonine (Cyclin dependent kinase 2)
Kinase-specific phospho-serine/threonine (Cyclin-dependent kinase)
Kinase-specific phospho-serine/threonine (Casein kinase 1)
Kinase-specific phospho-serine/threonine (Casein kinase 2)
Kinase-specific phospho-serine/threonine (Mitogen-activated protein kinase 1)
Kinase-specific phospho-serine/threonine (Mitogen-activated protein kinase 3)
Kinase-specific phospho-serine/threonine (Mitogen-activated protein kinases)
Kinase-specific phospho-serine/threonine (Protein kinase A)
Kinase-specific phospho-serine/threonine (Protein kinase B)
Kinase-specific phospho-serine/threonine (Protein kinase C)
Kinase-specific phospho-tyrosine (Proto-oncogenic tyrosine kinase)
Acetylation: N6-Acety-lysine
Methylation: N6-methyl-lysine
Methylation: Omega-N-methyl-arginine
Palmitoylation: S-palmitoyl-cysteine
Sulfation: Sulfo-tyrosine
SUMOylation: SUMOylated lysine
Glycosylation: O-GlcNAc serine/threonine

Table 3.1: Musite.net prediction models. General phosphorylation site prediction models are highlighted in red. Kinase-specific phosphorylation site prediction models are highlighted in green. Other post translational modification site prediction methods are highlighted in blue.

3.4 Result Tab

Each submission creates a new result tab with one or more results. Results are limited to one-hundred per submission to help prevent the server from being over-tasked. This also prevents a result page from being excessively long and difficult to manage. If a single accession or sequence is submitted then the results page displays a single result (see figure 3.1 top). If multiple accessions or sequences are submitted then results are organized in an accordion widget (see figure 3.4). This allows each result to be expanded or collapsed. The widget was modified to allow more than one result to be open at once so that results can be compared to each other. Results can also be rearranged by dragging them vertically up or down. This allows users to put results close together for easier comparison.

A panel at the top of a multiple results tab contains functions that are global to all results in the current tab. A global specificity slider changes the specificity of all results in the tab. Each individual result also has its own specificity slider. Two buttons in the global panel allow the opening or closing of all results in the tab.

Each individual result has an options panel with check-boxes that allow users to choose which features are visible. The available options are to show the protein sequence, specificity slider, graph and data table. By default, only the protein sequence and specificity slider are visible. The graph and data table are not features of the standalone Musite. The sequence, graph and data table give similar information, but in

different formats. All three update dynamically when the local or global specificity slider is changed.

A specificity slider dynamically changes the specificity threshold which determines what PTM sites are shown in results. Many tools similar to Musite.net maintain a fixed specificity threshold. The specificity slider allows more flexibility for researchers who may want to set the specificity threshold higher or lower than the default. The specificity slider uses a color range from yellow (lowest) to red (highest) to represent the specificity of a PTM. Predicted PTM sites in the protein sequence and data table are colored to match the color scale of the specificity slider (see figure 3.3).

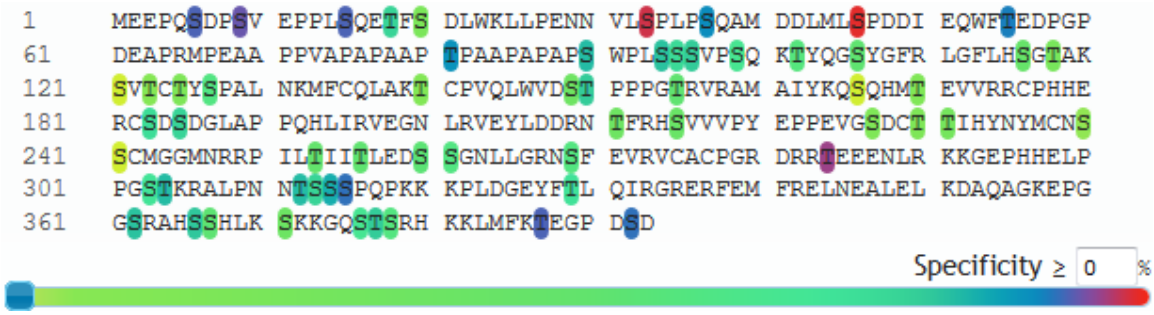

Figure 3.3: Musite.net specificity color scale.

The protein sequence highlights all predicted PTM sites with specificity above the selected threshold. The color scale for the highlighted amino acids is the same as the standalone version of Musite. Hovering over a highlighted amino acid shows a tooltip next to the mouse cursor. The tooltip contains the position and specificity of the highlighted amino acid.

The graph plots PTM sites with their position on the x- axis and their specificity on the y-axis. The scale of the graph automatically changes to minimize unused space. A legend in the top right shows the color of amino acids being plotted. The legend

becomes transparent when the mouse cursor is hovered over it so plotted points underneath the legend are easier to see. When users hover over a plotted point with the mouse cursor, a tooltip is shown with the same format as in the protein sequence.

The data table provides data for amino acid position, amino acid character, surrounding sequence, score, and specificity of PTM sites. The table can be ordered alphabetically by any of the columns. PTM sites in the data table are highlighted the same way as in the protein sequence.

The bottom panel allows results to be downloaded. Results can either be downloaded completely or filtered by specificity. Downloaded data for each protein starts with a header line and is tab delimited. Provided data includes amino acid position, amino acid character, surrounding sequence, score, and specificity. This is the same information provided in the data table.

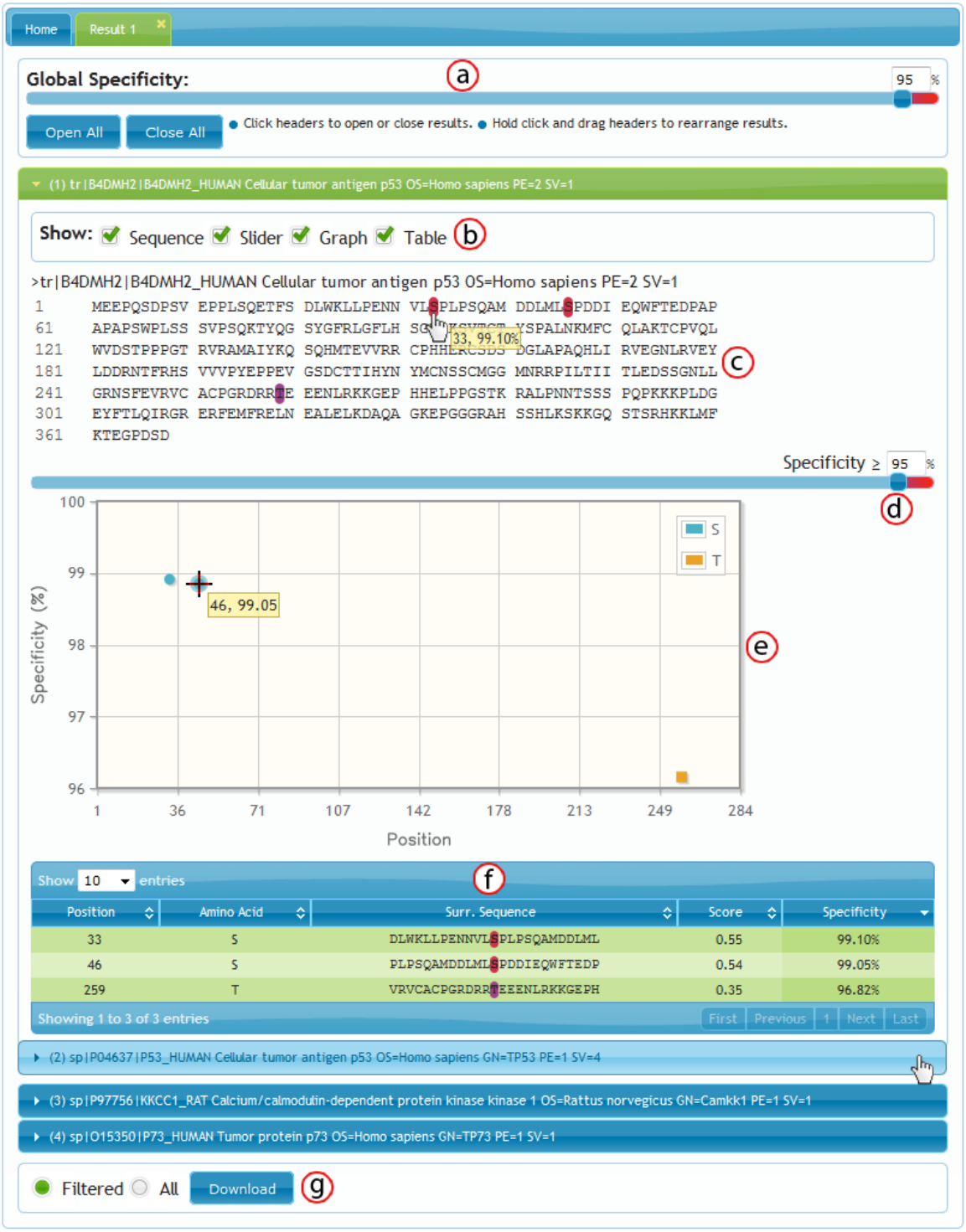

Figure 3.4: Musite.net result page. (a) Global options that affect all results in the selected tab. (b) Individual result presentation options. (c) Protein sequence with predicted PTM sites. Highlight color is based on specificity. (d) Specificity slider set to 95% by default. (e) Plot of predicted PTM sites by position and specificity. (f) Table of information on predicted PTM sites. (g) Download options.

3.5 Accessibility of Musite.net

Musite.net provides a RESTful web service API and direct URL submissions. The web service allows remote retrieval of serialized results. This allows results to be retrieved and saved by any software that can access the internet. The results can potentially be retrieved and imbedded within another web resource. Direct URL submissions allow prediction requests from web-links.

The web service is accessed through http://musite.net/service. Prediction results can be requested using an accession or protein sequence along with a prediction model. A list of all available prediction models can be requested by setting the "type" parameter to "getmodel". Setting the specificity parameter will only return results with a specificity equal to or above the set specificity. If the specificity parameter is not set then all results will be returned. can be set to override the default of 0.95. The results are returned in tab-delimited text format with each result on a new line (see figure 3.5).

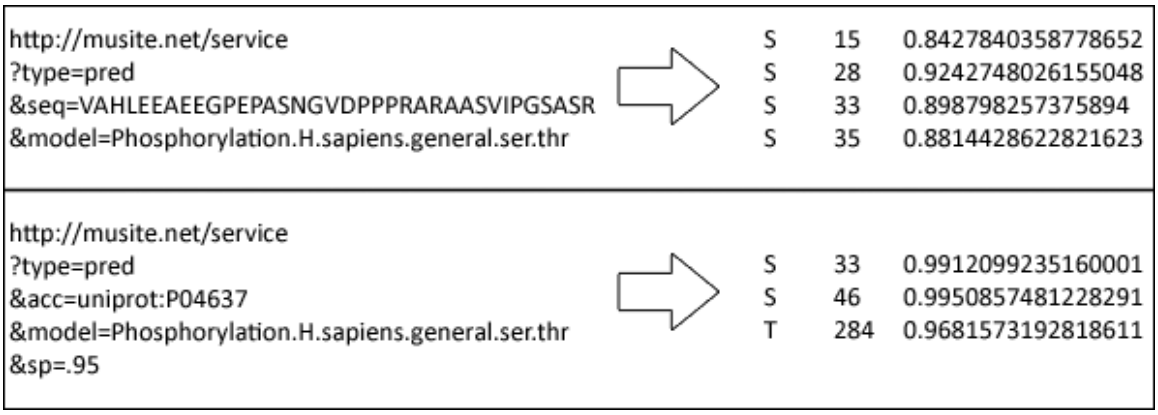

Figure 3.5: Musite.net example service API requests and their results.

Direct URL submissions allow users to go to a result page from a web link. This allows results to be bookmarked, sent through email, or referenced by other websites.

P^3DB makes use of this feature by allowing users to send a protein sequence to Musite.net for prediction. Submissions can be made using an accession or protein sequence along with a prediction model. For P^3DB, Researchers may find it useful to compare the known phosphorylation sites from experimental studies to the predicted sites of Musite.net. Musite.net will never be as accurate at experimental studies, but Musite.net does give a more general view of where phosphorylation sites may be located.

4. MODULAR TEMPLATE FILE STRUCTURE

Both the Musite.net and P^{3}DB design utilize a template file structure. There are many advantages to using templates over the classic method. The primary advantage is the separation of procedural code and formatting code. The formatting code of all documents is structured using Extensible Hypertext Markup Language (XHTML). XHTML provides more uniformity to a web document than older styles of HTML. XHTML makes websites more compatible in multiple browsers and mobile devices.

Templates consist of a processing file and a formatting file that are combined in a template engine. This allows web content to be created and maintained efficiently. Templates allow content to be broken into modular sections where each section can be changed without affecting the others. Each template section may be made of other templates or simply be static content. P^{3}DB and Musite.net use a simple, but effective template file structure. The sites consist of several templates. There is a single global template for the entire site that consists of three sections; the header, content block and footer.

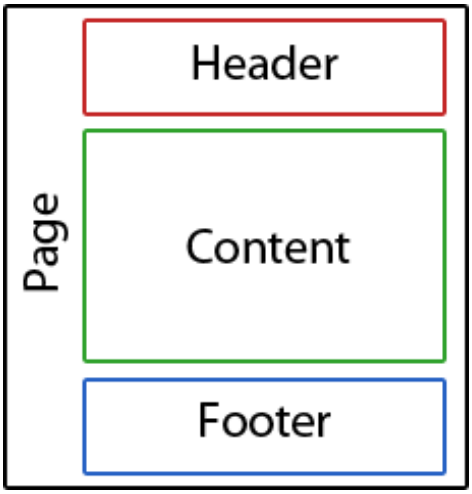

Figure 4.1: A page template containing a header, content, and footer template.

A global page template allows the header and footer to be the same on every page. The header contains all the information that needs to be at the top of an XHTML web page. This includes everything inside the <head> tag, such as the <doctype>, <meta>, <style>, and <script> tags. It also includes the opening <body> tag and the visible header which includes the title of the site, logo and navigation menu. The footer simply contains copyright information.

The content block contains the core information for each page. There is a unique template for the content of each page. Each content block is generated in two main steps. First, a processing script gathers all data that will be displayed on the requested page. The final processed data is in the form of data arrays or single variables. In the second step, the processed data is passed to the template engine which then imbeds the data into a formatting file (see figure 4.2). The formatting file may further manipulate the data, but only for formatting purposes, such as making tables.

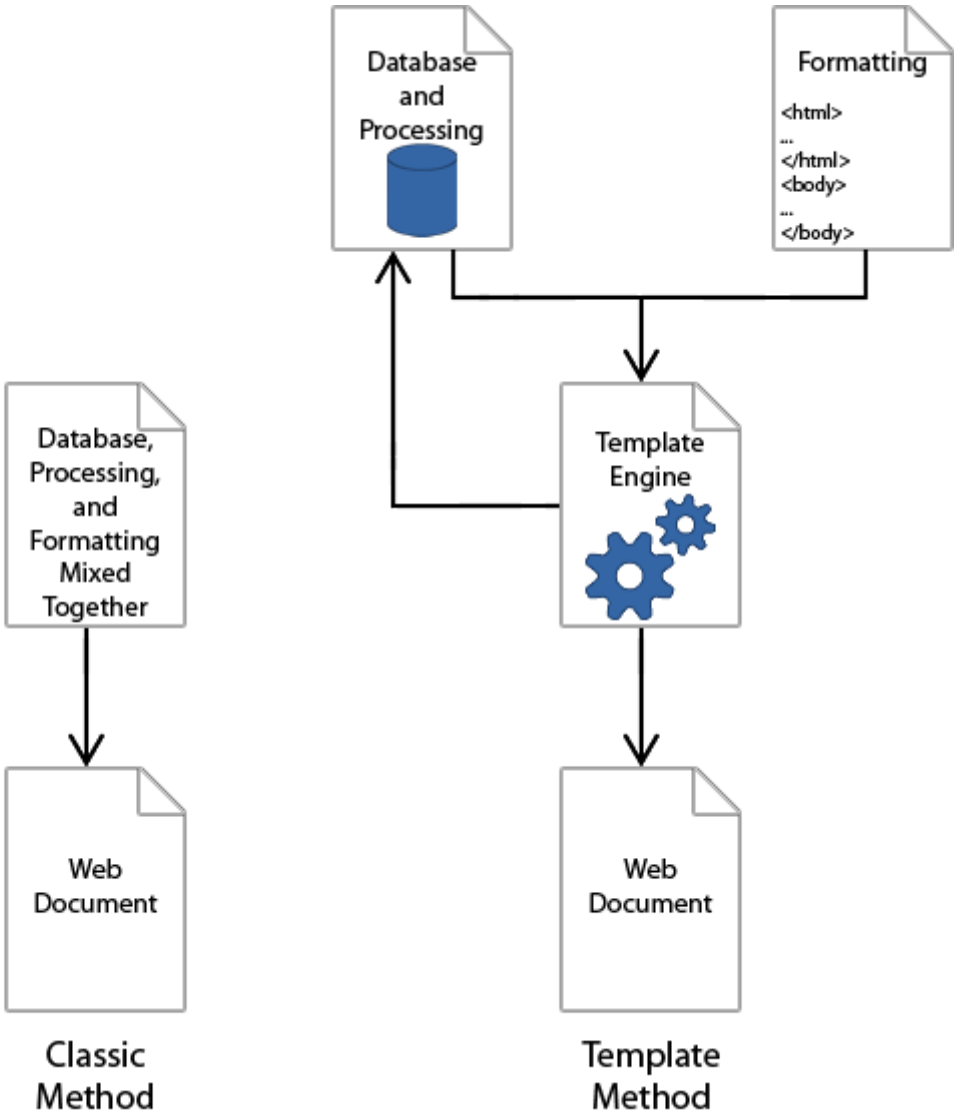

Figure 4.2: Flow chart comparison of the classic method and template method. A processing file and formatting file are combined in the template engine. A template can be part of another processing file recursively until the document is complete.

Maintaining uniformity throughout a website is important for the user experience. Page-to-page uniformity provides a professional look and is less confusing for the user. While simply displaying content is important, it is also important that the site maintains an intuitive interface where information can be found efficiently. Content can be interpreted differently depending on the layout it is presented in.

Simply copying an existing page and modifying it with new content is fast in the short term, but performing maintenance and adding new content becomes difficult. Templates create style standardization by allowing content to be modular. Similar content is coded once while appearing as many times as necessary throughout the site. One change to a template affects every location where that template is used. For example, the menu bar is coded in one location, yet it appears on every page of the website.

Separating procedural code from formatting code is a good programming practice in any context. This separation is especially useful in web coding because the formatting of web pages may change frequently. Modifying a web page is difficult when the formatting code is mixed with the procedural code. Templates provide an effective way to separate these two types of code. This separation makes it easy to change one type of code without affecting the other. For example, the style of a web-page can be changed completely without altering the procedural code in any way. Additionally, a page can have multiple styles, depending on user input, while using the same procedural code.

5. INTERFACE USING DYNAMIC WEB CODING

5.1 JavaScript and jQuery Features

JavaScript is a client-side scripting language used to create enhanced user interfaces and dynamic websites. jQuery is a JavaScript library that allows JavaScript code to be written much faster by simplifying JavaScript syntax. Writing separate code for multiple browsers is often necessary when writing JavaScript because of compatibility differences in browsers. jQuery simplifies browser compatibility by insuring that every jQuery function is compatible on all major browsers.

jQuery allows content to be accessed faster by eliminating the need to change between pages as often. jQuery also allows content such as draggable items which are nearly impossible to implement with only static web-pages.

All dynamic features in P^3DB and Musite.net are designed to be practical. Purely stylistic, or flashy, animations can be annoying to users. In P^3DB, when a table is too large for all content to fit horizontally, a shadow dynamically appears on the edge(s) where content is cut off. This is visually appealing, but this also intuitively suggests that the content is “under” the table and can be scrolled. Without this, the cut-off content may be easily overlooked by users.

Advanced JavaScript tools, such as Jmol, are very useful visualization tools. Jmol shows a three dimensional model of a protein. The model can be set to automatically rotate. The model can be zoomed and rotated manually with the mouse cursor [5].

Another advanced JavaScript tool is the spectrum visualization page. This page displays a spectrum graph with several interactive features.

Ideally a website is fully functional even if JavaScript is disabled in the browser. This is knows as non-obtrusive JavaScript. This ideology was used as much as possible while developing P^3DB to allow compatibility in a variety of running environments.

Musite.net is an exception to the non-obtrusive JavaScript ideology. Musite.net was designed as an extremely dynamic web-application and is heavily dependent on JavaScript. A completely different version of Musite.net would need to be developed to allow for Musite.net to run in an environment without JavaScript. However, the API of Musite.net may be used as a simplistic alternative if JavaScript must be disabled.

5.2 Asynchronous JavaScript and XML (Ajax) Features

Ajax is used to retrieve information from the server without reloading a page. This is done asynchronously meaning other scripts will continue to run while a response from the server is pending. Despite the name, Ajax is not required to be asynchronous, nor does the returned data need to be in XML format [6]. Ajax is used for several purposes by P^3DB and Musite.net.

The browse page of P^3DB uses Ajax for search pagination. For each protein in P^3DB, three-dimensional protein models are retrieved using Ajax and then displayed by Jmol. In general, any third party web resources can be retrieved using Ajax. P^3DB uses Ajax for a few enhancements. However, Musite.net uses Ajax for nearly every feature.

Musite.net uses Ajax to populate content for tabs. This is necessary because many tabs can be open at once. Generating previous tabs again each time a new tab is created would be difficult and impractical. In addition to tab creation, Musite.net uses Ajax to retrieve example text that is inserted into query fields.

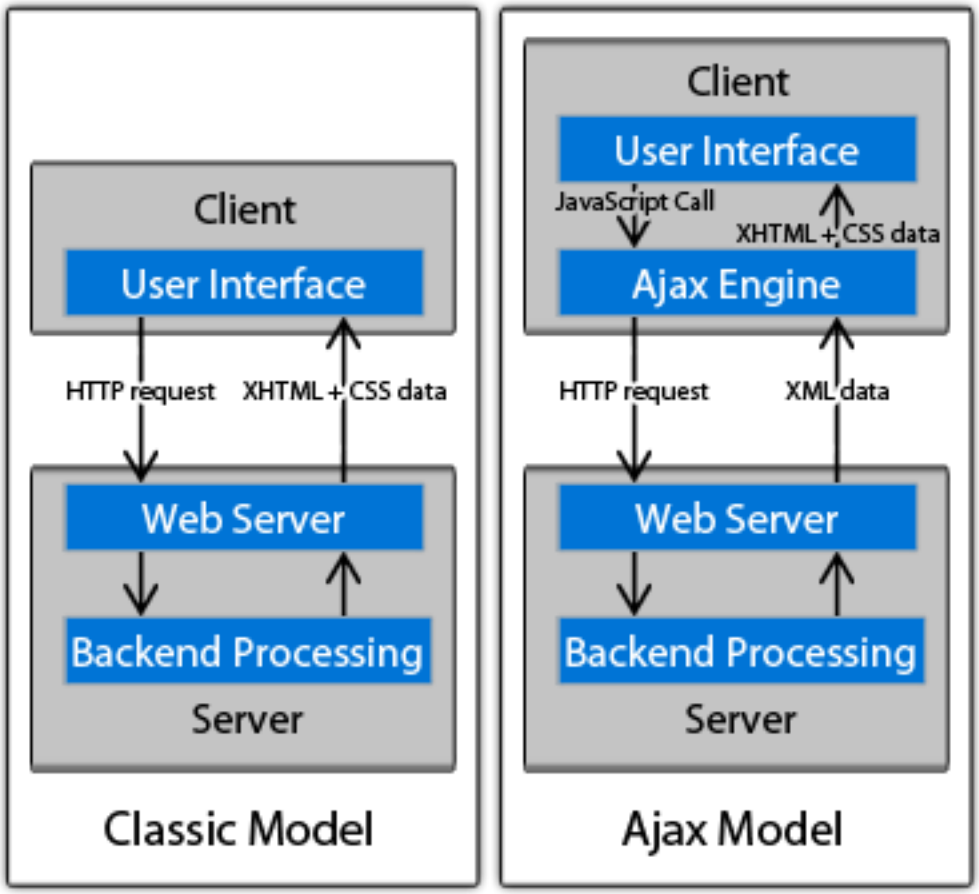

Figure 5.1: Flow chart showing the difference between the classic web application model and the Ajax web application model.

6. SUMMARY

Both P^3DB and Musite.net are useful resources for the bioinformatics community. Both take advantage of modern technologies to create an intuitive and visually appealing user interface. Future maintenance and development are designed to be easy with the code organized using a template file structure. Musite.net encourages community interaction with its API and direct URL submission. More data visualizations and others types of plant protein phosphorylation data are planned for P^3DB. In the future, Musite.net will allow users to change the prediction model dynamically without submitting a new prediction. P^3DB and Musite.net are intended to remain relevant tools for bioinformatics research with their current features and future improvements.

BIBLIOGRAPHY

1. Seo, J and Lee, KJ, *Post-translational modifications and their biological functions: proteomic analysis and systematic approaches.* J Biochem Mol Biol, 2004. **37**(1): p. 35-44.

2. PRIDE Inspector: a tool to visualize and validate MS proteomics data. http://www.ebi.ac.uk/pride/

3. Gao, J., Thelen, J.J., Dunker, A.K. and Xu, D. Musite, a tool for global prediction of general and kinase-specific phosphorylation sites. *Mol Cell Proteomics*, **9**, 2586-2600.

4. Gao, J and Xu, D, *The Musite open-source framework for phosphorylation-site prediction.* BMC Bioinformatics. **11 Suppl 12**: p. S9.

5. Jmol: an open-source Java viewer for chemical structures in 3D. http://www.jmol.org/

6. Ullman, Chris (March 2007). *Beginning Ajax*. wrox. ISBN 978-0-470-10675-4. Retrieved 24 June 2008.

7. Farriol-Mathis, N, Garavelli, JS, Boeckmann, B, Duvaud, S, Gasteiger, E, Gateau, A, Veuthey, AL, and Bairoch, A, *Annotation of post-translational modifications in the Swiss-Prot knowledge base.* Proteomics, 2004. **4**(6): p. 1537-1550.

8. Lee, TY, Huang, HD, Hung, JH, Huang, HY, Yang, YS, and Wang, TH, *dbPTM: an information repository of protein post-translational modification.* Nucleic Acids Res, 2006. **34**(Database issue): p. D622-627.

9. Gnad, F, Gunawardena, J, and Mann, M, *PHOSIDA 2011: the posttranslational modification database.* Nucleic Acids Res, 2011. **39**(Database issue): p. D253-260.

10. Gnad, F, Ren, S, Cox, J, Olsen, JV, Macek, B, Oroshi, M, and Mann, M, *PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites.* Genome Biol, 2007. **8**(11): p. R250.

11. Gao, J, Agrawal, GK, Thelen, JJ, and Xu, D, *P3DB: a plant protein phosphorylation database.* Nucleic Acids Res, 2009. **37**(Database issue): p. D960-962.

12. Bodenmiller, B, Campbell, D, Gerrits, B, Lam, H, Jovanovic, M, Picotti, P, Schlapbach, R, and Aebersold, R, *PhosphoPep--a database of protein*

phosphorylation sites in model organisms. Nat Biotechnol, 2008. **26**(12): p. 1339-1340.

13. Bodenmiller, B, Malmstrom, J, Gerrits, B, Campbell, D, Lam, H, Schmidt, A, Rinner, O, Mueller, LN, Shannon, PT, Pedrioli, PG, Panse, C, Lee, HK, Schlapbach, R, and Aebersold, R, *PhosphoPep--a phosphoproteome resource for systems biology research in Drosophila Kc167 cells.* Mol Syst Biol, 2007. **3**: p. 139.

14. Huttlin, EL, Jedrychowski, MP, Elias, JE, Goswami, T, Rad, R, Beausoleil, SA, Villen, J, Haas, W, Sowa, ME, and Gygi, SP, *A tissue-specific atlas of mouse protein phosphorylation and expression.* Cell, 2010. **143**(7): p. 1174-1189.

15. de Castro, E, Sigrist, CJ, Gattiker, A, Bulliard, V, Langendijk-Genevaux, PS, Gasteiger, E, Bairoch, A, and Hulo, N, *ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W362-365.

16. Gattiker, A, Gasteiger, E, and Bairoch, A, *ScanProsite: a reference implementation of a PROSITE scanning tool.* Appl Bioinformatics, 2002. **1**(2): p. 107-108.

17. Obenauer, JC, Cantley, LC, and Yaffe, MB, *Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs.* Nucleic Acids Res, 2003. **31**(13): p. 3635-3641.

18. Iakoucheva, LM, Radivojac, P, Brown, CJ, O'Connor, TR, Sikes, JG, Obradovic, Z, and Dunker, AK, *The importance of intrinsic disorder for protein phosphorylation.* Nucleic Acids Res, 2004. **32**(3): p. 1037-1049.

19. Blom, N, Gammeltoft, S, and Brunak, S, *Sequence and structure-based prediction of eukaryotic protein phosphorylation sites.* J Mol Biol, 1999. **294**(5): p. 1351-1362.

20. Blom, N, Sicheritz-Ponten, T, Gupta, R, Gammeltoft, S, and Brunak, S, *Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence.* Proteomics, 2004. **4**(6): p. 1633-1649.

21. Wong, YH, Lee, TY, Liang, HK, Huang, CM, Wang, TY, Yang, YH, Chu, CH, Huang, HD, Ko, MT, and Hwang, JK, *KinasePhos 2.0: a web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns.* Nucleic Acids Res, 2007. **35**(Web Server issue): p. W588-594.