A thesis submitted to The University of Manchester for the
Degree of PhD in the Faculty of Life Sciences

Informatics tools for the analysis and assignment of phosphorylation status
in proteomics

Dave Chi Hoo Lee
September 2014
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of Figures</td>
<td>7</td>
</tr>
<tr>
<td>List of Tables</td>
<td>8</td>
</tr>
<tr>
<td>Abstract</td>
<td>9</td>
</tr>
<tr>
<td>Declaration</td>
<td>10</td>
</tr>
<tr>
<td>Copyright</td>
<td>11</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>12</td>
</tr>
<tr>
<td>The Author</td>
<td>13</td>
</tr>
<tr>
<td>Abbreviations</td>
<td>14</td>
</tr>
<tr>
<td><strong>Chapter 1: Introduction</strong></td>
<td>16</td>
</tr>
<tr>
<td>1.1. Phosphorylation</td>
<td>17</td>
</tr>
<tr>
<td>1.2. Mass Spectrometry</td>
<td>20</td>
</tr>
<tr>
<td>1.3. Enrichment of phosphorylated peptides through liquid chromatography</td>
<td>21</td>
</tr>
<tr>
<td>1.3.1. Ion Exchange chromatography</td>
<td>22</td>
</tr>
<tr>
<td>1.3.2. Immobilised metal affinity chromatography</td>
<td>22</td>
</tr>
<tr>
<td>1.3.3. Metal-oxide affinity chromatography</td>
<td>23</td>
</tr>
<tr>
<td>1.4. Activation methods and alternatives more suited for fragmentation of phosphopeptides in tandem mass spectrometry</td>
<td>25</td>
</tr>
<tr>
<td>1.4.1. Alternative collision-based activation methods</td>
<td>28</td>
</tr>
<tr>
<td>1.4.2. Electron-transfer-dissociation activation</td>
<td>28</td>
</tr>
<tr>
<td>1.5. Peptide identification and localisation of phosphorylation sites</td>
<td>29</td>
</tr>
<tr>
<td>1.5.1. Database search-engines</td>
<td>30</td>
</tr>
<tr>
<td>1.5.2. Alternative sequence identification methods</td>
<td>30</td>
</tr>
<tr>
<td>1.5.3. Assessing statistical signification of phosphopeptide identifications</td>
<td>31</td>
</tr>
<tr>
<td>1.5.4. More advanced informatic approaches to improve identification</td>
<td>33</td>
</tr>
<tr>
<td>1.6. Site localisation</td>
<td>34</td>
</tr>
<tr>
<td>1.6.1. Site localisation algorithms</td>
<td>36</td>
</tr>
<tr>
<td>1.6.1.1. Probability-based localisers</td>
<td>36</td>
</tr>
<tr>
<td>1.6.1.2. Search-engine difference scores</td>
<td>38</td>
</tr>
<tr>
<td>1.6.1.3. The status of the false localisation rate</td>
<td>41</td>
</tr>
<tr>
<td>1.7. The prediction of phosphorylation sites</td>
<td>42</td>
</tr>
<tr>
<td>1.7.1. Kinase-specific and non-kinase-specific predictors</td>
<td>42</td>
</tr>
<tr>
<td>1.7.2. Popular phosphorylation-enriched characteristics</td>
<td>43</td>
</tr>
<tr>
<td>1.7.2.1. Kinase-substrate linear motifs</td>
<td>44</td>
</tr>
<tr>
<td>1.7.2.2. Protein disorder propensity</td>
<td>44</td>
</tr>
<tr>
<td>1.7.2.3. Conservation</td>
<td>47</td>
</tr>
</tbody>
</table>
1.7.3. Using machine-learning algorithms for prediction 48
1.7.3.1. Support vector machines 49
1.7.3.2. Artificial neural networks 49
1.7.3.3. Ensembl methods 50
1.7.3.4. Constructing a positive and negative set for predictor training 51
1.7.4. Using structural features for prediction 53
1.8. Assigning functional status of phosphorylation 54
1.8.1. Detection of functional sites through quantitative studies 56
1.9. Overview of thesis 59
1.10. References 62

Chapter 2: Benchmarking computational tools and consensus approaches for mass spectrometry-based phosphoproteomics 68

2.1 Abstract 69
2.2 Introduction 70
2.3 Materials and Methods 73
2.3.1. Synthetic phosphopeptide datasets 73
2.3.2. File formats 73
2.3.3. Database search-engines for sequence inference 73
2.3.4. Peptide-spectral-match and peptide-centric analysis 75
2.3.5. Computing statistically significant identification with QVality 75
2.3.5.1. Determining the true q-values of peptide spectral matches 75
2.3.6. Applying site localisation algorithms and scores 76
2.3.7. Criteria for assessing performance of sequence identification and phosphosite localisation 76
2.3.8. Consensus q-values across multiple search-engines and a voting approach for site localisation algorithms 76
2.4 Results 78
2.4.1. Assessment of search-engine performance on identifying phosphorylated sequences for different activation methods 78
2.4.1.1. True false discovery rate calculations deviate from the intended estimated rates 86
2.4.1.2. Consensus Heuristics improves phosphopeptide identification over any single search-engine under all activation methods 88
2.4.2. Phosphorylation site localisation 91
2.4.2.1. Comparative performance of site localisation algorithms under each activation approach 91
2.4.2.2. Consensus site localisation increases confidence in phosphosite Assignment 94
2.5 Discussion 96
2.5.1. A multiple search-engine strategy lowers error rates associated with search
2.5.2. A multiple site localisation strategy can improve localization of collision-derived spectra

2.6 Conclusions

2.7 References

Chapter 3: An investigation into discriminative features of biologically regulated phosphorylation sites in mammalian organisms

3.1. Abstract

3.2. Introduction

3.3. Materials and Methods

3.3.1. Datasets

3.3.1.1. Selection of true and functional phosphorylation sites from PhosphositePlus (Version 4th April 2014)

3.3.1.2. Acquiring functional sites from quantitative mass-spectrometry Datasets

3.3.1.3. Assessing the properties of less-functional sites

3.3.1.4. Negative set

3.3.2. Sequence-level analyses

3.3.2.1. Creating multiple sequence alignments from pre-defined eggNOG resource for conservation

3.3.2.2. Conservation calculations with rate4site

3.3.2.3. Disorder calculations with FoldIndex and DISOPRED

3.3.2.4. Secondary structure

3.3.2.5. Predicted accessible surface area

3.3.3. Structural analyses

3.3.3.1. Biological assembly units from the Protein Data Bank

3.3.3.2. Annotating structures with phosphorylation sites

3.3.3.3. FoldX suite

3.3.3.4. Positive electrostatic potential calculations

3.3.3.5. Accessibility calculations

3.3.3.6. Partitioning phosphorylation sites into interface and surface sites

3.4. Results

3.4.1. Predicted structural states reveal functional sites are more modest than the less-functional sites

3.4.2. Conservation analysis show less-functional sites are less conserved

3.4.3. Phosphorylation sites occupy more positively charged regions than non-phosphorylated sites over short distances

3.4.4. Total energy calculations of functional sites exceed the functionally unassigned at the interface but not negatives

3.5. Discussion

3.5.1. The role of physical susceptibility in determining functional phosphorylation
Chapter 4: Evaluating general and functional phosphorylation site prediction performance

4.1. Abstract

4.2. Introduction

4.3. Materials and Methods

4.3.1. Datasets

4.3.1.1. Phosphorylation resources

4.3.1.2. SILAC-derived phosphorylation sites

4.3.1.3. Defining positive and negative sets for training and testing

4.3.1.4. Independent tests

4.3.2. Machine Learning Procedures

4.3.2.1. Partitioning of data for feature selection, training and testing

4.3.2.2. Selection of physicochemical properties from AAindex for use as machine learning features

4.3.2.3. Selecting features for machine learning

4.3.3. Machine Learning

4.3.3.1. Optimisation of models with CARET package

4.3.3.2. Neural Network

4.3.3.3. Random Forest

4.3.3.4. Adaptive Boosting

4.3.3.5. Evaluating performance

4.3.4. Phosphorylation-characteristic features

4.3.4.1. Disorder calculations with FoldIndex and DISOPRED3

4.3.4.2. Conservation calculations with rate4site

4.3.4.3. Secondary structure

4.3.4.4. Predicted accessible surface area

4.3.5. Published phosphosite predictors

4.3.5.1. GPS

4.3.5.2. MUSITE

4.3.5.3. DISPHOS

4.3.5.4. PPSP

4.3.5.5. NetPhos

4.4. Results

4.4.1. Performance of machine learning models using combinatorial features on held-out test sets
4.4.2. Predictive performance of published predictors shows varied performance on functional datasets 158
4.4.3. Comparative evaluation of all predictors on functional sets 161
4.5. Discussion 166
4.5.1. Kinase-specific predictors are capable of avoiding less-functional sites 166
4.5.2. Kinase-specific predictors are also capable of detecting novel functional sites 166
4.5.3. Suggestions to improve functional prediction of phosphosites 167
4.6. Conclusion 169
4.7. References 170

Chapter 5: Conclusions 173
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1.1</td>
<td>Phosphorylation and de-phosphorylation of proteins</td>
<td>19</td>
</tr>
<tr>
<td>Figure 1.2</td>
<td>Sample separation by ion-exchange chromatography</td>
<td>24</td>
</tr>
<tr>
<td>Figure 1.3</td>
<td>Proposed nomenclature by Roepstorff for fragmentation of peptide sequences</td>
<td>27</td>
</tr>
<tr>
<td>Figure 1.4</td>
<td>Ambiguity in site assignment of phosphopeptides</td>
<td>35</td>
</tr>
<tr>
<td>Figure 1.5</td>
<td>'Schulz' diagram describing energy changes for binding</td>
<td>46</td>
</tr>
<tr>
<td>Figure 1.6</td>
<td>Example of a hyperplane separating two different classes (red and blue shapes) in support vector machines</td>
<td>52</td>
</tr>
<tr>
<td>Figure 1.7a</td>
<td>Labelling phase of a SILAC experiment</td>
<td>58</td>
</tr>
<tr>
<td>Figure 1.7b</td>
<td>Experiment phase of SILAC experiment</td>
<td>58</td>
</tr>
<tr>
<td>Figure 1.8</td>
<td>Overview of thesis</td>
<td>61</td>
</tr>
<tr>
<td>Figure 2.1</td>
<td>Phosphopeptide identifications for different activation methods and search engines</td>
<td>80</td>
</tr>
<tr>
<td>Figure 2.2</td>
<td>Peptide-centric precision-recall plots of search-engines under all activation methods</td>
<td>85</td>
</tr>
<tr>
<td>Figure 2.3</td>
<td>Comparison of estimated and true q-values for each search-engine under all activation conditions</td>
<td>87</td>
</tr>
<tr>
<td>Figure 2.4</td>
<td>(a - b). Venn diagrams of CID (a)- and ETD (b)-activated PSM-to-peptide-centric identifications</td>
<td>90</td>
</tr>
<tr>
<td>Figure 3.1</td>
<td>Physical features accessible surface area, disorder and secondary structure (coils) for each STY population</td>
<td>121</td>
</tr>
<tr>
<td>Figure 3.2</td>
<td>Rate4site conservation calculations for each STY population</td>
<td>122</td>
</tr>
<tr>
<td>Figure 3.3</td>
<td>Positive potential calculations with 5A, 10A, 25A and 40A radii around oxygen of hydroxyl-group on STY found at an interface or surface</td>
<td>124</td>
</tr>
<tr>
<td>Figure 3.4</td>
<td>Total potential energy calculations by FoldX for STYs at the interface and surface of protein structures</td>
<td>126</td>
</tr>
<tr>
<td>Figure 4.1</td>
<td>Workflow of how datasets were constructed for feature selection, training and testing for machine-learning</td>
<td>147</td>
</tr>
<tr>
<td>Figure 4.2</td>
<td>Recursive operating characteristic curves of machine learning-based predictors built with different features on hold-out tests</td>
<td>157</td>
</tr>
<tr>
<td>Figure 4.3</td>
<td>Summary predictive performance of four published phosphorylation site predictors against different datasets</td>
<td>160</td>
</tr>
</tbody>
</table>
List of Tables

Table 1.1. Summary of some available site localization algorithms 39
Table 1.2 Example multiple sequence alignment that highlights different scenarios and varying degrees of conservation 47
Table 2.1a. Phosphopeptide PSMs statistics for different search engines at a Q-value threshold of 0.05 81
Table 2.1b. Phosphopeptide PSM-to-peptide statistics for different search engines at a Q-value threshold of 0.05. 82
Table 2.2. Site localisation performance statistics for four different localisation algorithms 93
Table 2.3. Consensus site localization statistics applied to different activation methods 95
Table 3.1. Summary of datasets for assessment in this study. 111
Table 3.2. SILAC studies from which functional and less-functional phosphorylation sites were extracted from 112
Table 4.1. Summary of datasets for assessment in this study. 143
Table 4.2a. Summary of training sites used in MUSITE 143
Table 4.2b. Summary of functional datasets following removal of kinase-substrate annotated phosphorylation sites from the MUSITE training set. 144
Table 4.2c. Summary of functional datasets following removal of all sites overlapping with MUSITE training set. 144
Table 4.3. Summary of CARET procedure to build optimal predictive model with different machine learning algorithms 150
Table 4.4a. Performance statistics of machine learning-based predictors constructed from biological features (disorder, accessibility, secondary structure and conservation) on held-out test set 156
Table 4.4b. Performance statistics of machine learning-based predictors constructed from top-10 orthologous amino acid indices on functional datasets on held-out test set 156
Table 4.4c. Performance statistics of machine learning-based predictors using all features with GPS and MUSITE predictions (meta-predictors) on held-out test set 156
Table 4.5 Performance statistics of published predictors on functional datasets 163
Table 4.6a. Performance statistics of machine learning-based predictors constructed from biological features (disorder, accessibility, secondary structure and conservation) on functional datasets 164
Table 4.6b. Performance statistics of machine learning-based predictors constructed from top-10 orthologous amino acid indices on functional datasets 164
Table 4.6c. Performance statistics of machine learning-based predictors using all features with GPS and MUSITE predictions (meta-predictors) on functional datasets 165
Table 4.6d. Performance statistics of machine learning-based predictors using conservation and GPS predictions only as features in order to build a functional predictor 165
Abstract

Name: Dave Chi Hoo Lee
Institute: The University of Manchester
Degree title: Doctor of Philosophy
Thesis title: Informatics tools analysis for the analysis and assignment of phosphorylation status in proteomics
Date: September 2014

Presently, progress in the field of phosphoproteomics has been accelerated by mass spectrometry. This is not a surprise owing to not only the accuracy, precision and high-throughput capabilities of MS but also due to the support it receives from informaticians whom allow the automated analysis; making the task of going from a complex sample to a statistically satisfactory set of phosphopeptides and corresponding site positions with relative ease. However, the process of identifying and subsequently pinpointing the phosphorylation moiety is not straightforward and remains a challenging task. Furthermore, it has been suggested that not all phosphorylation sites are of equal functional importance, to the extent that some may even lack function altogether. Clearly, such sites will confound the efforts towards functional characterisation. The work in this thesis is aimed at these two issues; accurate site localisation and functional annotation.

To address the first issue, I adopt a multi-tool approach for identification and site localisation; utilising the different underlying algorithms of each tool and thereby allowing an orthogonal perspective on the same tandem mass spectra. Doing so enhanced accuracy over any single tool by itself. The power of this multi-tool approach stemmed from its ability to not predict more true positives but rather by removal of false positives.

For the second issue, I first investigated the hypothesis that those of functional consequence exhibit stronger phosphorylation-characteristic features such as the degree of conservation and disorder. Indeed, it was found that some features were enriched for the functional group. More surprisingly, there were also some that were enriched for the less-functional; suggesting their incorporation into a prediction algorithm would hinder functional prediction. With this in mind, I train and optimise several machine-learning algorithms, using different combinations of features in an attempt to (separately) improve general phosphorylation and functional prediction.
Declaration

No portion of the work in this thesis has been submitted in support of an application for another degree or qualification to this or any other university of institute of learning.
Copyright

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns any copyright in it (the “Copyright”) and s/he has given The University of Manchester the right to use such Copyright for any administrative, promotional, educational and/or teaching purposes.

ii. Copies of this thesis, either in full or in extracts, may be made only in accordance with the regulations of the John Rylands University Library of Manchester. Details of these regulations may be obtained from the Librarian. This page must form part of any such copies made.

iii. The ownership of any patents, designs, trade marks and any and all other intellectual property rights except for the Copyright (the “Intellectual Property Rights”) and any reproductions of copyright works, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property Rights and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property Rights and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://www.campus.manchester.ac.uk/medialibrary/policies/intellectual-property.pdf), in any relevant Thesis restriction declarations deposited in the University Library, The University Library’s regulations (see http://www.manchester.ac.uk/library/aboutus/regulations) and in The University’s policy on presentation of Theses.
Acknowledgements

Firstly, I would like to thank all of my supervisors for providing the opportunity for me to conduct this research and supporting me throughout the entirety of the PhD.

I would also like to thank members (present and past) from both the Eyers and Hubbard labs: Francesco Lanucara, Stephen W. Holman, Maria Belen-Gonzalez Sanchez, Christopher Gray, Craig Lawless, Paul Blakely and Andy Jarnuczak for their expertise, advice, knowledge and all-round willingness to aid me.

I would also like to thank Nick Gresham and members of the IT services who were vital in maintaining the local server. I am very grateful to the BBSRC and EPSRC for funding my research.

Of course, I thank family and friends for their support throughout the PhD.

Finally, I dedicate this work to my cousin Eva who lost her son to meningitis in April 2014.
The Author

Following my BSc (Hons) in Biochemistry at The University of Liverpool and MSc in Bioinformatics at The University of Manchester, I remained in Manchester and embarked on a PhD in Systems Biology.
Abbreviations

ADA Adaptive Boosting
ASA Accessible surface area
AUC Area under the curve
CARET Classification and Regression Training
CID Collision-induced dissociation
HCD Higher-energy c-trap dissociation/higher-energy collisional dissociation
HTP High-throughput
HTP+ Assumed true phosphosite determined by HTP-experiments from literature
ETD Electron transfer dissociation
ETDSA Electron-transfer dissociation with supplemental activation
FDR False Discovery Rate
FLR False Localisation rate
FN False Negative
FP False Positive
iPRG Proteome informatics research group
KL Kullback-Liebler
KSP Kinase-specific predictor
LC Liquid chromatography
LTP Low-throughput
MCC Matthews correlation coefficient
MD Mascot Delta
ML Machine learning
MS Mass spectrometry
MS/MS Tandem mass spectrometry
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSA</td>
<td>Multi-stage activation</td>
</tr>
<tr>
<td>nKSP</td>
<td>Non-kinase-specific predictor</td>
</tr>
<tr>
<td>nnet</td>
<td>Neural Network</td>
</tr>
<tr>
<td>P-Site</td>
<td>Phosphorylation site</td>
</tr>
<tr>
<td>NP-Site</td>
<td>Non-phosphorylation site</td>
</tr>
<tr>
<td>PBL</td>
<td>Probability-based localiser</td>
</tr>
<tr>
<td>PEP</td>
<td>Posterior error probability</td>
</tr>
<tr>
<td>PP</td>
<td>ProteinProspector</td>
</tr>
<tr>
<td>PPI</td>
<td>Protein-protein interaction</td>
</tr>
<tr>
<td>PPV</td>
<td>Positive predictive value</td>
</tr>
<tr>
<td>PSM</td>
<td>Peptide-spectrum-match</td>
</tr>
<tr>
<td>PSP</td>
<td>PhosphositePlus</td>
</tr>
<tr>
<td>PSP+</td>
<td>Manually curated functional phosphosites by PhosphositePlus</td>
</tr>
<tr>
<td>PTM</td>
<td>Post-translational modification</td>
</tr>
<tr>
<td>Q+</td>
<td>Assumed functional phosphosites from SILAC experiments</td>
</tr>
<tr>
<td>Q-</td>
<td>Assumed less-functional phosphosites from SILAC experiments</td>
</tr>
<tr>
<td>rf</td>
<td>Random Forest</td>
</tr>
<tr>
<td>ROC</td>
<td>Receiver operating characteristic</td>
</tr>
<tr>
<td>SED</td>
<td>Search-engine difference</td>
</tr>
<tr>
<td>SN</td>
<td>Sensitivity</td>
</tr>
<tr>
<td>SILAC</td>
<td>Stable isotope labelling by amino acids in cell culture</td>
</tr>
<tr>
<td>SLiP</td>
<td>Site Localisation In Peptide</td>
</tr>
<tr>
<td>Spec</td>
<td>Specificity</td>
</tr>
<tr>
<td>TN</td>
<td>True Negative</td>
</tr>
<tr>
<td>TP</td>
<td>True Positive</td>
</tr>
</tbody>
</table>
Chapter 1: Introduction

Phosphorylation is one of the major post-translational modifications within a biological system that is involved in dictating, either directly or indirectly, many cellular processes. This has made the task of annotating the phosphoproteome, which involves localising/pinpointing each site position and assigning a functional role, of great interest. Presently, the analytical technology leading the charge in phosphoproteomics is mass spectrometry (MS) which is capable of routinely collecting vast quantities of data with incredible precision and accuracy. The task of identifying phosphopeptides and subsequently localising the precise positions of phosphosites from MS-data is designated to the arsenal of software tools developed by a dedicated informatics community. However, despite best efforts by developers, accurate localisation is still challenging.

Furthermore, while localisation accuracy is important, it has also been suggested that not all sites necessarily have any functional consequence. This presents yet another issue for the wet-lab biologist; even if the sites collected are truly phosphorylated, which are the ones that should be selected for functional validation?

This chapter will examine the analytical and computational processes that are commonly employed in the field of MS-based phosphoproteomics and the subsequent task of functional assignment.

In this chapter, the details regarding MS-based informatics (identification and localisation of phosphopeptides) were submitted as a review by myself and two co-authors: my supervisor Simon J. Hubbard and Andrew R. Jones from the University of Liverpool.
1.1 Phosphorylation

Phosphorylation is a post-translational modification event where a phosphate group is transferred onto select residues of protein sequences. In eukaryotic systems, these privileged residues are predominantly serine (S), threonine (T) and tyrosine (Y) (with an estimated occurrence of ~90:~10:~1% respectively) (Junger & Aebersold, 2014) while in prokaryotes histidine (H) and aspartic (D) and glutamic (E) acids are also valid substrates (Kennelly & Potts, 1996).

This event itself is regulated by the dynamic balance in activity between kinase and phosphatase enzymes which attach and remove the moiety respectively (Fig. 1.1). The process of attaching a phosphate onto a substrate site is a careful and stringent process where the consequences of wrong placement could be dire. As such, to ensure careful phosphorylation of substrates, the catalytic domain of a kinase binds to substrates with high specificity; relying on complementary contacts between amino-acid side chains along the binding surfaces of the kinase and substrate (Ubersax & Ferrell, 2007).

Crucially, phosphorylation is deeply embedded in the cellular system architecture. Its role, either directly or indirectly, is regulatory where it acts to relay external stimuli to specific and carefully-evolved cascades of events that evoke appropriate biological responses. The prevalence of phosphorylation in signalling and regulatory processes has been widely cited to affect 30% of the proteome (Cohen, 2000) though this number could indeed be far larger due to the increasing volume and quality of data that is becoming available to the phosphoproteomic community (Junger & Aebersold, 2014).

Its integral role in mediating multiple crucial biological events (Mayr et al., 2001; Nishi et al., 2011; Schlessinger, 2000; Olsen et al., 2006; Willems et al., 1999) has led to major effort into methods and
technologies capable of fully elucidating the 'phosphoproteome'; the site-level resolution of phosphorylation of the proteome under a given condition. Given the advances in analytical capability in recent years and growing interest in mapping the mechanistic detail of intracellular signalling pathways, phosphoproteomics has become an active field with many groups attempting to find candidate targets for kinases and phosphatases of interest.

Although there are many individual studies based around antibodies to monitor and validate phosphorylation status of individual sites (Mandell et al., 2003), for high-throughput and genome-wide studies the analytical method of choice is typically mass spectrometry (Olsen et al., 2006; Bodenmiller et al., 2010). This has been driven by the ever improving instrumentation and associated analytical chemistry, in particular mass spectrometry augmented by MS-compatible phosphorylation enrichment techniques, which have made this strategy a key player in the field; enabling vast quantities of data to be produced with high (and constantly improving) resolution and precision, exemplified by some landmark studies (Olsen et al., 2006; Bodenmiller et al., 2010, Olsen et al., 2010; Wu et al., 2011). The increased prevalence of such studies has resulted in a deluge of data which in turn has supported the rapid expansion of content contained within phosphorylation databases (Diella et al., 2004; Heazlewood et al., 2008; Hornbeck et al., 2012; Saowski et al., 2013); providing the necessary data required for training and optimising numerous predictive informatic tools (Trost & Kusalik, 2011).
Figure 1.1. Phosphorylation and de-phosphorylation of proteins. Phosphorylation (red spot) of the protein (black star) is carried out by a kinase (1) while removal of the phosphate (2) is achieved with a phosphatase.
1.2 Mass spectrometry

The advent of mass spectrometry (MS) to the field of proteomics is a blessing due to its high-throughput capabilities with high precision and accuracy along with a supporting arsenal of informatics to process the vast quantities of data; therefore achieving a comprehensive and highly resolved snapshot of the system with ease.

There are many MS-based strategies available that enable the confident identification of proteins (and modifications therein) within complex mixtures. Hyphenated workflows such as ‘LC-(ESI)-MS/MS’ are particularly popular because each aspect can be optimised according to user specifications thereby allowing you to apply the high-powered attributes of MS to a variety of biological problems.

Most conventional MS workflows are ‘bottom-up’ meaning that they operate from a peptide-centric point of view. While the notion of using peptides for protein identification as oppose to the proteins themselves is counter-intuitive, the decision to use peptides (for now) is that of convenience with regards to MS-compatibility and, equally important, the informatics to analyse the subsequent data.

Briefly, using LC-MS/MS as an example, following enzymatic digestion (most commonly with trypsin) of proteins contained within a complex mixture (such as cell lysate), the resulting peptides are subjected to liquid chromatography (LC). LC acts upon some inherent (physicochemical) property of the peptides, allowing for a reduction in sample complexity by separating and concentrating non-identical and identical species respectively; leading to improved sequence identification.

Eluting peptides are made MS-compatible via electrospray ionisation (ESI), a process that transforms peptides into gaseous ions (Fenn et al., 1989) with specific mass-to-charge ratios (m/z) that are
measured with a detector.

The subsequent task of deducing the probable peptide sequence is achieved by tandem MS (MS/MS) where the gaseous ‘precursor’ (peptide) ions are subjected to collision events that cause them to fragment (in a relatively consistent and predictable way) and produce corresponding ‘product’ ions. These product ions return clues of the underlying sequence and together with the precursor information, enables you to correctly deduce the identity of the peptide.

These precursor and corresponding product ions are recorded and processed through informatics tools which return a set of candidate sequences according to the ions observed.

In phosphoproteomics, each core step (LC, MS/MS and informatics) has been modified from the conventional workflow above to deal with the typical issues surrounding the capture, behaviour and analysis of phosphorylation and therefore maximise identification and localisation of phosphopeptides and sites respectively. I discuss in more detail some of the confounding factors of phosphorylation with respect to each phase and describe methods that have been developed to address each issue.

1.3. Enrichment of phosphorylated peptides through liquid-chromatography

A major hurdle in phosphoproteomics is the low abundance of phosphorylated species relative to their cognate counterpart. In MS, this is problematic for a number of reasons, most notably, the low stoichiometric ratio makes it difficult to distinguish signal of true phosphorylated peptides from background noise and other low-intensity ions caused by chemical contaminants ( Aebersold & Goodlett, 2001). Therefore, all MS-based phosphoproteomic studies use an LC system designed to recover as much of the phosphopeptide population as possible.
In the case of phosphoproteomics, the obvious difference between phosphorylated and non-phosphorylated peptides is the presence of phosphate which decreases the charge state of the peptide. Therefore, to discriminate between the two and enrich the former, LC-based enrichment approaches exploit this property and the most popular at present for carrying this out are derived from ion-exchange (IEX) (Fig. 1.2) and metal-based techniques; immobilised metal affinity (IMAC) and metal-oxide affinity chromatography (MOAC).

1.3.1. Ion Exchange Chromatography

The presence of phosphates reduces the charge of a phosphopeptide relative to its native peptide form by approximately 1 per phosphate present under acidic conditions. This charge difference between phosphopeptides and their non-phosphorylated counterpart fits the pre-requisite for separation using ion-exchange chromatography (IEX).

Here, beads within the chromatographic column are derivatised with positively or negatively charged compounds. Samples running through the column are separated based on their interaction with the charges on these beads (electrostatic attraction encourages retention whilst repulsion reinforces elution). There are two general types of IEX: cationic and anionic where the beads within the column are negatively and positively charged respectively.

In context of enriching phosphopeptides, both anionic and cationic IEX (Gafken & Lampe, 2006) have been employed. It should be noted that a potential drawback to using anionic IEX is retention of non-phosphorylated peptides rich in acidic residues.

1.3.2. Immobilised Metal Affinity Chromatography

The next LC method commonly used in MS-based phosphoproteomics is immobilised metal affinity chromatography (IMAC). Here, metal ions (bound onto the chromatography column) interact with
their target functional moieties through the formation of co-ordinate covalent bonds (Dunn et al., 2010).

In context of phosphorylation, interaction with the phosphate group on phosphopeptides has been quite successful through the use of a variety of metals, the most common of which is Fe$^{3+}$. There are however a few issues surrounding the use of IMAC-based phosphorylation methods. The most relevant one is that, as with IEX, retention of highly acidic peptides can occur leading to ambiguity further downstream in data analysis (Dunn et al., 2010).

1.3.3. Metal-Oxide Affinity Chromatography

An alternative metal-based approach for phosphopeptide enrichment is to use metal-oxide affinity chromatography (MOAC). Here, binding takes place between the metal oxide and phosphate through bidentate ligand binding (two lone pairs of electrons are used) (Leitner, 2010).

MOAC is described as superior to IMAC for phosphopeptide enrichment because it displays a greater selectivity for phosphorylated peptides compared to acidic peptides leading to a much higher recovery (Pinske et al., 2004).

This higher selectivity was first noticed when DHB (2,5 – dihydroxy-benzoic acid) was incorporated in the loading buffer and subsequent washing steps (Larsen et al., 2005). It is hypothesised that the presence of DHB (and potentially other aromatic-containing carboxylic acids compounds) invoked the displacement of acidic non-phosphorylated peptides. However, this was only true in the case of MOAC as DHB was also removing bound phosphorylated peptides in IMAC. Therefore, the reason why MOAC is preferred over IMAC is not because MOAC binds fewer acidic non-phosphorylated peptides relative to IMAC but because it can remove them more efficiently with minimal loss of real phosphorylated peptides (Larsen et al., 2005).
Figure 1.2. Sample separation by ion-exchange chromatography. Sample contents are separated based on their interaction with beads embedded in the chromatography column (large blue negative spheres). Here, positively charged components will remain on column via electrostatic attraction with the negatively charged beads while the negatively charged components of the sample will elute due to repulsion from beads.
1.4. Activation methods and alternatives more suited for fragmentation of phosphopeptides in tandem mass spectrometry

Upon entry in the mass spectrometer, phosphopeptides are ionised to form gaseous ions with a specific mass-to-charge ratio \( m/z \) that is measured by a detector.

Identifying the phosphopeptide is achieved (as briefly introduced above) by MS/MS. During MS/MS, precursor ions are ‘activated’ whereby they are bestowed with energy through collisions with purposely selected molecules that cause the afflicted precursors to fragment in a particular way and produce product ions. The most critical aspect in this process is that the product ions produced are both consistent and reproducible with respect to the precursor ion and activation method applied. It is this fragmentation reproducibility that provides the basis for sequence identification by all informatic tools (Section 1.5).

The default type of activation method for many MS instruments is collision-induced dissociation (CID) where low-energy inert atoms are physically collided with target precursor ions. This event characteristically causes fragmentation of the CO - NH amide bond of peptides resulting in the formation of ‘b’- and ‘y’-type ions (Fig. 1.3). Unfortunately, identification of phosphopeptides is not well-suited for this activation method due to the labile nature of the phosphate group itself, which has a lower activation energy threshold than an amide bond and is frequently lost as a neutral species from the precursor ion during fragmentation in the gas phase. While this characteristic dominant phosphate neutral loss ion is useful in identifying that the precursor is indeed a phosphopeptide, it reduces the level of fragmentation in the peptide backbone, which then yields fewer sequence-informative ions therefore hampering the ability to distinguish the peptide sequence in subsequent informatic analysis. Furthermore, a loss of 98 Da is typically observed from
serine and threonine residues, whilst phosphotyrosine remains intact. Given that phosphorylated serine and threonine constitute the overwhelming majority of the phosphoproteome, this issue is highly problematic and has consequently prompted the search for alternative methods more suited phosphopeptides (Boersema et al., 2009).
Figure 1.3. Proposed nomenclature by Roepstorff for fragmentation of peptide sequences. Fragmentation along precise points of the peptide backbone have been designated their own specific code. Characteristic collision-induced dissociation results in the formation of b- and y-ions which fragment at CO-NH bond. Electron-transfer dissociation tend to result fragmentation of NH-CH bonds. The subscript number aside each fragment ion represents where in the peptide that ion derives from. B and C ion numbering is orientated along the peptide with respect to the N-terminal while their corresponding Y and Z are to the C-terminal. Figure adapted from Roepstorff (1984)
1.4.1 Alternative collision-based activation methods

Two relatively recent collision-based activation methods that are of utility to phosphoproteomics are multistage activation (MSA) (Schroeder et al., 2004; Ulintz et al., 2008) and High-energy Collisional Dissociation (HCD) (Olsen et al., 2006; Nagaraj et al., 2010).

In MSA, a precursor ion following loss of the phosphorylation group is purposely (and simultaneously) re-selected for fragmentation. Without the burden of the labile phosphate, potential energy from collision events can instead be directed towards the generation of sequence-informative ions for successful identification.

HCD is an activation method that achieves greater mass accuracy and resolution than any other collision-based technique along with the ability to access the full spectrum range and therefore identify more fragment ions with greater confidence (Olsen et al., 2006; Nagaraj et al., 2010). In context of phosphorylation, more energy is applied in HCD compared to conventional CID (Jedrychowski et al., 2011) which allows dissociation pathways with higher energy requirements than that of the neutral loss of phosphate, such as fragmentation of the amide bonds, to be accessed more readily. Both MSA and HCD have been shown to have positive benefits for phosphopeptide identification (Nagaraj et al., 2010).

1.4.2. Electron-transfer-dissociation activation

A real landmark in the field of phosphoproteomics was the introduction of electron-transfer dissociation (ETD) (Syka et al., 2004). This method involves the transfer of an electron to precursor (cat)ions via a radical anion which invokes the dissociation of amide bonds (Syka et al., 2004; Kim et al., 2012). The advantage over collision-based methods that ETD (and its later derivatives) provides is the capability to bypass the labile-biased dissociation of phosphorylation, allowing the
modification to remain intact and available for localisation by informatic tools. There was, however, a potential downside from an informatics perspective to the use of ETD; search-engines were not optimised to process this type of data.

All algorithms were originally developed with collision-based fragmentation methods in mind and were made ETD-compatible by adapting the algorithms to look for c- and z-type product ions produced by ETD. Unfortunately, the idiosyncrasies of ETD-derived data such as dominant unreacted and/or charge-reduced precursor ion peaks and ETD-exclusive neutral losses were unknown and therefore nullified the identification and subsequent localisation performance benefits of ETD (Good et al., 2009; Good et al., 2010).

Fortunately, many ETD-related behaviours have now been better characterised allowing notable improvements to be made in this area (Good et al., 2009; Good et al., 2010; Sadygov et al., 2009; Baker et al., 2010). As a result of the combined efforts of the MS-community, ETD-based methods are an excellent complementary approach to their collisional-counterparts in phosphoproteomic experiments (Molina et al., 2007; Swaney et al., 2007).

1.5. Peptide identification and localisation of phosphorylation sites

The basis of identification and the subsequent task of phosphosite localisation is the structural interrogation of characteristic product ions formed during activation. For peptide identification, elucidation from MS/MS spectra relies on informatic tools such as database-search-engines (Perkins et al., 1999; Craig & Beavis., 2003; Eng et al., 1994; Eng et al., 2013; Chalkley et al., 2008; Geer et al., 2004; Cox et al., 2011), de novo sequencers (Frank & Pevzner, 2005) and, to a lesser-extent, search tools based on spectral libraries (Hu & Lam, 2012).
1.5.1. Database search-engines

For the majority of data-dependent acquisition strategies in bottom-up proteomics where the identity of the peptides is not known *a priori*, database search-engines are the most widely used in phosphoproteomic analyses.

Briefly, these algorithms take a sequenced proteome and perform an *in silico* digestion, mimicking the enzyme used during digestion (such as trypsin), to generate the entire repertoire of peptides (and therefore their corresponding MS/MS) that could theoretically exist. Experimental MS/MS are directly compared against their most likely candidate theoretical peptides according to a range of factors including (but not limited to) precursor m/z, tolerance range of the MS instrument, instrument-type, activation method and intensity of the matched fragment ions. The best sequence candidate is chosen according to how well matched, again within tolerance thresholds, the experimental and candidate theoretical product ions are.

A limitation of database search-engines is their ability (or rather lack of) to identify novel sequences. Because identification by database search-engines is dependent upon the set of sequences (database), it would not be possible to detect any sequences that lie outside this set.

1.5.2. Alternative sequence identification methods (*de novo* sequencers and spectral libraries)

In the event that the proteome of the desired organism is not available, *de novo* sequencers are the next best option for sequencing MS/MS. These algorithms, through implementation of knowledge related to fragmentation, are computational embodiments of a manual annotator; they interrogate MS/MS in search of characteristic features that can lead to successful identification such as
fragmentation 'ladders' where the distance between each rung corresponds to the mass of a residue with/without a PTM (Frank & Pevzner, 2005).

Spectral libraries are the newest addition to the family of peptide identification algorithms (Lam, 2011). As mentioned above, MS/MS are reproducible and are a function of sequence and activation method applied and it is this reproducibility that forms the basis behind identification by spectral libraries; every species will have a characteristic MS/MS 'fingerprint'.

Spectral libraries consist of representative MS/MS which, like database-search-engines, are contrasted with the experimental MS/MS where the best matching is inferred to be the correct sequence. Unlike database-search-engines, spectral libraries are theoretically superior because of the increased discriminatory power derived from idiosyncratic gas-phase fragmentation behaviours contained within the characteristic fingerprints that cannot be replicated in a generic theoretical spectrum.

However, the limitation of spectral libraries lies in their assembly; not only do they require MS/MS with known identifications a priori but they also need a sufficient number of them to capture the characteristic features and therefore construct an accurate representative MS/MS. Many resources are available, namely NIST and PeptideAtlas (Deutsch et al., 2008) which have pre-made libraries available for use in addition to spectral library-related software to build your own in-house versions. Of note, dedicated phosphopeptide spectral libraries are available from PeptideAtlas.

1.5.3. Assessing statistical significance of phosphopeptide identifications

Candidate peptide spectrum matches (PSMs) generated from any sequence identification algorithm
will contain false matches by chance, and proteomics as a field has generated a variety of means to assess the levels of potential false discovery, specifically in sets of PSMs. Presently, the most widely used approaches compute a false discovery rate (FDR) or FDR-related statistic such as q-values (Kall et al., 2009).

FDR-based statistics allow the user to control for the expected number of false positive identifications. Empirical FDRs are computed through the target-decoy strategy where MS/MS are searched against the sequence database and a 'decoy' version of it (Elias & Gygi, 2007; Higdon et al., 2005). The formation of the 'decoy' set is most commonly done via direct reversal for simplicity, but other methods exist (e.g. randomly shuffled), and is either concatenated with the target database or run separately. The precise structure of how to build the appropriate FDR model has been subjected to vigorous assessment to find an optimal strategy (Higdon et al., 2005; Jeong et al., 2012) though no consensus exists so far.

Regardless of the FDR strategy employed, all MS-based phosphoproteomic experiments usually conduct these calculations using the entire set of available PSMs in the experiment. This poses a very important question to the validity of identifying the phosphopeptides (or any other PTM) which has recently been discussed in the literature (Fu & Qian, 2014). Here, it is postulated that because the aim of phosphoproteomic studies is to identify phosphopeptides, the FDR calculations should formally consider the sub-group containing the phospho-population. The core reasoning behind this hypothesis is that the underlying physicochemical properties of the phosphorylated peptides are different from the non-phosphorylated; leading to different characteristic fragmentation behaviours (such as dominant neutral loss of the phosphate described previously) and by extension phospho-specific and non-phospho-specific scoring distributions. Using the entire set of identifications will therefore lead to deviations from the true error rate; resulting in possible over- or underestimation
of false discovery rates, and so potentially loss of correct identifications or inclusion of false positives respectively. As of yet, such a consideration has not been implemented into the identification scheme of any of the popular software tools, and hence phosphopeptide dataset FDRs should be treated with care.

1.5.4. More advanced informatic approaches to improve identification

The arsenal of informatic tools available for analysing data from phosphoproteomic studies could be considered simultaneously as a curse and blessing. With so many tools and their own unique algorithms, one can acquire substantially different results from the same mass spectral data. This was shown in previous studies to affect protein identification in proteomic experiments (Searle et al., 2008), but the outcome of the iPRG 2010 study (http://www.abrf.org/index.cfm/group.show/ProteomicsInformaticsResearchGroup.53.htm) suggests it might be more severe in the phosphoproteomic realm. However, the variety of different underlying algorithms applied by each search engine offers a parsimonious way to take on this challenge where interrogating the same data from different, orthogonal perspectives provides a simple but robust solution. As noted, this concept has been demonstrated many times in traditional proteomics studies (Searle et al., 2008; Jones et al., 2009), and has also recently been shown to reduce of false positive identifications in phosphoproteomics by combining the output from multiple informatic tools (Soderholm et al., 2014).
1.6. Site Localisation

Site localisation is a far more challenging task compared to identification because it relies on the presence of specific fragment ions in the product ion spectrum that are characteristic of a given candidate site. Furthermore, the phosphate group must be intact on these fragments to unambiguously localise the site, which has already been noted as a potentially confounding problem. To make the problem even worse, site localisation becomes considerably more difficult when candidate sites are found in close proximity in the peptide sequence, generating fewer discriminatory ions. As a final testament to difficulty of the problem, even experienced manual curators can disagree with each other when given the same MS/MS spectrum and the known sequence, shown in Figure 1.4.
Figure 1.4. Ambiguity in site assignment of phosphopeptides. The phosphopeptide above generates a product ion spectrum from which it is challenging to unambiguously determine the true site determining ions. In this particular case, two b ions highlighted in green boxes are consistent with serine at position 7 in the peptide being modified, or alternately, the threonine at position 9 could be modified yielding a characteristic y9 ion (green box, lower panel). Experts inspecting the spectrum were divided on which is the most likely interpretation. The possibility that both peptides were present is also not excluded, since they would have the same precursor ion mass-to-charge value. This figure was adapted from the iPRG 2010 study which assessed the conformity of multiple participant groups on phosphoproteomic analysis when given the same phosphoproteomic dataset (http://www.abrf.org/index.cfm/group.show/ProteomicsInformaticsResearchGroup.53.htm)
1.6.1. Site localisation algorithms

There are two classes of localisation algorithms available to the public: probability-based localisers (PBLs) and search-engine difference scores (SEDs) (Chalkley & Clauser, 2012).

1.6.1.1. Probability-based localisers

The origin of many PBL tools stems from algorithms originally designed to process MS$^3$ mass spectra (Olsen & Mann, 2004), which were subsequently then applied to the PTM problem (Olsen et al., 2006). The algorithm designed by Olsen and Mann formulates the localisation problem as a binomial probability calculation, attempting to calculate a probability for each candidate phosphosite. This is derived from the following equation (Eq. 1.1a), where $k$ is the total number of intact phosphorylated ions successfully matched to theoretical ones for a candidate site, $n$ is the total number of ions possible and $p$ is the ‘peak depth’. The peak depth is the number of the top most intense fragment ions considered in each 100 $m/z$ unit bin across the spectrum and represents the probability of randomly matching a peak. This probability is subsequently logarithmically transformed into the ‘PTM Score’ (Eq. 1.1b).
\[ P(x) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} \]

**Equation 1.1a.** Binomial probability interpretation of annotating candidate phosphosites. Variables \( n \), \( k \) and \( p \) refer to total annotatable ions, ions successfully annotated and the probability of random matching respectively. In Olsen and Mann’s original model, \( p \) equates to 0.04.

\[ PTM \ Score = -10 \log_{10}(P(x)) \]

**Equation 1.1b.** Olsen and Mann’s ‘PTM Score’ (2004). A logarithmic transformation of the probability that the sites annotated from eq. 1.1a.

Following on from this, the Ascore algorithm continued to develop a probabilistic approach and has become arguably the most well-known gold-standard site localiser in the field (Beausoleil et al., 2006). Briefly, the Ascore is composed of two core-phases. The first follows Olsen's original model, except a cumulative binomial probability is calculated and peak-depth is selected automatically (to maximise the discrimination between true and false sites). The logarithmic transformation of the binomial calculations are termed 'Peptide Scores' instead of 'PTM score' used by Olsen (Olsen et al., 2004). The second applies the same basic workflow to the first, but with one critical adaptation that makes it more suitable for the localisation problem; by using solely the 'site-determining ions' (SDIs). These are ions that are exclusive to the phosphopeptide-isoform in question and therefore allow for unambiguous localisation. The Ascore is then computed as the difference in peptide score between the first- and second-ranked site candidates.
Since then there have been several localisation algorithms based or building upon this general principle, summarized in Table 1.1. This includes SLoMo (Bailey et al., 2009), whose most important contribution to this field was that it was the first ET-compatible localizer, highlighting the utility of alternate activation methods for phosphopeptide determination. Another algorithm is PhosphoRS (Taus et al., 2011) which develops the concept of peak depth determination further by allowing it to vary according to how sparsely/densely populated regions of a MS/MS spectrum are.

1.6.1.2. Search-engine difference scores

Assessment of peptide-spectral-matches (PSMs) to determine the most likely match is a ubiquitous process carried out by all search engines, considering inferred sequences in rank ordered lists to assign confidence to the candidates. A key principle embodied in the first automatic spectrum search tool, SEQUEST, has been exploited for phosphoproteomic localization purposes too; namely that the top hit should score significantly higher than the second ranked hit if it is truly correct. The higher the quality, the greater the score difference and more confident the identification (or in this case, localization). Search-engine difference (SEDs) scores are computed in the situation where multiple sites are possible for a given modification and the first and second ranked candidates are PTM isomers of each other.

Most of the tools available to compute this type of score are linked to particular search engines. Examples of SEDs for localisation include Mascot Delta (Savitski et al., 2011) and Site Localisation in Peptides (SLiP) (Baker et al., 2011), developed for the Mascot and ProteinProspector search engines respectively; both of which have been shown to offer good performance in distinguishing alternative phosphorylation sites based on the search engine scores (Savitski et al., 2011; Baker et al., 2011).
<table>
<thead>
<tr>
<th>Name</th>
<th>Activation methods supported</th>
<th>Notes &amp; Availability</th>
<th>Pre-requisites and/or dependencies</th>
<th>aReport alternate sites?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ascore</td>
<td>CID</td>
<td>First to implement site-determining ions. <a href="http://ascore.med.harvard.edu/">http://ascore.med.harvard.edu/</a> (Free)</td>
<td>Requires a pepXML containing PSM data and corresponding .dta files</td>
<td>No</td>
</tr>
<tr>
<td>SloMo/TuboSlomo</td>
<td>CID, ETD</td>
<td>Was the first PBL available for ETD-derived data <a href="http://massspec.bham.ac.uk/slomo/">http://massspec.bham.ac.uk/slomo/</a> (Free)</td>
<td>Requires a pepXML containing PSM data and corresponding .dta files</td>
<td>Yes (Top-2)</td>
</tr>
<tr>
<td>PhosphoRS</td>
<td>CID, ETD, HCD</td>
<td>Has been built and tested on CID-, ETD- and HCD-MS/MS</td>
<td>Version 1.0 requires a custom XML format containing both PSM information + ms/ms it was derived from. No specific search-engine that is preferred; as long as the necessary PSM details can be extracted.</td>
<td>Yes (All)</td>
</tr>
<tr>
<td>LuciPhor</td>
<td>CID, HCD</td>
<td>First algorithm to implement FLR estimate <a href="http://luciphor.sourceforge.net/">http://luciphor.sourceforge.net/</a> (Free)</td>
<td>Uses Transproteomic-pipeline (TPP) supported search engines (Mascot, X!Tandem and SEQUEST/COMET), processed via xinteract to pepXML file. Available under Linux OS, run from the command line.</td>
<td>Yes (Top-2)</td>
</tr>
<tr>
<td>MaxQuant PTMScore</td>
<td>All</td>
<td>Also includes site occupancy when quantification information is available, scored based on equation used in Olsen et al 2006 paper <a href="http://www.maxquant.org/downloads.htm">www.maxquant.org/downloads.htm</a></td>
<td>None</td>
<td>Yes (All)</td>
</tr>
<tr>
<td>Mascot Delta</td>
<td>All</td>
<td><a href="http://www.matrixscience.com/server.html">http://www.matrixscience.com/server.html</a></td>
<td>Mascot .dat files are required. Many groups have written code (including Mascot) to process .dat files.</td>
<td>Yes (All)</td>
</tr>
<tr>
<td>--------------</td>
<td>-----</td>
<td>------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------</td>
<td>----------</td>
</tr>
<tr>
<td>ProteinProspector (SLiP)</td>
<td>All</td>
<td><a href="http://prospector.ucsf.edu/prospector/mshome.htm">http://prospector.ucsf.edu/prospector/mshome.htm</a></td>
<td>Requires (free) registration on ProteinProspector webserver</td>
<td>Yes (All)</td>
</tr>
<tr>
<td>PhosphoScore</td>
<td>CID</td>
<td><a href="https://github.com/evanenter/ucsb/tree/master/school/CS167/main_project/code/PhosphoScore">https://github.com/evanenter/ucsb/tree/master/school/CS167/main_project/code/PhosphoScore</a></td>
<td>Specific to SEQUEST search-engine and explicitly requires .OUT (SEQUEST results) and .dta (peaklists)</td>
<td>No</td>
</tr>
<tr>
<td>PhosphoScan</td>
<td>CID</td>
<td>Available upon request to authors</td>
<td>Standalone tool</td>
<td>Yes</td>
</tr>
<tr>
<td>D-Score</td>
<td>All</td>
<td>Standardised localisation metric enabling comparison with other search engines. Proof-of-principle paper suggesting the applicability of universal deltas with PEP. No 'tool' is currently available.</td>
<td>Computation of posterior error probabilities for first- and second-ranked hits required for delta PEP</td>
<td>Yes (All)</td>
</tr>
<tr>
<td>PhosSA</td>
<td>CID, HCD</td>
<td><a href="http://helixweb.nih.gov/ESBL/PhosSA/">http://helixweb.nih.gov/ESBL/PhosSA/</a></td>
<td>Compatible with SEQUEST, Mascot search-engines and ProteomeDiscoverer</td>
<td>Yes (status assigned to all candidate peptides)</td>
</tr>
</tbody>
</table>

*a or at least second candidate sites is also provided. In principle, all candidate sites are reported by most PBL tools, usually up to and including a maximum of two sites per peptide.*
1.6.1.3. The status of the false localisation rate

An additional issue facing the proteomics practitioner undertaking phosphoproteomics experiments is when to believe a set of phosphosite assignments when ambiguity exists; i.e. in multi-site peptides. For identifications, the field has developed FDR-based approaches, but as Chalkley and Clauser pointed out, for site assignment we need a false localisation rate or FLR (Chalkey & Clauser, 2012). The FLR is the localisation equivalent of the FDR for peptide identification; a method to estimate, and therefore control, the proportion of falsely localised sites through a target-decoy strategy. However, it is not immediately obvious how to estimate such a rate. As noted by Chalkley and Clauser (Chalkey & Clauser, 2012), not only is the FLR sought after for this reason but also because it would allow a direct and fair comparison of different site localisation tools, and provide a universal metric to measure against (and potentially to integrate multiple tools in a principled way).

At present, there is no universally accepted method to determine the FLR. The key hurdle here being how to define the decoy population from which a background, null distribution of scores can be estimated, thereby enabling a FLR to be estimated. Currently, perhaps the best suggestion is to phosphorylate residues that do not carry this modification in nature. Here, Chalkley reasoned that appropriate decoy residues should have a similar frequency and close proximity to real phosphorylatable (STY) residues and suggests the use of proline and glutamic acid which correlate with serine and threonine respectively in the general context of these criteria (Chalkey & Clauser, 2012; Baker et al., 2011).
1.7. The prediction of phosphorylation sites

Prediction of protein phosphorylation sites is conducted by algorithms that have been optimised with carefully chosen biological and physicochemical features that are known to be enriched by phosphosites. Since the introduction of MS-based phosphoproteomics, both the number available and predictive accuracy of these predictors have been increasing (Trost & Kusalik, 2011) due to the rapidly expanding content contained within public repositories (Hornbeck et al., 2012, Diella et al., 2004, Stark et al., 2010; Heazlewood et al., 2008). Furthermore, because they have been developed using data from experiments whose objective is to find candidate functional sites, such as those modified by a given kinase, the characteristics of these sites ought to have been inherited. Critically, this makes predictors an invaluable resource when cross-referencing with a list of MS-determined phosphosites, since they can assist in finding those that adhere to the properties found in previously recorded sites; resulting in a subset of confidently localised and apparently biologically interesting sites, that can then undergo further functional validation.

1.7.1. Kinase-specific and non-kinase-specific predictors

There are two classes of phosphorylation predictors available: non-specific and kinase-specific predictors (Trost & Kusalik, 2011). Non-kinase specific predictors (nKSPs) are concerned with the phosphorylation status of candidate sites and are built on the basis that phosphosites have characteristic properties that distinguish them from the non-phosphosite population. One particular example is disorder propensity where its pioneering incorporation into a predictor led to superior performance over those without (Iakocheva et al., 2004). This stems from work which recognized that unfolded, disordered regions of proteins are frequently phosphorylated, which is consistent with one
potential role of phosphorylation in stabilizing the local structure of protein fold, and thereby affecting
its function. This could be by promoting the formation of local secondary structure such as alpha-helix,
which has been observed in such fundamental process as protein translation (Tait et al., 2010).

In contrast, kinase-specific predictors (KSPs) extend beyond predicting the phosphorylation status by
mapping predicted phosphosites to the effector kinase responsible. Prediction by KSPs is achieved
through using linear motifs (explained below in section 1.7.2.1) which computationally mimic kinase-
substrate interactions. Adherence of a candidate site and its surrounding residues to a pre-defined motif
indicates interface complementarity with the corresponding kinase, increasing the confidence that a
phosphorylation event can take place.

Development of nKSP algorithms is straightforward; requiring only a site's phosphorylation status. KSPs
algorithms are more challenging as they require sites that have been pre-annotated with effector
kinases in order to categorised and construct representative models of substrates for their respective
kinase family. The limitation here is the sparseness associated with this data in terms of kinase-classes
covered and unequal content for each class therefore limiting the number of predicted kinases and
affecting prediction accuracy respectively. Indeed, given the prevalence of intracellular
phosphorylation-mediated signalling cascades it is quite possible that indirect sites that are not directly
attributable to the parent kinase are mistakenly assumed to be direct targets. This could further
confound the predictors, and highlights the care required when designing and building training data sets
for machine learning in phospho-biology.

1.7.2. Popular phosphorylation-enriched characteristics

The success of any prediction algorithm depends on how well it can separate the positive and negative
population. There are a number of characteristics found to be enriched in real phosphorylation sites
that have subsequently been incorporated into prediction algorithms to improve their performance (Iakocheva et al., 2004; Gao et al., 2010; Dou et al., 2014). The following features are some household examples that are included in the development of many predictors.

### 1.7.2.1. Kinase-substrate linear motifs

A linear/sequence motif is a protein sequence with purposely placed residues that enable successful interaction events with other proteins (Diella et al., 2004). A prime example is the phosphate transfer event which requires complementary interaction between the surface of the effector kinase and residues surrounding the substrate site. This event is highly stringent to avoid unwittingly affecting the function of the protein, or indeed modification of an incorrect protein, in a non-intended way. Modelling this kinase-substrate specificity relationship through solved linear motifs is a powerful feature for predictors that forms the basis of all KSP predictors. From a KSP perspective, linear motifs have been used both by themselves (Obenaue et al., 2003) or as 'profiles' (position-specific-scoring-matrices) (Yaffe et al., 2001) that enable predictive models to be formed such as hidden markoff models (HMMs) (Huang et al., 2005).

### 1.7.2.2. Protein disorder propensity

It has been long assumed that protein function is predicated on the specific tertiary structure formed by the protein. It was therefore a surprise to find that, historically, in x-ray crystallography experiments aimed at elucidating protein structure that some regions of the proteins could not be determined, because they were apparently too structurally variable (Chouard, 2011). These regions are often known as 'disordered' regions and since their initial discovery have caused a paradigm shift in structural biology, leading away from the original dogmatic view that all proteins had a defined structure throughout their
entire polypeptide chain and hence had a corresponding, uniquely defined function (Chouard, 2011). Disordered regions have been found to be involved in a variety of functional processes, the most notable and relevant to this thesis is their ability to transition from a disordered-to-ordered conformation which is important for interacting with other partners (Dunker et al, 1998).

A thought experiment by Dunker and colleagues (Dunker et al, 1998) represented the mechanism by which disorder-to-order transitions mediate PPIs as a thermodynamic activation model (Fig. 1.5). In this model, the base state comprises of a disordered species and its ordered partner; with the former capable of existing in a finite number of conformations including the set complementary to a binding region on its partner. A PPI event takes place, and is successful, if the stabilising contacts between the ordered and disordered-to-ordered conformation are more favourable than the unbound states. In this thought experiment, it is postulated that the energy required to reach the appropriate 'solid'/ordered complementary state to the interacting partner corresponds to the degree of disorder in the disordered region.

Because phosphorylation has been found to be enriched in disordered regions, one mode by which it regulates function could be by influencing these transition interaction events. This notion prompted the inclusion of disorder for prediction (Iakoucheva et al., 2004) and has since remained a key feature in subsequent predictors.
Figure 1.5. 'Schulz' diagram describing energy changes for binding between one ordered protein (structures on left) and multiple partners (structures on right), each being increasingly dissimilar to the interface of the ordered protein (i-iii). Partners (i), (ii) and (iii) represent ordered, mixed degrees of order/disorder and disordered structures respectively. In order for (ii) and (iii) to bind, they must re-adjust themselves to enable complementary contact with the ordered partner interface where the amount of energy required increases as a function of base state disorder. Stripes and dots indicate order and disorder respectively.

Figure adapted from 'Protein Disorder and the evolution of molecular recognition: theory, predictions and observations' by Dunker and colleagues (1998)
1.7.2.3. Conservation

Conservation is a logical feature which assumes that residues of functional importance will be conserved in orthologues and widely applied throughout biological science to infer a candidate function relationship. The computation of conservation rates in this case for a given site relies on multiple-sequence alignments (MSA) built with orthologous protein sequences to your protein of interest (Table 1.2). The presumption here is that if a candidate site (S, T or Y) is conserved beyond expectation for any other reason, then it stands a good chance of having functional significance. Typically, residue frequency of the 20 standard amino acids in a MSA column dictates the degree of conservation: the greater the conservation, the fewer the number of mutations and *vice versa*.

<table>
<thead>
<tr>
<th>Orthologues</th>
<th>Site Position</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>a</td>
</tr>
<tr>
<td>1</td>
<td>D</td>
</tr>
<tr>
<td>2</td>
<td>D</td>
</tr>
<tr>
<td>3</td>
<td>D</td>
</tr>
<tr>
<td>4</td>
<td>D</td>
</tr>
<tr>
<td>5</td>
<td>D</td>
</tr>
<tr>
<td>6</td>
<td>D</td>
</tr>
<tr>
<td>7</td>
<td>D</td>
</tr>
<tr>
<td>8</td>
<td>D</td>
</tr>
<tr>
<td>9</td>
<td>D</td>
</tr>
<tr>
<td>10</td>
<td>D</td>
</tr>
</tbody>
</table>

*Table 1.2* Example multiple sequence alignment that highlights different scenarios and varying degrees of conservation. Columns a to f are ordered in descending degrees of conservation where the stereochemical properties become increasingly varied. Table was reproduced from review by Valdar (2002).
Methods to compute conservation initially relied on information-theory approaches (most notably Shannon’s Entropy) but have continuously evolved where each new method addresses the shortcomings of their predecessors, enabling increasingly accurate calculations.

Simpler information-theory approaches such as Shannon’s Entropy did not consider the chemical properties or entropies of distinct chemical groups where some have (more) similar geometry and chemical properties (Valdar, 2002) meaning that certain mutations were (more) acceptable. Many solutions were proposed to resolve this issue, for example, Mirny and Shakhnovich summate the individual Shannon entropies of distinct stereochemical groups (Valdar, 2002; Mirny & Shakhnovich, 1999).

Sequence redundancy was also an issue that adversely affects conservation calculations by biasing either the calculations directly or how the MSA is constructed. The latter would mean information from more distant orthologues may be lost more easily (Valdar, 2002). Weighting schemes using the phylogenetic distance between orthologous protein sequences have since been implemented in some methods (Pupko et al., 2004).

1.7.3. Using machine-learning algorithms for prediction

A predictive feature will have varying degrees of success in distinguishing phosphosites from non-phosphosites, with some more potent than others. A simple but elegant strategy to improve discriminatory power is to consider many features together. A single feature describes one aspect of phosphorylation but many presents a comprehensive view which can reveal subtle relationships between features that better distinguish the positive from the negative population; accomplishing more than any single feature alone. This can be encoded in a linear prediction system, where a combination of the features leads to a predictive metric to classify each test case. Equally, with complex interacting
biological features, non-linear systems may be necessary. Example machine-learning algorithms include support vector machines (SVMs), neural networks, randomForests and adaptive boosting.

1.7.3.1. Support vector machines

Support vector machines create a 'hyperplane', a boundary that partitions the positive and negative populations. Predictions using SVMs are based on which side of this hyperplane the predicted instance lies.

To select the optimal hyperplane and therefore the best predictor, the hyperplane is fine-tuned against support vectors (SVs); the closest positive and negative instances for a given candidate hyperplane which are the most difficult to classify correctly. The optimality criterion during SVM learning is to maximise the distance between hyperplane and SVs.

1.7.3.2. Artificial neural networks

Artificial neural networks (ANNs) are machine-learning algorithms inspired from neurons in the brain. The basis of ANNs and how they build predictive models is to discover how and which features interact to provide the most discriminatory power.

ANNs are formulated as graphs; a set of nodes that are connected with others using edges. Each node represents a function that processes and interprets the signals it receives. Should the input received be deemed sufficient then the node will ‘fire’ (if that is the operation of the node) an output signal to the following node it is connected to. Edges have designated rules on how nodes interact with each other. In neural networks, edges apply weights to these output signals prior to entry into the adjoining node. The topology of a neural network normally allocates nodes into 3 'layers' called the input, hidden and output layer (simpler networks can have just the input and output layers).
Nodes residing in the input layer accept the raw values from the set of features which are passed and fine-tuned by the hidden layer nodes before final adjustments are made by the output nodes that then return the final prediction value.

To optimise this network, ANNs re-adjust weights in order to enhance the features that do have discriminatory power over those that don’t; improving the predictive performance.

1.7.3.3. Ensemble methods

Some machine-learning algorithms take an ‘ensemble’ (collaborative) approach whose rationale is that it is challenging to build a single accurate predictive model that works universally but it is easier to generate many individual models which by themselves have modest accuracy but when formed together increase predictive performance (Schapire, 2003). There are two main categories of ensemble methods: 'bagging' and 'boosting'.

In bagging, a bootstrapping procedure is adopted where many training sets are created by resampling, with replacement, from the original set and a predictive model is built on each one. In boosting, each model is created with the same training set but with each iteration a weighting is applied to each training instance. This weighting represents the difficulty in classification; the harder it is to get a particular instance then the higher the weighting. This allows certain models to be built that concentrate on correctly classifying the more difficult cases.

The outcomes from each model are then combined together in some fashion to create a single representative outcome. In bagging, each model is given equal importance and so prediction for a given instance is the average of all the models whilst in boosting the weighted sum according to each classifier is taken for that instance.

The two algorithms I use later in my work (Chapter 3) are randomForests and Adaptive Boosting (ADA),
which fall under the bagging and boosting categories respectively.

### 1.7.3.4. Constructing a positive and negative set for predictor training

The most important aspect of training any algorithm for the purposes of prediction is arguably how to construct the positive and negative datasets. In many cases, this is relatively straightforward but this is not so simple when using experimental data with a high degree of expected error (high false positive rate). This is indeed the case with much of the data available in phosphoproteomics. While a positive definition is straightforward (a true phosphosite is one that has been experimentally validated, potentially by several independent studies), it is far more challenging to define a true negative.

The simplest definition of a negative is a site without any experimental evidence of phosphorylation. However, there could be a myriad of reasons that could explain its absence, the most plausible being that the correct conditions have not been met so the effector kinase is not active. Equally, the site might not be accessible, the peptide on which it sits is incompatible with mass spectrometry, and many others. 

Iakoucheva and colleagues (Iakoucheva et al., 2004) suggest the application of a BLASTP identity threshold in their training procedure where any initial negative sequence with a certain identity (30%) to any positives were excluded from the negative set. This criterion alleviates the issue of (near) identical sequences in close homologs that may be phosphorylated but have not yet been tested. As of yet, there is no preferred nor universally accepted method to create a negative set from phosphoproteomic data. In addition, the number of negatives will typically exceed the number of positives, often by a few orders of magnitude. This poses problems during the training procedure where the class imbalance may bias the predictive model into favouring the larger (negative) class. To resolve this issue, most published predictors use an equal (or near equal) number of negatives to positives to create (more) balanced datasets.
Figure 1.6 Example of a hyperplane separating two different classes (red and blue shapes) in support vector machines. Star shapes are support vectors used in deciding optimal position of hyperplane (dashed line) by optimising the distance between support vectors of both classes with respect to the hyperplane (arrowhead lines).
1.7.4. Using structural features for prediction

The complementary interaction between an effector kinase and substrate site is the result of specific conformations formed by each structure that allows the phosphate transfer reaction to take place. Structural-based features could be an invaluable resource as one can characterise this process with greater resolution and vigour than can be achieved by looking from a sequence-perspective alone; thereby providing a powerful source of information to augment predictors. Naturally, there have been studies that structurally characterise phosphorylation to find any properties that can be exploited.

In the work by Kitchen and colleagues (2008), phosphosites were found to locate in regions that electrostatically-stabilised a substrate site in the event it became phosphorylated and showed enrichment over non-phosphosites. Durek and colleagues (2009) evaluated specific kinase-substrate relationships where they look at the spatial arrangement of amino acids in search of structural-motifs (the 3-d upgrade of a linear-motif). It was observed that consistent improvements performance were made with their structural-motifs compared to published sequence-level predictors. Both of these examples could be envisioned to assist nKSP and KSP-type predictors respectively.

While structural information has been shown to offer discriminative power for a predictor (Blom et al., 1999), the overwhelming majority all phosphosite predictors do not make use of it and consider the prediction problem exclusively from the sequence-level, opting instead to remain using features such as disorder and conservation summarised above (Trost & Kusalik, 2011).

The main issue that prevents structural information being incorporated into all sequence-level phosphosite predictors is that phosphorylation typically occurs in disordered regions which are not easily structurally characterised; limiting the range of utility of structural features. However, with an increasing number of solved structures solved in data repositories (Berman et al., 2000) and dedicated structurally-annotated phosphorylation databases such as Phospho3d (Zanzoni et al., 2011) structural
features may become mainstream in the near future.

1.8. Assigning functional status of phosphorylation

Protein function can be described by the core three gene ontology categories: cellular component, molecular function and biological process (Lee et al., 2007). Throughout this thesis, I define a functionally important phosphosite as one that can influence any of these three aspects of protein function.

It is clear that with constantly improving experimental technologies and increasingly accurate phosphosite prediction algorithms, the number of phosphosites one is able to acquire inevitably increases. While this accomplishes a more comprehensive view of the phosphoproteome, it has been suggested that the functional importance of all sites is not equal (Lienhard, 2008). This presents a dilemma for the wet-lab biologist whose aim is functional validation of these sites: choice. How does one decide which phosphosites to focus on?

The consideration that there may be non-functional phosphosites was first raised in a letter by Lienhard (2008) who postulated that some phosphosites may be of little (or even no) functional consequence to their system through a series of counter-arguments for features that historically constitute functional importance: optimisation of ATP, low stoichiometry and site conservation.

Optimised ATP usage would be required for a system to survive but based on some basic assumptions of ATP production and average cellular content, there could be such an excess that non-intended phosphorylation (and by extension non-functional) could be afforded.
Phosphosite stoichiometry refers to the percentage a precise site within a protein population has been phosphorylated. Should there be a sufficient change in stoichiometry in response to a physiological stimulation (as measured by quantitative experiments) one can infer functionality. The precise threshold will vary from protein-to-protein but should it not be met then it is unlikely to invoke change and therefore could be deemed as non-functional. For this reason, Lienhard suggested that one possible characteristic of non-functional phosphorylation sites is low stoichiometry as it most likely presented little danger to normal cell functioning.

Also, it has been reported by Landry and colleagues (2009), who assessed the degree of conservation as a function of relative site stoichiometry and found that those of lower stoichiometry displayed a tendency to be less conserved; providing some support to the hypothesis that low-stoichiometry may relate to non-functional status.

Conservation is a powerful feature for measuring functionality where if a phosphosite is maintained in multiple orthologues then a logical inference is that it has been selected for by evolution and conserved because it serves some functional role in its system.

The argument made by Lienhard regarding the use of conservation is not against conservation itself but rather how it is interpreted. It is possible that poor conservation may be a result of other factors such as the existence of different pathways in orthologues. In this situation, a functional site in one organism will be predicted to be poorly conserved as it is not required (and therefore conserved) in its orthologues. Conversely, a truly non-functional site may be estimated as conserved because the orthologues used in its estimation were closely-related. In both cases, low and high conservation do not always necessarily mean functional or non-functional so caution should be advised as to how to interpret functionality (Lienhard, 2008).
However, while all phosphorylation sites may not be critical for function, it can be argued that no site can be considered non-functional. One suggestion is that phosphorylation is involved in the formation of negatively charged patches that electrostatically fine-tune binding events between phosphorylated proteins and their interacting partners (Serber & Ferrell., 2007). The electrostatic nature of such patches can be achieved if the phosphorylation sites occur within the same vicinity; reconciling the need for exact sites/positions to be conserved.

Throughout this thesis, I use the term ‘less’ as opposed to ‘non-’ functional because I cannot say with absolute certainty that a site is truly lacks any function.

1.8.1. Detection of functional sites through quantitative studies

To address the task of finding functional phosphosites on a system-wide scale, MS-compatible quantitative strategies (Mann, 2006) have been integrated into the conventional MS-based phosphoproteomic pipeline; combining the high-throughput capabilities, precision and accuracy attributes of MS with the ability to assign a measure of abundance to each identified phosphopeptide species.

Briefly, MS is described as being not ‘inherently’ quantitative because each peptide has its own unique physicochemical signature which is a vital component in dictating its precursor intensity signal. It is for this reason that different peptides cannot be directly compared with each other. MS-based quantification strategies rectify this issue by comparing the same molecule (therefore solving the conundrum above) via label-based or label-free quantitative strategies (Nikolav et al., 2012), the former being more relevant in phosphoproteomics.

A label is a purposely distinct modification that enables identical sequences to be distinguished from
each other. They are vital in comparative studies between the same biological systems under different conditions to observe what changes have been evoked and by how much. This is accomplished by computing the ratio between intensity measurements of the same, alternately-labelled (condition-specific), molecule which reflects the relative changes in abundance of said molecule between the different conditions.

Selection of sites whose change is believed to be of biological importance and not as a result of other factors (such as natural fluctuations) are normally chosen based around a static threshold (and often augmented with a secondary statistical threshold) (Bodenmiller et al., 2010; Dephoure et al., 2008).

This is ideal for functional phosphoproteomics where initial screening of functional phosphosites can be achieved through assessing the quantitative changes between different conditions, for example, a control and stimulated state where those displaying a (sufficient) response are inferred to have some functional role.

Many MS-quantitative labelling strategies have been developed but one of particular note is stable isotope labelling in cell culture (SILAC); a straightforward but powerful labelling strategy (Mann, 2006). Here, cell populations are grown in media containing labelled amino acids which, following a sufficient number of cell doubling cycles, progressively replace their native counterparts in the resulting cell progeny until the label is fully incorporated (Fig. 1.7a). These fully-labelled cell populations are mixed in equal concentrations, analysed by MS and ratios are computed using precursor intensities of identical (alternately-labelled) molecules to reveal relative changes in abundance (Fig. 1.7b).
Figure 1.7a. Labelling phase of a SILAC experiment. A cell culture is split into two populations (1); one remains unchanged (white) while the other is labelled (red). These cell cultures are doubled until the label in the labelled population has been fully integrated into the cell population (2).

Figure 1.7b. Experiment phase of SILAC experiment. The labelled population is subjected to a stimulus causing a specific response in the underlying cellular system. This is then compared to a control (unlabelled) population in equal concentrations to enable observation of changes in abundance between the two populations. Both figures were adapted from Ong & Mann (2007).
1.9. Overview of thesis

With an increasing number of phosphorylation sites from MS-based experiments and prediction algorithms, wet-lab based biologists are faced with two questions: which phosphosites are real and of those, which are also of functional importance (Fig. 1.8)?

1) Improving localisation accuracy of MS-based phosphoproteomic experiments

Correctly pinpointing phosphorylation sites from MS/MS has proven challenging due to disagreement between different informatic tools where, using the same MS/MS data and stringent statistical criteria, can yield different phosphopeptide identifications and phosphosite positions. This raises questions about the true phosphorylation status of a candidate substrate site and, more importantly, the quality of phosphoproteomic data contained within data repositories which are dominated by MS-derived data.

This forms the basis for my first results chapter (Chapter 2) where I focus on improving localisation accuracy to retain sites that are truly phosphorylated.

2) Predicting phosphosites that are functionally important

While localising the real phosphorylated sites is vital, another formidable problem has since appeared in phosphoproteomics; the notion that some sites may be of little functional consequence. Functional characterisation of phosphosites is an intensive process that requires considerable effort so avoiding those of less functional importance is desirable.

The prediction of functional phosphorylation sites is the subject of my second and third results chapters.

In my second results chapter (Chapter 3), I assess phosphorylation-characteristic features that are
used conventionally in predictive tools and examine whether there are any of possible utility in finding those that are functional whilst avoiding those that are not.

In my third results chapter (Chapter 4), I evaluate the performance of multiple published predictors on their ability to predict for functional and less-functional phosphosites. I compare their performance with predictors (using the features and combinations therein discussed in the second chapter) I train to determine which are suitable for general and, more importantly, functional prediction.
Figure 1.8 Overview of thesis. A hypothetical protein (blue stars) has been phosphorylated (coloured circles) following some appropriate event. Of these phosphosites, there will be some that are falsely localised due to identification and localisation errors (red) associated with informatic tools. Of those that are truly phosphorylated, some may be of little functional importance (yellow) while others are vital (green). The first phase of this thesis is aimed at the removal of the false positive phosphosites (b) and the second phase is the retention of those that are of functional importance (c).
1.10. References


Chapter 2: Benchmarking computational tools and consensus approaches for mass spectrometry-based phosphoproteomics

Mass spectrometry (MS)-based phosphoproteomics has been the main contributor of phosphorylation data to public repositories over the past decade and is responsible for the astonishing quantity of data we now have access to. These repositories are an invaluable source of information to the phosphoproteomic community so it is desirable that the contents within them are as accurate as possible.

However, it is well-known that the task of identification and localisation in MS-based phosphoproteomics is not straightforward. This is not only due to difficulty in analyzing phosphoproteomic data but it has also been found that different software tools can disagree with each other on the same MS/MS; raising questions as to whether a site is indeed real.

In this chapter, I conduct a benchmarking study to evaluate the performance of popular software tools that are routinely employed in the task of identification and localisation against a synthetic set of phosphopeptides. In addition, I also explore ways to improve upon these two tasks through combining the output of different tools.

The contents presented in this chapter are in preparation for submission.
2.1. Abstract

Phosphorylation is a key post-translational modification that mediates signal transduction in complex biological systems. Mass spectrometry (MS) is presently the most attractive strategy to study the phosphoproteome due to its high-throughput acquisition of data with accuracy and supporting informatics. Unfortunately, identification and subsequently pinpointing the precise site of phosphorylation is a challenging task. Furthermore, the performance of these computational tools responsible for analysing is seldom rigorously assessed. This poses serious concerns regarding the quality of phosphorylation data repositories whose main contributor commonly originate from MS.

In this study, the performance of popular computational tools employed for both identification and localisation are assessed using a set of synthetic phosphopeptides under different activation conditions. I find that from a conventional peptide-spectrum-match (PSM) point of view, all tools perform very well where they each are capable of achieving estimated error rates that are close to the true rate. However, upon converting from PSM-to-peptide/site (for identification and localisation respectively), the number of false positives sharply increase for all tools. In the face of this problem, I demonstrate that an ensemble approach that combines the output from multiple-tools can greatly diminish the false positive population for both PSM and PSM-to-peptide/site analyses; leading to statistically safer results.
2.2 Introduction

Phosphorylation site repositories (Hornbeck et al., 2012, Diella et al., 2004, Stark et al., 2010; Heazlewood et al., 2008) are a vital source of information for the phosphoproteomic community, for example, for the training and optimisation of prediction algorithms (Trost & Kusalik, 2011). Since the advent of mass spectrometry (MS)-based phosphoproteomics the size of these repositories has been rapidly increasing; from a few hundred sites over a decade ago (Blom et al., 1998) to well-over a hundred-thousand sites (Hornbeck et al., 2012).

There are two main aspects of MS-based phosphoproteomic data analysis: identification and localisation. Naturally, identification refers to the ability to deduce the identities of phosphopeptides and is achieved by either cross-referencing the acquired experimental MS/MS against other theoretical MS/MS or attempting to deduce the sequence directly from the MS/MS. The most common protocol to conduct identification is by database search-engines (Perkins et al., 1999; Craig & Beavis., 2003; Eng et al., 1994; Eng et al., 2013; Chalkley et al., 2008; Geer et al., 2004; Cox et al., 2011). The basis for identification by this type of algorithm is a direct comparison of acquired MS/MS against theoretical MS/MS generated from peptides with appropriate precursor ion mass-to-charge ratios (m/z) within a given error tolerance. The greater the extent of similarity between the acquired and theoretical MS/MS, the more likely that the theoretical sequence is also the identity of the peptide responsible for the acquired MS/MS.

The confidence of each peptide-spectrum match (PSMs) is marked by its designated score, which in turn can be cast as probabilities, False Discovery Rates (FDRs) or q-values depending on the statistical model the software offers you. Typically, a confidence threshold is applied to a set of candidate PSMs to limit them to a subset a fixed error rate (i.e. FDR < 0.05). Identification of phosphopeptides presents additional complexity to the problem as the phosphate group is usually labile in the gas phase (depending on which amino acid side group is modified) and hence there are three possible ions generated; intact fragments, a neutral loss of HPO₃ (80 Da), or a neutral loss of H₃PO₄ (98 Da) (Boersema et al., 2009).
Once a group of candidate phosphopeptides have been identified, the potentially more challenging step of unambiguously assigning the precise phosphosite can be attempted. Two main approaches have been proposed for site localisation; search-engine difference scores (SEDs) and probability-based localisers (PBLs) (Chalkley & Clauser, 2012). SED algorithms follow the logic that a greater difference in score between the top-site candidate and the second is indicative of the top being correct. Hence, the SED score is defined as the difference in search-engine score between phosphorylation isoforms where the top and alternate form are ranked sequentially. Examples include the Mascot Delta (MD) derived from Mascot Ion Scores (Savitski et al., 2011) and SLiP derived from ProteinProspector peptide scores (Baker et al., 2011). PBL algorithms are standalone and operate post-search-engine using MS/MS and their corresponding PSM sequence, but not the assigned scores. Instead, PBLs re-annotate the MS/MS and apply their own scoring function. While several PBLs now exist (Olsen et al., 2006; Beausoleil et al., 2006; Bailey et al., 2008; MacLean et al., 2008; Taus et al., 2011; Fermin et al., 2013) and each has their own unique scoring function, the probabilistic model commonly employed is the cumulative binomial model. Difference score based-PBLs, such as Ascore (Beausoleil et al., 2006) and SLoMo (Bailey et al., 2008), provide a local score derived from site-determining ions (SDIs); ions that can discriminate between the top-two candidate site isoforms. PhosphoRS applies an alternative approach where it uses the relative proportion of scores amongst candidates attributed to the top ranked candidate (Taus et al., 2011).

Unfortunately, even with such an array of tools on both fronts, the accurate detection of phosphorylation sites remains a challenging task. In 2010, the Proteome Informatics Research Group (iPRG) conducted a controlled study on phosphosite localisation tools available at that time (http://www.abrf.org/index.cfm/group.show/ProteomicsInformaticsResearchGroup.53.htm#943).

In this study, several groups were provided with the same set of MS/MS and asked to analyse and return the set of statistically significant identifications, and if possible, confidently localised sites. No restrictions on informatic tools and strategies were placed on the groups regarding how the data could be processed; only that groups returned the set of identified phosphorylated peptide sequences and localised sites which they deemed to be
true. The goal of the study was to assess the degree of conformity between groups with regards to how the data is handled and, more importantly, the identification and localisation outcomes. It was found that there was on average 57% agreement on the set of phosphorylated peptides identified between groups when considered on a pairwise basis. However, perhaps more worryingly, this level of agreement decreased substantially to ~38% consensus agreement when considering site localisation.

Another challenge when assessing site localisation data is that there is no widely accepted method of determining the false localisation rate (FLR) to allow user-controllable error rate selection. More specifically, the reason for this is difficulty lies in defining the null distribution from which one can estimate statistically significant localisation scores (Chalkley & Clauser, 2012). Presently, the best proposed method to estimate the FLR is through adaptation of the target-decoy strategy whereby non-standard phosphorylatable residues are theoretically phosphorylated and play the role of decoys (Baker et al 2012; Fermin et al., 2013). The FLRs for all these tools are determined via synthetic phosphopeptides.

With ambiguity stemming from the use of different tools and lack of a uniform FLR, it is no wonder why accurate localisation is challenging. This has worrying implications on the quality of data in public repositories storing MS-based phosphoproteomic data where even with stringent statistical cut-offs and rigorous analysis, the true number of false positives may well be underestimated.

In this chapter, using a publicly available synthetic phosphopeptide dataset where the site positions are known a priori (Savitski et al., 2011), I assess the performance of several popular computational tools utilised for both identification and localisation. My aim is to raise awareness of the informatic challenges attendant in phosphosite identification and offer suggestions regarding suitable analytical workflows. My results are consistent with previous studies, and I demonstrate how a simple consensus-tool based strategy can correctly identify and localise on par with the best algorithms but with greatly reduce error rates.
2.3 Materials and Methods

2.3.1. Synthetic phosphopeptide datasets

Mascot generic files (.mgf) generated by Savitski et al., 2010 were downloaded from the Kuster group FTP (ftp://ftp.lrz.de/transfer/proteomics/MDscore/). Those used in this study were the CID-, MSA-, HCD-, ETD- and ETDSA-MS/MS generated on an LTQ Orbitrap XL ETD instrument. This dataset consisted of 247 unique phosphosite species separated into five fractions to avoid overlap of isomeric species.

2.3.2. File formats

All analyses were performed where possible on database search-engine results that were acquired as, or converted to, .pep.xml format. ProteinProspector output was obtained natively as .pep.xml and X!Tandem (.xml) output was converted by Tandem2XML from the Trans-Proteomic Pipeline (TPP) suite (Keller et al., 2005). Mascot .dat files were used instead of pep.xml, as the necessary decoys required for statistical processing by QVality (Kall et al., 2009) could not be created using the Mascot2XML converter in the TPP suite. However, we were able to generate decoy PSMs for QVality in standard Mascot .dat files, which were independently processed.

2.3.3. Database search-engines for sequence inference

Mascot (v.2.2.06), ProteinProspector (v.5.10.1) and X!Tandem (v.13.09.01) were used. All database searches adhered to the parameters specified previously (Savitski et al. 2010). Briefly, precursor and product ion tolerance thresholds were set at 10 ppm and 0.5 Da respectively for all activation methods but HCD where the product ion tolerance was reduced to 0.02 Da. Instrument set-up (which defines the set of ions used for scoring) were ESI-
Ion-Trap for collision-based and ESI-ETD-Trap for ET-based methods on both Mascot and ProteinProspector webservers. For X!Tandem, $b$- and $y$- ions were selected for collision-based and $c$- and $z$- ions for ET-based methods. Carba-iodomethylation of cysteine residues was set as a fixed modification and oxidation of methionine, phosphorylation of serine, threonine and tyrosine and acetylation of N-termini of proteins were chosen as variable modifications.

The main deviation from Savitski's protocol here was the sequence database searched; the human international protein index (IPI) database. While this database could be added to our local Mascot server and X!Tandem search-engines, the ProteinProspector resource has a pre-defined set of databases for searching against, of which, the IPI was not amongst them.

To account for this and maximise conformity across all search engines, I select the human proteins within the Swiss-Prot (version 06/2013) database of ProteinProspector and acquired the equivalent set of proteins from UniprotKB. These versions were essentially identical, with 20,257 or 20,264 sequences from ProteinProspector's version and UniprotKB respectively. Both databases were verified to contain all 247 synthetic peptide sequences. In all cases, target and decoy database searches were run separately. Mascot and X!Tandem decoy searches were run against the reversed database whereas ProteinProspector was limited to a randomised database.

It should be noted in my searches that there were a small number of cases where the un-phosphorylated and/or incomplete forms of the synthetic peptide set were observed. These were removed and not taken into consideration throughout this study.
2.3.4. Peptide-spectral-match and peptide-centric analyses

All analyses in this study have been conducted from both a PSM-centric and PSM-to-peptide viewpoints. For the PSM-to-peptide analyses, PSMs were converted to peptides by taking the highest-scoring instance given the peptide sequence, modification status and fraction it belonged to.

2.3.5. Computing statistically significant identification with QVality

For all identifications, FDR-based q-values were estimated with QVality (Kall et al., 2009). This algorithm requires the distribution of scores generated from identifications against the forward (target) and reverse or random (decoy) databases, taking score sets derived from the independent search-engines. The scoring metric supplied to QVality here was a peptide score, transformed from individual search engine expectation values ($E_i$), defined as:

$$\text{PeptideScore}_i = -10 \log_{10}(E_i)$$

A q-value threshold of 0.05 was applied as the statistically significant cut-off in this chapter.

2.3.5.1. Determining the true q-values of peptide-spectral-matches

Access to a dataset of known synthetically-generated peptides supports the calculation of the “true” underlying false discovery rate (FDR). Here, I used only the target PSMs, defining correct (true) and incorrect (false) PSMs from a priori knowledge of the true sequence and phosphorylation state. FDRs were calculated as the ratio of false PSMs to true PSMs, when descending the list of sorted PSMs, assigning q-values to PSMs as the minimal false discovery rate at which a PSM appears in the list.
2.3.6. Applying site localisation algorithms and scores

This work compares both PBL and SED site localisation scores. SLoMo (Bailey et al., 2009) and PhosphoRS (Taus et al., 2011) PBL algorithms were used in this study. PhosphoRS was downloaded from the author’s webpage (http://cores.imp.ac.uk/protein-chemistry/download/) and SLoMo was obtained upon request from the authors (Dr Helen J. Cooper, personal communication). SLoMo difference and PhosphoRS peptide probabilities were used as localisation metrics. Mascot Delta and SLIP scores were computed from Mascot .dat and ProteinProspector .pep.xml files respectively. These were defined as the difference in Mascot Ion Score and ProteinProspector Peptide Score between the top and alternative site isoforms. Hence for the two SED approaches, a difference score of zero indicates no preference for any candidate phosphorylation site.

2.3.7. Criteria for assessing performance of sequence identification and phosphosite localisation

For peptide sequence identification, a PSM was considered true if both sequence and number of phosphorylation sites were identical to the known sequence in the synthetic peptide dataset. Correct localisation requires both sequence and phosphosite position to be correct. Oxidation status of methionine residues was ignored for the purposes of assessing PSM correctness. Assessment of phosphosite localisation was only conducted on the correct set of phosphopeptides that were confidently identified (0.05 q-value threshold computed by QVality) by both of the Mascot and ProteinProspector search-engines.

2.3.8. Consensus q-values across multiple search-engines and a voting approach for site localisation algorithms

In addition to evaluating the performance of published tools, I also implement a set of heuristic rules to assist in improving identification and localisation. For peptide identification, a consensus q-value was assigned to each PSM instance. Here, the geometric mean of q-values of identical PSMs (derived from the same identification,
MS/MS and fraction) across all three search engines were computed. PSMs originating from a single search-engine were removed and not considered, reasoning they would be less likely to be true, at the expense of sensitivity.

Consensus site localisation was conducted with two rules; for a given MS/MS spectrum, no single site localiser returned a score of zero, and at least one had a score greater than an assigned threshold value. Threshold values chosen reflect a false localisation rate of 0.05 originating from their respective publications. For cases where the appropriate cut-off could not be determined or safely assumed, a value of 13 was assigned which equates a p-value of 0.05. A voting approach was applied by counting the number of ‘votes’ each site localizer had for each site candidate.

Final empirical FDR statistics were computed for both PSM- and peptide/site-centric perspectives, for identification and localisation respectively, in the standard way, as the ratio of false positive to false and true positive discoveries.
2.4. Results

2.4.1. Assessment of search-engine performance on identifying phosphorylated sequences for different activation methods

For successful identification, the algorithms behind a search-engine should implicitly discriminate between sequence-informative ions and noise when comparing real spectra to theoretical ones. With this in mind, the majority of search-engines have been designed, tested and optimised on CID-activated MS/MS and then subsequently adapted to accommodate MS/MS derived from newer activation methods; namely, the ET-based (Syka et al., 2004). As such, an assessment on activation-specific performance of search-engines is important to determine the relative performance of the different search-engines for different activation methods for phosphopeptide identification. Savitski and colleagues conducted an initial investigation into this problem where they assessed the performance of the Mascot Delta (MD) site localisation metric under the above activation conditions on different instruments. Ultimately, they demonstrate both the ease and potency of MD for site localisation (Savitski et al., 2011).

Here, I assessed the performance of three search engines, Mascot, ProteinProspector and X!Tandem, with MS/MS derived from three collision-based activation methods; CID, MSA and HCD and two electron-transfer driven; ETD and ETDSA. In order to assess performance one needs to consider both the sensitivity (coverage) and accuracy (precision), since it is desirable to identify as many true phosphopeptides as possible at a low error rate, as well as to consider estimated and real False Discovery Rates (FDRs). It is worth noting here that given the variability in peptide ionisation and fragmentation properties in any given set, coupled to variability in mass spectrometry performance, complete coverage from this 247 phosphopeptide dataset is impossible. There are also other issues to note when assessing these results; one can limit the results to PSMs of phosphopeptides only, and not consider unmodified peptides, which will affect performance statistics (Fu & Qian, 2013). Finally, one can consider results from either a PSM- or peptide-centric perspective when estimating FDRs. The main
findings presented in this chapter follow the conventional approach of PSM-level statistics and taking into consideration all identified PSMs in order to follow a real-life scenario.

Starting with the collision-based methods, it is observed that the unique number of identifications made is ranked (from fewest to most) as CID, MSA and HCD for all search-engines (Fig. 2.1, Table 2.1a). Of the three search-engines, X!Tandem appears to slightly under-perform under CID activation conditions at a fixed q-value where it identifies fewer of the real positives along with a high number of incorrect identifications leading to high FDRs from both PSM-centric and PSM-to-peptide viewpoints.
Figure 2.1. Phosphopeptide identifications for different activation methods and search engines. Stacked bar plots at q-value 0.05 illustrating the number of correct, incorrect and missed identifications for each search-engine and activation method. Incorrect identifications have been further separated into non-phosphorylated and phosphorylated peptides. Of the collision-based methods, search-engines made the highest number of correct phosphopeptide identifications for HCD MS/MS. X!Tandem is slightly under-performing with fewer correct identifications under CID and MSA conditions. For the ET-based methods, search-engines identified more correctly for ETD while X!Tandem identified the least incorrect overall.
Table 2.1a. Phosphopeptide PSMs statistics for different search engines at a Q-value threshold of 0.05

PSM-level FDR statistics for different search engine combinations, based on q-value estimates computed by Quality at a 0.05 threshold are presented. In addition, the number and percentage of FPs that are non-phosphopeptides is also reported. Search engines considered are M: Mascot, X: X!Tandem, PP: ProteinProspector. ‘Combined’ refers to the superset of all PSMs identified by all three search engines (many of which will be supported by only a single search engine), and ‘AtLeastTwo” refers to PSMs supported by two or more search engines in agreement (same spectrum, same peptide sequence). In the last two cases, the estimated FDRs are not calculated as they are post-processed from individual search engine datasets.

<table>
<thead>
<tr>
<th>Activation Method</th>
<th>Search method</th>
<th>TP (Estimated)</th>
<th>FP (Estimated)</th>
<th>TP (Real)</th>
<th>FP (Real)</th>
<th>Non-phosphorylated FPs (%)</th>
<th>FDR (Estimated)</th>
<th>FDR (Real)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CID</td>
<td>Mascot</td>
<td>1158</td>
<td>92</td>
<td>1121</td>
<td>37</td>
<td>26 (70.3)</td>
<td>0.074</td>
<td>0.032</td>
</tr>
<tr>
<td></td>
<td>X!Tandem</td>
<td>886</td>
<td>66</td>
<td>832</td>
<td>54</td>
<td>34 (63)</td>
<td>0.069</td>
<td>0.061</td>
</tr>
<tr>
<td></td>
<td>PP</td>
<td>1251</td>
<td>91</td>
<td>1215</td>
<td>36</td>
<td>19 (52.8)</td>
<td>0.068</td>
<td>0.029</td>
</tr>
<tr>
<td></td>
<td>Combined</td>
<td>1248</td>
<td>216</td>
<td>1169</td>
<td>79</td>
<td>51 (64.6)</td>
<td>NA</td>
<td>0.063</td>
</tr>
<tr>
<td></td>
<td>AtleastTwo</td>
<td>1173</td>
<td>10</td>
<td>1153</td>
<td>20</td>
<td>15 (75)</td>
<td>NA</td>
<td>0.017</td>
</tr>
<tr>
<td>MSA</td>
<td>Mascot</td>
<td>1154</td>
<td>92</td>
<td>1090</td>
<td>64</td>
<td>39 (60.9)</td>
<td>0.074</td>
<td>0.055</td>
</tr>
<tr>
<td></td>
<td>X!Tandem</td>
<td>780</td>
<td>66</td>
<td>734</td>
<td>46</td>
<td>36 (78.3)</td>
<td>0.078</td>
<td>0.059</td>
</tr>
<tr>
<td></td>
<td>PP</td>
<td>1129</td>
<td>91</td>
<td>1086</td>
<td>43</td>
<td>25 (58.1)</td>
<td>0.075</td>
<td>0.038</td>
</tr>
<tr>
<td></td>
<td>Combined</td>
<td>1157</td>
<td>218</td>
<td>1071</td>
<td>86</td>
<td>56 (65.1)</td>
<td>NA</td>
<td>0.074</td>
</tr>
<tr>
<td></td>
<td>AtleastTwo</td>
<td>1080</td>
<td>8</td>
<td>1057</td>
<td>23</td>
<td>16 (69.6)</td>
<td>NA</td>
<td>0.021</td>
</tr>
<tr>
<td>HCD</td>
<td>Mascot</td>
<td>1208</td>
<td>85</td>
<td>1145</td>
<td>63</td>
<td>34 (54)</td>
<td>0.066</td>
<td>0.052</td>
</tr>
<tr>
<td></td>
<td>X!Tandem</td>
<td>1126</td>
<td>92</td>
<td>1068</td>
<td>58</td>
<td>25 (43.1)</td>
<td>0.076</td>
<td>0.052</td>
</tr>
<tr>
<td></td>
<td>PP</td>
<td>1203</td>
<td>75</td>
<td>1141</td>
<td>62</td>
<td>12 (19.4)</td>
<td>0.059</td>
<td>0.052</td>
</tr>
<tr>
<td></td>
<td>Combined</td>
<td>1299</td>
<td>216</td>
<td>1194</td>
<td>105</td>
<td>34 (32.4)</td>
<td>NA</td>
<td>0.081</td>
</tr>
<tr>
<td></td>
<td>AtleastTwo</td>
<td>1196</td>
<td>23</td>
<td>1165</td>
<td>31</td>
<td>12 (38.7)</td>
<td>NA</td>
<td>0.026</td>
</tr>
<tr>
<td>ETD</td>
<td>Mascot</td>
<td>973</td>
<td>58</td>
<td>938</td>
<td>35</td>
<td>13 (37.1)</td>
<td>0.056</td>
<td>0.036</td>
</tr>
<tr>
<td></td>
<td>X!Tandem</td>
<td>523</td>
<td>32</td>
<td>508</td>
<td>15</td>
<td>7 (46.7)</td>
<td>0.058</td>
<td>0.029</td>
</tr>
<tr>
<td></td>
<td>PP</td>
<td>981</td>
<td>58</td>
<td>930</td>
<td>51</td>
<td>15 (29.4)</td>
<td>0.056</td>
<td>0.052</td>
</tr>
<tr>
<td></td>
<td>Combined</td>
<td>1011</td>
<td>132</td>
<td>959</td>
<td>52</td>
<td>19 (36.5)</td>
<td>NA</td>
<td>0.051</td>
</tr>
<tr>
<td></td>
<td>AtleastTwo</td>
<td>939</td>
<td>3</td>
<td>930</td>
<td>9</td>
<td>4 (44.4)</td>
<td>NA</td>
<td>0.01</td>
</tr>
<tr>
<td>ETDSA</td>
<td>Mascot</td>
<td>704</td>
<td>43</td>
<td>673</td>
<td>31</td>
<td>23 (74.2)</td>
<td>0.058</td>
<td>0.044</td>
</tr>
<tr>
<td></td>
<td>X!Tandem</td>
<td>281</td>
<td>17</td>
<td>263</td>
<td>18</td>
<td>7 (38.9)</td>
<td>0.057</td>
<td>0.064</td>
</tr>
<tr>
<td></td>
<td>PP</td>
<td>918</td>
<td>56</td>
<td>877</td>
<td>41</td>
<td>16 (39)</td>
<td>0.057</td>
<td>0.045</td>
</tr>
<tr>
<td></td>
<td>Combined</td>
<td>774</td>
<td>110</td>
<td>716</td>
<td>58</td>
<td>33 (56.9)</td>
<td>NA</td>
<td>0.075</td>
</tr>
<tr>
<td></td>
<td>AtleastTwo</td>
<td>708</td>
<td>3</td>
<td>701</td>
<td>7</td>
<td>7 (100)</td>
<td>NA</td>
<td>0.010</td>
</tr>
</tbody>
</table>
Table 2.1b. Phosphopeptide PSM-to-peptide statistics for different search engines at a Q-value threshold of 0.05.

<table>
<thead>
<tr>
<th>Activation Method</th>
<th>Search method</th>
<th>TP (Estimated)</th>
<th>FP (Estimated)</th>
<th>TP (Real)</th>
<th>FP (Real)</th>
<th>Non-phosphorylated FPs (%)</th>
<th>FDR (Estimated)</th>
<th>FDR (Real)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CID</strong></td>
<td>Mascot</td>
<td>110</td>
<td>39</td>
<td>88</td>
<td>22</td>
<td>15 (68.2)</td>
<td>0.26</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>X!Tandem</td>
<td>114</td>
<td>53</td>
<td>78</td>
<td>36</td>
<td>23 (63.9)</td>
<td>0.32</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td>PP</td>
<td>114</td>
<td>47</td>
<td>90</td>
<td>24</td>
<td>12 (50.0)</td>
<td>0.29</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td>Combined</td>
<td>150</td>
<td>126</td>
<td>94</td>
<td>56</td>
<td>34 (60.7)</td>
<td>NA</td>
<td>0.37</td>
</tr>
<tr>
<td></td>
<td>AtleastTwo</td>
<td>99</td>
<td>9</td>
<td>91</td>
<td>8</td>
<td>6 (75.0)</td>
<td>NA</td>
<td>0.08</td>
</tr>
<tr>
<td><strong>MSA</strong></td>
<td>Mascot</td>
<td>139</td>
<td>48</td>
<td>98</td>
<td>41</td>
<td>25 (61.0)</td>
<td>0.26</td>
<td>0.30</td>
</tr>
<tr>
<td></td>
<td>X!Tandem</td>
<td>124</td>
<td>55</td>
<td>88</td>
<td>36</td>
<td>26 (72.2)</td>
<td>0.31</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td>PP</td>
<td>126</td>
<td>54</td>
<td>98</td>
<td>28</td>
<td>13 (46.4)</td>
<td>0.30</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td>Combined</td>
<td>166</td>
<td>139</td>
<td>97</td>
<td>69</td>
<td>43 (62.3)</td>
<td>NA</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>AtleastTwo</td>
<td>109</td>
<td>8</td>
<td>97</td>
<td>12</td>
<td>9 (75.0)</td>
<td>NA</td>
<td>0.11</td>
</tr>
<tr>
<td><strong>HCD</strong></td>
<td>Mascot</td>
<td>191</td>
<td>37</td>
<td>158</td>
<td>33</td>
<td>19 (57.6)</td>
<td>0.16</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>X!Tandem</td>
<td>188</td>
<td>51</td>
<td>155</td>
<td>33</td>
<td>14 (42.4)</td>
<td>0.21</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td>PP</td>
<td>198</td>
<td>57</td>
<td>156</td>
<td>42</td>
<td>9 (21.4)</td>
<td>0.22</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td>Combined</td>
<td>231</td>
<td>129</td>
<td>165</td>
<td>66</td>
<td>18 (27.3)</td>
<td>NA</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td>AtleastTwo</td>
<td>179</td>
<td>10</td>
<td>163</td>
<td>16</td>
<td>6 (37.5)</td>
<td>NA</td>
<td>0.09</td>
</tr>
<tr>
<td><strong>ETD</strong></td>
<td>Mascot</td>
<td>152</td>
<td>35</td>
<td>131</td>
<td>21</td>
<td>9 (42.9)</td>
<td>0.19</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td>X!Tandem</td>
<td>89</td>
<td>19</td>
<td>75</td>
<td>14</td>
<td>7 (50.0)</td>
<td>0.18</td>
<td>0.16</td>
</tr>
<tr>
<td></td>
<td>PP</td>
<td>153</td>
<td>37</td>
<td>126</td>
<td>27</td>
<td>9 (33.3)</td>
<td>0.20</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td>Combined</td>
<td>170</td>
<td>88</td>
<td>132</td>
<td>38</td>
<td>14 (36.8)</td>
<td>NA</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td>AtleastTwo</td>
<td>136</td>
<td>3</td>
<td>130</td>
<td>6</td>
<td>2 (33.3)</td>
<td>NA</td>
<td>0.04</td>
</tr>
<tr>
<td><strong>ETDSA</strong></td>
<td>Mascot</td>
<td>133</td>
<td>34</td>
<td>108</td>
<td>25</td>
<td>17 (68.0)</td>
<td>0.20</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td>X!Tandem</td>
<td>74</td>
<td>17</td>
<td>59</td>
<td>15</td>
<td>7 (46.7)</td>
<td>0.19</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>PP</td>
<td>145</td>
<td>44</td>
<td>116</td>
<td>29</td>
<td>10 (34.5)</td>
<td>0.23</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>Combined</td>
<td>161</td>
<td>91</td>
<td>111</td>
<td>50</td>
<td>28 (56.0)</td>
<td>NA</td>
<td>0.31</td>
</tr>
<tr>
<td></td>
<td>AtleastTwo</td>
<td>113</td>
<td>3</td>
<td>109</td>
<td>4</td>
<td>4 (100.0)</td>
<td>NA</td>
<td>0.04</td>
</tr>
</tbody>
</table>
Of the collision-based conditions, it is clear that all search-engines identified significantly more of the synthetic phosphopeptide population from HCD conditions. Not only are more correctly identified, it is also observed that the number of incorrect identifications is on par with CID and MSA. However, of the incorrect identifications, it appears that a larger proportion of these are phosphopeptides (Fig. 2.1; Table 2.1a & Table 2.1b).

Comparison of Mascot and ProteinProspector, from a PSM-centric perspective on this particular dataset, shows that ProteinProspector offers similar performance for both estimated and true FDRs for all collision-based methods. However, from a PSM-to-peptide view ProteinProspector tends to make more incorrect identifications compared to Mascot (Table 2.1b) and also made the highest number of incorrect phosphopeptide identifications under HCD.

Overall, given the similar precision-recall curve profiles and fewer incorrect identifications, the search-engines most suitable for CID and MSA conditions, according to this synthetic peptide dataset, are Mascot and ProteinProspector (Figure 2.2). All three search-engines performed equally well under HCD conditions.

Electron-transfer driven methods are orthologous to collision-based methods, and this class of fragmentation helps circumvent the issue of phosphate neutral loss that predominates CID-activated MS/MS, thereby increasing the allocation of ion current towards sequence-characteristic fragments and so improving both identification and localisation performance. In this study, ET-based methods outperformed CID and MSA by correctly identifying more phosphopeptides and fewer incorrect. HCD acquired more correctly phosphorylated peptides compared to the ET-based methods but ETD made fewer incorrect identifications at the same statistical threshold.

The main trends from collision-based analyses are replicated; Mascot and ProteinProspector are superior to X!Tandem for both ETD and ETDSA. At the PSM-level Mascot is outperforming ProteinProspector under ETD conditions where the latter makes more incorrect identifications (Table 2.1a). The calculated FDRs for ETDSA shows while almost equal, ProteinProspector correctly identifies far more PSMs which results in more correct
identifications at the PSM-to-peptide level. However, this is at the cost of more incorrect which raises the error rate showing that ProteinProspector is slightly underperforming compared to Mascot (Table 2.1b).

This data also highlights the well-known issue that peptide FDRs calculated from PSM FDRs are normally much larger (Jeong et al., 2012), yielding real FDRs between ~20-30% (Table 2.1b). This trade-off is apparent in the precision-recall profiles for these search-engines in Figure 2.2, which suggests there is little to choose between Mascot and ProteinProspector for ETD in terms of phosphopeptide identification, though apparently ProteinProspector handles ETDSA data slightly better.

Overall, Mascot and ProteinProspector perform consistently well throughout and in nearly all cases X!Tandem underperforms to some degree on this dataset. It should be cautioned that this measure of X!Tandem’s performance may be specific to this phosphopeptide dataset, but nevertheless, this is a representative test of how most users would run these software tools and therefore highlights the difficulty in optimising searches for phosphoproteomic identification.
Figure 2.2. Peptide-centric precision-recall plots of search-engines under all activation methods. Mascot and ProteinProspector have similar performance throughout and consistently outperform X!Tandem under all but HCD conditions.
2.4.1.1. True false discovery rate calculations deviate from the intended estimated rates

The ability to control the estimated number of false positives using an FDR method has become a *de facto* standard when acquiring statistically acceptable identifications. This is a key issue as failure to do so will lead to false negative, or worse, false positive identifications. Having access to a synthetic set here allows the true rate to be computed; allowing an assessment on if, and to what extent, the deviation of the intended from the true rate is.

From a PSM-level (Table 2.1a) perspective, each search-engine for all collision-based methods resulted in near or underestimation of the true rate at an empirically estimated threshold of 5% (Table 2.1a & Fig. 2.3), demonstrating the effectiveness of target-decoy strategies. However, because of the potential redundant nature associated with PSM-level calculations whereby multiple high-quality identical species can contribute to improving the identification performance at the same FDR threshold, it is important to re-consider the statistically acceptable PSMs from a PSM-to-peptide point of view.

Here, it is found that converting of PSMs into peptides and recalculating the true FDR led to a substantially increased error rate; reaching between 17% and 30% (Table 2.1b).

For the ET-driven methods, the same observation is made where true FDRs are underestimated from a PSM-level perspective (Fig. 2.3) but upon conversion to peptides and recalculation the true peptide FDRs are above 5%. However, unlike the collision-based, this increased error rate is lower and ranged between 13% and 20%; supporting the notion that ET-based methods have improved identification capabilities (in addition to localisation) in phosphoproteomics than their collision-method counterparts (Molina *et al.*, 2007; Swaney *et al.*, 2006).
Figure 2.3. Comparison of estimated and true q-values for each search-engine under all activation conditions. Overall, from a PSM-perspective, the estimated rate adheres close to the true rate for most search-engine and activation condition combination within statistically interesting ranges (0 – 0.05). Mascot and ProteinProspector perform the best of the three search-engines whose estimated identification rates adhere the closest to the true rates.
2.4.1.2. Consensus Heuristics improves phosphopeptide identification over any single search-engine under all activation methods

The results presented demonstrate that for any activation method and search-engine applied, the estimated statistical rate (from a PSM-perspective) is near or underestimates the true error rate. However, what is also revealed is that when considering from a PSM-to-peptide perspective, this error rate increases dramatically; resulting in more false positives for all experimental and informatic permutations above with this phosphopeptide dataset.

One popular approach to address this issue is to employ multiple search-engines, for example, as part of a proteomic pipeline (Jones et al., 2009; Keller et al., 2005; Searle, 2008). The key principle here is that the true identity of a PSM is more convincing when there is agreement between different search-engines, especially if each had their own unique algorithm thereby providing an orthologous point of view on the same MS/MS.

The potential improvement from such an approach is underpinned by the data shown Figure 2.4, comparing the overlap in identifications made by individual search-engines. The majority of true (correct) phosphopeptide identifications are consistently made by two or all three search-engines, as exemplified by the CID and ETD cases shown in Figure 2.4. Similarly, as expected, a substantial number of incorrect identifications originate from a single search-engine. Eliminating these from further consideration removes a substantial fraction of false positive results. Indeed, as observed in Figure 2.4, most of the false positives are also non-phosphopeptides and additional removal of these leads to a low real FDR. For example, in CID only 2 false phosphopeptides are identified at a PSM q-value < 0.05 by two or more search engines, whilst 86 true identifications were made in equivalent fashion, yielding a true peptide FDR of 2.3%.

Although several probabilistic and more sophisticated approaches exist for the integration of multiple search engines results (Jones et al., 2009; Keller et al., 2002), I demonstrate here that the same proof-of-principle holds too for phosphopeptide identifications. I calculated the geometric mean q-value for all PSMs matched by two or more search engines, retaining consensus PSMs with q-values up-to 0.05. These are shown as the “AtLeastTwo” rows in Table 2.1a and Table 2.1b for each activation method. In addition, I include the full consensus approach
(“Combined”) where single search engine PSMs were not excluded. The ‘AtLeastTwo’ approach consistently reports a low number of incorrect identifications whilst retaining a large number of the correct true PSMs relative to any single search-engine under all activation conditions and imposed cut-offs (Table 2.1b, Fig. 2.4).
Figure 2.4 (a - b). Venn diagrams of CID (a)- and ETD (b)-activated PSM-to-peptide-centric identifications. The Venn diagrams show the number of sequence-unique peptide identifications reported by individual search engines and combinations thereof, for correct, incorrect, and incorrect (phosphorylated only) peptide populations. The search engines shown correspond to Mascot (M), ProteinProspector (PP) and X!Tandem (X). In the correct population, a considerable number of true phosphopeptides are detected by multiple search-engines as opposed to the incorrect set where both the total and phosphorylated-only incorrect identifications are almost exclusive for individual search engines.
2.4.2. Phosphorylation site localisation

Once a set of confidently identified phosphopeptides is available, the next task is to assign the location of the phosphosite. This is trivial for single site peptides (i.e. with a single serine) but frequently there are multiple candidate sites that are in close proximity to each other. There are two main classes of algorithms designed specifically for localising site positions; search-engine derived scores (SEDs) and probability-based site localisers (PBLs).

To assess solely localisation performance, the set of correctly identified PSMs for each activation method both Mascot and ProteinProspector passing the 0.05 FDR-based q-value threshold were retained. These constrained sets of correct PSMs supports a direct comparison between the MascotDelta (MD) and SLiP algorithms (SEDs) and PhosphoRS and SLoMo (PBLs).

2.4.2.1. Comparative performance of site localisation algorithms under each activation approach

Each activation approach will employ a different mechanism to induce the fragmentation of peptide precursor ions into their constituent product ions. Naturally this means that an algorithm (identification or localisation) will underperform on MS/MS derived from a method it was not designed or tested on. SEDs inherit the efforts made by their parent search-engines which aim to account for idiosyncrasies associated with the instrument and activation methods that generate the MS/MS being tested. PBLs on the otherhand are independent of the search-engine output and instead apply their own, arguably simpler, approaches to peak annotation and scoring.

Here, I assessed the performance and suitability of PBLs and SEDs under each activation condition and apply appropriate cut-offs to generate FDR statistics, aiming to identify favourable pairings of localisation algorithm and activation methods. Insight into this question would allow one to make a more informed decision into which localiser to use.
Amongst the SED-based localisers tested on this synthetic peptide dataset, ProteinProspector’s SLiP score outperforms the Mascot Delta (MD) under all activation methods except for MSA and HCD, with lower error rates at both PSM- and, more importantly, site-levels. The performance difference is most evident for ET-based activation, especially ETD, where SLiP achieves over a four-fold lower FDR at the PSM- and peptide-levels.

Comparison of PBL-based algorithms shows that more sites are localised by PhosphoRS than SLoMo under all activation methods. However, while PhosphoRS is capable of localising more sites, the cost is at a much greater FDR at both PSM and site-levels. SLoMo outperforms PhosphoRS for ET-based approaches, where it achieves an FDR of 0 for both ETD and ETDSA.

Deciding on whether it would be preferable to use one localiser type over the other is less clear. With the exception of the SEDs confidently localising fewer PSMs under CID, SEDs and PBLs have a comparable performance between each other throughout from both PSM- and site-levels.

In summary, when the correct sequence has been identified, the site localisers perform well in assigning the correct site of phosphorylation in ambiguous cases. There is no clear winner between algorithm types but with respect to activation methods, ProteinProspector’s SLiP and SLoMo algorithms would be highly recommended for ET-derived experiments.
Table 2.2. Site localisation performance statistics for four different localisation algorithms

This table contains localisation error rates from PSMs following a 0.05 q-value identification threshold, removal of known incorrect phosphopeptides according synthetic dataset and a tool-determined 5% false localisation rate. Mascot Delta (MD) and Site Localisation in Peptides (SLiP) represent the search-engine difference algorithms while SLoMo and PhosphoRS constitute the probability-based localisers (PBL). The contents of those passing sites were partitioned into correct and incorrect localisations and further processed to acquire the corresponding non-redundant site-centric values. Localisation error rates were computed from both PSM- and site-centric perspectives using these values as a FDR where the incorrect localisations represent the FPs.

<table>
<thead>
<tr>
<th>Activation Method</th>
<th>Site Localiser</th>
<th>Correct (PSM)</th>
<th>Incorrect (PSM)</th>
<th>Correct (Site)</th>
<th>Incorrect (Site)</th>
<th>FDR(PSM)</th>
<th>FDR(Site)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CID</td>
<td>MD</td>
<td>473</td>
<td>7</td>
<td>58</td>
<td>5</td>
<td>0.015</td>
<td>0.079</td>
</tr>
<tr>
<td></td>
<td>SLiP</td>
<td>409</td>
<td>3</td>
<td>56</td>
<td>2</td>
<td>0.007</td>
<td>0.034</td>
</tr>
<tr>
<td></td>
<td>SLoMo</td>
<td>538</td>
<td>9</td>
<td>62</td>
<td>4</td>
<td>0.016</td>
<td>0.061</td>
</tr>
<tr>
<td></td>
<td>PhosphoRS</td>
<td>567</td>
<td>22</td>
<td>64</td>
<td>9</td>
<td>0.037</td>
<td>0.123</td>
</tr>
<tr>
<td>MSA</td>
<td>MD</td>
<td>468</td>
<td>5</td>
<td>70</td>
<td>5</td>
<td>0.011</td>
<td>0.067</td>
</tr>
<tr>
<td></td>
<td>SLiP</td>
<td>413</td>
<td>6</td>
<td>62</td>
<td>5</td>
<td>0.014</td>
<td>0.075</td>
</tr>
<tr>
<td></td>
<td>SLoMo</td>
<td>512</td>
<td>10</td>
<td>71</td>
<td>5</td>
<td>0.019</td>
<td>0.066</td>
</tr>
<tr>
<td></td>
<td>PhosphoRS</td>
<td>532</td>
<td>36</td>
<td>74</td>
<td>15</td>
<td>0.063</td>
<td>0.169</td>
</tr>
<tr>
<td>HCD</td>
<td>MD</td>
<td>638</td>
<td>10</td>
<td>117</td>
<td>7</td>
<td>0.015</td>
<td>0.056</td>
</tr>
<tr>
<td></td>
<td>SLiP</td>
<td>567</td>
<td>29</td>
<td>106</td>
<td>13</td>
<td>0.049</td>
<td>0.109</td>
</tr>
<tr>
<td></td>
<td>SLoMo</td>
<td>337</td>
<td>6</td>
<td>73</td>
<td>3</td>
<td>0.017</td>
<td>0.039</td>
</tr>
<tr>
<td></td>
<td>PhosphoRS</td>
<td>685</td>
<td>23</td>
<td>121</td>
<td>12</td>
<td>0.032</td>
<td>0.09</td>
</tr>
<tr>
<td>ETD</td>
<td>MD</td>
<td>449</td>
<td>12</td>
<td>87</td>
<td>10</td>
<td>0.026</td>
<td>0.103</td>
</tr>
<tr>
<td></td>
<td>SLiP</td>
<td>473</td>
<td>3</td>
<td>90</td>
<td>2</td>
<td>0.006</td>
<td>0.022</td>
</tr>
<tr>
<td></td>
<td>SLoMo</td>
<td>404</td>
<td>0</td>
<td>80</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>PhosphoRS</td>
<td>461</td>
<td>1</td>
<td>90</td>
<td>1</td>
<td>0.002</td>
<td>0.011</td>
</tr>
<tr>
<td>ETDSA</td>
<td>MD</td>
<td>344</td>
<td>13</td>
<td>77</td>
<td>10</td>
<td>0.036</td>
<td>0.115</td>
</tr>
<tr>
<td></td>
<td>SLiP</td>
<td>359</td>
<td>5</td>
<td>77</td>
<td>5</td>
<td>0.014</td>
<td>0.061</td>
</tr>
<tr>
<td></td>
<td>SLoMo</td>
<td>286</td>
<td>0</td>
<td>70</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>PhosphoRS</td>
<td>350</td>
<td>2</td>
<td>77</td>
<td>2</td>
<td>0.006</td>
<td>0.025</td>
</tr>
</tbody>
</table>
2.4.2.2. Consensus site localisation increases confidence in phosphosite assignment

As with identification, I investigate whether a reduction in false positive localisations via multi-tool approach can also be achieved. Because there is presently no metric by which each localisation score can be directly compared, I instead applied a simple consensus ‘voting’ approach; using the localisation outcome of the four site algorithms.

A heuristic was applied where a PSM with any of the algorithms returning a score of zero was removed. It was reasoned that although the different localisation scores are not directly comparable, a score of zero is a universal signal suggesting ambiguity in site assignment. A second rule was to accept a PSM if a single site localiser assigned a score greater than it’s empirically-determined 0.05 FLR. Localisations that met both these criteria were given a single vote which were tallied in a PSM- or site-centric fashion and used to generate FDR statistics (Table 2.3).

Indeed, this voting scheme achieves notably lower FDRs, comparable to the best performing localiser under that activation method (Table 2.3). At the site-centric level, the FDR of the consensus method outperformed every localiser under collision-based methods, which were notable for the highest number of incorrect localisations using the independent algorithms.
Table 2.3. Consensus site localization statistics applied to different activation methods.

Localisation results from individual site localisation analyses were processed through a voting heuristic where each algorithm 'votes' its confidently localised sites. Votes were tallied and subjected to localisation error rate calculations and partitioned into their derivative correct and incorrect from PSM- and site-centric perspectives as performed previously with individual localisations.

<table>
<thead>
<tr>
<th>Activation method</th>
<th>Number of votes</th>
<th>Correct (PSM)</th>
<th>Incorrect (PSM)</th>
<th>Correct (Site)</th>
<th>Incorrect (Site)</th>
<th>FDR (PSM)</th>
<th>FDR (Site)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CID</td>
<td>2</td>
<td>563</td>
<td>6</td>
<td>65</td>
<td>3</td>
<td>0.011</td>
<td>0.044</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>488</td>
<td>4</td>
<td>58</td>
<td>2</td>
<td>0.008</td>
<td>0.033</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>329</td>
<td>1</td>
<td>47</td>
<td>1</td>
<td>0.003</td>
<td>0.021</td>
</tr>
<tr>
<td>MSA</td>
<td>2</td>
<td>538</td>
<td>6</td>
<td>71</td>
<td>4</td>
<td>0.011</td>
<td>0.053</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>477</td>
<td>2</td>
<td>67</td>
<td>2</td>
<td>0.004</td>
<td>0.029</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>324</td>
<td>0</td>
<td>53</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>HCD</td>
<td>2</td>
<td>669</td>
<td>8</td>
<td>117</td>
<td>6</td>
<td>0.012</td>
<td>0.049</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>533</td>
<td>1</td>
<td>105</td>
<td>1</td>
<td>0.002</td>
<td>0.009</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>298</td>
<td>1</td>
<td>68</td>
<td>1</td>
<td>0.003</td>
<td>0.014</td>
</tr>
<tr>
<td>ETD</td>
<td>2</td>
<td>471</td>
<td>3</td>
<td>91</td>
<td>2</td>
<td>0.006</td>
<td>0.022</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>451</td>
<td>1</td>
<td>86</td>
<td>1</td>
<td>0.002</td>
<td>0.011</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>378</td>
<td>0</td>
<td>77</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ETDSA</td>
<td>2</td>
<td>367</td>
<td>3</td>
<td>77</td>
<td>3</td>
<td>0.008</td>
<td>0.038</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>339</td>
<td>2</td>
<td>75</td>
<td>2</td>
<td>0.006</td>
<td>0.026</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>252</td>
<td>0</td>
<td>67</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
2.5. Discussion

2.5.1 A multiple search-engine strategy lowers error rates associated with search identification

Ignoring optimisation of parameters, the selection of an appropriate search-engine is made difficult due to the different algorithmic strategies each employ ranging from not only how theoretical candidates are assigned, for example, cross-correlation by COMET (Eng et al., 2013) but the computation of statistical metrics, for example, theoretical calculations using the N of within-mass-tolerance candidates (Perkins et al., 1999) or distribution-based linear tail fitting to compute expectation values (Chalkley et al., 2008). It is therefore not surprising that regardless of the search-engine used, the true statistical rate does not equate the intended rate. This makes the key question of deciding the most appropriate search-engine one that is nearly impossible and certainly subjective. More importantly, the true rate increases when converting PSM-to-peptide. A viable solution to this problem is to apply a multiple search-engine strategy which was demonstrated here to have the vital capacity of minimising the number of false positive identifications.

2.5.2. A multiple site localisation strategy can improve localization of collision-derived spectra

Once a selection of statistically convincing phosphorylated sequences have been acquired and undergone stringent processing to restrict the number of false positives, one again has a large selection of site localisation tools available at one’s disposal. I have compared the two core types of site localisation algorithms: search-engine difference (SED) and probability-based localisation (PBLs) scores and found overall that once the correct sequence has been found (for this dataset), localisation tends to be successful. This was even more noticeable for the ET-based activation methods relative to the collision-based counterparts which can be explained by the mode of fragmentation where the former circumvents the notorious issue of neutral loss.
Given the success above in lowering error rates associated with identification, a similar consensus-based approach composed of simple heuristic rules and a voting scheme were applied to further improve localisation.

It was found that this voting scheme improved FDRs relative to using any site localisation algorithm alone for collision-derived experiments which incidentally contained more incorrect localisations than their ET-derived counterparts. The effect of the voting method did not result in vastly improved localisation rates for the latter but this is likely due to the already low number of false localisations rendering the false positive removal property of the voting approach ineffective. Overall, the detection and removal of incorrect site localisations can be achieved through multi-localiser validation as was the case with sequence identification.
2.6. Conclusions

Using a set of synthetic phosphopeptides subjected to different activation methods, I have benchmarked the performance of several computational tools in context of identification and localisation. I find that while the error rate at the PSM-level is close to the intended rate, the rate increases upon conversion of PSM-to-peptide.

For identification, analysis of HCD-derived MS/MS captures most of the phosphopeptide space with a comparable number of incorrect identifications as CID and MSA. For the electron-based activation methods, both outperform CID and MSA. Against HCD, neither ETD nor ETDSA could identify as many unique phosphopeptides but they do make fewer incorrect identifications and so achieve amongst the best/lowest error rates.

Analysis using search-engines independently and in combination showed that no single activation method identified the full-set of 247 unique phosphopeptide sequences. In localisation, some algorithms are more suited for different activation-methods than others. Of note were ProteinProspector’s SLIP and the SLoMo algorithms for ET-based approaches.

I find that the integration of multiple tools presented a more potent approach to increase the number of correct identifications and localisations over any single tool in isolation. This is attributed to the different underlying algorithmic procedures (with regards to both interpretation of the MS/MS to decide which peaks should be used and subsequent scoring method for said peaks) leading to differences in identification and localisation outcomes as a function of the tool (and parameters therein) used. However, while it is clear that combining the outcomes of each tool increases the number of identifications and localisations made, this is true for not only real phosphopeptides (and sites) but also incorrect results. Critically, these caveats were circumvented through the application of simple heuristic rules such as removal of any search-engine unique matches and removal of zero localisation scoring instances, which had the benefit of universally removing false positives at both the PSM and peptide-level.
In conclusion, MS-based phosphorylation analysis is, and will remain, a challenging task where output contains an unintended number of false positives. It is impossible to rectify this by optimizing the performance of a single tool for every potential experimental condition. Instead, I find that a cross-validation/ensemble approach involving multiple different computational tools is a simpler and more viable alternative and can achieve confident and reliable results by constricting the false positive population for both identifications and localisations, alleviating the higher-than-intended false positive rates consistently seen throughout this study.
2.7. References


Proteome Research, 3, 958-964.


Chapter 3: An investigation into discriminative features of biologically regulated phosphorylation sites in mammalian organisms

The work in Chapter 2 demonstrated the benefits of a multi-tool approach in both identification and localisation by reducing the false positive population. However, while striving to achieve perfect localisation accuracy is without doubt a worthwhile initiative, a critical question that has not been addressed is assigning some measure of functionality to phosphosites. Until now, it has been assumed that all phosphorylation events play an evolutionarily pre-defined role in the cellular system but it has recently been postulated that some may pose little, or even no, functional consequence. The presence of such sites that may not invoke downstream responses are detrimental to wet-lab biologists whom may choose said sites for study in functional validation experiments.

Clearly, the best way to provide assistance and avoid such a scenario is to address this issue of predicting functional importance.

In this chapter, I assess the predictive capabilities of conventional phosphorylation-characteristic features (from sequence-and structural-level perspectives) towards identifying the (potentially) biologically important phosphosites to more effectively aid bioscientists in their work.
3.1. Abstract

With the possibility that some phosphosites lack function and that currently no predictors are capable of confidently providing some measure of function, the accidental selection of such sites appears to be inevitable.

In this study, I generate several datasets (each with assigned levels of functional importance) and evaluate them against known phosphorylation-characteristic properties (which are conventionally used in general phosphosite prediction). My hypothesis here is that the functional and less-functional groups of my datasets will distinguish themselves from each other according to these properties; providing the basis for a functional predictor.

In examining these properties, I find that the less functional population are less conserved than the functional groups. Furthermore, I also find that the less functional display some enrichment towards phosphorylation susceptibility-related features such as protein disorder. This may impact of the utility of already published phosphosite predictors that have incorporated these features into their underlying prediction schemes.

These results suggest that some properties aid while others maybe detrimental in searching for functional sites. Knowing them will allow for a more logical way to construct an effective functional predictor.
3.2. Introduction

The ultimate goal of functional phosphoproteomics is to assign function for each localised site and thereby enable the system to be exploited and manipulated. The sheer volume of data generated by phosphorylation-based mass spectrometry approaches has resulted in the colossal expansion of content contained within public phosphorylation data repositories which have proven to be an invaluable resource for the development of phosphosite prediction algorithms (Trost & Kusalik, 2011). While this deluge of data is welcomed by the informatics community, it is important to ask whether we are any closer to achieving confident functional annotation with site-level resolution.

The consideration that some phosphorylation sites may contribute little, or even no, functional consequence to their system was first raised in a letter by Lienhard (2008) through a series of counter-arguments of features that historically constitute functional importance: optimisation of ATP, low stoichiometry and site conservation (Chapter 1, section 1.8). The consequence of less-functional phosphorylation sites will directly impact the use of phosphosite prediction algorithms which are an indispensable tool for wet-lab biologists where they can be applied post-experiment as confirmatory support and select sites that may be biologically interesting.

Assuming they do not form a minute proportion of the content contained within public repositories, a plausible conclusion is that utilising the entire repertoire of phosphosites would succeed in creating a highly sensitive predictor capable of correctly detecting nearly all true phosphorylation sites at the expense of focussing on the 'more' functional ones. What is needed is a way to find successfully predict those that are on the more functional end of the scale; thereby alleviating the above problem.

The performance of a phosphosite predictor is dependent on how well the set of built-in features can discriminate a site that is phosphorylated from one that is not; the more discriminant the features between these two states, the better the capabilities of the predictor. Already, many descriptive
features of phosphorylation have been assessed in the literature and these include disorder propensity (Linding et al., 2003; Prilusky et al., 2005; Iakoucheva et al., 2004), conservation (Tan et al., 2009) and motif-context (Yaffe et al., 2001; Huang et al., 2005).

As expected, incorporation of such features into prediction algorithms does indeed lead to superior performance than those without (Iakoucheva et al., 2004; Gao et al., 2010; Dou et al., 2014). However, in all cases these features have been assessed for their ability to separate phosphorylated from non-phosphorylated sites; as far as I am aware there has been no large scale study that has assessed whether a distinction exists between the (more) functional and less-functional sites.

Addressing this problem will have tremendous consequence in future generations of prediction algorithms.

In this chapter, I investigate different sequence- and structural-perspective computational methods that have been reported in the literature to discriminate phosphorylated from non-phosphorylated species. My key aim to address here is to assess whether there is a difference between phosphosites of varying degrees of functional confidence using the features under assessment and, more importantly, if they can be utilised in a predictive context such as machine learning.
3.3. Materials and Methods

3.3.1. Datasets

3.3.1.1. Selection of true and functional phosphorylation sites from PhosphositePlus (Version 4th April 2014)

PhosphositePlus (PSP) is a phosphorylation repository that consists of sites derived predominantly from MS-based experiments in mammalian organisms; with human and mouse as the main contributors (Hornbeck et al., 2012). In this study, I use only phosphosites originating from human and mouse. This resource contains phosphosites derived from both literature, determined by both low- (LTP) and high-throughput (HTP) means, and in-house phosphosites by Cell Signalling Technology (CST). Every phosphosite and how it was determined is also accompanied with the number of times it was recorded. Because phosphosites derived from HTP studies are cautioned to have more than desired false positive localisations, these records present a way of addressing the possibility of incorrect site assignment. My criterion of a true phosphosite is one that has been either determined by LTP or at least twice if it was determined only from HTP-means. My rationale is that the sites within PSP have been curated to follow a consistent localisation false positive rate (FPR) of 5%, therefore the FPR for a true site that has been recorded more than once would equate to $0.05^n$ where n is the number of occurrences.

My dataset of true phosphosites composed of LTP- and HTP-published sites are denoted as ‘LTP’ and ‘HTP+’.

PSP also provides a set of sites whose functional role have been manually curated. These phosphosites have been determined via LTP methods and form my gold-standard functional sites and are designated as 'PSP+'.

3.3.1.2. Acquiring functional sites from quantitative mass-spectrometry datasets

Quantitative MS strategies coupled with phospho-enrichment has enabled a HTP approach to pinpoint functional phosphosites (Mann, 2006). This dataset is composed of phosphosites derived only from SILAC-based phosphorylation studies conducted on human and mouse organisms to conform to my PSP dataset. The key characteristics of these datasets are summarised in Table 3.2.

SILAC was chosen for two reasons; first, it is an established quantitative method which has been applied to many studies (http://silac.org/research_articles). The second, and equally important, is that the overwhelming majority of SILAC studies process data through the MaxQuant analytical pipeline (Cox & Mann, 2008). A consistent informatics protocol will provide some resistance to variation as a result of different lab groups, experimental protocols etc; making the process of comparing different studies more robust.

From these datasets, I acquire a set of potentially functional phosphosites through application of stringent identification, localisation and quantitative criteria.

A 1% identification threshold was used. For most studies, this was measured as the posterior error probability (PEP), implemented by the Andromeda search-engine part of MaxQuant but for those prior to the release of Andromeda (Cox et al., 2011), Mascot was used as the identification search engine (Cox & Mann, 2008). In these cases, I use a Mascot Ion Score of 25 (p value less than 1%).

For localisation, datasets provide either the MaxQuant PTM probability or an Ascore value (Beausoleil et al., 2006). For both I use the author recommended thresholds of 75% and 13 (p value of 0.05) for MaxQuant's PTM probability and Ascore respectively.

Of the sites that remain, two quantitative filters were applied. A primary filter acts at the peptide-level to find sites deemed worthy of being assigned as functional. It was assumed that those with a minimum Log2 1.5 fold-change (Bodenmiller et al., 2011) was a credible definition.
The secondary, if the information is available, operates at the protein-level where there can be no more than a $\log_2 0.58$ change; to ensure that quantitative change observed at the peptide-level is attributed to site stoichiometry and not protein abundance.

The quantitative-assumed functional sites are denoted as ‘Q+’.

### 3.3.1.3. Assessing the properties of less-functional sites

In this work, I also assess the properties of sites found in quantitative studies that I deem to be potentially less functional. The criterions imposed here are a maximum positive $\log_2 0.58$ threshold (which is deemed sufficiently low to say the peptide is not regulated under the condition being examined) (Dephoure et al., 2008) and restricting phosphosites to proteins that have been found to contain a functional phosphosite.

However, these alone cannot be used to directly classify the less-functional population due to several caveats associated with defining the less-functional set; the most prominent being that a site failing to reach an assumed functional threshold in one condition (and is therefore assumed to be non-functional) may do so under another condition. Also different sites on the same protein can lead to alternate function (Olsen et al., 2008).

Because there is presently no definition or criterion to determine which sites are less functional according to quantitative criterion, I take a pragmatic approach where I take each phosphosite sequence window (a 21-mer with 10 residues either side of the phosphosite) and cross-reference these against the 21-mers of every single true site of my PSP dataset.

While most have sites not been assigned a specific function, I assume that the majority of these published sites in PSP are the product of experiments whose aims are to test conditions that invoke truly functional sites. Therefore, many may indeed be functional but have yet to be fully characterized.
Because PSP is the most up-to-date resource containing phosphosites acquired from varied experiments, conditions and objectives, it provides the most comprehensive picture of the functional phosphoproteome currently available. As such, this cross-referencing will remove many functional sites (as well as the less-functional ones) leaving me with novel sites that exhibit low phosphorylation change that may truly be less-functional.

3.3.1.4. Negative set

A standard definition for a negative site is to choose an STY with no supporting experimental evidence of phosphorylation. However, a phosphosite may not have been annotated because the phosphopeptide itself was not detectable. A plethora of factors could contribute to this lack of detectability such as a potentially lowered ionisation efficiency relative to its cognate form.

To address this issue, I use the PeptideAtlas resource (Deutsch et al., 2008) which stores a comprehensive set of peptides observed through previous MS-based proteomic experiments. Under the assumption that phosphorylation confounds the detection; if the unmodified form is not found then it is unlikely for the phosphorylated version to be detected. The safest option in this situation would therefore be removal of that site.

Here, I acquire a set of observed tryptic peptides from 18 PeptideAtlas builds that met a satisfactory peptideprophet PSM FDR of 0.01 resulting in ~830,000 unique tryptic peptides.

I create an initial negative set, using the standard definition, consisting of peptides derived from the same proteins as the phosphorylated sets. BLASTP runs were conducted between the initial negative and PeptideAtlas peptides where only those that fully aligned and matched perfectly were retained and formed the final negative set.
Table 3.1. Summary of datasets for assessment in this study.

<table>
<thead>
<tr>
<th>Assumed Functional Status</th>
<th>Total Numbers</th>
<th>STY numbers</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative</td>
<td>210087</td>
<td>104460 (S)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>74533 (T)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>31094 (Y)</td>
<td></td>
</tr>
<tr>
<td>HTP+</td>
<td>59219</td>
<td>41039 (S)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>8563 (T)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>7114 (Y)</td>
<td></td>
</tr>
<tr>
<td>Unassigned (potentially functional)</td>
<td>12689</td>
<td>7831 (S)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2542 (T)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2316 (Y)</td>
<td></td>
</tr>
<tr>
<td>LTP</td>
<td></td>
<td></td>
<td>PhosphositePlus</td>
</tr>
<tr>
<td>PSP+</td>
<td>7337</td>
<td>4291 (S)</td>
<td></td>
</tr>
<tr>
<td>Functional</td>
<td></td>
<td>1446 (T)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1600 (Y)</td>
<td></td>
</tr>
<tr>
<td>Q+</td>
<td>3140</td>
<td>2642 (S)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>463 (T)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>35 (Y)</td>
<td></td>
</tr>
<tr>
<td>Q-</td>
<td>1027</td>
<td>807 (S)</td>
<td></td>
</tr>
<tr>
<td>(Potentially) less functional</td>
<td></td>
<td>208 (T)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>12 (Y)</td>
<td>SILAC studies</td>
</tr>
</tbody>
</table>
Table 3.2. SILAC studies from which functional and less-functional phosphorylation sites were extracted

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Species</th>
<th>Cell-type</th>
<th>Normalised against protein?</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biggelaar (2014)</td>
<td>H. sapiens</td>
<td>BOEC (Blood outgrowth endothelial cells)</td>
<td>No</td>
<td>Temporal phosphoproteome of thrombin stimulated BOECs</td>
</tr>
<tr>
<td>Opperman (2012)</td>
<td>H. sapiens</td>
<td>HTERT-PRE (Telomerase-expressing human retinal pigment epithelial)</td>
<td>No</td>
<td>Elucidating Plk (Polo-like kinase) 1 targets by comparing wild-type and a mutant that was purposely inhibited with phosphatase inhibitor.</td>
</tr>
<tr>
<td>Robitaille (2013)</td>
<td>Hela</td>
<td>No</td>
<td>Elucidation of substrates of mTOR through knocking out genes (Raptor and Rictor) for essential components of mTOR (1 and 2) complexes</td>
<td></td>
</tr>
<tr>
<td>Pan (2008)</td>
<td>Mus. musculus</td>
<td>Liver (HePa 1-6)</td>
<td>No</td>
<td>Phosphatase-inhibited vs wild-type in Liver cells</td>
</tr>
<tr>
<td>Weintz (2010)</td>
<td>Mus. musculus</td>
<td>Macrophage</td>
<td>No</td>
<td>Assessing dynamic changes of phosphoproteome in macrophages when stimulated by LPS</td>
</tr>
<tr>
<td>Zanivan (2013)</td>
<td>Skin</td>
<td>No</td>
<td>Investigation of phosphoproteomic changes in skin carcinogenesis</td>
<td></td>
</tr>
</tbody>
</table>
3.3.2. Sequence-level analyses

3.3.2.1. Creating multiple sequence alignments from pre-defined eggNOG resource for conservation

EggNOG is a phylogenetic resource where, through an automated BLAST-based approach, clusters or 'groups' of orthologous (OG) proteins have been built (Powell et al., 2012). I apply stringent criteria to establish appropriate alignments that provide robust conservation calculations. Here, I consider only core-classified proteins (which are essential for establishing OGs) that are mammalian-derived and not annotated as predicted or hypothetical. Mammalian taxonomic accession identifications were taken from eggNOG.

My procedure to generate the pre-requisite multiple-sequence alignments for conservation calculations is as follows:

1) A BLASTP search is run between the phosphorylated proteins against the appropriate species-specific protein sequences provided by eggNOG to find identical proteins. This is to account for differences in sequence of identical proteins that may arise from using different protein database versions.

2) Protein matches with 95% identity or better and the same sequence length as the target were deemed identical matches. This was considered to be suitably stringent and conservative criteria to eliminate false positives and retain sufficient information.

3) Each query protein match was then appended to its associated OG. In every instance, the identical matched proteins from the BLASTP search was removed from the MSA, to avoid redundancy, and re-aligned using the MAFFT alignment algorithm under default conditions. The
MAFFT algorithm was chosen in accordance to the alignment procedure used eggNOG itself (Powell et al., 2012).

All BLASTP searches were run using default statistical parameters.

### 3.3.2.2. Conservation calculations with rate4site

Phylogenetic-based conservation calculations were conducted with the rate4site (Pupko et al., 2004) algorithm. Rate4site was downloaded from http://www.tau.ac.il/~itaymay/cp/rate4site.html and calculations were conducted under default settings. The phylogenetic tree was built by rate4site from my post-processed eggNOG MSAs.

### 3.3.2.3. Disorder calculations with FoldIndex and DISOPRED

Both sequence and machine-learning disorder calculations were performed. For sequence-level disorder, an in-house implementation of the FoldIndex algorithm (Prilusky et al., 2005) was used to estimate disorder propensity. All calculations are performed on 21-mer windows with an STY in the centre (indexed as 0) and 10 residues either side. Those that do not abide this window size (because they were located towards the protein termini) were not computed.

In FoldIndex (eq. 3.1) H is the mean hydrophobicity (in accordance to the Kyte/Doolittle scale) and R is the absolute mean charge where residues that can be charged at a physiological pH of 7.0 are assumed to be arginine, histidine and lysine (positive) and aspartatic acid and glutamatic acid (negative). Because FoldIndex is a measure of protein folding, values were multiplied by -1 for a disorder-centric interpretation of predictions.
$FI = 2.785(H) - (R) - 1.151$

**Equation 3.1.** Prilusky’s FoldIndex algorithm for measuring folding propensity of a protein sequence. H and R are mean hydrophobicity and absolute mean charge respective. FoldIndex predictions were conducted over 21-mer windows.

Machine-learning based-disorder calculations were conducted with DISOPRED (v3.0) which was downloaded from the PSIPRED server hosted at University College London (UCL). Databases searches required for DISOPRED were conducted against the Uniprot/Swissprot (06/2014) that had been filtered to mask low-complexity regions using the pFilt script provided with the executables. DISOPRED disorder values were extracted from .diso files where any score between 0 – 0.49 were predicted to be ordered and 0.5 – 1 were disordered (personal communication with Dr Federico Minneci, a member of the team currently maintaining the PSIPRED server).

### 3.3.2.4. Secondary structure

The PSIPRED (v. 3.3) workflow is a ML-based predictor of protein secondary structure and available from the PSIPRED server at UCL along with DISOPRED3. Like DISOPRED3, PSIPRED was searched against the low-sequence complexity masked Uniprot/Swissprot database (06/2014). Secondary structure predictions were taken from .ss2 files.

### 3.3.2.5. Predicted accessible surface area

RVP-net is a neural network-based predictor that predicts accessible surface area (ASA) from protein sequence (Ahmad *et al.*, 2003). The algorithm takes in individual protein sequences and returns predictions.

Executables were acquired following personal communication with their creator Prof. Shandar Ahmad.
3.3.3. Structural analyses

3.3.3.1. Biological assembly units from the Protein Data Bank

The Protein Data Bank (PDB) is a repository containing a collection of structures of protein and nucleic acid biomolecules. X-ray crystal structures of the 'biological' unit assemblies, which are conformations predicted to be the native form following post-processing to remove artefactual crystal contacts, were downloaded in August 2013 via an rsync bash script provided by the PDB at: ftp://ftp.rcsb.org/pub/pdb/software/.

3.3.3.2. Annotating structures with phosphorylation sites

Structural annotation of phosphosites consisted of three phases:

1) Heuristic BLASTP between proteins of interest against those in PDB to short-list the entire phosphorylation dataset to regions containing phosphosites.

2) Sequence alignment to assess sequence fidelity of mapped region.

3) Structural checking to ensure local region of substrate site is of sufficient structural quality to conduct subsequent analyses on.

Because many phosphosites are located in regions not amendable to x-ray crystallography, namely regions of disorder, a BLASTP search was conducted between my dataset of phosphorylated protein sequences against non-redundant PDB protein sequences. Matches with an identity of at least 90% were deemed satisfactory and retained. This heuristic efficiently finds preliminary PDB structures likely to contain my phosphosites.

Structural redundancy within the preliminary structures was addressed with the PIESCES culling server.
(Wang & Dunbrack, 2003) using a maximum similarity identity threshold of 90%, 3.0 Å resolution or less (better) and an R-free value of up to 0.3 to acquire a set of non-identical and structurally sound structures.

Sequence alignments of 15-mers (-/+ 7 residues with the phosphosite at position 0) against their mapped structural sequence were conducted in R with matchPattern of the Biostrings package. A single mismatch was allowed on the flanking regions of the phosphosite but not the site itself.

To ensure structural quality, the amide bonds -/+1 with respect to the mapped phosphosite were checked to ensure they did not exceed a 3-dimensional (Euclidean) distance of 1.35 Å. Finally, only those found in a complex were retained.

3.3.3. FoldX suite

Potential energy calculations of structure upon computational mutations were conducted with the FoldX algorithm (v.30b). Mutations considered here were the phosphorylation of serine, threonine and tyrosine residues. All energy changes were calculated and models built with the buildmodel function. The buildmodel function was used here instead of the sitemutate function provided by FoldX due to a problem with generating TPO structures.

3.3.3.4. Positive electrostatic potential calculations

Positive/basic electrostatic potential estimations for each phosphosite were computed using the Phospho_patches algorithm from Kitchen and co-workers (2011). Phospho_Patches was run locally with fortran and perl scripts created and provided by Dr Jim Warwicker.
3.3.3.5. Accessibility calculations

All accessibility calculations were conducted with the NACCESS (Hubbard, 1992). Standard accessible surface area (SASA) values were acquired from default NACCESS.

3.3.3.6. Partitioning phosphorylation sites into interface and surface sites

Difference in accessibility (DACCESS) calculations were performed to locate the interface regions. Here, I computed the difference in surface accessible solvent areas (SASA) between complex and chain forms of a biological unit where any positive difference are caused by interactions within the complex; indicating interface regions.

Interface and surface situated sites were defined here as sites with positive and zero DACCESS values respectively.
3.4. Results

3.4.1. Predicted structural states reveal functional sites are more modest than the less-functional sites

Physical accessibility, degree of disorder, and coiled secondary order state (Fig. 3.1) are inter-related characteristics that are determining factors for the susceptibility of a phosphosite to an effector kinase. It is observed here that the distributions of these three properties for HTP- and LTP-based sites were different where the functionally unassigned LTP-based sites were more folded, less accessible and less coiled than the HTP-based sites (Fig. 3.1). The subset of functional sites belonging to each of these throughput methods also follow the same general pattern; PSP+ (LTP-based) were more folded and less accessible than the Q+ (HTP-based) sites.

However, the most startling difference is observed when comparing the functional groups against the less-functional which exhibit a greater degree of disorder, accessibility and coiling (Fig. 3.1). These differences were also deemed statistically significant according to Mann-Whitney tests (p-values of 0.01, $1.4 \times 10^{-25}$ and $1.9 \times 10^{-14}$ for accessibility, disorder and coiling respectively). This suggests that the less-functional set are more susceptible to (potentially unintentional) phosphorylation events. More importantly, from a predictor-context, using these three features maybe counterproductive in finding the functional population.

3.4.2. Conservation analysis show less-functional sites are less conserved

A feature that is integral to function is one that will persist throughout evolution. It is observed that the functional (Q+ and PSP+) and functionally-unassigned phosphosites (HTP+ and LTP) are more conserved (supported by Mann-Whitney tests where all had p-values of 0 except Q+ which was 0.006) than the non-phosphosites (Negative), conforming to previous studies (Gnad et al., 2007).
Critically, it is also observed that the less-functional phosphosites are drastically less conserved than any other group (including the negatives). This staggering enrichment demonstrates the power in using conservation as a feature to select functional phosphosites (by avoiding the less functional).
Figure 3.1. Physical features accessible surface area, disorder and secondary structure (coils) for each STY population. For all features, the HTP+ and LTP phosphosites share a similar distribution with their functional equivalents (Q+ and PSP+ respectively). The Q− phosphosites exhibit greater accessibility, disorder and coiling relative to the functional groups suggesting they are more susceptible to phosphorylation. Accessibility, disorder and secondary structure predictions were derived from RVP-net, DISOPRED and PSIPRED respectively.
Figure 3.2. Rate4Site conservation calculations for each STY population. Both the functionally unassigned and functional phosphorylation populations are more conserved than the negative according to Mann-Whitney tests. The less-functional group (Q-) is significantly less conserved than any other population.
3.4.3. Phosphorylation sites occupy more positively charged regions than non-phosphorylated sites over short distances

One mode of action by which phosphorylation mediates its role is through charge-charge interactions between the phosphate group and the functional side-chains in the immediate environment. Kitchen and colleagues (Kitchen et al., 2008) conducted a study into electrostatic stabilisation of a phosphate by looking at the degree of positive charge within the vicinity of structurally annotated phosphosites and non-phosphosites; demonstrating this feature can discriminate between the two populations. Using different sized radii with respect to the hydroxyl group of STYs, I find that over short distances (5 Å and 10 Å) my results agree with those published by Kitchen where the immediate environment surrounding the phosphosites is more positive than the non-phosphosites. Upon increasing this distance (25 Å and 40 Å), this discriminatory signal weakens (Fig. 3.3). Upon partitioning the structurally annotated phosphosites according to functional status (using PSP+), it is observed that those with empirically determined functional roles are located in more positive regions than the unassigned population. This may suggest utility in finding those that are functional but results were found not to be statistically significant (p-values of 0.38 and 0.10 for 5 Å and 10 Å respectively).

No SILAC-determined sites (Q+ and Q-) could be structurally annotated here. A reasonable explanation behind this is because they are located in regions of greater disorder which by default are more difficult to elucidate with x-ray crystallography.
Figure 3.3. Positive potential calculations with 5A, 10A, 25A and 40A radii around oxygen of hydroxyl-group on STY found at an interface or surface. Functional (PSP+), unknown and negative sites are green, yellow and red respectively. There is appears to be an enrichment for the functional group that decreases as the radii increases. While this would suggest that functional phosphosites are located in immediate electrostatic stabilising environments, Mann-Whitney tests indicate they were not statistically significant.
3.4.4. Total energy calculations of functional sites exceed the functionally unassigned at the interface but not negatives

Computational modeling algorithms allow predictions in binding energy changes upon altering the original structure thereby highlighting areas of potential interest (Guerois et al., 2002). Using the popular FoldX suite (Schymkowitz et al., 2005), Nishi and colleagues (Nishi et al., 2011) assess the effect of phosphorylation at protein-protein interfaces for homo- and hetero-complexes by inserting phosphate groups in silico at annotated sites. In their assessment, they find that 30 – 40% of their data exhibit significant (> +1kcal/mol) changes in binding energy against the native states.

My results conform to those of Nishi where the addition of phosphate at an interface will generally result in a positive binding energy delta suggesting destabilisation of the complex (Nishi et al., 2011).

In addition, my work also demonstrates that the functional phosphosites exert greater delta values compared to those with unassigned status (Fig. 3.4). Of these, 68% (63/93) and 46% (131/285) of the functional and unassigned-functional sites show greater than +1 kcal/mol.

Intriguingly, 57% (53/93) and 44% (96/285) of functional and functionally-unassigned sites elicit changes greater than +2 kcal/mol respectively. This threshold marks binding 'hotspots'; sections of an interface containing the main instigators of binding events that are responsible for a large proportion of binding energy (Tuncbag et al., 2009). The larger proportion of functional phosphosites in these areas may suggest one mode of action they have is the disruption of PPIs.
Figure 3.4. Total potential energy calculations by FoldX for STYs at the interface and surface of protein structures. Functional, functionally-unassigned and negative sites are green, yellow and red respectively. The functional set at the interface show large positive potential energy changes in excess of 2 kcal/mol (marking binding hotspots), suggesting that functional phosphorylations cause significant disruption to protein-protein interactions.
3.5. Discussion

At the moment phosphosite predictors are concerned with the accurate prediction of phosphosites. However, with the possibility that some sites have more definable function than others and with the increasing quantity of phosphorylation site data being amassed, selecting those that are most functional and worthy of further vigorous study to elucidate their biological role becomes increasingly difficult.

In this study I have collected three sets of phosphosites: those believed to be of functional importance, those with unassigned status and those that are assumed to be less-functional. I have assessed all populations with respect to features known to enrich for phosphorylation over non-phosphorylated residues to assess whether they have a more pronounced enrichment for the functional population; thereby allowing the development of a functional site predictor.

3.5.1. The role of physical susceptibility in determining functional phosphorylation sites

Disordered regions dictate protein-protein interactions (PPIs) by their ability to transition into a more ordered state (Dunker et al., 1998). This process was modelled from a thermodynamic perspective by Dunker and colleagues where the transition from disorder to a complementary ordered state with a binding partner depends on the energy required for the transition to take place and how stable the resulting interacting complex is. This model postulates that the greater the base disorder, the more energy is required for a (sufficiently) complementary conformation to be reached and consequently the less exothermic the whole process.

My results revealed that the assumed less-functional phosphopeptides had a greater degree of disorder than the functional set.

In context of Dunker's model, the lower degree of disorder associated with the functional sets suggests that these regions may have base conformations closer to (and so can more easily achieve) the
necessary conformation for interaction to take place with partner molecules. Phosphorylating such locations would presumably have a profound effect on native function, for example, by influencing the disorder-to-order transition step.

Another aspect of Dunker’s thermodynamic PPI model is a description of balance between 'specificity' and 'affinity'. Specificity in this model refers to the ability to interact with partners and affinity describes the strength of these interactions. The relationship between these two is anti-correlated; where high-specificity comes with the cost of low affinity and vice versa.

Logically, increasing the degree of disorder will allow access to a greater variety of structural conformations and in turn provide the opportunity to interact with more potential partners. However, this also raises the base disorder state which, according to the thermodynamic model, decreases the favourability of the interaction and therefore weakens their existence together.

This balance provides a potential explanation as to the origins of my assumed less functional sites in my set where their increased susceptibility (corresponding to an increased specificity and lowered affinity) allow for low-levels of (unintentional) phosphorylation to occur.

### 3.5.2. The selection of appropriate features for building a functional phosphosite predictor

The success of a prediction algorithm will depend on how well the features you describe your problem with separate the different classes you have. Predicted assignment (classification) and confidence in said prediction (regression) by machine-learners will examine the positive and negative classes in an attempt to establish rules that best separate them.

For phosphorylation, these features are most notably those associated disorder, physicochemical properties surrounding the site, linear motifs and conservation.
Beginning from a sequence-level perspective, I find that the less-functional phosphosites are enriched for conventional physical property features (disorder, accessibility and coiled secondary structure) but severely depleted in terms of conservation.

All non-kinase-specific predictors (nKSPs) predict phosphorylation state based on the properties of the residues surrounding a site. Disorder is an infamous example (Iakoucheva et al., 2002) and other related properties (accessibility and secondary structure) have since been incorporated in predictors thereafter. Given that my results demonstrate that all phosphosites are enriched for such properties relative to the negative set it is clear why they were chosen as discriminatory features. However, the greater enrichment observed in the less functional population presents a problem in context of a functional predictor. Because the negative population is clearly less disordered than any of the phosphorylation sets, the machine-learning will recognise this relationship and therefore assign those with a higher degree of disorder with an equally high prediction of being phosphorylated.

A reasonable conclusion here is that the less-functional will be predicted on par with or (worst case scenario) will be superior to those truly with a functional role; a devastating outcome for my cause and, more importantly, any wet-lab biologist.

Conservation on the other hand shows promise with regards to predictive potential where the less-functional sites are less conserved than any other site (including the negative set) assessed.

From a structural perspective, the process of structurally annotating substrate sites onto crystal structures greatly reduces the starting material by almost 1000-fold. This unfortunately diminishes the value of structural features for my project where I ultimately want to find discriminatory features and train a predictor. Furthermore, the SILAC-based sets could not be mapped (likely due to a potent mixture of low numbers and increased disorder) so it is difficult to comment with regards to functionality.
However, the results do reveal interesting characteristics of functional phosphorylation that was not possible by the sequence-level analyses.

Here, the FoldX energy calculations predict that functional phosphosites situated in the interface tend to be located in binding 'hotspots'; the regions of the interface that contribute a significant proportion of the binding energy of the interaction event. A suggestion made above is that functional phosphosites could influence the PPI event and this result provides one mode by which it is achieved; by being situated at these binding 'hotspots' they are at an optimal place to influence the interaction. Judging from the positive energy changes (which suggests unfavourable endothermic changes), this influence is in the direction of disruption; functional phosphosites appear to be hindering PPIs from taking place?
3.6. Conclusion

In this chapter, I have subjected my phosphorylation datasets (each assigned with varying degrees of functional importance) to a battery of algorithms that predict phosphorylation-characteristic properties. The aim here was to assess whether these features not only distinguish general phosphosites from non-phosphosites but also if there was any signal that discriminates the functional from less-functional.

The results of my assessment suggest that of my features, conservation is the most suitable for a function-driven phosphosite predictor. This is not only due to the less-functional sites being less conserved than all other phosphorylation groups (which fulfils the criterion above) but because these sites are also less conserved than the non-phosphosites. This severe separation in conservation should assist in detecting those that are functional.

However, not only did I find a feature that has the potential to predict functional sites, I also find features that may hinder this aim too. Disorder, accessibility and secondary structure are favourite features of non-kinase-specific phosphosite predictors. Here, they appear to favour the less-functional sites which, from a predictor-perspective, would mean that these sites will be predicted as well or better than truly functional sites.

In conclusion, certain phosphorylation-characteristic features can improve or worsen our chances of building a predictor capable of detecting functional phosphosites. The key to constructing a functional predictor lies in their appropriate use.
3.7. References


Chapter 4: Evaluating general and functional phosphorylation site prediction performance

In the previous chapter, I find that some phosphorylation-characteristic features have the capacity to enrich for the functional subset of sites while others are likely to achieve the opposite. While knowing this is vital for constructing an effective functional predictor, the knowledge that some features are counterproductive in constructing a functional predictor presents a concern for many already published predictors whom use these features; does this mean these predictors are more prone to detecting potentially less functional sites than those without?

In this chapter I aimed to train general phosphorylation site predictors with prediction power that are on par (or better) than other published predictive tools by using (and supplementing) known phosphorylation-enriched features (many of which were assessed in my previous chapter). I also endeavor in creating a functional predictor based on knowledge regarding the set of functionally productive and counterproductive features discovered in the previous chapter. Finally, I evaluate the performance of all predictors on my functionally assigned datasets from the previous chapter to assess the performance of published predictors on this functional prediction problem and whether my own predictors trained with these productive features are capable of avoiding the less-functional set.
4.1. Abstract

Currently, phosphosite predictors have one goal in mind: ‘is that site phosphorylated?’ None have been assessed for their predictive ability towards functional and less-functional sites. This poses a concern for the wet-lab biologist where the tool they might be using could be more successful in predicting those that are less-functional.

There were two aims in this chapter.

The first was to generate phosphosite predictors using machine-learning schemes that incorporate a variety of different features and assess which have possible utility for general prediction. Here, I find that my meta-predictors which use the entire array of features and are supplemented with prediction outcomes from published predictors have superior performance to all others.

My second aim was to evaluate the performance of both publicly available and my own ML-derived predictors on my functionally-assigned set of phosphosites in light of my results in Chapter 3 where some features were found to be enriched for the less-functional set. It is important to investigate whether the predictors that utilise these features have also inherited the ability to more easily predict these less-functional sites. For the public predictors, it unfortunately does appear to be the case where the non-specific-kinase predictors (nKSPs) do show good prediction of this set. However, in evaluating the kinase-specific-predictors (KSPs), I find that these are a little more resilient against these sites; suggesting that these predictors may be advised for use over the nKSPs.

For my own predictors, I observe a similar conclusion where predictors trained with certain features (namely the physicochemical-based and meta-predictors) were, despite good general predictive performance, typically less-effective from a functional perspective where they would struggle to differentiate between the functional and less-functional datasets.
In response to this, I attempt to make a functional predictor by using only conservation and GPS output (both of which were shown to not enrich for the less-functional sites). This subsequently resulted in a predictor that exhibited the lowest predictive performance towards detecting the less-functional group; indicating that this potent combination has promise in discriminating against these sites.
4.2. Introduction

With the ever increasing number of localised phosphosites one can attain from a typical MS-based phosphoproteomic experiment, wet-lab biologists are mindful as to which phosphosites they can select to devote time and effort into functionally characterising. However, with potentially thousands of confidently localised sites, this task becomes proportionally more challenging. For assistance, one could use the array of phosphorylation site predictors available.

Phosphosite predictors operate by incorporating phosphorylation-characteristic features such as linear-motifs (Huang et al., 2005), disorder propensity (Iakoucheva et al., 2004), conservation (Biswas et al., 2010) or combinations of the above features (Gao et al., 2010; Gnad et al., 2010; Dou et al., 2014) etc to distinguish true substrate sites from non-phosphosites. All are built with the contents of phosphorylation databases (Hornbeck et al., 2012, Diella et al., 2004, Stark et al., 2010; Heazlewood et al., 2008) where their optimisation criterion is to maximise separation between P-sites and NP-sites. Such predictors are an invaluable resource in the selection of in vivo determined sites because they can be used as secondary confirmation that a site adheres to the biological properties of those that have been recorded in said phosphorylation databases, thereby raising expectations that the sites acquired contain a biological signal and are worthy of functional pursuit. Furthermore, with the staggering volume of data that has been deposited in repositories (well-over 100,000 validated sites), the accuracy of these predictors are constantly increasing. Still, one of the most common challenges that remains for bioinformaticians working in phosphoproteomics is the correct prediction of likely sites of modification and hence there has been active development of site prediction tools.

However, while prediction accuracy is important, there are very few predictors that are capable of providing some measure of function. This is an important task to address because it has been considered that not all sites identified are necessarily strictly functional which will hamper efforts in attempting to
understand the biological mechanisms of a system by drawing our attention away from those sites that are functional.

In response to this dilemma, there has been activity in developing predictors capable of addressing this issue and fortunately some are now available. Two examples are CPhos (Zhao et al., 2012) and a webserver by Niu and colleagues (2012), both of which measure and select potentially functional sites based on their degree of conservation. CPhos uses Shannon’s Entropy from a site and window perspective on multiple-sequence alignments (MSA) to generate a motif conservation score while Niu's webserver is a suite of published tools whose conservation is estimated with the rate4site algorithm (Pupko et al., 2002). While neither tool provide a confidence measure for functional status such as a probability value (rather they are rank ordered according to conservation) this is certainly a step in the right direction and will hopefully inspire other developers to create a tool for functional prediction.

Until such predictors become available, an important question to ask now is how well the arsenal of present predictors perform in predicting functional phosphosites. Similarly, I considered there to be room for improvement in performance overall, in parallel with my goal of identifying truly functional sites.

In this chapter, I had two major goals:

First, using a variety of machine-learning schemes, I assess which different combinations of (biological, physicochemical and predictor output) features are effective for general phosphorylation site prediction.

Second, I evaluate how well both published and my own (general- and function-orientated) ML-derived predictors perform on the functionally annotated set of phosphorylation sites (PSP+, Q+ and Q-) from the previous chapter.
4.3. Materials and Methods

4.3.1. Datasets

4.3.1.1. Phosphorylation resources

Phosphosites in this study were acquired from the mammalian-centric phosphorylation resource PhosphositePlus (PSP) (April 2014) (Hornbeck et al., 2012) following stringent thresholds to acquire high-quality phosphosites from the literature. To construct a confident dataset for building and testing my ML protocol, I use any phosphosite that has been determined via LTP- or have been verified at least twice by HTP-methods.

4.3.1.2. SILAC-derived phosphorylation sites

SILAC-based phosphorylation studies on human and mouse were acquired from (http://silac.org/research_articles) (Table 3.2 in Chapter 3).

I applied stringent identification and localisation thresholds, set at a 1% posterior error probability (PEP) and 0.75 MaxQuant localisation probability respectively, to retain a set of confidently identified and localised P-sites. These sites were then subjected to quantitative filtering to find assumed functional (Q+) and less-functional (Q-) sites. The less-functional sites were generated by first taking the set of low-change phosphosites and removal of any site that was found to match any of those that were functional or published in the PSP resource. The rationale behind using published sites as well is that the vast majority are likely to be the products of studies whose goals were aimed at elucidating and therefore understanding the phosphoproteome as a function of condition. As such, the majority of sites should have some degree of functional importance. Furthermore, because of the vast span of experiments (each assessing the phosphoproteome under a different condition) contained within PSP, these sites should comprehensively cover much of the total functional phosphoproteome and therefore address
the most difficult question in defining functional and non-functional sites; has the right condition been met to invoke phosphorylation of the site?

4.3.1.3. Defining positive and negative sets for training and testing

The definitions to construct the positive and negative datasets are detailed in the methods of Chapter 3. Briefly, I used only PSP sites where my positive set was defined as the set of phosphosites that have been detected at least twice to restrict the number of potential false positive localisations associated with HTP-experiments. The negative set consists of peptides with no evidence of phosphorylation but have been deemed detectable according to the PeptideAtlas repository which stores a collection of confidently identified tryptic peptides at a false discovery rate threshold of 0.01. This approach addresses the situation where a true phosphosite is missed, and therefore assumed to be not phosphorylated and classed as a negative under a conventional definition, because it was undetectable. Total numbers of non-overlapping sequences in all positive and negative populations have been summarised (along with the STY compositions) in Table 4.1.

4.3.1.4. Independent tests

Four datasets were constructed to assess the predictive capabilities of my ML-based predictors. The first is composed of the phosphosites that have been put aside from my ML training regime. These consist of one-third of the total amount of data used (the other two-thirds being allocated for feature selection and training) and is used for assessing the general predictive capabilities of the trained ML models. Further details are found below (Fig. 4.1).

The remaining three datasets have been constructed specifically for assessing function.
The first set is composed of functional sites, designated after manual curation from PSP and are considered the gold-standard set. These have been denoted throughout as ‘PSP+’. The second and third are the Q+ and Q- sets explained above.

However, because the published predictors I later assess against these datasets: DISPHOS (Iakouchva et al., 2004), NetPhos (v2.0) (Blom et al., 1999), PPSP (Xue et al., 2006), GPS (Xue et al., 2008) and MUSITE (Gao et al., 2010) have been trained on various phosphorylation repositories such as PSP which contain phosphosites from published studies, there is a possibility that some phosphosites from my acquired SILAC sets may have been part of the training regime for these predictors. This will bias performance evaluations as predictors are predicting sites that they have been trained on.

To ensure a fair assessment of predictive capabilities, these already-used phosphosites need to be removed.

To achieve this, I take the published training phosphorylation sites from the latest predictor assessed in this work (MUSITE) (Table 4.2a) under the assumption that it is the best candidate for containing overlapping training sets with earlier published predictors. This training set also contains kinase-substrate information which allows for specific removal of KSP training sites.

When assessing the predictive performance of KSPs and my functional ML predictor (which utilises GPS output), I remove only kinase-substrate annotated phosphorylation sites from my functional datasets (Table 4.2b). For nKSPs and my ML predictor containing MUSITE output, I remove all MUSITE training phosphorylation sites (Table 4.2c). For my ML predictors using only biological or my top-10 orthologous physicochemical features, no filter is applied.
Table 4.1. Summary of datasets for assessment in this study.

<table>
<thead>
<tr>
<th>Assumed Functional Status</th>
<th>Total Numbers</th>
<th>STY numbers</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative</td>
<td>210087</td>
<td>104460 (S)</td>
<td>74533 (T)</td>
</tr>
<tr>
<td>HTP+</td>
<td>59219</td>
<td>41039 (S)</td>
<td>8563 (T)</td>
</tr>
<tr>
<td>LTP</td>
<td>12689</td>
<td>7831 (S)</td>
<td>2542 (T)</td>
</tr>
<tr>
<td>PSP+</td>
<td>7337</td>
<td>4291 (S)</td>
<td>1446 (T)</td>
</tr>
<tr>
<td>Q+</td>
<td>3140</td>
<td>2642 (S)</td>
<td>463 (T)</td>
</tr>
<tr>
<td>Q-</td>
<td>1027</td>
<td>807 (S)</td>
<td>208 (T)</td>
</tr>
</tbody>
</table>

Table 4.2a. Summary of training sites used in MUSITE. Sites have been further partitioned to those that have kinase-substrate annotations to allow filtering specifically for kinase-substrate predictors.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Total Numbers</th>
<th>STY Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kinase-substrate annotated sites</td>
<td>2793</td>
<td>1629 (S)</td>
</tr>
<tr>
<td>Total training sites</td>
<td>34193</td>
<td>25317 (S)</td>
</tr>
</tbody>
</table>
Table 4.2b. Summary of functional datasets following removal of kinase-substrate annotated phosphorylation sites from the MUSITE training set.

<table>
<thead>
<tr>
<th>Assumed Functional Status</th>
<th>Total Numbers</th>
<th>STY numbers</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSP+</td>
<td>4474</td>
<td>2670 (S)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>871 (T)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>933 (Y)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>PhosphositePlus</td>
<td></td>
</tr>
<tr>
<td>Q+</td>
<td>3035</td>
<td>2566 (S)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>443 (T)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>26 (Y)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>SILAC studies</td>
<td></td>
</tr>
<tr>
<td>Q-</td>
<td>1027</td>
<td>807 (S)</td>
<td></td>
</tr>
<tr>
<td>(Potentially) less-functional</td>
<td></td>
<td>208 (T)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>12 (Y)</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.2c. Summary of functional datasets following removal of all sites overlapping with MUSITE training set.

<table>
<thead>
<tr>
<th>Assumed Functional Status</th>
<th>Total Numbers</th>
<th>STY numbers</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSP+</td>
<td>3672</td>
<td>2146 (S)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>747 (T)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>779 (Y)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>PhosphositePlus</td>
<td></td>
</tr>
<tr>
<td>Q+</td>
<td>1593</td>
<td>1313 (S)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>263 (T)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>17 (Y)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>SILAC studies</td>
<td></td>
</tr>
<tr>
<td>Q-</td>
<td>931</td>
<td>741 (S)</td>
<td></td>
</tr>
<tr>
<td>(Potentially) less-functional</td>
<td></td>
<td>179 (T)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>11 (Y)</td>
<td></td>
</tr>
</tbody>
</table>
4.3.2. Machine Learning Procedures

4.3.2.1. Partitioning of data for feature selection, training and testing

Three datasets, each containing a third of the total positive population along with an equal number of negatives to create balanced datasets, are constructed for three stages in the ML procedure: feature selection, training and independent testing. Balancing is necessary to avoid biased predictive models that favour the larger class. All peptide sequences were randomly selected and pre-processed to remove any overlapping and/or redundant peptide sequences between the positive and negative instances and all permutations of dataset and ML stage were conducted (Fig. 4.1).

4.3.2.2. Selection of physicochemical properties from AAindex for use as machine learning features

From a sequence-level perspective, the information that distinguishes phosphosites from non-phosphosites is encoded directly in the amino acids surrounding the substrate site, for example, disorder, accessibility and secondary structure predictions.

To expand upon this and improve the predictive power for general phosphosites, I utilise the physicochemical properties contained within these surrounding amino acids via the AAindex resource which stores ‘amino-acid indices’. These indices contain information pertaining to the physicochemical properties for each of the 20 naturally occurring amino acids (Kawashima & Kanehisa, 2000; Kawashima et al., 2008) and is therefore suitable for my needs. Currently there are 544 AA indices via the seqinr library in R (Charif & Lobry 2007).

To assign a representative index value to each positive and negative site, I take the 21-mer sequence window (+/- 10 residues relative to the site) and compute the mean index value of the residues of the window.
4.3.2.3. Selecting features for machine learning

Because a large number of indices within AAIndex are highly redundant (often being updated versions of previous iterations) I apply a feature selection scheme used by the ConSEQUENCE algorithm (Eyers et al., 2011) that addresses this redundancy and maximises the degree of orthogonality between features to improve ML performance.

Briefly, the Kullback-Liebler divergence (Eq. 4.2) is computed between positive and negative populations according to their representative index and ranked in descending order of divergence with a minimum threshold of an absolute divergence of 0.7.

Starting with the first ranked, the spearman correlation coefficient is iteratively calculated for the subsequent ranked indices until an absolute coefficient of 0.5 or less is reached. This index is marked and used to calculate the SCC (replacing the initial first ranked index) with the subsequent feature. This process is repeated until the end of the index list is reached. The top ten-ranked non-redundant features surviving these criterion were taken as features for machine learning.

\[
KL\text{-}dist(p,q) = \sum_k p_k \log_2 \left( \frac{p_k}{q_k} \right)
\]

**Equation 4.1.** Kullback-liebler distance

\[
KL\text{-}divergence = KL\text{-}dist(p,q) + KL\text{-}dist(q,p)
\]

**Equation 4.2.** Kullback-liebler divergence. The divergence is the sum of distances where distributions q and p switch as background.
Figure 4.1. Workflow of how datasets were constructed for feature selection, training and testing for machine-learning. Feature selection in this scheme refers to the selection of the top-10 orthologous amino acid indices from AAindex. Phosphosites from PhosphositePlus (PSP) (1) were randomly partitioned into three equal sized subsets and balanced with an equal number of negative instances (2). Each permutation of these subsets (3) was used to train three (random forest, adaptive boosting and neutral network) corresponding machine learning-based phosphosite predictors.
4.3.3. Machine Learning

All machine-learning was conducted through the CARET package (Kuhn, 2008); a standardised and friendly-to-implement interface with multiple machine-learning (ML) algorithms available in R. Here, the randomForest (rf), neural network (nnet) and adaptive boosting (ada) were chosen as the base learners to train optimal P-site predictive models for their ability to deal with complex non-linear problems.

4.3.3.1. Optimisation of models with CARET package

To select the optimal ML model, I use the optimisation functions offered by the CARET (Classification and Regression Training) package in R. These functions perform cross-validation (CV) calculations across a range of user-defined parameters, each specific to its respective ML model and returns a performance metric (Table 4.3). The criterion I use to select the optimal model was to maximise the 'kappa' statistic, a re-adjusted accuracy measurement according to the expected accuracy. When the optimal parameters are chosen, CARET creates a final model using the entire training set.

4.3.3.2. Neural Network

Neural networks are a ML approach where classification is achieved by building and optimising a complex interconnected network to find which features or combination of features best predict whether a site is phosphorylated. Available via the nnet package in R. Variation in number of hidden layer nodes and decay rate were optimised with 200 iterations/epochs. All calculations were conducted using the default (sigmoidal) transfer function.
4.3.3.3. Random Forest

A random forest is an ensemble approach to the standard decision-tree based classifier. RFs randomly use subsets of provided features to generate a large number of trees which each contribute a single vote and the classification outcome is an accumulation of these votes.

The 'randomForest' package is an R implementation of the Brieman RandomForest fortran package. Default conditions under classification mode was used here. The default number of trees is 500. The number of branches 'mtry' was kept as sqrt(features) in the parameter tuning for caret.

4.3.3.4. Adaptive Boosting

Adaptive boosting is the second ensemble method used in this study. Boosting algorithms work by generating many classification models, weighting each one according to performance accuracy and combining their predictions to generate a final prediction. The rationale here is that the final prediction is a more accurate (or at least no worse) representation than any single of the models alone. The adaptive boosting algorithm applies weights according to how difficult it is to correctly label samples during training; forcing the classifiers to focus on those that were hard to get right (Freund & Schapire, 1996).

4.3.3.5. Evaluating performance

To assess performance of each prediction, I compute the specificity (Spec), sensitivity (Sens), positive predictive value (ppv), matthew correlation coefficient (MCC), and the balanced precision-recall F-measure (F). Metrics displayed in performance tables are taken at the maximal F-measure (Gou et al., 2014). All performance metrics and ROC plots in this chapter were computed with the ROCR (version 1.05) package.
Equations 4.3. Performance measures for evaluating predictors. Spec, Sens, F and MCC refer to specificity, sensitivity, F-score and Matthew correlation coefficient. The F-measure is a harmonic mean of precision (P) and recall (R) that has been weighted according to the alpha variable. Here, the alpha variable was left as the default value of 0.5 (as stated in the ROCR library) which gives both P and R equal weighting in the equation.

Table 4.3. Summary of CARET procedure to build optimal predictive model with different machine learning algorithms.
4.3.4. Phosphorylation-characteristic features

All phosphorylation-characteristic features from the previous chapter were incorporated into my ML scheme. These features were considered in different combinations to assess their effectiveness for general and functional prediction.

Brief explanations of each feature are provided here and more details can be found in the methods section of Chapter 3.

4.3.4.1. Disorder calculations with FoldIndex and DISOPRED3

DISOPRED3 calculations were conducted against the Uniprot/Swissprot database (06/2014) that had been filtered to mask low-complexity regions. DISOPRED3 was downloaded from the PSI-PRED server at University College London. Disorder predictions were extracted from the resulting .diso files.

FoldIndex (Prilusky et al., 2005) was computed though an in-house script that takes in 21-mer windows with an STY in the centre (indexed as 0) and 10 residues either side.

4.3.4.2. Conservation calculations with rate4site

Phylogenetic-based conservation calculations were conducted with the rate4site (Pupko et al., 2004) algorithm on pre-processed eggNOG multiple sequence alignments (MSA). The processing procedure retains only mammalian species whose proteins were not annotated as 'hypothetical' or 'by prediction'.

4.3.4.3. Secondary structure

PSIPRED (v3.3) (Jones, 1999) was downloaded from from the PSIPRED server at UCL. My phosphorylated protein sequences were searched against a Uniprot/Swissprot database (06/2014) that was pre-
processed to mask regions of low complexity. Secondary structure predictions were taken from .ss2 files.

4.3.4.4. Predicted accessible surface area

Predicted accessible surface area (ASA) values were computed from protein sequences with the rvp-net-big algorithm (Ahmad et al., 2003). This algorithm takes in single protein sequences in fasta format. Binaries for these executables were acquired following personal communication with their author Prof. Shandar Ahmad.

4.3.5. Published phosphosite predictors

The performance of phosphosite predictors against my functional datasets was carried out with a small number of widely-used predictors. This assessment includes examples of the two main classes of predictors: kinase- and non-kinase-specific predictors (KSPs and nKSPs) which are concerned with deducing the kinase-substrate relationship (KSR) and general phosphorylation status of a candidate site respectively. Here, GPS and PPSP are KSPs while DISPHOS and NetPhos are nKSPs. Because MUSITE has been constructed for both general predictions and forming KSRs, I designate it as a ‘Hybrid’ predictor.

4.3.5.1. GPS

GPS (version 2.1.2) was downloaded from http://gps.biocuckoo.org/down.php. All calculations were run with the 'All' threshold activated which assigns each STY a predictive score using all predictive kinases classes and sub-classes. The 'High' thresholds for each class, corresponding to a 2% and 6%
false positive rates (FPRs) for ST and Y kinases respectively, were used to compute a delta score for each site. Of these, the most positive delta is taken to represent that site.

### 4.3.5.2. MUSITE

Pre-calculated MUSITE predictions on Human and Mouse Uniprot/Swiss-prot records were downloaded from http://musite.sourceforge.net/predictions.php. Corresponding proteins were found between our dataset and those pre-calculated in MUSITE by cross-referencing and ensuring perfect sequence matches between the two. MUSITE scores were extracted from these sequences.

### 4.3.5.3. DISPHOS

DISPHOS calculations were performed through SCANSITE3 webserver (http://scansite3.mit.edu/) which accepts multiple protein sequences in fasta format. Following analysis by SCANSITE3, additional DISPHOS analyses are made available. Because the SCANSITE3 predictor was not being used itself, all settings were left as default: ‘mammalian-kinase motifs’ and ‘moderate’ sensitivity. The DISPHOS output returns sequence-windows and the corresponding prediction score. In addition, sequence windows were marked (*) by DISPHOS if they were part of the dataset used to train the algorithm. All such instances were removed and took no part in my performance evaluation.

### 4.3.5.4. PPSP

Predictions by PPSP were made on its server using the full array of kinases with the 'Balanced' (Default) performance selected. Multiple sequences were uploaded onto webserver in fasta format. PPSP ‘Risk Difference’ delta scores were taken and used to represent that predictor during performance evaluation.
4.3.5.5. NetPhos

NetPhos predictions were performed using the Netphos webserver found at: http://www.cbs.dtu.dk/services/NetPhos/. Multiple sequences were uploaded in fasta format. Prediction scores were taken for performance evaluation.

Because DISPHOS, PPSP and Netphos predictions here were acquired from web-servers, my datasets could not be used in their entirety. Instead, random subsets of consisting of equal numbers of positive and negative instances for S, T and Y (if possible) were selected as representative data points per dataset (Table 4.1). For the Q+ and Q- datasets, only S were considered due to the low number of T and Y residues.

All scripts were written in R version 3.0.2. Large computational tasks were handled on a local linux server.
4.4. Results

4.4.1. Performance of machine learning models using combinatorial features on held-out test sets

The discriminatory power of a predictor stems from the set of features that it uses. New predictors augment their predictive capabilities with well-known features such as disorder and conservation. Here, I have utilised physicochemical characteristics, biological properties and borrow power from other predictors as features in my machine-learning scheme, arranging them in different combinations to assess which produce the best general predictive models (Fig. 4.2 & Table 4.4a - c). Comparisons of each combination with GPS and MUSITE predictions were included. Predictions by webserver-exclusive predictors were excluded due to the relatively lower numbers of predictions compared to GPS and MUSITE.

Of the different combinations, my ML-based predictors constructed from the ConSequence approach (Eyers et al., 2011) (yielding the top-10 orthologous AAindex indices) outperforms those made of the phosphorylation-enriched biological features (disorder, accessibility, secondary structure and conservation rates), with AUCs averaging at ~0.78 (Table 4.4b) and ~0.74 (Table 4.4a) respectively. As expected, the meta-predictors which use both feature types and supplements them with both GPS and MUSITE predictions results in superior performance and achieves AUCs ~0.86 for all ML-based predictors (Table 4.4c).

In every instance, there was little performance difference between any of my predictors as a function of the ML used for their construction.
Table 4.4a. Performance statistics of machine learning-based predictors constructed from biological features (disorder, accessibility, secondary structure and conservation) on held-out test set

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Machine Learner</th>
<th>Spec</th>
<th>SN</th>
<th>ppv</th>
<th>MCC</th>
<th>F</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hold-out Test</td>
<td>ADA</td>
<td>0.48</td>
<td>0.84</td>
<td>0.62</td>
<td>0.35</td>
<td>0.71</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>Neural Network</td>
<td>0.49</td>
<td>0.84</td>
<td>0.62</td>
<td>0.36</td>
<td>0.72</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>Random Forest</td>
<td>0.44</td>
<td>0.87</td>
<td>0.61</td>
<td>0.35</td>
<td>0.72</td>
<td>0.74</td>
</tr>
</tbody>
</table>

Table 4.4b. Performance statistics of machine learning-based predictors constructed from top-10 orthologous amino acid indices on functional datasets on held-out test set

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Machine Learner</th>
<th>Spec</th>
<th>SN</th>
<th>ppv</th>
<th>MCC</th>
<th>F</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hold-out Test</td>
<td>ADA</td>
<td>0.57</td>
<td>0.82</td>
<td>0.66</td>
<td>0.40</td>
<td>0.73</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>Neural Network</td>
<td>0.60</td>
<td>0.81</td>
<td>0.68</td>
<td>0.42</td>
<td>0.74</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>Random Forest</td>
<td>0.57</td>
<td>0.84</td>
<td>0.66</td>
<td>0.42</td>
<td>0.74</td>
<td>0.79</td>
</tr>
</tbody>
</table>

Table 4.4c. Performance statistics of machine learning-based predictors using all features with GPS and MUSITE predictions (meta-predictors) on held-out test set

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Machine Learner</th>
<th>Spec</th>
<th>SN</th>
<th>ppv</th>
<th>MCC</th>
<th>F</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hold-out Test</td>
<td>ADA</td>
<td>0.68</td>
<td>0.88</td>
<td>0.74</td>
<td>0.57</td>
<td>0.80</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>Neural Network</td>
<td>0.70</td>
<td>0.88</td>
<td>0.74</td>
<td>0.58</td>
<td>0.80</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>Random Forest</td>
<td>0.71</td>
<td>0.86</td>
<td>0.74</td>
<td>0.58</td>
<td>0.80</td>
<td>0.87</td>
</tr>
</tbody>
</table>
Figure 4.2. Recursive operating characteristic curves of machine learning-based predictors built with different features on hold-out tests. Three combinations were considered: a) Biological features (disorder, accessibility, secondary structure, and conservation), b) Top-10 orthologous amino acid indices (derived from ConSequence approach) and c) a combination of a), b) and GPS and MUSITE predictions. These are arranged in ascending order of predictive performance. GPS and MUSITE are included to enable comparisons with published predictors. In all cases, my predictors outperform GPS. With the biological features alone (a), my predictors underperform compared to MUSITE. When using the physicochemical properties derived from the amino acid indices, my predictor performance increases and is on par with MUSITE. When all features and predictions from GPS and MUSITE are used together, I get superior performance over all other combinations and published predictors. The performance of my predictors appears to be unaffected by the machine-learning algorithm used to build them.
4.4.2. Predictive performance of published predictors shows varied performance on functional datasets

With over 40 different P-site predictors, each with their own approach to building a predictive model, it is difficult to choose which are the most suitable. Moreover, there has not yet been any assessment conducted on the performance of predictors towards functional sites.

Briefly, all predictors can be categorised into two different classes; kinase-specific (KSPs) and non-kinase-specific predictors (nKSPs) whose predictions are based on pre-defined kinase-substrate relationships and enriched properties at/or surrounding the substrate site respectively. Because KSPs are built explicitly using linear motif information they are highly accurate but may struggle with detecting novel sites. NKSPs on the other-hand are more sensitive than KSPs which overcomes the prediction of novel sites but the trade-off for this increased sensitivity is a decrease in specificity where they may be more vulnerable to predicting false positives.

Here, the performance of five widely-used predictors was assessed: GPS 2.1 and PPSP 1.0 (KSPs); DISPHOS 1.3 and netphos 2.0 (nKSPs) and MUSITE (hybrid between KSP and nKSP) with my three different functional datasets (PSP+, Q+ and Q-) where analyses reveal differences in performance as a function of predictor class (Fig. 4.3). Due to the low number of threonine and tyrosine found in the quantitative sets, all predictors were assessed using serine to allow a fairer comparison.

For the nKSPs, both perform well on the quantitative sets (and notably DISPHOS on the Q-). The KSPs demonstrate similar performance for the PSP+ and quantitative sets with the former being marginally superior. For the quantitative sets, GPS predict the Q+ almost as well as the PSP+ while for PPSP, the Q+ performed on par with the Q-.

However, the key observation that separates nKSPs from KSPs (more clearly between DISPHOS and GPS respectively) is that the former is reveals some enrichment towards the Q- set. It was remarked in the
previous chapter that using features that pertain to physical properties (disorder, accessibility and secondary structure) could be counterproductive in finding functional sites because they showed enrichment for my less functional Q- set. The observed prediction towards the Q- set by nKSPs appears to suggest that inclusion of these features and their ability to enrich the Q- have carried forward into the predictors.
Figure 4.3. Summary predictive performance of four published phosphorylation site predictors against different datasets. Predictors have been separated according to their classes: kinase-specific and non-kinase-specific predictors. DISPHOS and NetPhos (v2) are non-kinase-specific while GPS and PPSP are the kinase-specific predictors. The datasets compared are the non-phosphorylated (negative, white), gold-standard known functional sites through manual curation (PSP+, green), and assumed functional and less-functional (Q+, yellow and Q-, red respectively) from quantitative studies. The non-kinase specific predictors demonstrate good performance on the quantitative sets (notably the Q-) while the KSPs are perform well on the functional sets. Interestingly, the nKSPs tend to more strongly predict the less-functional groups. MUSITE, a hybrid predictor using nKSP and KSP information, exhibits similar performance traits to nKSPs where the quantitative sets are predicted well with focus on the less-functional; a trend that maybe rooted in the inclusion of features such as disorder.
4.4.3. Comparative evaluation of all predictors on functional sets

Next, the performances of the published and my general ML-based predictors against the functional groups were assessed. In addition, I also assess the performance of a potentially functional ML-based predictor, trained using the features that have been shown in my analyses to be depleted for the less-functional set; conservation and linear-motif prediction via GPS.

Comparisons are conducted using standard performance statistics (sensitivity, specificity, AUC etc) which normalises the different scoring metrics used therefore allowing direct comparability. None of these phosphosites (PSP+, Q+ and Q-) were included in the training and optimisation process to avoid creating biased models.

Again, due to the low numbers of threonine and tyrosine residues in the Q- set, all prediction assessments were conducted against serine to allow for comparisons between both published and my own ML-based predictors.

Beginning with the published predictors (Table 4.5), the nKSPs and KSPs have similar performances with average AUCs of ~0.71 and ~0.70 respectively while MUSITE has an AUC of 0.68 against the PSP+ set.

In the Q+ analyses, with the exception of DISPHOS and MUSITE, all predictors display a fall or have equivalent performance to the PSP+ set.

In the Q- analyses, the nKSPs appear to better predict the less-functional sites with DISPHOS achieving its best performance here with an AUC of 0.78 while MUSITE has the second highest AUC of 0.73.

Critically, the KSPs have the poorest predictive performances here with GPS and PPSP each achieving AUCs of 0.64 indicating that these types of predictors have the best chance to avoid less-functional sites.
For my ML-based predictors (Table 4.6a - d), the biological-, physicochemical-based, meta- and functional predictors exhibited similar predictive performance with AUCs of ~0.72, ~0.75, ~0.76 and ~0.71 respectively against the PSP+ set.

Against the Q+ set, the meta-predictors did not display superior predictive performance as expected but was similar to the physicochemical-based predictors (both with AUCs of ~0.74) while the biological- and functional predictors demonstrated the lowest AUCs (~0.68 and ~0.66 respectively).

Finally, my predictors predict the Q- sites with AUCs of ~0.71, ~0.76 and ~0.72, ~0.59 for biological-, physicochemical, meta- and functional-predictors respectively. It appears that the inclusion of physicochemical-properties derived surrounding residues maybe responsible for this improved performance towards the less functional set. Critically, by augmenting GPS predictions with conservation; a decrease in predictive power towards the Q- set relative to any other predictor is achieved while maintaining a reasonable performance (with the exception of the neural network-based predictor) for the functional sets. This suggests that supplementing motif-based phosphorylation site prediction with conservation has the ability to avoid the less-functional phosphosites and can form the basis of a functional predictor.
Table 4.5 Performance statistics of published predictors on functional datasets. Each statistic is taken at the optimal (maximum) F-score. ‘PSP+’ and ‘Q+’ are functional while ‘Q-’ are deemed less functional sites. ‘KSP’ and ‘nKSP’ are kinase-specific and non-kinase-specific predictors and ‘hybrid’ are a mixture between the two.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Predictor type</th>
<th>Predictor</th>
<th>Sp</th>
<th>SN</th>
<th>Ppv</th>
<th>MCC</th>
<th>F</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSP+</td>
<td>nKSP</td>
<td>DISPHOS</td>
<td>0.28</td>
<td>0.94</td>
<td>0.57</td>
<td>0.29</td>
<td>0.71</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Netphos</td>
<td>0.30</td>
<td>0.87</td>
<td>0.56</td>
<td>0.22</td>
<td>0.68</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>KSP</td>
<td>PPSP</td>
<td>0.32</td>
<td>0.86</td>
<td>0.56</td>
<td>0.22</td>
<td>0.68</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GPS</td>
<td>0.56</td>
<td>0.75</td>
<td>0.63</td>
<td>0.31</td>
<td>0.68</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td>Hybrid</td>
<td>MUSITE</td>
<td>0.30</td>
<td>0.86</td>
<td>0.51</td>
<td>0.20</td>
<td>0.64</td>
<td>0.68</td>
</tr>
<tr>
<td>Q+</td>
<td>nKSP</td>
<td>DISPHOS</td>
<td>0.49</td>
<td>0.86</td>
<td>0.63</td>
<td>0.38</td>
<td>0.73</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Netphos</td>
<td>0.24</td>
<td>0.91</td>
<td>0.55</td>
<td>0.21</td>
<td>0.68</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>KSP</td>
<td>PPSP</td>
<td>0.29</td>
<td>0.89</td>
<td>0.56</td>
<td>0.23</td>
<td>0.69</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GPS</td>
<td>0.29</td>
<td>0.89</td>
<td>0.55</td>
<td>0.21</td>
<td>0.68</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>Hybrid</td>
<td>MUSITE</td>
<td>0.38</td>
<td>0.90</td>
<td>0.58</td>
<td>0.33</td>
<td>0.70</td>
<td>0.71</td>
</tr>
<tr>
<td>Q-</td>
<td>nKSP</td>
<td>DISPHOS</td>
<td>0.54</td>
<td>0.87</td>
<td>0.65</td>
<td>0.43</td>
<td>0.75</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Netphos</td>
<td>0.12</td>
<td>0.96</td>
<td>0.52</td>
<td>0.16</td>
<td>0.68</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>KSP</td>
<td>PPSP</td>
<td>0.04</td>
<td>1</td>
<td>0.51</td>
<td>0.12</td>
<td>0.67</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GPS</td>
<td>0.17</td>
<td>0.96</td>
<td>0.51</td>
<td>0.22</td>
<td>0.67</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>Hybrid</td>
<td>MUSITE</td>
<td>0.47</td>
<td>0.87</td>
<td>0.56</td>
<td>0.36</td>
<td>0.68</td>
<td>0.73</td>
</tr>
</tbody>
</table>
**Table 4.6a.** Performance statistics of machine learning-based predictors constructed from biological features (disorder, accessibility, secondary structure and conservation) on functional datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Machine Learner</th>
<th>Spec</th>
<th>SN</th>
<th>Ppv</th>
<th>MCC</th>
<th>F</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ADA</td>
<td>0.34</td>
<td>0.90</td>
<td>0.64</td>
<td>0.30</td>
<td>0.75</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>Neural Network</td>
<td>0.34</td>
<td>0.91</td>
<td>0.64</td>
<td>0.30</td>
<td>0.76</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>Random Forest</td>
<td>0.38</td>
<td>0.88</td>
<td>0.66</td>
<td>0.32</td>
<td>0.75</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>ADA</td>
<td>0.34</td>
<td>0.88</td>
<td>0.51</td>
<td>0.25</td>
<td>0.64</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td>Neural Network</td>
<td>0.36</td>
<td>0.88</td>
<td>0.52</td>
<td>0.28</td>
<td>0.66</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>Random Forest</td>
<td>0.34</td>
<td>0.90</td>
<td>0.52</td>
<td>0.28</td>
<td>0.66</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>ADA</td>
<td>0.44</td>
<td>0.87</td>
<td>0.50</td>
<td>0.33</td>
<td>0.64</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td>Neural Network</td>
<td>0.47</td>
<td>0.88</td>
<td>0.51</td>
<td>0.36</td>
<td>0.65</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>Random Forest</td>
<td>0.38</td>
<td>0.91</td>
<td>0.48</td>
<td>0.32</td>
<td>0.63</td>
<td>0.70</td>
</tr>
</tbody>
</table>

**Table 4.6b.** Performance statistics of machine learning-based predictors constructed from top-10 orthologous amino acid indices on functional datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Machine Learner</th>
<th>Spec</th>
<th>SN</th>
<th>Ppv</th>
<th>MCC</th>
<th>F</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ADA</td>
<td>0.24</td>
<td>0.93</td>
<td>0.60</td>
<td>0.24</td>
<td>0.73</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>Neural Network</td>
<td>0.32</td>
<td>0.90</td>
<td>0.62</td>
<td>0.29</td>
<td>0.73</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>Random Forest</td>
<td>0.32</td>
<td>0.90</td>
<td>0.62</td>
<td>0.28</td>
<td>0.73</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>ADA</td>
<td>0.40</td>
<td>0.90</td>
<td>0.60</td>
<td>0.34</td>
<td>0.72</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>Neural Network</td>
<td>0.44</td>
<td>0.88</td>
<td>0.60</td>
<td>0.35</td>
<td>0.72</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>Random Forest</td>
<td>0.45</td>
<td>0.87</td>
<td>0.62</td>
<td>0.36</td>
<td>0.72</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>ADA</td>
<td>0.44</td>
<td>0.91</td>
<td>0.62</td>
<td>0.39</td>
<td>0.73</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>Neural Network</td>
<td>0.44</td>
<td>0.92</td>
<td>0.62</td>
<td>0.41</td>
<td>0.74</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>Random Forest</td>
<td>0.54</td>
<td>0.87</td>
<td>0.65</td>
<td>0.43</td>
<td>0.74</td>
<td>0.77</td>
</tr>
</tbody>
</table>
Table 4.6c. Performance statistics of machine learning-based predictors using all features with GPS and MUSITE predictions (meta-predictors) on functional datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Machine Learner</th>
<th>Spec</th>
<th>SN</th>
<th>Ppv</th>
<th>MCC</th>
<th>F</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSP+</td>
<td>ADA</td>
<td>0.24</td>
<td>0.94</td>
<td>0.64</td>
<td>0.28</td>
<td>0.76</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>Neural Network</td>
<td>0.29</td>
<td>0.94</td>
<td>0.64</td>
<td>0.29</td>
<td>0.76</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>Random Forest</td>
<td>0.34</td>
<td>0.9</td>
<td>0.66</td>
<td>0.3</td>
<td>0.76</td>
<td>0.76</td>
</tr>
<tr>
<td>Q+</td>
<td>ADA</td>
<td>0.46</td>
<td>0.86</td>
<td>0.56</td>
<td>0.34</td>
<td>0.68</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>Neural Network</td>
<td>0.5</td>
<td>0.86</td>
<td>0.57</td>
<td>0.36</td>
<td>0.68</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>Random Forest</td>
<td>0.52</td>
<td>0.82</td>
<td>0.57</td>
<td>0.36</td>
<td>0.68</td>
<td>0.72</td>
</tr>
<tr>
<td>Q-</td>
<td>ADA</td>
<td>0.32</td>
<td>0.92</td>
<td>0.40</td>
<td>0.28</td>
<td>0.56</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>Neural Network</td>
<td>0.52</td>
<td>0.80</td>
<td>0.44</td>
<td>0.30</td>
<td>0.57</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>Random Forest</td>
<td>0.64</td>
<td>0.72</td>
<td>0.50</td>
<td>0.35</td>
<td>0.59</td>
<td>0.73</td>
</tr>
</tbody>
</table>

Table 4.6d. Performance statistics of machine learning-based predictors using conservation and GPS predictions only as features in order to build a functional predictor

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Machine Learner</th>
<th>Spec</th>
<th>SN</th>
<th>Ppv</th>
<th>MCC</th>
<th>F</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSP+</td>
<td>ADA</td>
<td>0.44</td>
<td>0.84</td>
<td>0.68</td>
<td>0.32</td>
<td>0.76</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>Neural Network</td>
<td>0.05</td>
<td>0.99</td>
<td>0.60</td>
<td>0.11</td>
<td>0.74</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>Random Forest</td>
<td>0.40</td>
<td>0.86</td>
<td>0.68</td>
<td>0.29</td>
<td>0.76</td>
<td>0.73</td>
</tr>
<tr>
<td>Q+</td>
<td>ADA</td>
<td>0.23</td>
<td>0.92</td>
<td>0.50</td>
<td>0.21</td>
<td>0.64</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>Neural Network</td>
<td>0.28</td>
<td>0.87</td>
<td>0.50</td>
<td>0.18</td>
<td>0.63</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>Random Forest</td>
<td>0.24</td>
<td>0.93</td>
<td>0.50</td>
<td>0.23</td>
<td>0.65</td>
<td>0.67</td>
</tr>
<tr>
<td>Q-</td>
<td>ADA</td>
<td>0.15</td>
<td>0.96</td>
<td>0.40</td>
<td>0.18</td>
<td>0.56</td>
<td>0.59</td>
</tr>
<tr>
<td></td>
<td>Neural Network</td>
<td>0.20</td>
<td>0.90</td>
<td>0.40</td>
<td>0.12</td>
<td>0.55</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>Random Forest</td>
<td>0.22</td>
<td>0.92</td>
<td>0.41</td>
<td>0.18</td>
<td>0.57</td>
<td>0.59</td>
</tr>
</tbody>
</table>
4.5. Discussion

4.5.1. Kinase-specific predictors are capable of avoiding less-functional sites

From my evaluations of different predictors against functionally assigned datasets, there is evidence for a difference in performance between the KSP- and nKSP-type predictors for the less-functional sites. For KSP-type predictors, the less-functional phosphosites are the worst predicted out of the phosphorylation groups tested. Conversely, the nKSPs appear to predict the less-functional as well (or better) than the functional groups.

The performance of the nKSPs conform to the results in my previous chapter where I found that features pertaining to the surrounding area around substrate sites may be counterproductive due to enrichment for the less-functional sites. This is apparent when inspecting the Q- performance measures of both DISPHOS (which uses disorder information and other local properties) and my predictors trained using only the top-10 orthologous physicochemical features (Table 4.6b).

In addition, MUSITE is a hybrid method that supplements predictions with features such as disorder and consequently performs well for the less-functional sites; further supporting the suggestion that the inclusion of these features are detrimental for functional prediction.

From these results, it would seem that KSPs are the logical choice of predictor in the pursuit of functional phosphosites due to them being inherently less-inclined towards strongly predicting less-functional sites.

4.5.2. Kinase-specific predictors are also capable of detecting novel functional sites

The Q+ set represent those sites that have assumed functional importance based on their satisfactory quantitative change upon stimulation. More importantly, because these are the
product of experiments aimed probing the phosphoproteome under different conditions, it is this set where functionally novel sites are more likely to reside; making these sites candidates for further functional characterisation.

Because the training of KSP-type predictors require linear motifs that have been pre-annotated with their effector kinase, a potential drawback is a lower sensitivity to novel substrate sites whose linear motifs were not considered (or lacking) in the training set.

My results demonstrate that this is not the case where GPS displayed better prediction for the Q+ over the Q- and almost as well as the PSP+ set. However, while the performance statistics of PPSP appear to suggest that PPSP has good predictive ability towards the Q+ over the Q-, it is apparent that their scoring distributions largely overlap; suggesting that it is difficult to discern between the Q+ and Q- using this particular KSP. This is likely due to the range of linear motifs considered by both predictors where PPSP is much more limited (~60 motifs) compared to GPS (~400); increasing the overall sensitivity of GPS.

MUSITE also performed well at detecting the novel Q+ sites. This is likely attributed to its KNN feature which is an automated way of finding novel linear motifs and thereby circumventing the pre-defined motif limit of conventional KSPs.

4.5.3. Suggestions to improve functional prediction of phosphosites

In this chapter I have trained several phosphosite predictors using different combinations of features. For general phosphorylation prediction, the most powerful incorporates all conventional phosphorylation and physicochemical features along with the predictions from GPS and MUSITE predictors. Unsurprisingly, with regards to functional assessment I find that features describing the properties of the surrounding residues result in better prediction of
the less-functional (Q-) dataset; conforming to the findings in Chapter 3 and the performances of the nKSPs here.

Clearly, the only features that can be used to develop a functional predictor are conservation and linear motifs (from KSP analyses) which demonstrated a smaller enrichment for the less functional group. When used together to train predictive models, I find that their combination resulted in a reduced ability (signified by the lowest AUC relative to any other predictor) to predict the less-functional sites. This suggests it is likely that these two features will play a major part in future renditions of functional predictors. One possible suggestion to further improve upon these features would be to estimate conservation from the viewpoint of a linear-motif.

Presently, it is protocol to compute conservation of a particular site via either a site-centric approach (using solely the site itself) or to employ a window-based method which uses a defined number of residues neighbouring the site (Capra & Singh, 2007). A linear-motif approach should be more accurate relative to these two methods as it would focus on the key residues that define the motif for kinase attack.

However, developing linear-motifs targeted by kinases relies on pre-annotated data which is unfortunately sparse. In human, there are estimated to be approximately 500 kinases and many of these have yet to be fully characterised. An elegant solution that could overcome this issue was implemented in the MUSITE predictor where sequences are clustered together. Here, the rationale is that true substrates will have more similar motifs to each other than to the non-substrates. Production of these clusters would allow profiles to be built and therefore PSSMs which we can focus the conservation calculations on.
4.6. Conclusion

In this study, I have evaluated the performance of different published predictor algorithms and my own (developed through various combinations of features) against different sets of phosphorylation sites to address the question of which are the most suitable for general and, more importantly, functional phosphosite prediction. In comparing kinase- and non-kinase-specific predictors (KSPs and nKSPs), my analyses suggest that nKSPs maybe more susceptible to detecting less-functional sites. The reason for this is attributed to features that describe the physicochemical surroundings of a substrate site which were shown in the previous chapter to enrich for this subset of sites. On the other hand, the KSPs appear to be more able in discriminating against the less-functional set which suggests they are more suitable for functional prediction.

For my own predictors, I find that using the entire array of features results in superior general predictive performance to any other feature combination and also two popular prediction tools (GPS and MUSITE). However, these predictors also yield a high predictive performance towards the less-functional set; making it difficult to discern the functional from the less functional sites. Critically, for functional prediction, my predictor trained on only conservation and GPS predictions was the most promising in avoiding the less-functional; signifying the start of creating an effective functional predictor.
**4.7. References**


in Arabidopsis thaliana and a plant-specific phosphorylation site predictor. Nucleic Acids Research, 36, D1015-D1021.


STARK, C., SU, T., BREITKREUTZ, A., LOURENCO, P., DAHABIEH, M.


Chapter 5: Conclusions

The work in this thesis has presented some approaches that may assist in addressing key issues surrounding the accurate annotation of phosphorylation sites. More specifically, I concentrate on two problems associated with the field: ambiguous site localisation of MS-based phosphoproteomic data and prediction of functional phosphosites. Both problems are highly relevant as underestimating them will lead astray the wet-lab biologist in their pursuit of functionally characterizing the role of phosphosites.

The work presented in this thesis demonstrates how more successful phosphorylation localisation can be achieved through supplementing phosphopeptide identifications with even one additional informatics tool. The subsequent task of finding functional phosphorylation sites is difficult but with an appropriate predictor, one can increase their chance of fulfilling this goal.

There were two key aspects in this work:

1) Improving site localisation of data from MS-based phosphoproteomic experiments
   A) Benchmarking the performance of widely-used tools in phosphoproteomic analysis with regards to both identification and localisation (Chapter 2)
   B) Application of a multi-tool approach to rectify incorrect site localisations due to ambiguity between tools (Chapter 2)

2) Improving prediction of functional phosphosites
   A) Assessing common phosphorylation-characteristic features to observe whether any enrichment exists for the functional subset (Chapter 3)
   B) Evaluating the performance of available prediction algorithms on datasets of varying
degrees of functionality (Chapter 4)

C) Utilizing phosphorylation-characteristic features to train predictors for general and functional phosphosite prediction (Chapter 3 and 4)

The annotation of phosphorylation status is gradually becoming easier thanks to the efforts made by all those working in the phosphoproteomic field. However, as exemplified by the iPRG 2010 study, there is ambiguity across different informatics tools used in phosphoproteomic analysis which makes the task of correctly pinpointing phosphosites from MS-based phosphoproteomic data challenging. While it is surprising to find that different tools perform so differently given their underlying goal is the detection and localisation of phosphosites, the bigger problem is that the large selection of tools is a double-edge sword to wet-lab biologists. Indeed, there are a variety of tools that can be used but which are the most appropriate?

This is an important question to ask because each tool varies from another with respect to their type (notably database search-engines, de novo sequencers and spectral libraries), algorithms for distinguishing and selecting the set of true peaks from noise and the scoring scheme applied. Ultimately, this variation is responsible for a major informatic limitation of many MS-based phosphoproteomic studies where phosphopeptides (and sites if possible) are found by one tool but not another and vice versa.

My work in Chapter 2 explores this question where I analyze synthetic phosphorylation data generated under different activation conditions and benchmark a variety of algorithms for identification and localisation. As expected, some of the activation methods outperform others. Most notably, HCD which was able to capture the largest fraction of the total synthetic phosphorylation population. Critically, there are some tools that are indeed more suitable for certain activation methods, for example, ProteinProspector performed well for electron-transfer (ET)-derived MS/MS
owing to the efforts made in implementing ET-idiosyncratic fragmentation behavior into the annotation and scoring scheme.

Surprisingly, initial evaluation of how well both search-engines and site localisers perform demonstrate that both are perfectly adequate from a conventional peptide-spectrum-match (PSM) point of view (achieving real FDRs close to or better than the intended rate) where real false discovery rates (FDRs) close to their intended rate or better were achieved. However, problems arose when acquiring the unique set of phosphopeptides and sites (PSM-to-peptide/site) derived from these statistically satisfactory PSMs. Re-computing the FDR using this set resulted in a substantial increase with regards to both identification and localisation. As the majority of MS-based phosphoproteomic experiments are conducted from a PSM perspective to calculated statistics and then extract the resulting (unique) phosphopeptides and sites, could this be one potential explanation as to how false positives are finding their way into public data repositories?

Chapter 2 advocates the use of a multi-tool approach (for both identification and localisation) where cross-referencing with even one additional tool is capable of substantially reducing the number of false positives acquired. This was more effective for identification though this could be considered the vital step as the correct phosphopeptide is first required otherwise attempting site localisation becomes moot.

However, despite the advantages of multiple informatics tools, there are also associated caveats that prevent their widespread adoption into analytical workflows. Without a doubt, the most prominent issue is their installation where due to various reasons such as specific operating-systems, software versions and pre-requisite software packages a tool will fail to work; hindering any lab with little or no informatics support.

Following the correct localisation of phosphorylation sites comes the undoubtedly more challenging task of functional annotation. The reason for this is not simply due to the possibility that some sites
may lack function but rather because there are very few predictors that have the ability to predict functional sites; they are mostly concerned with the general prediction of phosphosites.

To delve into this relatively new aspect of phosphoproteomics, we must first determine whether there are any differences in the functional subset which can then be used as the basis for constructing a scheme to train and optimize a functional phosphosite predictor. Generic phosphorylation site prediction relies on the use of phosphorylation-characteristic features that separate those that are phosphorylated from those that are not so naturally, these are the prime candidate features to evaluate.

Through assessment of a variety of commonly used features in Chapter 3, it is revealed that there are both productive and counter-productive features that aid and hinder the search of functional sites.

Expectedly, conservation was found to be the most crucial feature in discriminating between those with and without functional importance. Furthermore, the degree of conservation associated with the less-functional phosphosites was less than non-phosphorylated sites; strongly advocating its use in a prediction-context. But perhaps the most interesting result was that features pertaining to phosphorylation susceptibility showed some enrichment for the less-functional subset. Because features such as protein disorder have become a staple feature in some predictors, a logical assumption is that they predict those less-functional as well (or even better) than those with function.

This in turn would have a profound impact upon experimental biologists who intend to use these in silico tools either to focus upon a select number of sites to then experimentally investigate or validate those that have been found. Successful prediction of those that are indeed less-functional would not be desirable and hinder our progress.

This was the focus of Chapter 4 where I evaluate the performance of published kinase- and non-kinase-specific predictors (KSPs and nKSPs) and find there is evidence to support the hypotheses made above.
With this in mind, I develop and assess the predictive capabilities of my own predictors using powerful machine-learning approaches; each built using different subsets of features to find the most suitable combination(s) towards the functional prediction problem.

Here, GPS (a KSP) predictions and conservation, which individually did not display enrichment towards the less-functional phosphosites, were used to create an ML-model that demonstrated a diminished ability to predict these phosphosites compared with any other predictor evaluated in this work.

A way to further debilitate our ability to predict the less-functional sites would be to improve upon the conservation calculations by incorporating the conservation of the positions that form the linear motif as opposed to using only the phosphosite position itself. Clearly, if a site is truly phosphorylated and functional, the linear motif would be conserved to ensure successful enzymatic interaction with the appropriate kinases and phosphatases and therefore allow the site to continue its biological role.

Overall, this computational study has explored methods that have the capability to first improve phosphorylation site annotation from MS-based data and second, predict those that may be worth functional pursuit. This first goal was achieved through a simple but powerful multi-tool approach that was able to remove the false positive instances. While the second part was not resolved, I am confident that I have uncovered some key characteristics of functional and less-functional sites that will bring us closer to generating functional predictors.