TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Genomorientierte Bioinformatik

Structure-Based Analysis of Tissue-Specific Post-translational Modifications

NERMIN PINAR KARABULUT

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender: Prof. Dr. Aurélien Tellier
Prüfer der Dissertation:

1. Prof. Dr.rer.nat. Dmitrij Frischmann
2. Prof. Dr.rer.nat. Jürgen Cox (University of Copenhagen)

Die Dissertation wurde am 07.04.2016 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 20.06.2016 angenommen.

Nermin Pinar Karabulut: *Structure-Based Analysis of Tissue-Specific Post-translational Modifications,* A dissertation in Bioinformatics, © January 2016

SUPERVISOR:
Dmitrij Frishman

Life is not easy for any of us. But what of that? We must have perseverance and above all confidence in ourselves. We must believe that we are gifted for something and that this thing must be attained.

— Marie Curie

Dedicated to my mother Nuray Tümer.

ABSTRACT

Ever since the posttranslational modifications (PTMs) were discovered and associated to evolution of many diseases, both experimental and computational studies have gained pace to better understand the mechanism behind PTMs. This thesis aims to investigate two kinds of PTMs – acetylation and phosphorylation – in a tissue-specific manner by utilizing the sequence and structural characteristics contained in their environments. In the first part, we present a comprehensive tissue-based analysis of sequence and structural features of lysine acetylation sites (LASs). We show that acetylated substrates are characterized by tissue-specific motifs both in linear amino acid sequence and in spatial environments. We further demonstrate that the general tendency of LASs to reside in ordered regions and, specifically, in α-helices, is also subject to tissue specific variation. In line with previous findings we show that LASs are generally more evolutionarily conserved than non-LASs, especially in proteins with known function and in structurally regular regions. On the other hand, as revealed by metabolic pathway analysis, LASs have diverse cellular functions in different tissues and are frequently associated with tissue-specific protein domains. In the second part, we present the first comprehensive analysis of global and tissue-specific sequence and structure properties of phosphorylation sites utilizing recent proteomics data. We identified tissue-specific motifs in both sequence and spatial environments of phosphorylation sites. Target site preferences of kinases across tissues indicate that, while many kinases mediate phosphorylation in all tissues, there are also kinases that exhibit more tissue-specific preferences which, notably, are not caused by tissue-specific kinase expression. We also demonstrate that many metabolic pathways are differentially regulated by phosphorylation in different tissues. The findings obtained from these two parts of the thesis may imply the existence of tissue-specific enzymes and proteases regulating posttranslational modifications. In the last part of the thesis, we present the first tissue-specific phosphorylation site prediction approach, TSPhosPred (Tissue-Specific Phosphorylation Prediction) based on the feature set consisting of sequence-based and structure-based environment characteristics of phosphorylation sites as well as functional annotations. Experimental structures along with predicted structures are also utilized, and yield an improved accuracy over existing tools in both cross-validation and independent testing. Supportively, the cross-tissues prediction strengthens the necessity and the significance of tissue-specific models to obtain improved prediction of phosphorylation sites.

ZUSAMMENFASSUNG

Seit der Entdeckung von Posttranslationaler Modifikation (PTM) und ihrer Bedeutung in vielen Krankheiten ist die Anzahl an experimentellen und theoretischen Studien, die sich mit den Mechanismus hinter den PTM beschäftigen gestiegen. In dieser Doktorarbeit analysieren wir Acetylierung und Phosphorylierung auf Gewebespezifische Sequenzen und Strukturen. Im ersten Teil der Arbeit zeigen wir eine umfassende Sequenz- und Struktur-Analyse von Gewebespezifischen Lysin Acetylierung Stellen (LASs). Wir zeigen, dass acetylierte Substrate durch gewebespezifische Sequenz und Struktur Motive gekennzeichnet sind. Des weiteren zeigen wir, dass LASs bevorzugt in strukturell geordneten Regionen und Alpha-Helixen vorkommen. Übereinstimmend mit vorhergehenden Studien zeigen wir, dass LASs stärker konserviert sind als nicht LASs, vor allem in Proteinen mit bekannter Funktion und in strukturell regelmäßigen Regionen. Die Analyse von Stoffwechselwegen zeigten, dass LASs eine Vielzahl an gewebespezifische Funktionen haben die häufig mit gewebespezifische Proteindomänen assoziiert ist. Der zweite Teil enthält eine umfassende Analyse von globalen und gewebespezifischen Sequenzen und Strukturen von Phosphorylierungsstellen. Wir identifizierten gewebespezifische Motive sowohl in der Sequenz als auch der Struktur von Phosphorylierungsstellen. Die beobachtete gewebespezifische Präferenz von Kinase Zielen zeigte, dass diese gewebespezifische Präferenz einiger Kinasen von deren Genexpression unabhängig ist. Wir zeigten auch, dass viele Stoffwechselwege durch gewebespezifische Phosphorylierung reguliert sind. Die Ergebnisse aus beiden Teilen der Doktorarbeit implizieren die Existenz von Enzymen und Proteasen die gewebespezifische Phosphorylierung regulieren. Im letzten Teil der Doktorarbeit stellen wir die erste gewebespezifischen Phosphorylierungsstellen vorhersage Methode TSPhosPred (Tissue-Specific Phosphorylation Prediction) vor. TSPhosPred nutzt zur gewebespezifischen Phosphorylierungsstellen Vorhersage die Phosphorylierungsstellen spezifische Sequenz- und Struktur-eigenschaften. Durch die Kombination von vorhergesagten und experimentell validierten Strukturen erreicht TSPhosPred bessere Genauigkeit als bereits bekannte Methoden. Die gewebespezifischen Kreuzvalidierung validierte die Bedeutung und Notwendigkeit eines gewebespezifischen Modelles zur verbesserten Vorhersage von Phosphorylierungsstellen.

PUBLICATIONS

Some ideas and figures have appeared previously in the following publications:

N.P. Karabulut and D. Frishman. Tissue-Specific Sequence and Structural Environments of Lysine Acetylation Sites. *Journal of Structural Biology* 191(1): 39-48, 2015.

N.P. Karabulut and D. Frishman. Sequence- and Structure-Based Analysis of Tissue-Specific Phosphorylation Sites. *Journal of Proteomics*, submitted, 2016.

N.P. Karabulut and D. Frishman. Prediction of Tissue-Specific Phosphorylation Sites by Integrating Sequence- and Structure-Based Features, manuscript in preparation.

Poster: N.P. Karabulut, S. Tyanova, D. Frishman. Sequence and Structure Analysis of Tissue-Specific Lysine Acetylation Sites. *22nd Annual International Conference on Intelligent Systems for Molecular Biology (ISMB)*, Boston, USA, 2014.

Poster: N.P. Karabulut, S. Tyanova, D. Frishman. Tissue-Specific Sequence and Structural Environments of Lysine Acetylation Sites. *IT For Life Science @ Bayer*, Leverkusen, Germany, 2014.

ACKNOWLEDGMENTS

I wish to express my sincere appreciation and gratitude to my supervisor Dmitrij Frishman for giving me the opportunity to conduct my PhD thesis at the TU München, his support, and worthwhile guidance through my Ph.D. study.

Besides my supervisor, I am also thankful to Prof. Dr. Aurelien Tellier, who was so kind to immediately agree on becoming the chair of my examination committee, and to Dr. Jürgen Cox, who again so kindly agreed on becoming the second examiner of my examination committee. I am really grateful to them.

I gratefully acknowledge the support of the TUM Graduate School's Thematic Graduate Center Regulation and Evolution of Cellular Systems (RECESS) at the TU München. And special thanks to Claudia Luksch for her assistance and help during the RECESS program.

Furthermore, I would like to express my gratitude to my colleagues: Peter Hönigschmid for his expert advice on my machine learning project; Stefka Tyanova for introducing me to post-translational modifications when I first started my Ph.D. study, and for keeping her advices on my project; Florian Goebels for being my office-mate, and sharing his knowledge and experience with me all the time; Usman Saeed for always being open to discuss every problem and help for the grid issues; Drazen Jalsovec for his IT administrative support and especially for helping on grid problems; and to the entire Frishman lab for the friendship we had together.

I am thankful to Florian Goebels and Simon Goebels for the German translation of Abstract of this thesis.

I am especially grateful to Leonie Corry - The Angel - for helping me out with everything one can imagine. I will always be appreciated for her helpfulness, especially when I first moved to Germany. My new life in Germany would have been so hard if I did not know her.

Last but not least, a very big acknowledgement my family and friends definitely deserve. Especially, I thank to my lovely mother for her great love and always being there in every kind of situation. And a very special thanks to my beloved friend Nazire for everything she has brought to my life, and making this Ph.D. life more joyful and meaningful.

CONTENTS

I INTRODUCTION 1

1 INTRODUCTION 3

1.1 Experimental Methods 4

1.2 Acetylation 6

1.3 Phosphorylation 7

1.4 Computational Methods 8

1.5 Thesis Motivation and Outline 10

II TISSUE-SPECIFIC SEQUENCE AND STRUCTURAL ENVIRONMENTS OF LYSINE ACETYLATION SITES 11

2 TISSUE-SPECIFIC SEQUENCE AND STRUCTURAL ENVIRONMENTS OF LYSINE ACETYLATION SITES 13

2.1 Materials and Methods 13

2.1.1 Data collection and preprocessing 13

2.1.2 Sequence (1D) environments of acetylated and reference (non–acetylated) lysine residues 14

2.1.3 Lysine acetylation sites with known 3D structure 15

2.1.4 Statistics 16

2.1.5 Three-dimensional (3D) environments of acetylated and reference (non-acetylated) lysine residues 16

2.1.6 Conservation analysis of lysine acetylation sites 16

2.1.7 Structural features of lysine acetylation sites 17

2.1.8 Analysis of structural folds and functional domains 17

2.1.9 KEGG pathway analysis 17

2.1.10 Abundance of KAT paralogs 17

2.2 Results and Discussion 18

2.2.1 Global and tissue-specific sequence motifs of lysine acetylation sites 18

2.2.2 Global and tissue-specific sequence motifs of lysine acetylation sites in proteins with known 3D structure 21

2.2.3 Spatial environments of lysine acetylation sites 21

2.2.4 Evolutionary conservation of lysine acetylation sites 22

2.2.5 Tissue-specific structural properties of lysine acetylation sites 24

2.2.6 Proteins containing acetylated lysines are involved in tissue-specific biological pathways 25

2.3 Conclusion 27

III SEQUENCE- AND STRUCTURE-BASED ANALYSIS OF TISSUE-SPECIFIC PHOSPHORYLATION SITES 29

3 SEQUENCE- AND STRUCTURE-BASED ANALYSIS OF TISSUE-SPECIFIC PHOSPHORYLATION SITES 31

3.1 Materials and Methods 31
3.1.1 Datasets of phosphorylated and reference (non-phosphorylated) sites 31
3.1.2 Identification of sequence motifs 33
3.1.3 Obtaining 3D structures of phosphorylated proteins 33
3.1.4 Statistics 34
3.1.5 Spatial (3D) environments of phosphorylated and reference (non-phosphorylated) serine/threonine/-tyrosine residues 34
3.1.6 Structural properties of phosphorylation sites 34
3.1.7 Analysis of structural folds and functional domains 34
3.1.8 KEGG pathway analysis 35
3.1.9 Kinase analysis 35
3.1.10 Tissue-specific expression of kinases 35
3.2 Results and Discussion 36
3.2.1 Analysis of sequence motifs of global phosphorylation sites 36
3.2.2 Tissue-based analysis of sequence motifs of phosphorylated sites 37
3.2.3 Tissue-based analysis of phosphorylation sequence motifs in proteins with known 3D structure 40
3.2.4 Tissue-specific spatial motifs of phosphorylation sites 41
3.2.5 Structural properties of phosphorylation sites 42
3.2.6 Phosphorylated proteins take part in tissue-specific biological pathways 43
3.2.7 Kinases target tissue-specific phosphorylation sites 46
3.3 Conclusions 49
IV PREDICTION OF TISSUE-SPECIFIC PHOSPHORYLATION SITES BY INTEGRATING SEQUENCE- AND STRUCTURE-BASED FEATURES 51
4 PREDICTION OF TISSUE-SPECIFIC PHOSPHORYLATION SITES BY INTEGRATING SEQUENCE- AND STRUCTURE-BASED FEATURES 53
4.1 Materials and Methods 53
4.1.1 Data collection and preprocessing 53
4.1.2 Training and independent test sets 54
4.1.3 Feature extraction 56
4.1.4 Machine learning 62
4.1.5 Prediction performance assessment 63
4.1.6 Comparing with existing tools 63
4.2 Results and Discussions 64
4.2.1 Predictive performance of sequence environments surrounding phosphorylation sites 64
4.2.2 KNN scores as features 66

4.2.3 Influence of spatial amino acid content and structural environment of phosphorylation sites on prediction 67
4.2.4 The contribution of functional annotations on phosphorylation prediction 67
4.2.5 Predictive performance of the random forest model on tissue-specific phosphorylation sites 78
4.2.6 Prediction on independent test data 81
4.2.7 Cross-tissues performance evaluation on independent testing 82
4.2.8 Performance comparison with the existing prediction tools 82
4.3 Conclusions 84
V SUMMARY 87
5 SUMMARY 89
VI APPENDIX 91
A APPENDIX 93
BIBLIOGRAPHY 175

LIST OF FIGURES

Figure 1 Evolution of post–translational modifications leading to proteome complexity. 4
Figure 2 Summary of mass spectrometry (MS)–based proteomics for the lysine acetylation. 5
Figure 3 Number of LASs and non-LASs from the LAS1D and LAS3D datasets in different tissues. 15
Figure 4 The comparison of 1D and 3D environments of global lysine acetylation sites. 18
Figure 5 Two sample logo analysis of LASs from the LAS1D dataset in different tissues. 19
Figure 6 Sequence (1D) and structural (3D and *pure* 3D) environments of LASs from the LAS3D dataset represented by two sample logos and circular plots, respectively. 23
Figure 7 KEGG pathway analysis of the acetylated proteins from the LAS3D dataset. 26
Figure 8 A graphical summary of the followed methodology for phosphorylation sites. 32
Figure 9 The comparison of 1D and 3D environments of global serine phosphorylation sites. 36
Figure 10 Two sample logo analysis of PSSs from the PS1D-70 dataset in different tissues. 38
Figure 11 Sequence (1D) and structural (3D and *pure* 3D) environments of PSSs, PTSs, and PYSs from the PS3D-90 dataset represented by two sample logos and circular plots, respectively. 39
Figure 12 KEGG pathway analysis of the serine phosphorylated proteins from the PS1D-70 dataset. 43
Figure 13 Tissue preferences of kinases targeting serine phosphorylated sites in the PS1D-70 dataset. 45
Figure 14 Comparison of the Ser/Thr kinase expression across tissues. 47
Figure 15 Tripartite graph showing interactions between serine phosphorylation motifs, kinases and tissues. 48
Figure 16 Comparison of KNN scores between serine phosphorylation and non-phosphorylation sites depending on tissues in the PS1D data. 64
Figure 17 Comparison of KNN scores between serine phosphorylation and non-phosphorylation sites depending on tissues in the PS3D data. 65
Figure 18 Comparison of KNN scores between threonine phosphorylation and non-phosphorylation sites depending on tissues in the PS1D data. 65

Figure 19 Comparison of KNN scores between threonine phosphorylation and non-phosphorylation sites depending on tissues in the PS3D data. 66
Figure 20 ROC curves of serine phosphorylation prediction models obtained from 10-fold cross-validation in different tissues without using the KNN score. 68
Figure 21 ROC curves of serine phosphorylation prediction models obtained from independent testing in different tissues without using the KNN score. 69
Figure 22 ROC curves of serine phosphorylation prediction models obtained from 10-fold cross validation in different tissues with using the KNN score. 70
Figure 23 ROC curves of serine phosphorylation prediction models obtained from independent testing in different tissues with using the KNN score. 71
Figure 24 ROC curves of threonine phosphorylation prediction models obtained from 10-fold cross-validation in different tissues without using the KNN score. 72
Figure 25 ROC curves of threonine phosphorylation prediction models obtained from independent testing in different tissues without using the KNN score. 73
Figure 26 ROC curves of threonine phosphorylation prediction models obtained from 10-fold cross-validation in different tissues with using the KNN score. 74
Figure 27 ROC curves of threonine phosphorylation prediction models obtained from independent testing in different tissues with using the KNN score. 75
Figure 28 Two sample logo analysis of LASs in different tissues in the LAS1D dataset. 94
Figure 29 Two sample logo analysis of LASs from the LAS1D dataset in brain. 95
Figure 30 Two sample logo analysis of LASs from the LAS1D dataset in brown fat. 95
Figure 31 Two sample logo analysis of LASs from the LAS1D dataset in heart. 96
Figure 32 Two sample logo analysis of LASs from the LAS1D dataset in intestine. 96
Figure 33 Two sample logo analysis of LASs from the LAS1D dataset in kidney. 97
Figure 34 Two sample logo analysis of LASs from the LAS1D dataset in liver. 97
Figure 35 Two sample logo analysis of LASs from the LAS1D dataset in lung. 98
Figure 36 Two sample logo analysis of LASs from the LAS1D dataset in muscle. 98
Figure 37 Two sample logo analysis of LASs from the LAS1D dataset in pancreas. 99

Figure 38 Two sample logo analysis of LASs from the LAS1D dataset in perirenal fat. 99
Figure 39 Two sample logo analysis of LASs from the LAS1D dataset in skin. 100
Figure 40 Two sample logo analysis of LASs from the LAS1D dataset in spleen. 100
Figure 41 Two sample logo analysis of LASs from the LAS1D dataset in stomach. 101
Figure 42 Two sample logo analysis of LASs from the LAS1D dataset in testis fat. 101
Figure 43 Two sample logo analysis of LASs from the LAS1D dataset in testis. 102
Figure 44 Two sample logo analysis of LASs from the LAS1D dataset in thymus. 102
Figure 45 1D environment of LASs from the LAS3D dataset in different tissues. 103
Figure 46 3D environment of LASs from the LAS3D dataset in different tissues. 104
Figure 47 Pure 3D environment of LASs from the LAS3D dataset in different tissues. 105
Figure 48 Secondary structure analysis of LASs and the residues surrounding them from the LAS3D dataset in different tissues. 106
Figure 49 SCOP class analysis of LASs and the residues surrounding them in different tissues in the LAS3D dataset. 107
Figure 50 Global and tissue-specific occurrence of acetylated proteins from the LAS3D dataset in protein domains. 107
Figure 51 The comparison of 1D and 3D environments of global PTSs. 108
Figure 52 The comparison of 1D and 3D environments of global PYSs. 109
Figure 53 Two sample logo analysis of PSSs in different tissues in the PS1D-70 dataset. 110
Figure 54 Two sample logo analysis of PSSs from the PS1D-70 dataset in blood. 111
Figure 55 Two sample logo analysis of PSSs from the PS1D-70 dataset in brain. 111
Figure 56 Two sample logo analysis of PSSs from the PS1D-70 dataset in brainstem. 112
Figure 57 Two sample logo analysis of PSSs from the PS1D-70 dataset in cerebellum. 112
Figure 58 Two sample logo analysis of PSSs from the PS1D-70 dataset in cortex. 113
Figure 59 Two sample logo analysis of PSSs from the PS1D-70 dataset in heart. 113
Figure 60 Two sample logo analysis of PSSs from the PS1D-70 dataset in intestine. 114

Figure 61 Two sample logo analysis of PSSs from the PS1D-70 dataset in kidney. 114

Figure 62 Two sample logo analysis of PSSs from the PS1D-70 dataset in liver. 115

Figure 63 Two sample logo analysis of PSSs from the PS1D-70 dataset in lung. 115

Figure 64 Two sample logo analysis of PSSs from the PS1D-70 dataset in muscle. 116

Figure 65 Two sample logo analysis of PSSs from the PS1D-70 dataset in pancreas. 116

Figure 66 Two sample logo analysis of PSSs from the PS1D-70 dataset in perirenal fat. 117

Figure 67 Two sample logo analysis of PSSs from the PS1D-70 dataset in spleen. 117

Figure 68 Two sample logo analysis of PSSs from the PS1D-70 dataset in stomach. 118

Figure 69 Two sample logo analysis of PSSs from the PS1D-70 dataset in testis. 118

Figure 70 Two sample logo analysis of PSSs from the PS1D-70 dataset in thymus. 119

Figure 71 Two sample logo analysis of PTSs in different tissues in the PS1D-70 dataset. 120

Figure 72 Two sample logo analysis of PYSs in different tissues in the PS1D-70 dataset. 121

Figure 73 Two sample logo analysis of PSSs in different tissues in the PS3D-90 dataset. 122

Figure 74 Two sample logo analysis of PTSs in different tissues in the PS3D-90 dataset. 123

Figure 75 Two sample logo analysis of PYSs in different tissues in the PS3D-90 dataset. 124

Figure 76 3D environment of PSSs from the PS3D-90 dataset in different tissues. 125

Figure 77 3D environment of PTSs from the PS3D-90 dataset in different tissues. 126

Figure 78 3D environment of PYSs from the PS3D-90 dataset in different tissues. 127

Figure 79 Pure 3D environment of PSSs from the PS3D-90 dataset in different tissues. 128

Figure 80 Pure 3D environment of PTSs from the PS3D-90 dataset in different tissues. 129

Figure 81 Pure 3D environment of PYSs from the PS3D-90 dataset in different tissues. 130

Figure 82 Disorder region analysis of PSSs and the residues surrounding them from the PS1D-70 dataset in different tissues. 131

Figure 83 Disorder region analysis of PTSs and the residues surrounding them from the PS1D-70 dataset in different tissues. 132

Figure 84 Disorder region analysis of PYSs and the residues surrounding them from the PS1D-70 dataset in different tissues. 133
Figure 85 Secondary structure analysis of PSSs and the residues surrounding them from the PS3D-90 dataset in different tissues. 134
Figure 86 Secondary structure analysis of PTSs and the residues surrounding them from the PS3D-90 dataset in different tissues. 135
Figure 87 SCOP class analysis of serine phosphorylation sites and the residues surrounding them in different tissues in the PS3D-90 dataset. 136
Figure 88 KEGG pathway analysis of the threonine phosphorylated proteins from the PS1D-70 dataset. 137
Figure 89 KEGG pathway analysis of the tyrosine phosphorylated proteins from the PS1D-70 dataset. 137
Figure 90 Global and tissue-specific occurrence of serine phosphorylated proteins from the PS3D-90 dataset in protein domains. 137
Figure 91 Global and tissue-specific occurrence of threonine phosphorylated proteins from the PS3D-90 dataset in protein domains. 138
Figure 92 Global and tissue-specific occurrence of tyrosine phosphorylated proteins from the PS3D-90 dataset in protein domains. 138

LIST OF TABLES

Table 1 Experimental methodologies for identifying post-translational modifications. 6
Table 2 Data summary of lysine acetylation sites. 14
Table 3 Data summary of phosphorylation sites. 32
Table 4 Data summary of phosphorylation sites used in classification. 54
Table 5 Data summary of phosphorylation sites used in training depending on tissue and residue types in both PS1D and PS3D datasets. 55
Table 6 The summary of features used to train each prediction model. 57
Table 7 The summary of k values chosen for each prediction model in different tissues. 60
Table 8 Classification performance analysis of 10-fold cross validation for each prediction model in different tissues with/without using the KNN score. 78

Table 9 Classification performance analysis of independent testing for each prediction model in different tissues with/without using the KNN score. 81
Table 10 Cross-tissues prediction performance of serine phosphorylated sites based on independent testing in different tissues using the *seqBinaryEnc* category without the KNN score. 83
Table 11 Performance comparison with existing tools. 84
Table 12 Summary of sequence motifs associated with LASs in the LAS1D dataset. 143
Table 13 Summary of sequence motifs associated with LASs in the LAS3D dataset. 145
Table 14 Tissue-specific evolutionary conservation analysis of LASs and non-LASs. 146
Table 15 Accessibility and B-factor analysis of LASs in different tissues. 150
Table 16 Sequence motifs of all phosphorylation sites at 70% sequence similarity with the $p < 0.000001$ significance threshold. 159
Table 17 Two sample logo comparison of phosphorylation sites between our study and the study of Lundby *et al.* in brain. 160
Table 18 Two sample logo comparison of phosphorylation sites between our study and the study of Lundby *et al.* in testis. 161
Table 19 Accessibility and B-factor analysis of PSS in different tissues in the PS3D-90 dataset. 165
Table 20 Accessibility and B-factor analysis of PTS in different tissues in the PS3D-90 dataset. 169
Table 21 Accessibility and B-factor analysis of PYS in different tissues in the PS3D-90 dataset. 173

Part I

INTRODUCTION

1 INTRODUCTION

Protein post-translational modification (PTM) is a mechanism occurring after the translation is completed by ribosomes. It generally refers to the covalently addition of a functional group to a protein that alters the chemical makeup and function of the protein, and eventually leads to different biological outcomes in response to requirements of the cell. PTMs play important roles in protein signaling (Morrison et al., 2002), cellular differentiation (Grotenbreg and Ploegh, 2007), protein degradation (Geiss-Friedlander and Melchior, 2007), localization (Sirover, 2012), and regulations of gene expression (Wang et al., 2015) and protein-protein interactions (Duan and Walther, 2015). PTM cross-talk where the PTM of a protein can also regulate the PTMs of other proteins may also occur that leads to more aspects of protein functions. Modifications are most often regulated by enzymes including kinases, acetyltransferases, glycosyltransferases etc., whereas reversible modifications (removal of functional groups and reverse of the biological activity) are carried by proteases such as phosphatases, deacetylases, glycosidases and so on. Many proteins harbor post-translational modifications, and many domains within proteins are even modified on more than one residue. Moreover, regulatory enzymes also go under auto-modification, yielding a large interconnected network. This complex network carries high importance since the abnormal regulation of PTMs is connected to evolution of many diseases, such as cancer (See Figure 1 for proteome complexity).

To date, more than 400 different PTM types on more than 90000 PTM sites have been identified by experimental analysis (Khoury, Baliban, and Floudas, 2011). These modifications include phosphorylation, acetylation, glycosylation, ubiquitination, methylation, lipidation and so on where phosphorylation and acetylation are the two of most studied PTM types. Identification of PTMs harbors a great insight in understanding the mechanism behind it; however, there still exist many technical challenges in the development of specific detection and purification methods, and these methods are costly and labor-intensive. Alternatively, many computational methods have been improved for *in silico* identification of modification sites. In the remaining part of the Introduction section, we (i) introduce the existing experimental approaches in identifying PTMs, (ii) place a greater focus on phosphorylation and acetylation, which are the PTM types subjected to study in this thesis, (iii) give a background information about computational approaches based on phosphorylation, and (iv) describe the thesis motivation that leaded us to conduct this research.

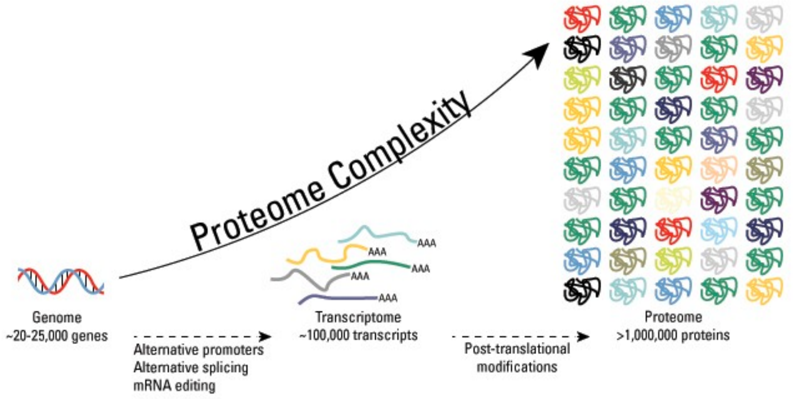
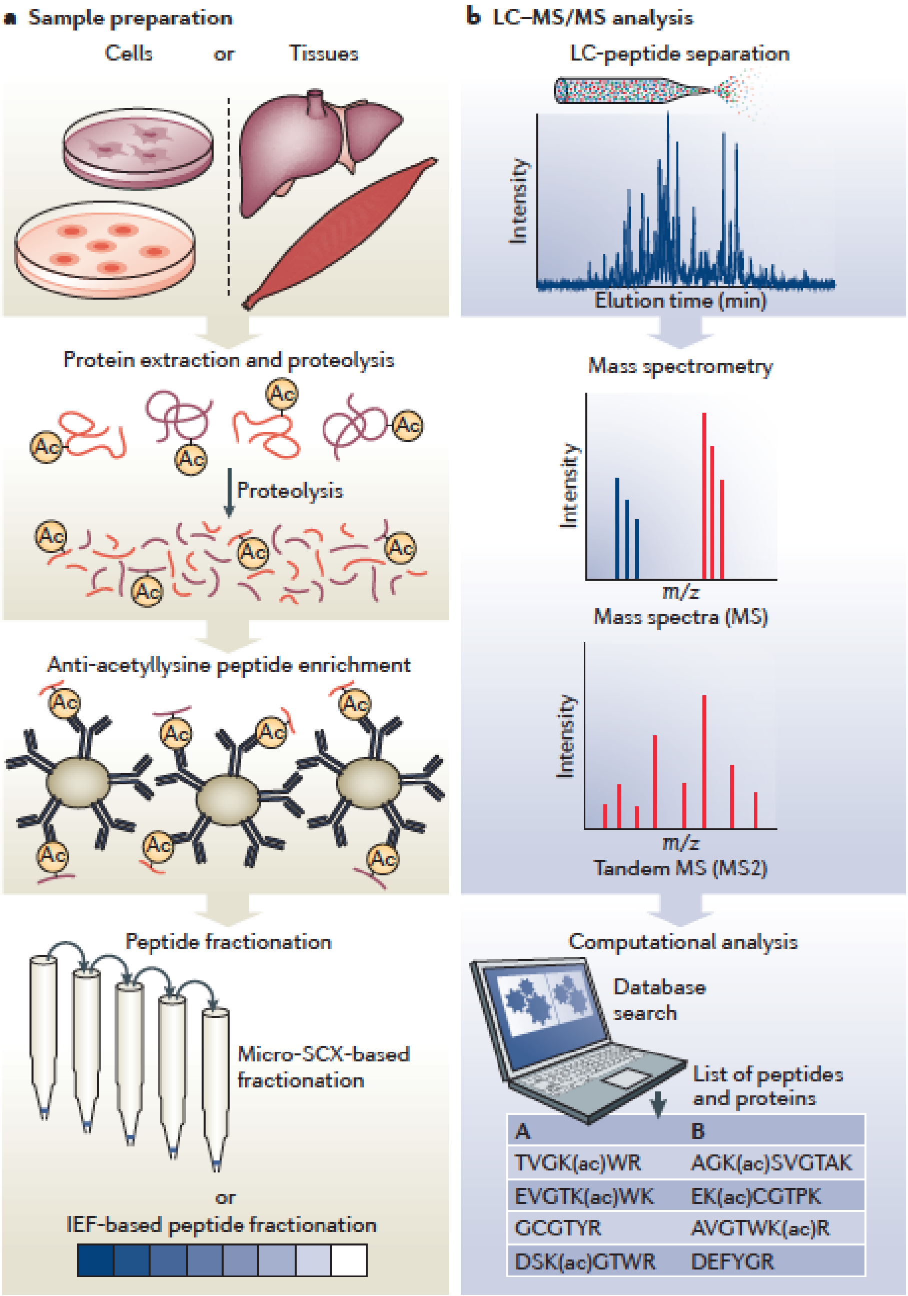


Figure 1: The evolution of post-translational modifications leading to proteome complexity. Human genome consists of 20000 - 25000 genes, whereas in the transcriptome level around 100000 transcripts were processed, and alternative splicing occurs where appropriate. However, the myriad of different post-translational modifications substantially increases the size of proteome, which comprises over one million proteins; hence, the complexity of proteome increases relatively (Figure taken from (*ThermoFisher Scientific*)).

1.1 EXPERIMENTAL METHODS

There are several methodologies currently used to identify PTMs (see Table 1). The most conventional ones are *in vitro* PTM reaction assays using Western blot analysis, radioactive isotope-labeled substrates, and peptide and protein microarrays (Zhao and Jensen, 2009). However, they all contain bottlenecks, and are labor-intensive even though they are useful. In the case of identification of protein methylation and acetylation using radio-isotopes of carbon (C) and hydrogen (H), for instance, modified proteins are hardly efficiently detected as C and H are weak radio emitters. Western blot analysis is another approach where modification-specific antibodies are used to identify the presence of a protein in a sample. Although this method is quite efficient in detecting small amounts of proteins, especially immunogenic responses from infectious agents, it is not sufficient for complex samples, and it depends on the prior knowledge of the position and type of specific modifications, as well as the availability of antibodies (Chandramouli and Qian, 2009). Protein or peptide microarrays, on the other hand, is a high-throughput method where large number of proteins can be monitored in parallel. Kinase assays, for instance, are commonly used for peptide arrays to screen phosphorylation sites. However, the identified candidate proteins subsequently require validation (Zhao and Jensen, 2009).

Recent advancements in mass spectrometry-based (MS-based) proteomics over the last years have led to the identification of many PTMs in almost any kind (Figure 2). The approaches for protein characterization in MS-based proteomics can be classified into two: (i) top-down proteomics where intact proteins are directly analyzed to detect modifications, (ii) bottom-up proteomics, which includes



A	B
TVGK(ac)WR	AGK(ac)SVGTAK
EVGTK(ac)WK	EK(ac)CGTPK
GCGTYR	AVGTWK(ac)R
DSK(ac)GTWR	DEFYGR

Figure 2: Summary of mass spectrometry (MS)-based proteomics for the lysine acetylation. **A**. The representation for the sample preparation. Proteins are extracted from tissues, and digested into peptides. Eventually, many peptides are obtained, but a small amount of them is only acetylated as shown with yellow circles. The acetylated peptides are then enriched to reduce the complexity where acetyllysine antibodies selectively bind to acetylated peptides. Peptide fractionation approaches can also be applied to reduce the complexity. **B**. The representation for the MS analysis of the acetylated peptides. Nanoflow liquid chromatography (LC) is used to separate peptides from a reversed-phase column. After the electrospray ionization, the mass-to-charge ratio of peptide ions (MS spectra) is measured in the mass spectrometer. Finally, the existence and position of post-translational modifications, and the abundance of peptides are found by computationally processing MS spectra results (Figure taken from (Choudhary et al., 2014)).

Method	In vitro/in vivo	Advantages	Disadvantages
Radioactive isotope labeling	In vitro or in vivo	Reagents accessible	Inconvenience or hazard low sensitivity
Western blotting	In vitro or in vivo	Good affinity	Moderate sensitivity
Peptide/protein array	In vitro	Rapid, global scale	Possibly non-specific, low sensitivity, requires verification
MS–proteomics	In vitro	Specific, global scale	Need enrichment methods

Table 1: Experimental methodologies for identifying post-translational modifications (Table taken and modified from (Zhao and Jensen, 2009)).

the digestion of proteins into peptides (Choudhary et al., 2014). In bottom-up proteomics, after proteins are extracted from their cellular environments, they are digested to peptides by proteases - for instance the commonly used protease, trypsin (Olsen and Mann, 2013). The modified peptides are further extracted from the pool of all peptides by using liquid chromatography (LC) and ionized in the electrospray source. A particular position is assigned for the modification type of interest using the In High Performance Liquid Chromatography (HPLC) technique. Finally, eliminated peptides are subjected to mass spectrometer where mass spectra and fragmentation spectra are measured and recorded. Even though MS-based proteomics has overcome some drawbacks of conventional methods mentioned above, it still harbors limitations. For instance, overuse of trypsin for digesting peptides only enable the identification of trypsin-accessible peptides. Alternative proteases, on the other hand, lead to less specific and costly outcomes (Choudhary et al., 2014). The lower abundance of modified peptides also makes the identification from their fragmentation spectra difficult (Olsen and Mann, 2013).

1.2 ACETYLATION

Lysine acetylation is a reversible posttranslational modification (PTM), which involves the transfer of an acetyl group to the epsilon-amino group of a lysine residue of the substrate protein. This modification was previously known to target only histones, but more recently a broad spectrum of proteins was identified as acetylated and deacetylated by lysine acetyltranferases (KATs) and lysine deacetylases (KDACs), respectively, underscoring the important role played by lysine acetylation in diverse cellular processes including the regulation of subcellular localization, protein stability, enzymatic activity, nucleic acid binding, and protein-protein interactions. Studies of lysine acetylation mechanisms moved into the scientific limelight ever since their association with major diseases, such as cancer, was discovered.

Recent advancements in high-resolution mass spectrometry-based proteomics have led to identification of thousands of lysine acetylation sites (LASs) (Henriksen et al., 2012), rendering possible proteome-wide in silico analyses of their sequence context as well as theoretical predictions of LASs (Basu et al., 2009; Hou et al., 2014; Lu et al., 2011; Shao et al., 2012; Suo et al., 2012). Currently available data reveal significant diversity of amino acid sequences surrounding lysine acetylation sites, making it difficult to derive consensus acetylation motifs. This diversity might be due to the broad variety of KATs and KDACs encoded, for example, in the human and mouse genomes (22 KATs and 18 KDACs) as well as to non-enzymatic lysine acetylation (Choudhary et al., 2014). Most of the LASs known today have not yet been associated to their cognate KATs and KDACs due to the technical challenges in detecting KAT- and KDAC-specific acetylation sites by high-throughput in vitro acetylation assays. To close this gap, Li et al. made a commendable effort in manually assigning 384 known LASs to three selected KAT families (Li et al., 2012), which, however, is still a far cry from close to 5000 experimentally confirmed LASs known from literature as of 2012.

Beyond linear sequence motifs, it has been hypothesized that the local structural environments of lysines can influence their predisposition to be recognized by KATs. Indeed, Kim et al. (Kim et al., 2006) found that in mouse proteins, acetylated lysines prefer α-helical conformation, avoid disordered regions, and typically reside on protein surface. At the same time Okanishi et al. (Okanishi et al., 2013), while confirming the tendency of acetylated lysines to be exposed, did not find any relationship between acetylation propensity and local secondary structure in Thermus thermophilus. Both studies were performed on rather limited datasets of acetylation sites. Recent availability of much larger proteome-wide acetylation assays warrants a deeper look into the role of structure in shaping the substrate spectrum of KATs.

1.3 PHOSPHORYLATION

Protein phosphorylation is a reversible posttranslational modification (PTM) that represents the most common PTM type in eukaryotes, and plays a crucial role in many essential cellular processes, including cellular signaling, metabolism, differentiation, regulation of protein activity and subcellular localization (Roskoski, 2015). Protein phosphorylation and de-phosphorylation are controlled by more than 500 protein kinases and more than 100 phosphatases, respectively, which, in their turn, are regulated by phosphorylation, yielding a complex picture of interconnected signaling pathways. As many of these pathways are disease-related, understanding the mechanisms of phosphorylation has become a high priority for drug design.

Quantitative mass spectrometry-based phosphoproteomics have resulted in a massive amount of serine/threonine/tyrosine phosphorylation sites. However, methods to experimentally identify kinase substrates are still costly and laborious that a substantial amount of experimentally identified phosphorylation sites is still lack of experimentally annotated kinase family. PhosphoSitePlus includes 209000 phosphorylation sites where only 13751 of them (6.6%) were identified with corresponding kinases (Imamura et al., 2014). As a result, many studies made *in silico* attempts to derive consensus sequence motifs depending on kinase family (Chen et al., 2011; Damle and Mohanty, 2014; Gnad, Gunawardena, and Mann, 2011; Miller et al., 2008; Obenauer, Cantley, and Yaffe, 2003). Miller et al. introduced NetPhorest, which is an atlas of sequence motifs for phosphorylation sites targeted by 179 protein kinases and 104 phospho-binding domains. It further classifies the non-annotated experimentally identified phosphorylation sites to related kinases and phospho-binding domains. This atlas also contributes to understanding of different characteristics of phosphorylation signaling. Damle et al. built a network using experimentally identified kinase-substrate pairs and domains of phosphoproteins, revealing novel patterns for domain preferences of kinases. This network showed that many of the kinases phosphorylate only a few proteins domains, whereas only a small number of kinases phosphorylate a broad spectrum of protein domains. Although many of these studies emphasized on substrate-specificity across kinases where this substrate-specificity depends on sequence surroundings of phosphorylation sites, Chen et al. used motifs to derive more biological information, and proposed that phosphorylation distribution is dependent on cellular compartment type (Chen et al., 2014). Accordingly, cellular compartment-specific sequence motifs for phosphorylation were extracted, and experimentally identified phosphorylation sites were subsequently classified into corresponding cellular compartments.

Studies also showed that spatial amino acid content surrounding phosphorylation sites along with structural preferences play also an important role in kinase active site (Durek et al., 2009; Iakoucheva et al., 2004; Su and Lee, 2013; Tyanova et al., 2013). Durek et al. mapped experimentally identified phosphorylation sites onto three-dimensional structures, and categorized based on associated kinase. The spatial environments of phosphorylation sites were characterized in both global and kinase-specific manners, and further incorporated along with sequence information in prediction of phosphorylation sites. The preceding study by (Su and Lee, 2013) conducted a similar analysis as Durek et al., but with a more comprehensive dataset. Tyanova et al., on the other hand, introduced a different aspect for the analysis of structural properties of phosphorylation sites. Rather than the static way, the dynamic properties of phosphorylation sites with structural features were investigated at six time points of the cell division cycle. This study showed that phosphorylation sites take part in different functions depending on the need at different time scales, and the tendency of phosphorylation sites regulated at different time points of the cell division cycle is associated to structural environments of those phosphorylation sites.

1.4 COMPUTATIONAL METHODS

Protein phosphorylation, as we mentioned in Section 1.3, is a complex interconnected network carrying high importance since the abnormal regulation of phosphorylation is connected to evolution of diseases, such as cancer. The identified number of phosphorylation sites to date, however, could not show the same pace as its importance in cellular processes due to the fact that experimental methods, i.e. mass spectrometry (MS)-based phosphoproteomics, are expensive and labor-intensive. As a consequence, computational approaches have been substantially studied to elucidate more phosphorylation sites, contributing to understanding of the mechanism behind phosphorylation.

Based on the targeting area, available predictors can be classified into four: (i) Kinase-specific predictors, which are based on the idea that each kinase family targets different subset of substrates depending on the sequence amino acid content around phosphorylation sites (Blom et al., 2004; Fan et al., 2014; Gao and Xu, 2010; Li, Du, and Xu, 2010; Suo et al., 2014; Xue et al., 2010), or sequence content in combination with structural characteristics of phosphorylation sites (Blom, Gammeltoft, and Brunak, 1999; Durek et al., 2009; Hjerrild et al., 2004; Linding et al., 2008; Saunders et al., 2008; Su and Lee, 2013). These predictors take a protein sequence and the type of the kinase as inputs, and calculate the probability of each candidate site (serine/threonine/tyrosine residues) in the query protein phosphorylated by the given kinase. (ii) Organism-specific predictors, where not only human phosphorylation sites, but also phosphorylation sites in other species (Durek et al., 2010; Gao and Xu, 2010; Trost and Kusalik, 2013) have also been predicted. (iii) Subcellular-specific predictors, which utilize the information on localization of phosphorylation sites in subcellular compartments (Chen et al., 2014). (iv) Global predictors, which distinguish globally phosphorylated phosphorylation sites from non-phosphorylated counterparts. This kind of predictors calculates the probability of a candidate site in the query sequence to be phosphorylated by any existing kinase (Dou, Yao, and Zhang, 2014; Zhao et al., 2012) (see also reviews from (Trost and Kusalik, 2011; Xue et al., 2010)).

The above-mentioned studies have achieved great performance in phosphorylation site prediction, but they harbor some drawbacks. Namely, the regulation mechanism behind phosphorylation depending on different tissues has been shown, and the existence of tissue-specific kinases and phosphatases has been proposed in some parts of this thesis and previous studies. These findings decrease the robustness of models current predictors generate. On the other hand, most of the existing predictors only utilized the sequence features surrounding phosphorylation sites. However, it has been shown in this thesis and previous studies that phosphorylation sites also harbor structural characteristics (Durek et al., 2009; Su and Lee, 2013; Tyanova et al., 2013). The redundancy elimination has also been performed at very generous thresholds where predictors using sequence features would yield bias results. The prediction performance on phosphorylation prediction eventually remained not accurate and sufficient – high specificity, but low sensitivity.

1.5 THESIS MOTIVATION AND OUTLINE

The enzymes that catalyze PTM events have different expression levels in different tissues and cellular compartments. Comprehensive studies of protein glycosylation (Kaji et al., 2012), phosphorylation (Lundby et al., 2012b) and acetylation (Lundby et al., 2012a) revealed thousands of differentially modified sites, opening up the possibility that PTM sites may possess substantially different sequence and spatial properties across tissues, depending on which particular enzyme catalyzes a particular modification event. The existence of compartment-specific sequence signatures for phosphorylation (Chen et al., 2014; Wijk et al., 2014) and lysine acetylation (Choudhary et al., 2009; Kim et al., 2006; Lundby et al., 2012a; Shao et al., 2012) has already been firmly established, whereas their tissue-specific preferences still remain unexplored. This thesis focuses on tissue-specific sequence and structural preferences of acetylation and phosphorylation sites in Chapter 2 and Chapter 3, respectively. In Chapter 4, we present the first tissue-specific phosphorylation site prediction approach, TSPhosPred (Tissue-Specific Phosphorylation Prediction), which aims to address the drawbacks of current phosphorylation site predictors. We believe that this thesis will enlarge the horizon of phosphorylation and acetylation, and contributes to understanding of the complex evolution of post-translational modifications.

Part II

TISSUE-SPECIFIC SEQUENCE AND STRUCTURAL ENVIRONMENTS OF LYSINE ACETYLATION SITES

2 TISSUE-SPECIFIC SEQUENCE AND STRUCTURAL ENVIRONMENTS OF LYSINE ACETYLATION SITES

Lysine acetylation is a reversible post–translational modification that regulates a broad spectrum of biological activities across various cellular compartments, cell types, tissues, and disease states. While compartment–specific trends in lysine acetylation have recently been investigated, its tissue-specific preferences remain unexplored. Here we present the first comprehensive tissue-based approach analyzing the sequence and structural features of lysine acetylation sites (LASs) based on the recent experimental data of (Lundby et al., 2012a). We assessed the extent of evolutionary conservation of LASs and its dependence on functional and structural properties of proteins by comparing rat, mouse, and *C.elegans* acetylomes. We further investigated tissue-specific functional roles and domain preferences of acetylated proteins.

2.1 MATERIALS AND METHODS

2.1.1 *Data collection and preprocessing*

The dataset used in our analysis contains 15474 lysine acetylation sites (LASs) in 4541 proteins identified by high-resolution tandem mass spectrometry in 16 rat tissues: brain, heart, muscle, lung, kidney, liver, stomach, pancreas, spleen, thymus, intestine, skin, testis, testis fat, perirenal fat, and brown fat (Lundby et al., 2012a). For each lysine-acetylated peptide in each tissue we obtained information about the UniProt (Consortium, 2014) IDs of the best-matching proteins (one or more), the sequence position of the acetylated site, and the intensity values (summed up extracted ion current of all isotopic clusters associated with the peptide in the corresponding tissue).

In order to find the best-matching UniProt ID for each acetylated peptide we applied the following procedure: (i) All fragments were excluded from consideration. (ii) If there was only one UniProt ID associated with an acetylated peptide, and its sequence position and the sequence of the corresponding full-length protein in the UniProt database were known, then we directly used that protein. (iii) Otherwise, we aligned all pairs of proteins and then chose the pair having the maximum sequence identity out of all pairs sharing at least 90% sequence identity. The idea behind this approach is to find those UniProt proteins corresponding to the given peptide that show at least some consistency in terms of their overall primary structure. If no pair of proteins associated with the given peptide showed more than 90% sequence identity, this peptide was excluded from consideration. (iv) Finally, out of two aligned best–matching proteins we retained the longer one. We obtained 10626 acetylation sites on 3541

Datasets	Description	Number
Initial dataset	LASs	10626
	Proteins	3541
Structure-based dataset	LASs	2566
	Proteins	856
LAS1D (non-redundant sequence-based)	LASs (positive set)	9868
	Non-LASs (negative set)	94362
LAS3D (non-redundant structure-based)	LASs (positive set)	2218
	Non-LASs (negative set)	8777

Table 2: Data summary of lysine acetylation sites.

proteins, each of them having only one best-matching UniProt ID. The decrease in the number of acetylation sites is due to not satisfying the above criteria, not finding the sequence of the corresponding full-length protein in the UniProt database, or not finding a lysine residue in the specified sequence position of the finally obtained protein.

2.1.2 *Sequence (1D) environments of acetylated and reference (non–acetylated) lysine residues*

The positive dataset of tissue-specific LASs consisted of all lysine acetylated sites displaying non-zero intensity values in the corresponding tissue. The negative (reference or non-LASs) set was generated by extracting all lysine residues not annotated as acetylated by Lundby et al. (Lundby et al., 2012a) and relating them to those tissues in which the protein harboring the reference site also has at least one experimentally observed LAS. Then, we generated 21-mer sequences (from position -10 to position +10) surrounding each site in both positive and negative datasets and performed homology reduction on these 21-mers using CD-HIT (Li and Godzik, 2006) at the 90% identity threshold. Note that some of acetylation and reference sites occur in more than one tissue. The resulting dataset, which we call LAS1D, is composed of non-redundant 21-mer sequences corresponding to 9868 LASs and 94362 non-LASs (Table 2). The distribution of LASs and non-LASs in different tissues is given in Figure 3.

We used the Two Sample Logo method (Vacic, Iakoucheva, and Radivojac, 2006) for differential analysis of 21-mer occurrence in different tissues, using the corresponding LASs and non-LASs as positive and negative sample inputs, respectively. For example, LASs observed in brain were compared to non-LASs in brain. Amino acids were colored using the WebLogo defaults, and t-test with a cut-off p-value of 0.05 was used to select significantly enriched residues. The Motif-X online tool (Chou and Schwartz, 2011) was used to extract motifs from the 21-mer sequences of LASs, using LAS and non-LAS as the foreground and background datasets, respectively.

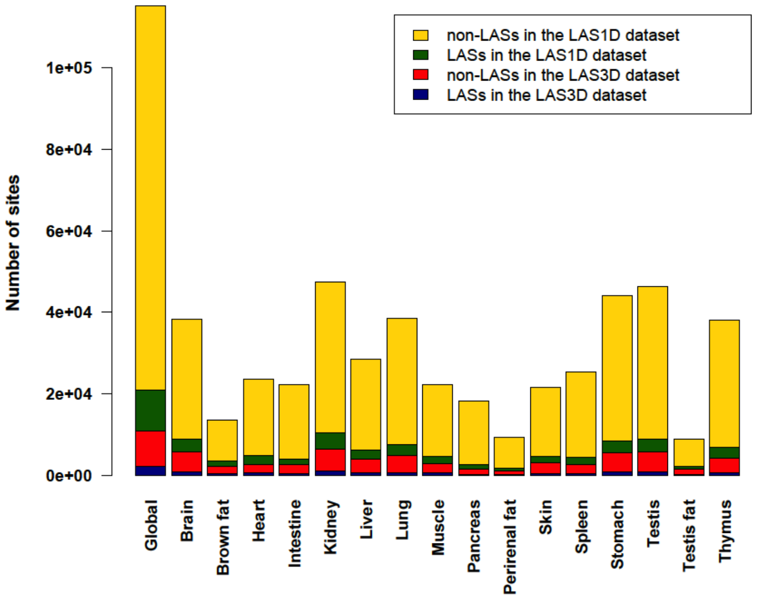


Figure 3: Number of LASs and non-LASs from the LAS1D and LAS3D datasets in different tissues.

2.1.3 *Lysine acetylation sites with known 3D structure*

In order to analyze the properties of spatial (3D) environments of LASs we collected a dataset of proteins with known atomic structure containing lysine residues annotated as acetylated by Lundby et al. (Lundby et al., 2012a). Using the amino acid sequences of acetylated proteins as queries we extracted the total of 1689 related 3D structures from the Protein Data Bank (Berman et al., 2000) based on BLAST-P (Camacho et al., 2009) hits with E-value <0.001 and sequence identity >90%. We did not require the alignments to be global and to cover the total length of the compared proteins as this would lead to a dramatic reduction of our structural dataset. Instead, we selected the alignments that cover the ±50 residue environment of the acetylation sites with higher than 80% identity with the candidate structure. Using this procedure we obtained 2566 acetylation sites in 856 protein structures after excluding low–resolution structures ($\geqslant 3$Å).

The structure-based positive and negative LASs datasets were generated as described above for sequence-based data. Homology reduction was again performed on 21-mer sequences surrounding LASs and non-LASs at the 90% identity threshold. The resulting dataset, which we call LAS3D, contains 2218 LASs and 8777 non-LASs in proteins with known structures (see Table 2 and Figure 3).

2.1.4 *Statistics*

Statistical analyses were performed using the R environment (Team, 2009) and custom Java programs. We used the non-parametric two-sample Kolmogorov-Smirnov test and the Fisher test to assess the significance of the differences between numerical and categorical datasets, respectively. Relative frequency of a certain property (e.g. conservation) of LASs and their sequence neighborhoods observed in a given tissue was compared to that of non-LASs and their sequence neighborhoods observed in the same tissue. We used the non-parametric Kruskal-Wallis test to perform multiple comparisons between expression profiles of KAT paralogs across tissues.

2.1.5 *Three-dimensional (3D) environments of acetylated and reference (non-acetylated) lysine residues*

Spatial amino acid environments of LASs in the LAS3D dataset were determined by calculating the occurrence of 20 different amino acid types within the radial distances of 2 to 12 Å from the acetylated lysine residue in accordance with the previous studies analyzing the spatial environment of phosphorylation sites (Durek et al., 2009; Su and Lee, 2013). Distances between amino acid residues were defined based on the minimal distance between any pair of atoms belonging to these residues. In order to isolate the influence of spatial structure from 1D sequence motifs we also defined *pure* 3D amino acid environments of LASs by excluding from consideration those amino acids already present in the sequence vicinity of LASs, as defined in the previous section. In both cases the Fisher exact test was employed to assess the significance of the differences between LASs and non-LASs in each tissue, and these differences were efficiently visualized using our in-house software tool. For each radial distance ranging from 2 to 12 Å (in increments of 1Å) and for each amino acid type we calculated (i) the significance (p-value) of the amino acid at that position using Fisher exact test, and (ii) the odds ratio of the amino acid at that position by dividing the normalized occurrence of acetylated amino acids to that of non-acetylated amino acids.

2.1.6 *Conservation analysis of lysine acetylation sites*

Using an approach similar to the one given in (Weinert et al., 2011) we extracted Caenorhabditis elegans orthologs of acetylated proteins contained in the LAS1D dataset from the InParanoid database (Ostlund et al., 2010) and compared the evolutionary conservation of LASs and non-LASs based on Needleman-Wunsch alignments (Needleman and Wunsch, 1970) between acetylated protein sequences and their C. elegans counterparts. We assessed the conservation by comparing the frequency of conserved LASs among all LASs to the frequency of conserved non-LASs among all non-LASs. Note that for this analysis we used all C. elegans orthologs of mouse and rat acetylated proteins, including those that are not acetylated. Statistical significance of the differences between the conservation of LASs and non-LASs was calculated using the Fisher exact test.

2.1.7 *Structural features of lysine acetylation sites*

The surface accessibility of LASs and non-LASs in the LAS3D dataset along with their sequence surroundings was calculated using NACCESS (Hubbard and Thornton, 1993). We used the absolute (rather than relative) accessibility scores of amino acid side chains produced by NACCESS that are larger than zero. The rationale for this choice is that in contrast to LASs, non-LASs often reside in the core of the protein and considering such buried non-LASs could lead to biased results.

We used DisEMBL (Linding et al., 2003) to predict disordered/unstructured regions within protein sequences. A LAS/non-LAS in the LAS1D dataset was considered to reside in a disordered region if it was predicted by DisEMBL to be located in a region associated with either loops/coils, or hot loops, or missing coordinates. Secondary structure assignments were obtained from the DSSP database (Joosten et al., 2011).

2.1.8 *Analysis of structural folds and functional domains*

We investigated structural folds of lysine acetylated proteins in each tissue in the LAS3D dataset according to the *class* and *protein domain* levels of the SCOP database (Murzin et al., 1995) hierarchy. At the *protein domain* level false discovery rate control was performed for multiple hypothesis correction in each tissue, all p-values were adjusted, and the significance threshold after the correction $p < 0.05$ was used. At the structural class level, the significance threshold $p < 0.01$ was used.

2.1.9 *KEGG pathway analysis*

We identified enriched pathways across tissues in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa et al., 2006) using the best-matching UniProt identifiers of each LAS and non-LAS in the LAS3D dataset (see above). False discovery rate control was performed for multiple hypothesis correction in each tissue, all p-values were adjusted, and the significance threshold after the correction $p < 0.01$ was used.

2.1.10 *Abundance of KAT paralogs*

For each experimentally identified human KAT (Li et al., 2012) we found the mouse ortholog as well as its paralogs using the KEGG database. Protein expression levels of paralogs across tissues were obtained from the PaxDb database (Wang et al., 2012).

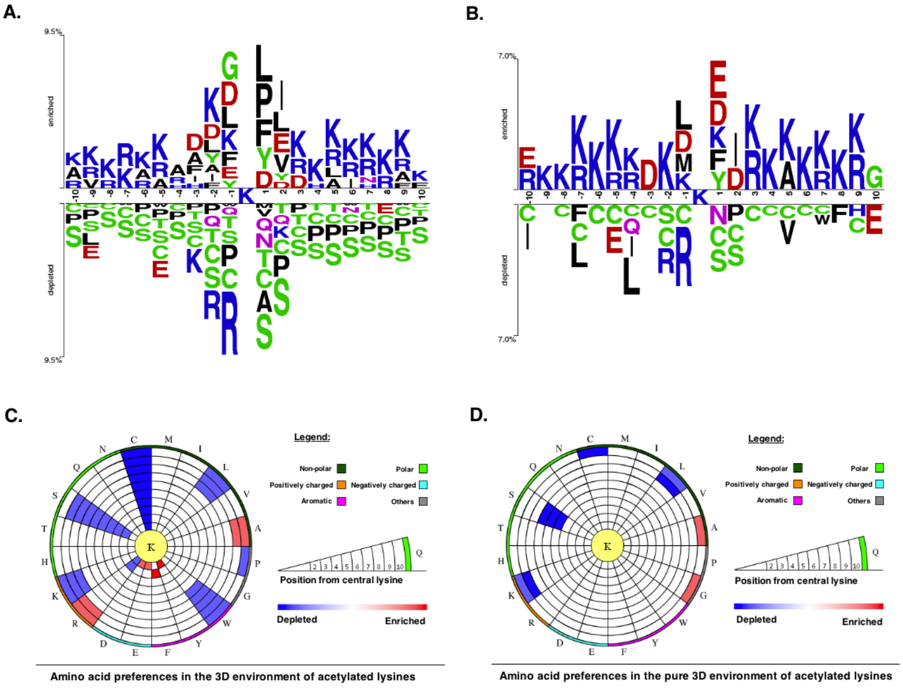


Figure 4: The comparison of 1D and 3D environments of global lysine acetylation sites. (A) Two sample logo analysis of global LASs in the LAS1D dataset. (B) Two sample logo analysis of global LASs in the LAS3D dataset. (C) 3D environments of LASs in the LAS3D dataset. (D) *Pure* 3D environments of LASs in the LAS3D dataset.

2.2 results and discussion

2.2.1 *Global and tissue-specific sequence motifs of lysine acetylation sites*

It has been previously reported that LASs have compartment-specific sequence motifs (Choudhary et al., 2009; Kim et al., 2006; Lundby et al., 2012a; Shao et al., 2012). Here we first investigate both global and tissue-specific acetylation trends at the sequence level based on the LAS1D dataset. Across all tissues, amino acids with bulky side chains are enriched at positions from -3 to +2 with respect to the acetylated lysine (Figure 4A), as reported before (Choudhary et al., 2009; Hou et al., 2014; Lundby et al., 2012a; Suo et al., 2012, 2013). In accordance with the previous studies (Maksimoska et al., 2014) we also find that glycine is strongly enriched at position -1, while non-polar and hydrophobic isoleucine (I), valine (V) and leucine (L) residues are preferred at positions +1 and +2. Negatively charged residues frequently occur at positions from -3 to +3, while the wider sequence context of LASs exhibits a strong preference to positively charged residues.

LASs in individual tissues generally follow the global trends discussed above, but some tissue-specific trends are also clearly observed. For instance, brain and muscle (Figure 5A and Figure 5B) harbor LASs having asparagine (N) residue at the position +3 whereas pan-

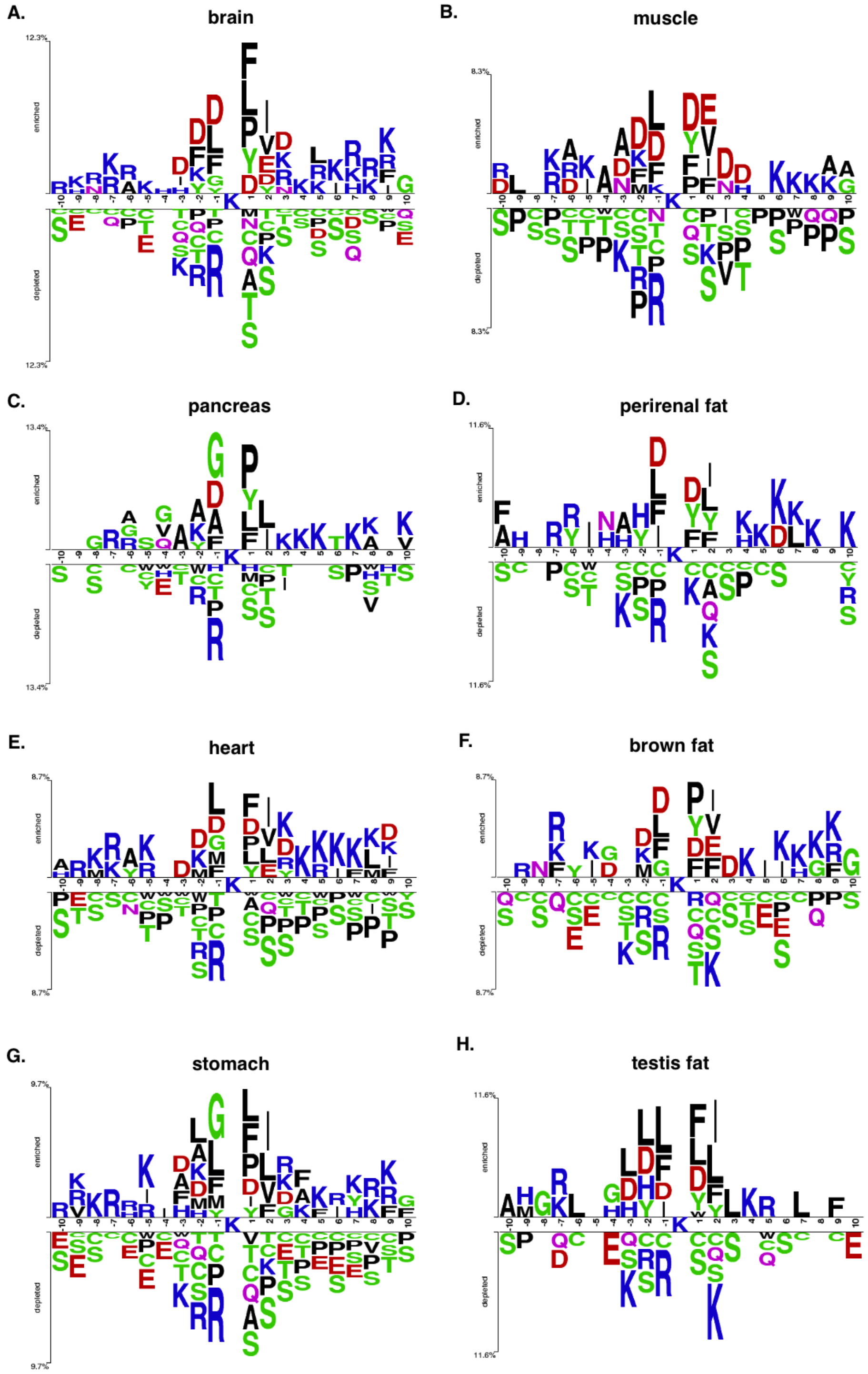


Figure 5: Two sample logo analysis of LASs from the LAS1D dataset in brain (A), muscle (B), pancreas (C), perirenal fat (D), heart (E), brown fat (F), stomach (G) and testis fat (H). See data on other tissues given in Figure 28 - Figure 44.

creas and perirenal fat (Figure 5C and Figure 5D) include LASs having glutamine (Q) and asparagine residues, respectively, highly enriched at position -4. LASs in heart and stomach (Figure 5E and Figure 5G) show a strong preference for methionine (M) residue at both positions -1 and -2, while in brown fat and muscle (Figure 5F and Figure 5B) LASs need methionine only at position -2. Histidine (H) residues are strongly preferred by LASs in testis fat at positions -2, -3 and -4 (Figure 5H). In the downstream region of the acetylated lysines negatively charged residues occur less frequently in thymus, pancreas, perirenal fat and spleen (See Figure 28 – Figure 44 for more detailed graphs for these and other tissues).

Besides the two-sample logo analysis, we also used the Motif-X software to delineate tissue-specific sequence motifs in the LAS1D dataset that significantly deviate from the global sequence pattern (Table 12). For instance, the motifs I-AcK and I-X-X-AcK are only associated with brain-specific LASs. Similarly, the motif E-X-AcK-Y is not observed in any tissues except for intestine. We therefore conclude that tissue-specific sequence motifs are not just random subsets of global motifs, but rather reflect the required environment for acetylation in each tissue.

Acetylation is regulated both enzymatically, by lysine acetyltranferases (KATs), lysine deacetylases (KDACs) and bromo-domain- containing acetyllysine binders, and non-enzymatically (Choudhary et al., 2014). While the existence of compartment-specific KATs is still being debated (Lundby et al., 2012a; Sadoul et al., 2011), our findings may imply the existence of tissue-specific KATs and KDACs. We were not able to detect significant differences in the abundance of paralogs of experimentally identified KATs across different tissues, which implies that tissue-specific motifs are not a result of tissue-specific KAT expression. On the other hand, previous studies have proposed that even though KATs might share a conserved substrate-binding site, different non-catalytic subunits of KATs may cause the diversity in substrate sequences (Berndsen et al., 2008; Clements et al., 2003; Poux and Marmorstein, 2003). Thus, the tissue-specific substrate sequences reported here may be suggestive of the existence of tissue-specific non-catalytic subunits of KATs. Moreover, the fact that chaperones associated with KATs influence their substrate specificity (Berndsen et al., 2008; Fillingham et al., 2008; Recht et al., 2006) raises the opportunity that this influence may be exercised in a tissue dependent manner. On the other hand, we speculate that non-enzymatic acetylation might also vary from tissue to tissue depending on the concentration of metabolites (acetyl-CoA, acetyl-phosphate etc.) and the pH level (Choudhary et al., 2014), leading to diverse substrate sequences. Such diversity could conceivably be caused by the following reasons: (i) absence of the recognition site by KATs, resulting in random sequences being favored for deprotonation of amino groups by acetyl-CoA, (ii) the requirement for specific lysine environment for CoA donation of acetyl-CoA, (iii) regulation of enzymatic lysine acetylation by non-enzymatic acetylation (crosstalk), such that if non-enzymatic acetylation does not occur due to the low abundance of a metabolite, the regulated acetylation of another lysine would be obstructed and the substrate sequence of the corresponding KAT would be underrepresented in that tissue.

2.2.2 *Global and tissue-specific sequence motifs of lysine acetylation sites in proteins with known 3D structure*

We conducted a separate analysis of lysine acetylation sequence motifs in the LAS3D dataset, which only contains proteins with known 3D structure. Since this dataset is obviously depleted in disordered regions and hence has a somewhat different amino acid composition, the corresponding lysine acetylation motifs exhibit somewhat different residue preferences compared with the full LAS1D dataset. The enrichment of the disorder promoting glycine residue, for instance, is not observed at position -1 of the global LAS signature (Figure 4B). On the other hand, in some tissues, including brain, kidney and testis the sequence neighborhoods of LASs are enriched in positively and negatively charged residues, while in pancreas LASs require arginine only at the amino acid position -7 (Figure 45). In pancreas, LASs in the LAS3D dataset exhibit a strong preference for negatively charged aspartic acid (D) residue at the amino acid position +3 and negatively charged glutamic acid (E) at the amino acid position -2, which is in strong contract to the tendencies found for pancreas based on the entire LAS1D dataset (see above). Brain is characterized by the frequent occurrence of negatively charged residues between amino acid positions -2 to +3. LASs in intestine are special in that they are associated with enriched glycine (G) and methionine (M) residues at positions -1 and -8, respectively, whereas none of the LASs observed in other tissues have such preferences. Interestingly, as opposed to global LAS signatures, LASs in stomach and testis fat do not harbor any negatively charged residues, while LASs in kidney and thymus show a strong preference for polar asparagine (N) residue at amino acid position -3.

2.2.3 *Spatial environments of lysine acetylation sites*

In the spatial surroundings of LASs across all tissues there is a strong depletion of cysteine (C), which is also avoided in the sequence motifs discussed above. In the close proximity of global acetylation sites (around 2-3 Å away), strong enrichment of hydrophobic, aromatic, low flexibility and order-promoting tyrosine (Y) and phenylalanine (F) residues is observed (Figure 4C). Another prominent trend is strong enrichment of positively charged, surface exposed, highly flexible and disorder promoting arginine residue at larger distances (10 to 11 Å). Interestingly, enrichment of positively charged residues in the 3D environment of global LASs is not as strong as in the 1D sequence neighborhood.

Tissue-specific spatial environment analysis reveals some additional statistical trends. LASs in brain and stomach (Figure 6A and Figure 6B) have a strong preference for glutamic acid and methionine residues, respectively, whereas perirenal fat and spleen (Figure 6C and Figure 6D) harbor LASs whose spatial environment is enriched in histidine (H). Thus, patterns of amino acid usage around LASs are generally consistent between 1D and 3D environments, although some of the tendencies found at the primary structure level, such as the enrichment of negatively charged residues in brown fat (Figure 6E) and the enrichment of glycine and tyrosine residues in intestine (Figure 6F), are not observed in the 3D environments (see Figure 46 for more detailed graphs for all tissues).

Amino acid residues participating in local sequence motifs around LASs are also situated in their spatial vicinity. In order to disentangle the effects of sequence neighbors from those of spatial neighbors we conducted a separate analysis of *pure* structural environments (see Section 2.1). While only a weak enrichment of alanine (A) and glycine residues is observed in global *pure* 3D environment trends (Figure 4D), tissue-specific preferences are more strongly pronounced. One of the striking examples is the strong enrichment of tyrosines at a distance between 4 Å and 7 Å in spleen (Figure 6D). Similarly, LASs residing in brown fat have a preference for hydrophobic alanine and valine (V) residues in their *pure* 3D environments, which is not observed when sequence context is also considered (Figure 6E). Spleen harbors LASs enriched in tyrosine and isoleucine (I) residues in their *pure* 3D environment, whereas tyrosine residues in the 1D environment, and both tyrosine and isoleucine residues in the 3D environment are not enriched (Figure 6D). We thus find that, beyond sequence motifs, there are statistically significant patterns of residue preferences in the spatial environment of LASs, which implies that the substrate-specificity of KATs and KDACs may be characterized by both of sequence and spatial environments of LASs (see Figure 47 for more detailed graphs for all tissues).

2.2.4 *Evolutionary conservation of lysine acetylation sites*

Previous studies, in which the conservation of lysine acetylation in *Drosophila melanogaster* or in humans was compared to that in nematodes and zebrafish (Weinert et al., 2011), indicate that LASs are significantly more conserved than non-LASs. With an approach similar to the one given in (Weinert et al., 2011), we created sequence alignments between rat and mouse proteins from the LAS1D dataset and their *C. elegans* orthologs obtained from the InParanoid database (Ostlund et al., 2010). We then compared the conservation frequency of global rat and mouse LASs and non-LASs to the conservation frequency of these sites in the *C. elegans* counterparts. As expected, we find that acetylated lysine residues (conservation frequency of 45.7%) are more conserved than non-acetylated lysine residues (conservation frequency of 41.3%) (p-value < 6.17 x 10-7), although at the tissue level this trend was only significant in brain (48.12% and 40% of LASs and non-LASs with p-value < 0.0028), lung (49.03% and 44.54% of LASs and non-LASs with p-value < 0.0061) and thymus (50.87%

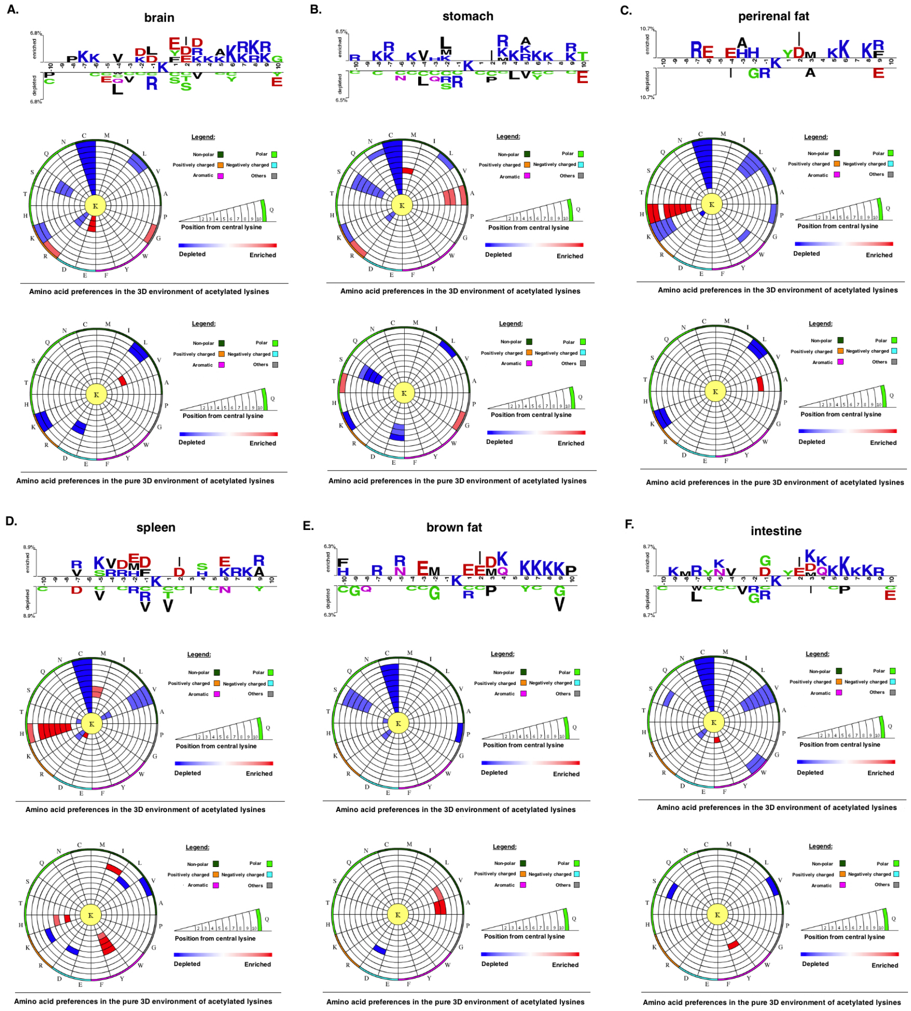


Figure 6: Sequence (1D) and structural (3D and *pure* 3D) environments of LASs from the LAS3D dataset represented by two sample logos and circular plots, respectively. Data for four tissues is shown: brain (A), stomach (B), perirenal fat (C), spleen (D), brown fat (E) and intestine (F) (see Figure 45, Figure 46 and Figure 47 for all tissues). The circular amino acid propensity plots were produced by our in-house software tool. Color code for amino acid residue enrichment in the circular plots: (i) red – enriched, if p-value < 0.05 and odds ratio > 1, (ii) blue – depleted, if p-value <0.05 and odds ratio <1, and (iii) white –neither enriched nor depleted, if p-value $\geqslant 0.05$.

and 45.31% of LASs and non-LASs with p-value < 0.0011). We speculate that the weaker evolutionary conservation of acetylated positions at the tissue level is due to the tissue-specific variation of protein abundance levels, as was already shown for phosphoproteins (Levy, Michnick, and Landry, 2012). We subsequently compared the conservation of LASs and non-LASs separately in irregular/regular regions, ordered/disordered regions, and functional/unknown proteins, motivated by the previous reports that the conservation of phosphorylation sites strongly depends on these key factors (Levy, Michnick, and Landry, 2012). We find that according to the KEGG database 59.85% of all conserved global LASs reside in proteins with known functions, whereas for conserved global non-LASs this percentage is much lower, at 50.16% (p < 9.35 x 10-14). This observation is also significant in all tissues except liver and pancreas (Table 14). On the other hand, both global LASs and non-LASs are more frequently and equally conserved in disordered and regular regions (57.19% and 58.37%, respectively for disordered regions; 60.43% and 60.95%, respectively, for regular regions). We therefore conclude that acetylation sites follow the evolutionary trends similar to phosphorylation processes in that they are more conserved in proteins with known function and in structurally regular regions.

2.2.5 *Tissue-specific structural properties of lysine acetylation sites*

In line with the previous reports (Kim et al., 2006; Rojas et al., 1999; Suo et al., 2012), both global LASs and the residues surrounding them tend to be consistently more solvent exposed than non-acetylated lysines and their 1D environment by about 10% on average (see Table 15). No tissue-specific solvent exposure preference was observed. Structural preferences of lysine acetylation sites display tissue-specific character. For instance, LASs and the residues surrounding them in all tissues except for thymus, spleen and pancreas reside significantly more often in ordered regions than in disordered regions, as global LASs also do. On the other hand, no enrichment of LASs for disordered regions in any tissue was observed.

In all tissues except pancreas, testis fat, muscle and perirenal fat both LASs and the residues in their close proximity tend to reside in α-helices more often than in other types of secondary structure (Figure 48). Beyond the enrichment in α-helices, we also find that LASs in stomach tend to avoid loops whereas LASs in testis are depleted in β-sheets. In addition, in many tissues residues adjacent to LASs on the C-terminal side are depleted in β-sheets. No structural preferences of LASs could be observed in testis fat.

In terms of SCOP fold preferences global LASs and non-LASs show the same behavior in that they are mainly found in all-α proteins (Figure 49). The same trend exists in many individual tissues except for muscle and brown fat where acetylation sites are significantly enriched in (α+β) proteins. On the other hand, LASs in heart, skin and testis are significantly depleted in (α/β) proteins. Based on the analysis of B-factors we also find that LASs preferentially occur in more rigid regions of protein structures. LASs and the residues surrounding them have smaller B-factors than non-LASs do (Table 15).

2.2.6 *Proteins containing acetylated lysines are involved in tissue-specific biological pathways*

It has been firmly established that the structural environment of phosphorylated residues is a key factor determining kinase specificity and ultimately the functional role of phosphorylation processes (Su and Lee, 2013; Tyanova et al., 2013). We therefore investigated whether lysine acetylated proteins are specialized for various functions in different tissues. It is worth reminding that in all cases we compare the enrichment or depletion of acetylated proteins in specific pathways relative to non-acetylated proteins *in each individual tissue* (see Section 2.1), so the trends presented below cannot be the consequence of mere presence or absence of certain functions in those tissues. In accordance with previous studies (Henriksen et al., 2012; Kim et al., 2006; Lu et al., 2011; Patel, Pathak, and Mujtaba, 2011), we find that global lysine acetylated proteins are involved in energy generation processes including the TCA cycle, fatty acid metabolism, and glycine/serine/threonine metabolism. However, there are also significant differences between global and tissue-specific lysine acetylated proteins in terms of their biological pathway preferences (Figure 7). Similar to the trends in globally acetylated proteins, lysine acetylated proteins in all tissues except for spleen, pancreas, testis fat and liver take part in the citrate cycle (TCA cycle) pathway. On the other hand, proteins involved in RNA transport show significant enrichment in acetylation sites only in brain. Acetylation appears to play a role in the terpenoid backbone biosynthesis pathway only in liver. We also identified tissue-specific acetylation patterns in several disease pathways. It appears that only in heart acetylated proteins are involved in the regulation of Type II diabetes mellitus, whereas Epstein-Barr virus infection is only regulated by testis-specific acetylated proteins.

It has been shown in several studies that acetylated proteins not only take part in nuclear processes, such as regulation of gene transcription, but are also involved in signaling pathways. On the other hand, it has also been previously proposed that differential lysine acetylation between human liver and leukemia cells might be cell type-dependent (Choudhary et al., 2009; Patel, Pathak, and Mujtaba, 2011). Our findings indicate the existence of biological processes that are affected by acetylation in a tissue-specific manner. Compounds modulating KAT-, KDAC- and bromo domain-containing proteins show promise as potential drugs against cancer, cardiac illness, diabetes, and neurodegenerative disorders (Choudhary et al., 2014; Patel, Pathak, and Mujtaba, 2011). Our observation that acetylated proteins are involved in a number of tissue-specific disease pathways, combined with the speculation about the existence of tissue-specific KATs, KDACs and bromo domain-containing proteins discussed above, may serve as an indication that small molecules and drugs can be designed to target tissue-specific disease pathways.

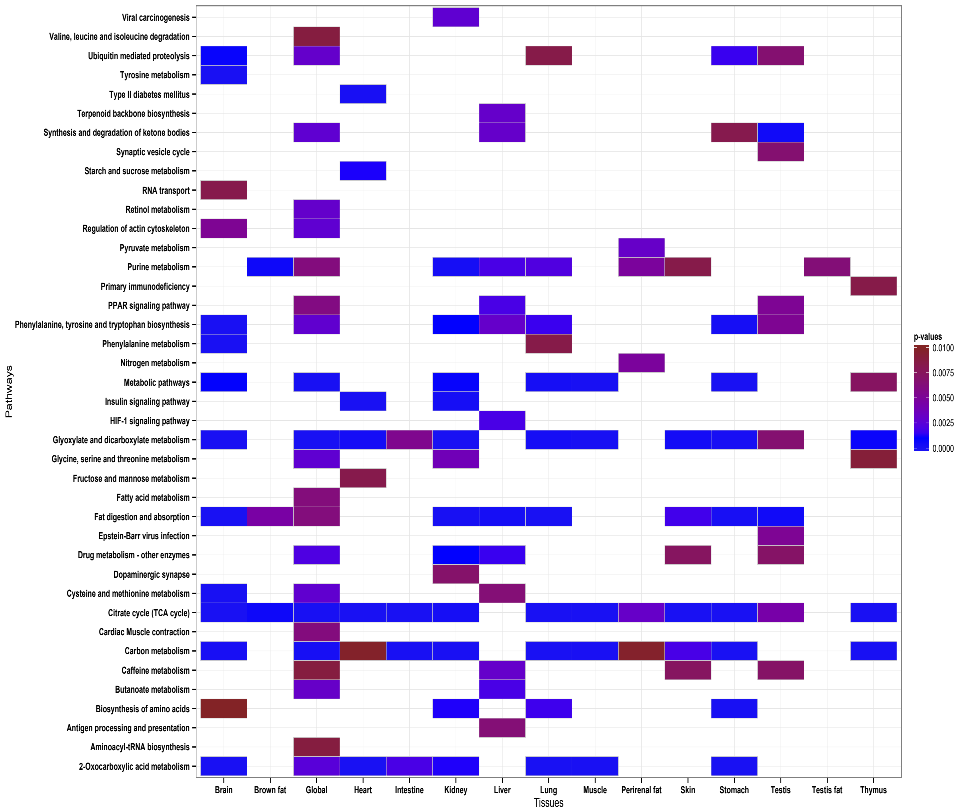


Figure 7: KEGG pathway analysis of the acetylated proteins from the LAS3D dataset. Pathways with a corrected p-value < 0.01 in each tissue are considered significant. Spleen and pancreas are not shown in the figure since acetylated proteins in spleen and pancreas are not enriched in any pathways.

Unsurprisingly, tissue-specific preferences regarding biological pathways are accompanied by the variation in occurrence of acetylation sites within key protein domains such as aldehyde reductase (dehydrogenase), creatine kinase C-terminal domain, mitochondrial class kappa glutathione S-transferase, and phosphoglucose isomerase (Figure 50). Proteins containing the zeta isoform domain are exclusively acetylated in brain. The zeta type protein kinase C (PKC) is expressed in brain where it is involved in mitogenic signaling, cell proliferation, cell polarity, inflammatory response, and maintenance of long-term potentiation and memories. While it has been reported that two specific sites of this kinase need to be phosphorylated for its full activation (Consortium, 2014), our findings suggest that some of its sites may also need to be acetylated. Class mu GST domain-containing acetylated proteins are enriched testis. Mu belongs to the cytosolic superfamily of Glutathione S-transferases (GSTs), which catalyze the joining of glutathione (GSH) to xenobiotic substrates for detoxification. GSTs also play role in the cell signaling processes. Overexpression of GSTP1-1 – an isozyme of the mammalian GST family - has been associated to cancer. GSTP1-1 is a very important drug target and its acetylation in testis warrants further investigation. Proteins containing malate dehydrogenase (a component of the TCA cycle, see above) and annexin v domains (whose function is still unknown) are acetylated in many tissues.

2.3 CONCLUSION

In this work we present evidence that lysine acetylation sites display tissue-specific preferences for certain residues both in their linear amino acid sequence and in spatial environments. We further demonstrate that LASs are generally more evolutionarily conserved than non-LASs, the trend that is especially pronounced in proteins with known function and in structurally regular regions. The occurrence of LASs and the residues surrounding them in disordered regions and regular secondary structures also displays a tissue-specific character. These findings imply the existence of tissue-specific KATs and KDACs able to differentiate between various types of local structural environments beyond mere amino acid content in sequence and in spatial environments. Lysine acetylated proteins are specialized for various functions in different tissues, and this specialization is supported by tissue-specific key domain preferences. Since compounds modulating KAT-, KDAC- and bromo domain-containing proteins are potential drugs against many diseases, including cancer, the speculation about the existence of tissue-specific KATs, KDACs and bromo domain-containing proteins indicates the need for tissue-specific drug target designs.

Part III

SEQUENCE- AND STRUCTURE-BASED ANALYSIS OF TISSUE-SPECIFIC PHOSPHORYLATION SITES

3 SEQUENCE- AND STRUCTURE-BASED ANALYSIS OF TISSUE-SPECIFIC PHOSPHORYLATION SITES

Phosphorylation is the most widespread and studied reversible post-translational modification, and play a role in the regulation of almost every cellular activity. Sequence and structural properties of global phosphorylation sites, as well as of those specific for individual cellular compartments, have been previously investigated. By contrast, tissue-specific preferences of phosphorylation sites at the sequence and structure level remain largely unexplored. In this study we performed a comprehensive tissue-specific analysis of the sequence and structural environments of phosphorylation sites as well as globally phosphorylated sites, employing a recent experimental data by Lundby *et al.* (Lundby et al., 2012b) that provides new insights into the underlying mechanism of phosphorylation. We demonstrate that substrate recognition by kinases is guided by both sequence and structural features of phosphorylation sites in a tissue specific manner. Based on the known kinase-substrate associations we demonstrate that many kinases are active in a tissue-specific manner, an effect which is apparently not caused by tissue-specific expression of kinase genes. We also examined differential functional roles and domain preferences of phosphorylation sites across tissues (see Figure 8 for summary).

3.1 MATERIALS AND METHODS

3.1.1 *Datasets of phosphorylated and reference (non-phosphorylated) sites*

We used the dataset from of 31480 phosphorylation sites in 7280 proteins identified by high-resolution tandem mass spectrometry in 14 rat tissues: brain (dissected into cerebellum, cortex and brainstem), heart, muscle, lung, kidney, liver, stomach, pancreas, spleen, thymus, intestine, testis, perirenal fat, and blood (Lundby et al., 2012b). The UniProt (Consortium, 2014) IDs of the best-matching proteins (one or more), the sequence position of the phosphorylation site, and the intensity values (summed up extracted ion current of all isotopic clusters associated with the peptide in the corresponding tissue) were gathered for each phosphorylation site in each tissue.

The best-matching UniProt ID for each phosphorylated peptide was identified as described previously in Chapter 2. Briefly, we aligned all pairs of proteins associated with a given peptide and chose the longer protein of the pair having the maximum sequence identity out of all pairs. We obtained 17917 phosphorylation sites on 5443 proteins, each of them having only one best-matching UniProt ID. This dataset contains 14661 phosphorylated serine sites (PSSs), 2832 phosphorylated threonine sites (PTSs), and 424 phosphorylated tyrosine sites (PYSs) (Table 3). The decrease in the number of phosphorylation

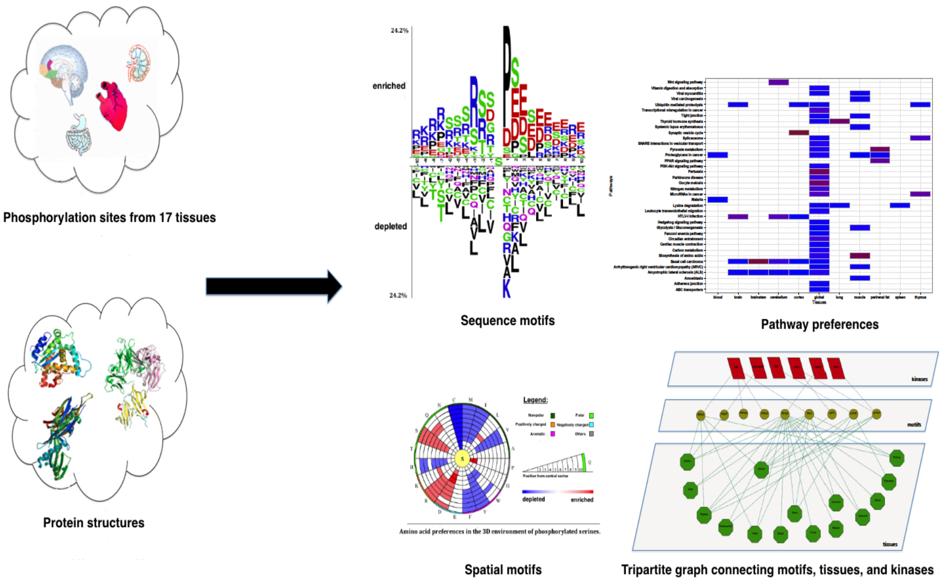

Phosphorylation sites from 17 tissues
Protein structures
enriched
depleted
Sequence motifs
Pathway preferences
Legend:
Amino acid preferences in the 3D environment of phosphorylated serines.
depleted
enriched
kinases
motifs
tissues
Spatial motifs
Tripartite graph connecting motifs, tissues, and kinases

Figure 8: A graphical summary of the followed methodology for phosphorylation sites.

Datasets	S[a]	T[b]	Y[c]	Total[d]
Initial dataset[e]	14661/348754	2832/215735	424/100083	17917/664572
PS1D-70[f]	9254/128578	1594/82698	249/38598	11097/249874
Structure-based dataset[g]	729/7377	232/6564	69/4384	1030/18325
Structure-based and solvent accessible dataset[h]	489/5941	181/5482	53/3981	723/15404
PS3D-90[i]	423/4162	140/3790	46/2804	609/10756

Table 3: Data summary of phosphorylation sites. [a] Number of serine phosphorylation sites/non-phosphorylation sites. [b] Number of threonine phosphorylation sites/non-phosphorylation sites. [c] Number of tyrosine phosphorylation sites/non-phosphorylation sites. [d] Total number of phosphorylation sites/non-phosphorylation sites including all residue types. [e] Initial dataset directly obtained from the study of Lundby *et al.* (Lundby et al., 2012b). [f] Sequence-based dataset after sequence redundancy reduction on peptides at the 70% identity level. [g] The structure-based dataset where phosphoproteins were mapped on PDB structures. [h] The structure-based dataset including solvent accessible phosphorylation/non-phosphorylation sites. [i] Structure-based dataset after the sequence redundancy reduction on peptides at the 90% identity level.

sites may be due to failure in finding either the best-matching UniProt ID, or a serine (S)/threonine (T)/tyrosine (Y) residue in the specified sequence position, or the sequence of the corresponding full-length protein in the UniProt database.

We also generated a negative (reference or non-PSSs, non-PTSs, non PYSs) dataset by extracting all S/T/Y residues not annotated as phosphorylated by Lundby *et al.* and matching them to those tissues in which the protein containing the reference site had at least one experimentally observed PSS/PTS/PYS. Then, the 21-mer sequences (from -10 to +10) surrounding each site in both positive and negative datasets were extracted, and homology reduction on these 21-mers were performed using CD-HIT (Li and Godzik, 2006) at the 70% identity threshold. Some of phosphorylation and reference sites are found in more than one tissue. The statistics of the resulting datasets, PS1D-70, can be found in Table 3 (11097 phosphorylation sites in 5286 proteins).

3.1.2 *Identification of sequence motifs*

We used the Two Sample Logo method (Vacic et al., 2006) to find enriched and depleted residues in the 21-mer sequences occurring in different tissues, using the associated PSSs/PTSs/PYSs and non-PSS/non-PTSs/non-PYSs as positive and negative sample inputs, respectively. For instance, PSSs/PTSs/PYSs found in kidney were compared to non-PSS/non-PTSs/non-PYSs in kidney. The Motif-X online tool (Chou and Schwartz, 2011) was employed to detect sequence motifs from 21-mer sequences where PSSs/PTSs/PYSs and non-PSS/non-PTSs/non-PYSs were used as the foreground and background datasets, respectively.

3.1.3 *Obtaining 3D structures of phosphorylated proteins*

We applied the same procedure as we described in Chapter 2 to collect 3D structures of phosphorylated proteins. We extracted the total of 2079 related 3D structures from the Protein Data Bank based on BLAST-P hits. After requiring at least 80% identity within the sequence segment spanning ±50 sequence positions around the phosphorylated residue, we obtained 1030 phosphorylation sites in 399 protein structures with the resolution better that 3Å.

Once the structure-based positive and negative PSSs/PTSs/PYSs datasets were generated as described above for the sequence-based data, homology reduction was again performed at the 90% identity level, taking into account only solvent accessible phosphorylation sites. The resulting dataset, which we call PS3D-90, contains 609 phosphorylation sites (423 PSSs, 140 PTSs and 46 PYSs) and 10756 non-phosphorylation sites (4162 PSSs, 3790 PTSs and 2804 PYSs) in 332 proteins with known structures (Table 3). The number of PSSs, PTSs and PYSs in each tissue can be found in Table 19, Table 20 and Table 21, respectively. Since the number of PYS per tissue in the PS3D-

90 dataset is too low, we generally avoided tissue-based analyses of PYSs in the PS3D-90 dataset.

3.1.4 *Statistics*

The R environment (Team, 2009) was used for statistical analyses. For the numerical data we used the non-parametric two-sample Kolmogorov Smirnov test, whereas the Fisher exact test was applied for the categorical data. The comparisons were made within each tissue where the occurrence of a particular property of phosphorylation sites was compared to that of non-phosphorylation sites.

3.1.5 *Spatial (3D) environments of phosphorylated and reference (non-phosphorylated) serine/threonine/tyrosine residues*

By calculating the occurrence of 20 different amino acid types within the radial distances of 2 to 12 Å from the phosphorylated S/T/Y residue, and excluding amino acids already present in the sequence vicinity of PSSs/PTSs/PYSs, 3D and *pure* 3D environments of phosphorylation sites in the PS3D-90 dataset were determined, respectively. Distances between amino acid residues were defined based on the minimal distance between any pair of atoms belonging to these residues. For both type of environment the Fisher exact test was performed to assess the significance of the differences between PSSs/PTSs/PYSs and non-PSS/non-PTSs/non-PYSs in each tissue, and we used our in-house software tool to visualize these differences efficiently. We applied the procedure defined in Chapter 2 to find the propensity of each amino acid type at each radial distance ranging from 2 to 12 Å (in increments of 1Å).

3.1.6 *Structural properties of phosphorylation sites*

We extracted structural features of phosphorylation sites as described in Chapter 2. Briefly, we utilized NACCESS (Hubbard and Thornton, 1993) to calculate the surface accessibility of phosphorylation sites and their sequence environments. DisEMBL (Linding et al., 2003) was used to predict disordered regions in each phosphorylated protein sequence. We gathered the secondary structure annotations from the DSSP database (Joosten et al., 2011). Note that the number of PYSs associated with known secondary structures was not sufficient for comparison; therefore, they were excluded from this analysis.

3.1.7 *Analysis of structural folds and functional domains*

Structural folds of phosphorylated proteins from the PS3D-90 dataset in each tissue were examined according to the *class* and *protein domain* levels of the SCOP hierarchy (Murzin et al., 1995). At the structural *class* level, the significance threshold $p < 0.05$ was used, whereas at the *protein domain* level false discovery rate control was performed for multiple hypothesis correction in each tissue, and the significance threshold after the correction $p < 0.05$ was used after all p-values were adjusted. Note that the numbers of PTSs and PYSs associated with known SCOP folds were not sufficient for comparison; therefore, they were excluded from this analysis.

3.1.8 *KEGG pathway analysis*

Using the best-matching UniProt identifiers of each PSSs/PTSs/PYSs and non-PSSs/non-PTSs/non-PYSs in the PS1D-70 dataset, pathways obtained from Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa et al., 2006) were analyzed across tissues, and enriched pathways were detected. False discovery rate control was performed for multiple hypothesis correction in each tissue, and the significance threshold after the correction $p < 0.01$ was used after all p-values were adjusted.

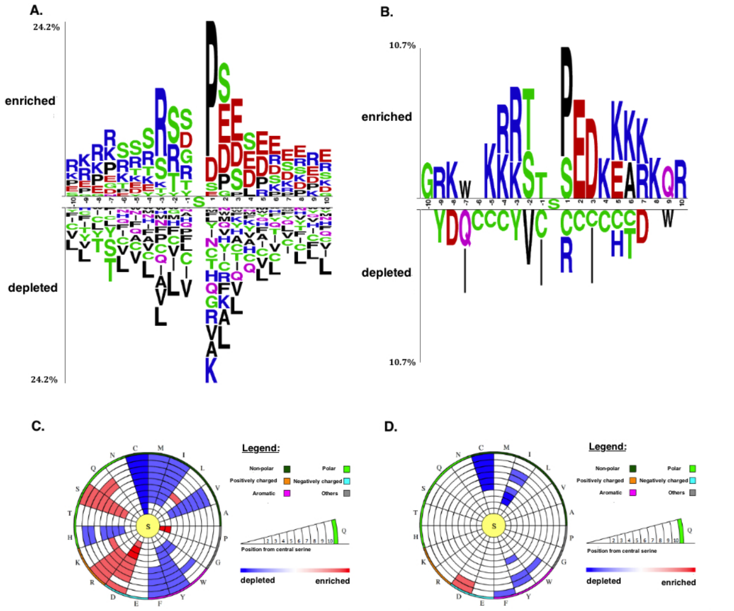
3.1.9 *Kinase analysis*

In order to analyze enriched kinases across tissues, we used the substrate-matched kinase data given by Lundby *et al.* (Lundby et al., 2012b) where phosphorylated protein sequences were matched against motifs of known kinases using the PHOSIDA Motif Matcher toolkit (Gnad, Gunawardena, and Mann, 2011). This toolkit finds the matches of 33 previously identified kinase motifs around each phosphorylation site in a particular protein sequence. The list of matched kinases is as follows: AKT, ATM, ATR, AURORA, AURORA-A, CAMK2, CDK1, CDK2, CHK1, CHK2, CK1, CK2, ERK/MAPK, GSK3, NEK6, PKA, PKD, PLK, PLK1, and RAD53. Based on these assignments, matched kinases of 8836 out of 11097 phosphorylation sites could be determined for the PS1D-70 dataset. We compared the counts of PSSs/PTSs/PYSs in each tissue phosphorylated by each kinase to the occurrence of non-PSSs/non-PTSs/non-PYSs in the corresponding tissue. False discovery rate control was performed for multiple hypothesis correction in each kinase class, all p-values were adjusted, and the significance threshold after the correction $p < 0.01$ was used.

Relationships between kinases, tissues and motifs were visualized by means of a tripartite graph as implemented in Cytoscape (Shannon et al., 2003). Only the motifs that are significantly enriched in each tissue (Table 16) and the corresponding kinase motifs provided in Phosida (Gnad, Gunawardena, and Mann, 2011) were considered while drawing the graph.

3.1.10 *Tissue-specific expression of kinases*

For each kinase we identified the corresponding rat UniProt ID from the PhosphoSitePlus database (Hornbeck et al., 2012), which contains experimentally identified kinase-substrate data. If no rat information was found, we attempted to find mouse and human UniProt



Amino acid preferences in the 3D environment of phosphorylated serines.

Amino acid preferences in the pure 3D environment of phosphorylated serines.

Figure 9: (A) Two sample logo analysis of global PSSs in the PS1D-70 dataset. (B) Two sample logo analysis of global PSSs in the PS3D-90 dataset. (C) 3D environments of global PSSs in the PS3D-90 dataset. (D) *Pure* 3D environments of global PSSs in the PS3D-90 dataset.

IDs, in this order of preference. If a query kinase could not be identified in the rat, mouse or human proteomes based on the PhosphoSitePlus database, it was excluded from further analysis. Protein expression levels of kinases across tissues were obtained from the PaxDb database (Wang et al., 2012). Since no tissue-specific expression data for rat proteins is provided in PaxDb, we used PaxDb data for the mouse orthologs of rat or human kinases obtained from the KEGG database. For some of the kinases considered in this study PhosphoSitePlus contains information for multiple isoforms (for instance, GSK3A and GSK3B isoforms for GSK3). In such cases, the expression of each isoform was analyzed separately. To assess the existence of tissue-specific kinase expression, a one-sample t-Test was used. The expression value of a particular kinase in a particular tissue was compared to expression values of the same kinase in all tissues. All p-values were adjusted and the significance threshold after the correction $p < 0.01$ was used.

3.2 results and discussion

3.2.1 *Analysis of sequence motifs of global phosphorylation sites*

We first examined global phosphorylation trends at the sequence level based on the PS1D-70 dataset. In general our findings are in line previous studies (Chen et al., 2014; Schwartz and Gygi, 2005; Villen et al., 2007; Zhao et al., 2012). Proline (P) residues are enriched at position +1 with respect to globally phosphorylated serine and threonine sites (Figure 9A and Figure 51). Negatively charged glutamic acid (E) and aspartic acid (D) residues as well as polar serines are also enriched in the upstream regions of phosphorylated serine, threonine and tyrosine (Y) sites (Figure 9A, Figure 51A, and Figure 52A). There is also a strong trend for charged lysine (K) and arginine (R) residues to be enriched in the sequence neighborhood of phosphorylated serines and threonines, except for the positions from +1 to +4.

3.2.2 *Tissue-based analysis of sequence motifs of phosphorylated sites*

Previous studies have shown that beyond the general trends, phosphorylation sites exhibit kinase-specific sequence and spatial motifs as well as compartment-specific sequence motifs (Chen et al., 2014; Durek et al., 2009; Su and Lee, 2013). The availability of experimentally identified tissue-specific phosphorylation sites has now enabled us to examine phosphorylation trends across tissues. Our analysis of the PS1D-70 dataset revealed clear tissue-specific preferences. For instance, only PSS in perirenal fat have phenylalanine (F) residues at position +1 (Figure 10A), whereas only PSS in pancreas and testis have a preference for glutamine (Q) residues at position -2 (Figure 10B and Figure 10C). Only in cortex, glycine (G) residues are enriched at position -3 with respect to central phosphorylated serine sites (Figure 10D). Histidine residues are only enriched at position +6 in blood (Figure 10E), whereas PSS in stomach and liver show preferences for asparagine (N) residues at position +3 (Figure 10F and Figure 10G). In contrast to the global trends, none of the tissue-specific phosphorylation sites show any enrichment for serine residues at position +1 (See Figure 53 – Figure 70 for more detailed graphs for these and other tissues). Tissue-specific trends for PTSs are also quite prominent. For instance, only PTS in cortex show a preference for methionine (M) residues at position -6, whereas asparagine residues are strongly preferred at positions -7 and -8 only in muscle (Figure 71). Glutamine residues are enriched for PTSs in blood at position -7.

We utilized the Motif-X software to identify enriched sequence motifs that are exclusively tissue-specific and cannot be detected when analyzing global trends (Table 16). For instance, the motifs pS-P-E, pS-P-X-X-E, E-X-X-X-X-X-X-X-X-pS-P, E-X-X-X-X-pS, pS-X-X-X-X-X-X-X-X-D, pS-X-X-X-X-X-X-X-X-E, E-X-X-X-X-X-X-X-X-pS and pS-X-X-X-X-X-X-X-K are only associated with brain-specific PSS. On the other hand, the motif pS-X-X-X-X-D is only observed in PSS in liver. These observations indicate that individual tissues harbor specific sequence environments for phosphorylation.

Lundby *et al.* have already investigated sequence motifs of phosphorylation sites in two tissues - brain and testis - but in their work comparison was performed between tissue-specific and global phosphorylation sites. Using global sequence signatures of phosphorylation sites as a background dataset for studying tissue-specific motifs may fail to reveal potential systematic biases existing in the foreground dataset and can lead to random and non-informative motif signals (Yao et al., 2013). In this work we investigate the enrichment or depletion of phosphorylation sites relative to their non-phosphorylated

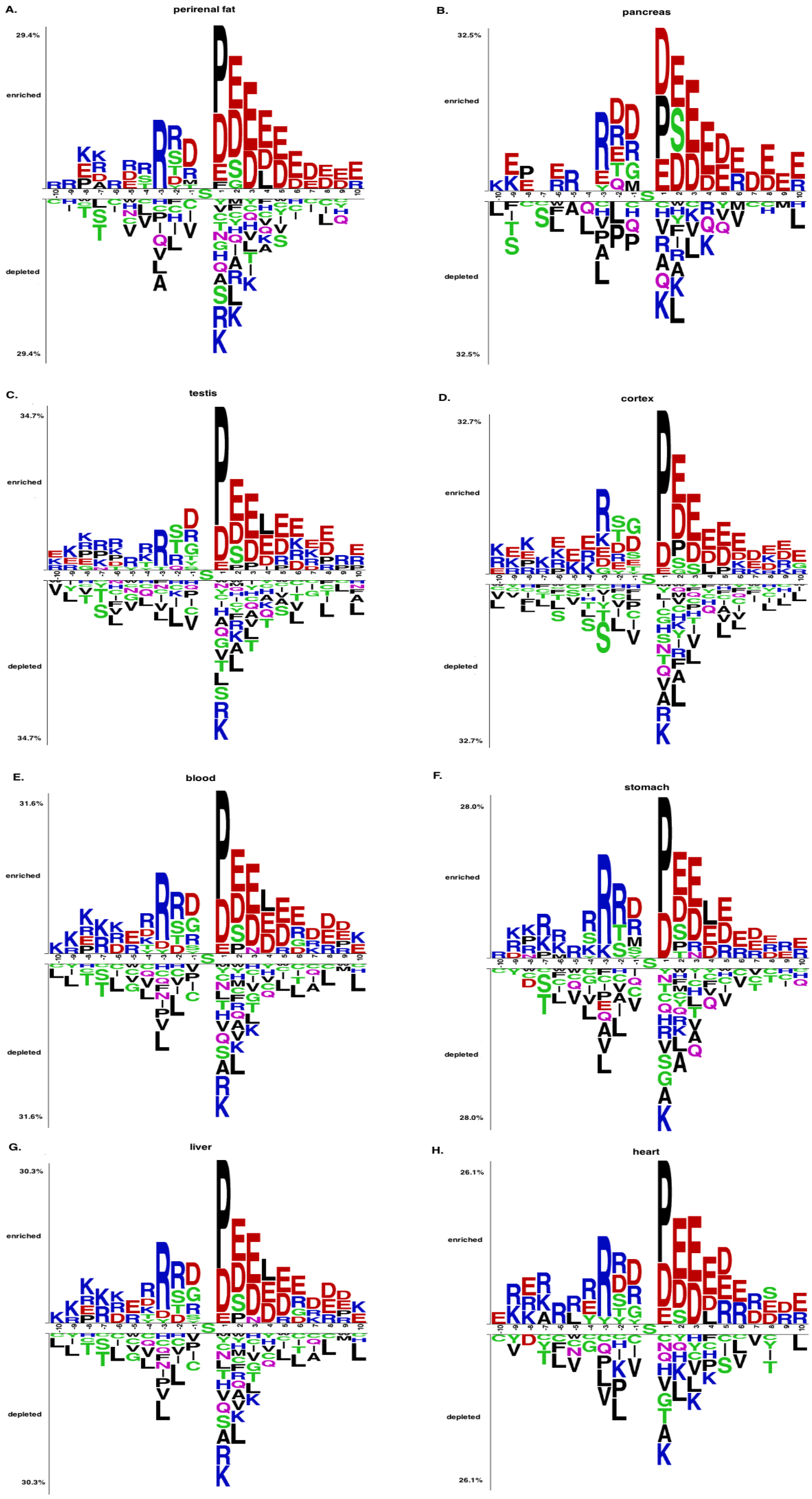
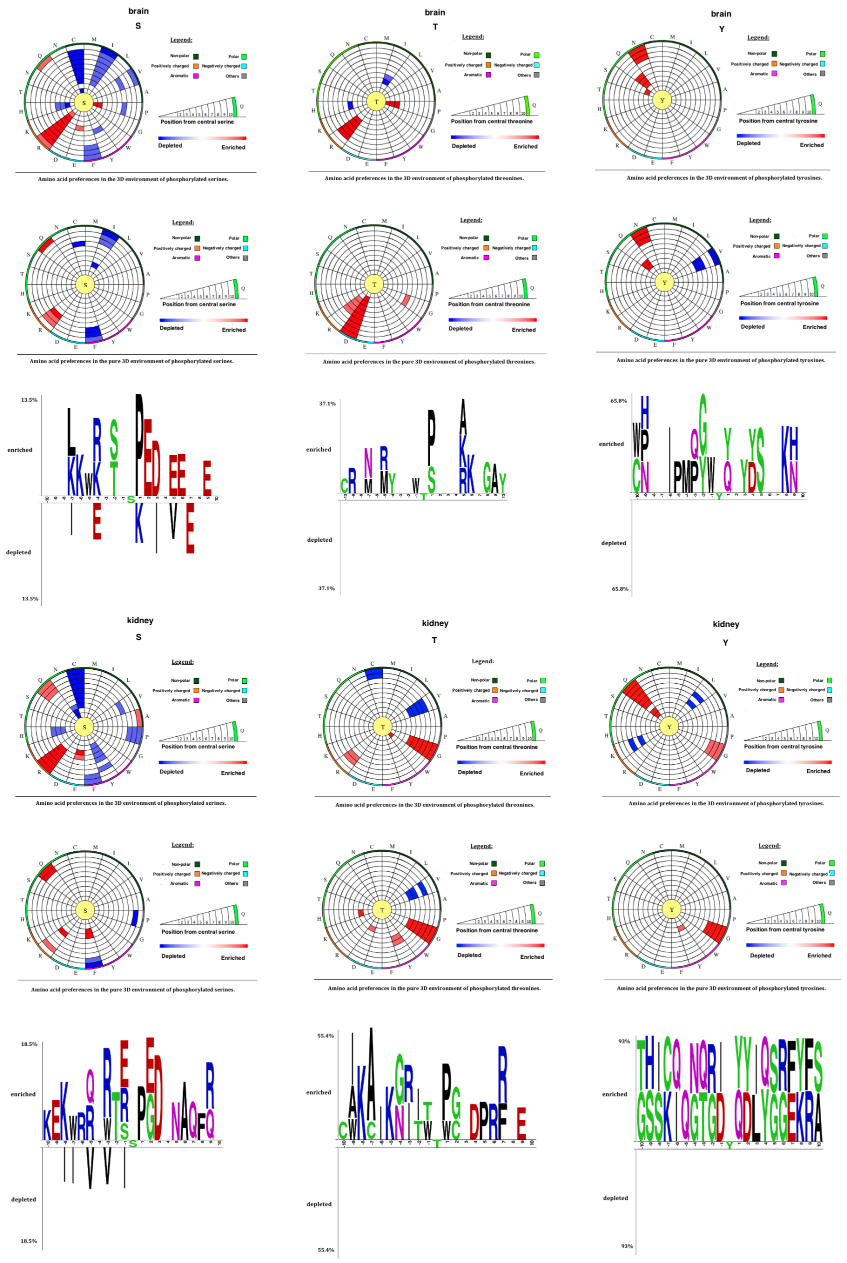


Figure 10: Two sample logo analysis of PSSs from the PS1D-70 dataset in perirenal fat (A), pancreas (B), testis (C), cortex (D), blood (E), stomach (F), liver (G) and heart (H). See data on other tissues given in Figure 53 - Figure 70.



Amino acid preferences in the 3D environment of phosphorylated serines.

Amino acid preferences in the 3D environment of phosphorylated threonines.

Amino acid preferences in the 3D environment of phosphorylated tyrosines.

Amino acid preferences in the pure 3D environment of phosphorylated serines.

Amino acid preferences in the pure 3D environment of phosphorylated threonines.

Amino acid preferences in the pure 3D environment of phosphorylated tyrosines.

Figure 11: Sequence (1D) and structural (3D and *pure* 3D) environments of PSSs, PTSs, and PYSs from the PS3D-90 dataset represented by two sample logos and circular plots, respectively. Data for brain and kidney is shown (see the figures in Appendix A for all tissues). The circular amino acid propensity plots were produced by our in-house software tool. Color code for amino acid residue enrichment in the circular plots: (i) red – enriched, if p-value < 0.05 and odds ratio > 1, (ii) blue – depleted, if p-value < 0.05 and odds ratio < 1, and (iii) white –neither enriched nor depleted, if p-value $\geqslant$ 0.05.

counterparts in each individual tissue, thus ensuring that the trends found are not due to the tissue-dependent variation of protein abundance or individual biological roles of proteins. In other words, enriched motifs identified through this approach are solely due to phosphorylation processes occurring in a particular tissue and not any other tissue-specific properties. We detected 15.4% and 17.1% more enriched/depleted residues in the sequence environment of phosphorylation sites in brain and testis, respectively, based on two sample logos, even though some trends for certain residues at particular positions overlap in both studies (58.8% and 57.5% in brain and testis, respectively) (see Table 17 and Table 18). Note that we applied separate two sample logo analyses for serine and threonine sites, whereas Lundby et al. combined both residues. Furthermore, discriminative motif analysis using the Motif-X software yielded qualitatively different tissue-specific motifs compared to the motifs obtained in the previous study by Lundby *et al.* One of the most significant motifs we identified - E-X-X-X-X-pS - is only observed in brain but not in any other tissue (Table 16), whereas residue E in the corresponding position has not been even found enriched using the former method (X and pS here represent any residue and phosphorylated serine, respectively). We found that the residue G is enriched at position -1 in testis and the positively charged residues R and K are enriched at almost all positions in downstream regions of phosphorylated serines in both brain and testis, whereas none of these effects could be observed using global phosphorylation sites as a background set. Similarly, residue P is enriched in all tissues at position +1 in our analysis; however, Lundby *et al.* found P to be enriched at that position in brain, but not in testis. We also found the motif pS-X-X-R to be highly specific for testis, while the residue R is in fact depleted in the corresponding position in the previous study.

High specificity of kinase action is an important requisite for exquisite regulation of signal transduction processes (Kobe et al., 2005). In accordance with this notion, compartment-specific kinases have been proposed in a previous work (Chen et al., 2014), whereas the existence of compartment- and tissue-specific acetyltransferases (KATs) has also been discussed in Chapter 2 and (Lundby et al., 2012a; Sadoul et al., 2011). The substrate sequence specificity among tissues identified in this work suggests the existence of tissue-specific kinases and phosphatases.

3.2.3 *Tissue-based analysis of phosphorylation sequence motifs in proteins with known 3D structure*

As a pre-requisite for the investigation of structural trends (see below), we performed a separate analysis of phosphorylation sequence motifs in the subset of proteins possessing a known 3D structure (dataset PS3D-90). The results obtained are markedly different from the ones derived from the PS1D-70 dataset due to the fact that structurally characterized proteins are depleted in disordered regions, which leads to a different amino acid composition. Specifically, negatively charged residues, as well as disorder-related glycine and serine residues are less pronounced in this dataset. For instance, aspartic acid and glutamic acid are not observed at position +1 of the global and tissue-specific PSSs (Figure 9B, Figure 11 and Figure 73). Lysine and alanine (A) residues are enriched at positions +4 and +6 of the global PSS in the PS3D-90 dataset, whereas they are not enriched at the same positions of the global PSS in the PS1D-70 dataset, where lysine residues are even depleted at position +4. Glutamic acid residues are depleted at positions -4 and +7 for structurally known PSS in brain, whereas they are enriched at these positions in the PS1D-70 dataset. PSS in heart have strong preferences for alanine and proline residues at positions -4 and -3, respectively, whereas such preferences could not be observed in the PS1D-70 dataset, and proline residues are even depleted at position -3. Asparagine and glutamine residues are generally enriched in the upstream regions of structurally known tissue-specific sites, whereas there are mainly depleted in the PS1D-70 dataset.

3.2.4 *Tissue-specific spatial motifs of phosphorylation sites*

Previous studies found only minor differences between the 3D structural surroundings of phosphorylated and non-phosphorylated sites, whereas stronger trends were observed when taking into account kinase preferences (Durek et al., 2009). In this study we detected more significant spatial motifs associated both with global and tissue-specific PSSs. In accordance with (Durek et al., 2009), arginine, proline, leucine and serine residues are enriched in the spatial environment of global PSSs, but we also found aspartic acid, lysine and glutamine residues to be enriched (see Figure 9C). Frequent occurrence of aspartic acid residues has also been observed in 1D and 3D environments of PSSs in previous studies (Lundby et al., 2012b; Su and Lee, 2013).

Similar to the trend in the sequence motifs discussed above cysteine (C) residues are strongly avoided in the spatial surroundings of PSS in all tissues. Strong enrichment of proline and aspartic acid residues is observed in close proximity of global PSS (around 2-5 Å away, see Figure 9C), which again parallels the trends in sequence motifs. However, the enrichment of leucine residues, at a distance range of 6 to 7 Å, is not found in the sequence environment. On the other hand, no enrichment of glutamic acid is observed in the 3D environment even though glutamic acid is highly enriched in the sequence motifs of global PSS.

We also found tissue-specific trends in the spatial environments of phosphorylation sites (Figure 76, Figure 77 and Figure 78). PSS in brain have a strong preference for arginine residues, and it is much more predominant than the enrichment in the 1D environment. Histidine and cysteine residues are strongly depleted in the 3D environment of PSS in kidney, whereas no preference for these residues is observed in the 1D environment. Thus, patterns of amino acid usage around phosphorylated sites are generally consistent between 1D and

3D environments, although some of the tendencies found at the spatial environment are not observed in the 1D environment alone.

In order to disentangle the influence of local amino acid content from the effects caused by spatial proximity, we performed a separate analysis of *pure* structural environments (see Section 3.1). While only a weak enrichment of certain amino acid residues is observed in global *pure* 3D environments of PSSs (aspartic acid, Figure 9D) and PTSs (aspartic acid and glycine, Figure 51), tissue-specific preferences are more clear cut (Figure 79, Figure 80 and Figure 81). For instance, aspartic acid residues at a distance between 3 Å and 12 Å with respect to PTSs are strongly enriched in brain, which is not observed when sequence context is also considered. Similarly, PTS residing in kidney have a preference for histidine only when their *pure* 3D environments are considered. These findings imply that there are statistically significant patterns in spatial residue preferences of phosphorylation sites in addition to significant sequence patterns. Previous studies have shown that the spatial context plays a role in the recognition of substrates by kinases (Durek et al., 2009; Su and Lee, 2013). Here we found that the amino acid composition in 3D varies in a tissue-dependent manner, which implies that both sequence and spatial environments of phosphorylation sites may play a role in determining the substrate-specificity of kinases and phosphatases across tissues.

3.2.5 *Structural properties of phosphorylation sites*

It has been previously shown that phosphorylation sites generally reside in unstructured and disordered regions (Huttlin et al., 2010; Kreegipuu, Blom, and Brunak, 1999). We assessed the disordered region preferences of phosphorylation sites in the PS1D-70 dataset, where disordered regions were predicted from phosphorylated sequences. We found that PSSs and the residues surrounding them follow the same tendency in all individual tissues (Figure 82), while PTSs display the preference for disordered regions in all tissues except for muscle (Figure 83).

Based on the experimental structures of phosphorylated proteins, we also analyzed structural properties of phosphorylation sites in the PS3D-90 dataset. However, we were only able to assess the global preferences of phosphorylated sites since the datasets of proteins with known 3D structure phosphorylated in a tissue specific manner were not large enough to obtain statistically significant results.

Global PSSs, PTSs and PYSs along with the residues surrounding them are consistently and significantly more solvent exposed than non-PSSs, non-PTSs and non-PYSs and their 1D environments by about 30.73% ($p < 2.2$ x 10-16), 27.05% ($p < 6.9$ x 10-6) and 30.02% ($p < 1.5$ x 10-3) on average, respectively (see Table 19, Table 20 and Table 21), which is in line with previous reports (Durek et al., 2009; Kim et al., 2006; Suo et al., 2012). We also found that PSSs preferentially occur in more flexible regions of protein structures, as assessed by the B-factor analysis. Globally phosphorylated serine sites and the residues surrounding them in a very close proximity have

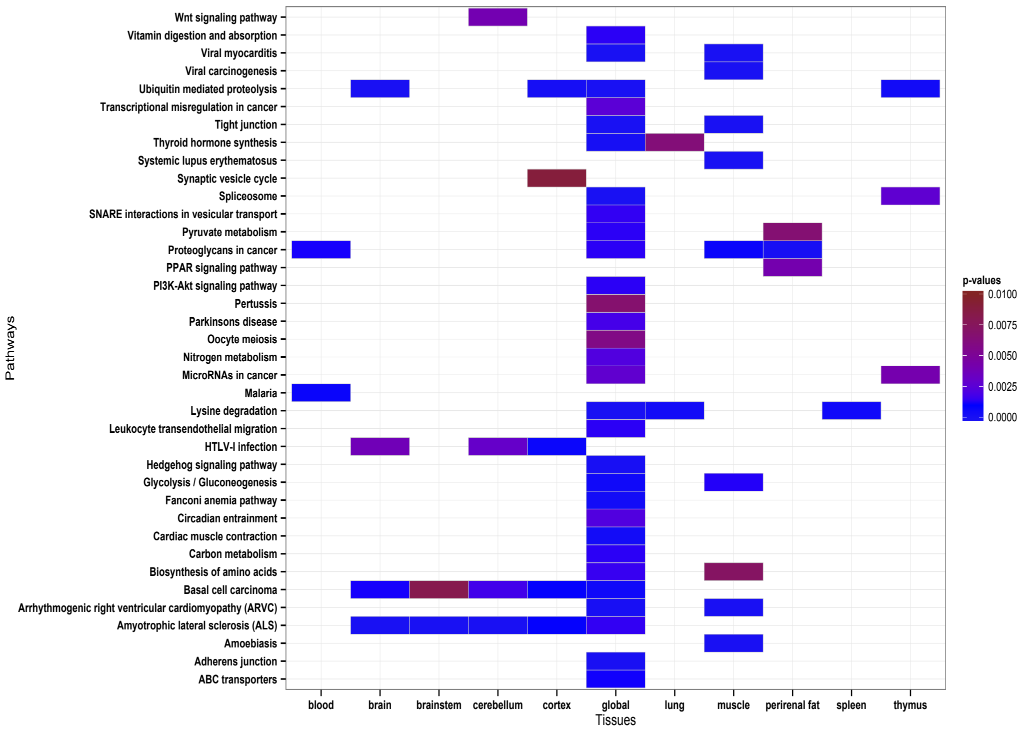


Figure 12: KEGG pathway analysis of the serine phosphorylated proteins from the PS1D-70 dataset. Pathways with a corrected p-value < 0.01 in each tissue are considered as significantly enriched. Insignificant results are represented by white color. Only the tissues with at least one significantly enriched pathway are shown.

larger B-factor values than their non-phosphorylated counterparts do (Table 19). Globally phosphorylated PSSs and PTSs in the PS3D-90 dataset display a clear tendency to reside in loops (Figure 85 and Figure 86).

3.2.6 *Phosphorylated proteins take part in tissue-specific biological pathways*

As previous studies we also have shown in Chapter 2 that the structural environment of phosphorylation and acetylation sites plays an important role in determining kinase/acetyltransferase specificity and ultimately the functional role of phosphorylation and acetylation processes (Su and Lee, 2013; Tyanova et al., 2013). Hence, we conducted a tissue-based analysis for phosphorylated proteins to assess their specialization for various functions in different tissues. Following the work in Chapter 2, we compare phosphorylated proteins in specific pathways to non-phosphorylated counterparts in each individual tissue (see Section 3.1), which makes this analysis independent from the mere presence or absence of certain functions in the respective tissues. Crucial roles of phosphorylated proteins in membrane transport, metabolism, signaling pathways, and their disease related pathways are in general well-known. We found that globally phosphorylated serine sites are linked to various processes, including ABC transporters activities, adherens junction formation, cardiac muscle contraction system, energy generation processes of glycolysis/gluconeogenesis, pyruvate metabolism and lysine degradation, hedhedog signaling pathway, leukocyte transendothelial migration system, tight junction, and ubiquitin mediated proteolysis (see full list in Figure 12). Proteins carrying serine phosphorylation are associated with disease processes and macromolecules affecting tumor progress, including the regulation of arryhythmogenic right ventricular cardiomyopathy (ARVC) and basel cell carcinoma diseases, and proteoglycans in cancer.

Serine phosphorylated proteins display distinctly different biological pathway preferences both in a global and tissue-specific manner. For example, proteins involved in synaptic vesicle cycle show significant enrichment in PSSs only in cortex. Only proteins phosphorylated on serine residues in muscle play a role in regulation of diseases including amoebiasis, systematic lupus erythematosus and viral carcinogenesis, whereas only phosphorylation in blood appears to be involved in the regulation of malaria disease, which is a parasitic protozoans-caused disease transmitted by the biting of mosquitoes (WHO, 2015).

Globally phosphorylated proteins on threonine residues are involved in energy metabolism and disease related pathways including viral myocarditis, viral carcinogenesis and tight function processes. Tissue-specific preferences can only be detected for phosphoproteins in muscle that they also take part in the regulation of diseases such as amoebiasis, arrhythmogenic right ventricular cardiomyopathy (ARVC) and systematic lupus erythematosus (Figure 88). These findings show that biological processes are regulated by phosphorylation in a tissue dependent manner. Kinases are one of the most important drug targets for a number of diseases - including cancer, hypertension, Parkinson's disease, and autoimmune diseases (Roskoski, 2015). The fact that phosphoproteins exhibit tissue-specific preferences in the regulation of disease pathways implies that designing drugs targeting tissue-specific disease pathways may be a promising avenue towards improved and more specific therapeutic effects, as suggested previously in Chapter 2.

In order to assess the association between biological pathways and functional domains harboring phosphorylated sites, we first analyzed statistical preferences at the top level of the SCOP hierarchy, structural class. Except for the slight enrichment of cerebellum-, brain- and blood-specific PSSs in all-α proteins, we could not detect any significant trends (Figure 87). At the second level of the SCOP hierarchy, which reflects structural folds, we identified a number of significant preferences for PSSs, which parallel tissue-specific biological pathway preferences described above. Protein domains found to be phosphorylated in a tissue-specific manner include fructose-1,6-bisphosphate aldolase domain, which belongs to the glycolysis pathway, and creatine kinase C-terminal domain, which is a key enzyme responsible for the intracellular energetic homeostasis of vertebrate excitable tissues and for catalysis of the reversible transfer of the high-energy phos-

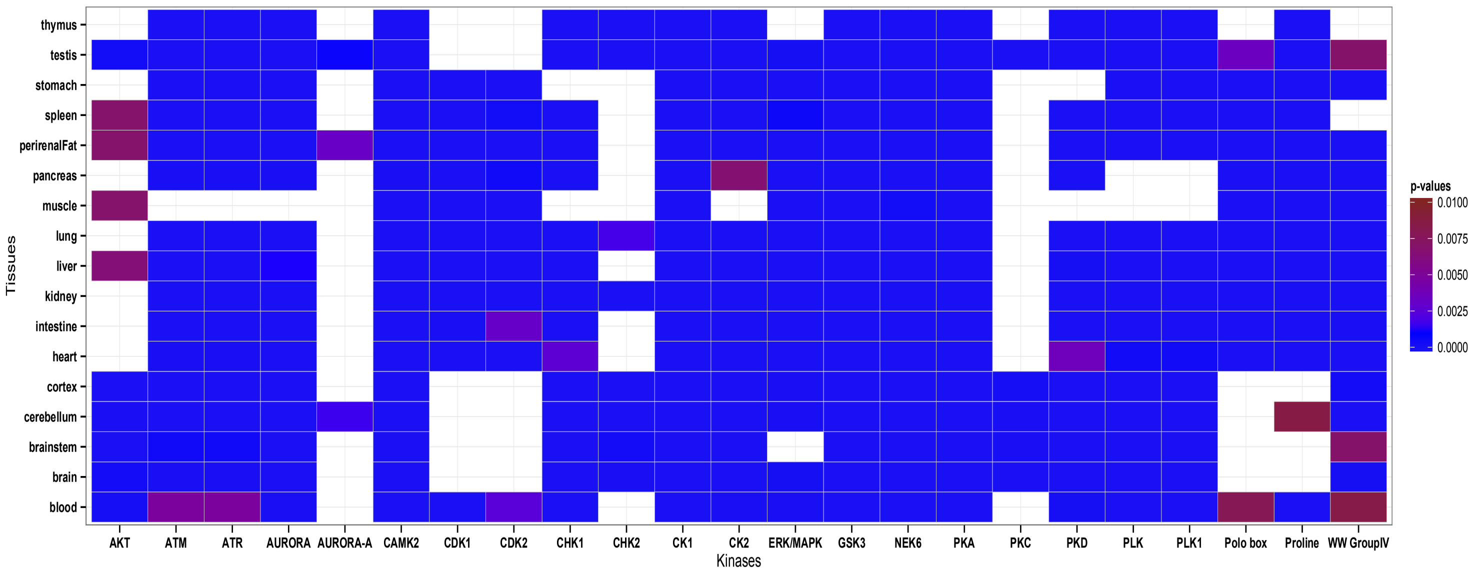


Figure 13: Tissue preferences of kinases targeting serine phosphorylated sites in the PS1D-70 dataset. Tissues with a corrected p-value < 0.01 in each kinase class are considered significantly enriched for the corresponding kinase classes. Insignificant results are represented by white color.

phate between the ATP/ADP and creatine/phosphocreatine systems (Chen et al., 2012) (Figure 90). The beta-chain of hemoglobin is phosphorylated on serine residues only in blood, which may imply the relevance of phosphorylation for the oxygen transport function. The cytosolic class alpha glutathione S-transferase (GST) domain resides in proteins phosphorylated on serine residues only liver where the GST domain catalyzes the joining of glutathione to xenobiotic substrates for detoxification. GSTs also plays a role in cell signaling and the overexpression of GSTP1-1 – an isozyme of the mammalian GST family - has been associated to cancer (Laborde, 2010), making it a potential drug target and warranting an experimental investigation of its phosphorylation. Globally phosphorylated proteins carrying threonine phosphorylation harbor the same domains as serine phosphorylated proteins do, but their tissue-specific preferences differ (Figure 91). For instance, only phosphoproteins in thymus include the histone H2A domain, which gains function in DNA repair after serine phosphorylation of one of its variants (Jakob et al., 2011). We therefore propose that threonine phosphorylation in thymus might also generate new functions of the H2A proteins. The adipocyte lipid-binding protein (ALBP), which is a carrier protein in fatty acid metabolism and has roles in many diseases (Baxa et al., 1989), creatine kinase N-domain, which is an important enzyme in energy-consuming processes, and cAMP-dependent PK catalytic subunit domain, which is essential for phosphorylation of some proteins (Manni et al., 2008), are phosphorylated on threonine residues only in perirenal fat.

3.2.7 *Kinases target tissue-specific phosphorylation sites*

Echoing what we have performed in Chapter 2, in which we proposed the existence of tissue-specific lysine acetyltranferases (KATs) and lysine deacetylases (KDACs), the findings presented here imply the existence of tissue-specific protein kinases and phosphatases. The wealth of information on kinase-substrate association enabled us to examine kinase classes in different enriched tissues. Lundby *et al.* matched known sequence motifs of kinases to identified sequence motifs of tissue-specific phosphorylation and observed sequence motifs recognized by different kinases specific across tissues. Based on the association between phosphorylation sites and kinases provided by Lundby *et al.*, we examined tissue-specific target site preferences of kinases. Many kinases, including CK1, GSK3, NEK6, and PKA, mediate serine/threonine phosphorylation in all tissues, while some other kinases show more tissue-specific preferences (Figure 13). For instance, phosphorylation sites targeted by AURORA-A are only prominent in cerebellum, perirenal fat and testis, whereas PKC mediated phosphorylation is particularly pronounced in brain (as well as in brainstem, cerebellum and cortex) and testis. Similar to the study by Huttlin *et al.* (Huttlin et al., 2010) where tissue-specific protein expression and phosphorylation were shown to be uncorrelated, we compared the expression of each kinase in all tissues and found that the observed tissue-specific kinase preferences in phosphorylation are not caused

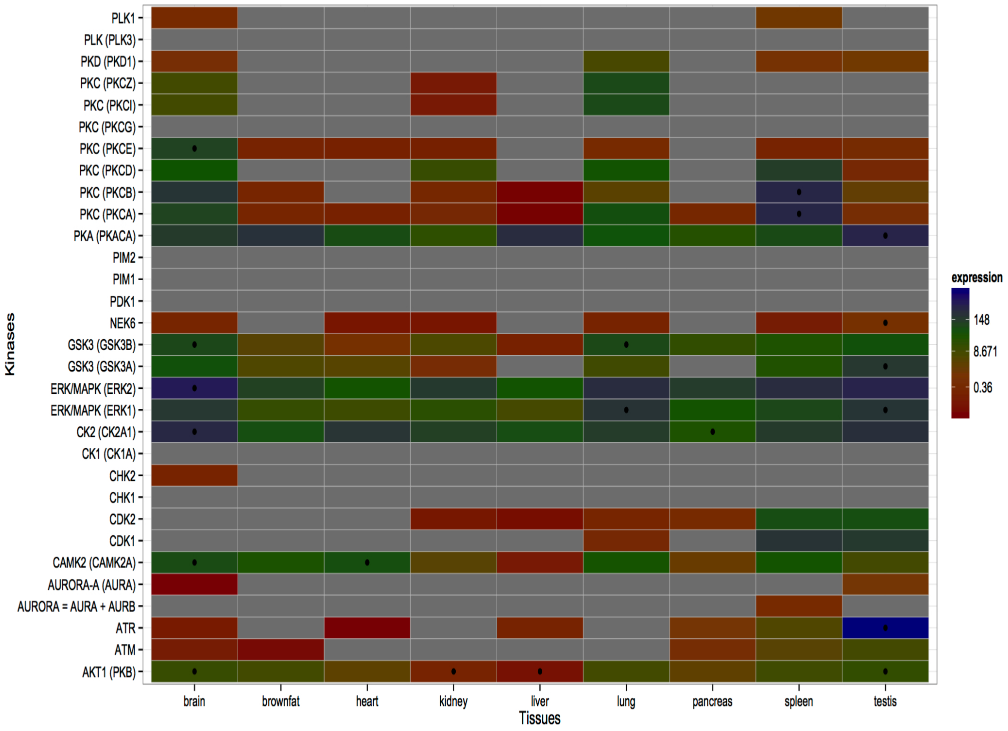


Figure 14: Comparison of the Ser/Thr kinase expression across tissues. Expression values are colored in log scale. Kinases with a corrected p-value < 0.01 in each tissue are considered significant and are indicated with black circles. Grey rectangles represent kinases whose expression values could not be found in the downloadable dataset of the PaxDb database. Note that only the expression values of tissues drawn here could be found in the PaxDb database. The median value 8.671 is assumed as the midpoint, calculated from all expression values in this heatmap.

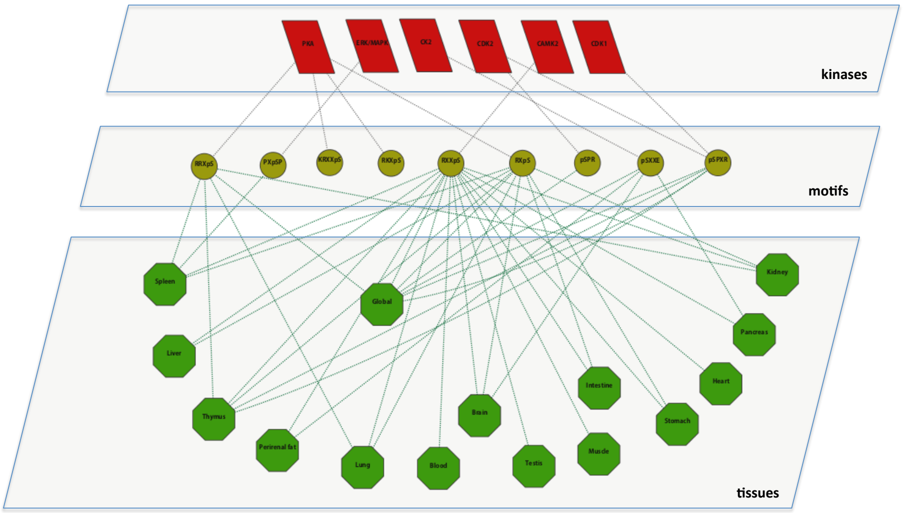


Figure 15: Tripartite graph showing interactions between serine phosphorylation motifs, kinases and tissues. The red parallelograms, yellow circles and green octagons represent kinases, motifs and tissues, respectively. Phosphorylated serine residues in motifs are represented with *pS*. Threonine and tyrosine phosphorylation sites are not shown in the network since no match between consensus kinase motifs and enriched tissue motifs was found.

by tissue-specific kinase expression. For instance, even though PKC mediated phosphorylation is only prominent in testis and brain (also including brainstem, cerebellum and cortex), the expression levels of PKCA (an isoform of PKC) are quite similar in all tissues except for spleen where it is more strongly expressed. Conversely, we found phosphorylation sites targeted by AKT1 in many tissues including spleen and liver (Figure 13), whereas AKT1 itself is less expressed in liver versus other tissues, and its expression level is similar in spleen and other tissues (Figure 14).

To better understand the kinase-tissue association, we built a tripartite graph joining kinases, motifs and tissues as nodes (Figure 15). Edges between kinases and motifs indicate that a consensus motif for the corresponding kinase is provided in the PHOSIDA database (Gnad, Gunawardena, and Mann, 2011). Edges between motifs and tissues mean that a given tissue contains phosphorylation sites enriched for the corresponding motifs in their sequence environment (as given in Table 16). Note that this graph is only a very small currently available subset of the general network connecting motifs, kinases and tissues. Nevertheless, we discovered some remarkable trends, such as the enrichment of the motif RXXpS in all tissues. This motif serves as a hub motif on the network, but only the kinase CAMK2 targets the phosphorylation sites containing the motif RXXpS in their environments. Phosphorylation sites in liver are enriched for the motif RXpS in addition to the motif RXXpS, and as a result it can be inferred that the kinases PKA and CAMK2 are highly active in liver. The motif PXpSP, is only targeted by the kinase ERK/MAPK in spleen. In pancreas the motifs pSXXE and RXXpS are enriched, which are required by the kinases CK2 and CAMK2. The kinases CDK1 and CDK2 target the motif pSPXR both globally and in a thymus-specific fashion.

3.3 CONCLUSIONS

Here we present a comprehensive study of phosphorylation sites at the sequence and structure level in 14 rat tissues based on the proteomics data recently published by Lundby *et al.* (Lundby et al., 2012b). We show that phosphorylation sites display tissue-specific preferences for certain residues in their linear amino acid sequence. Primary sequence motifs of phosphorylation sites in two tissues - brain and testis - have already been investigated by Lundby *et al.*, but in their work the comparison was performed between tissue-specific and global phosphorylation sites. By contrast, in this work we compare the enrichment or depletion of phosphorylation sites to their non-phosphorylated counterparts in each individual tissue, which allowed us to uncover some previously unnoticed trends. Beyond the known tendency of phosphorylation sites and the residues surrounding them to reside in disordered regions and irregular secondary structures, we also identified tissue-specific preferences for certain residues in their spatial environments. In addition to the previously described tissue-specific sequence motifs targeted by kinases (Huttlin et al., 2010; Lundby et al., 2012b), our findings would seem to indicate that tissue-specific spatial motifs in the substrates also play a role in kinase targeting. We also demonstrate that while many kinases mediate phosphorylation in all tissues, there are also kinases that operate in a tissue-specific manner. Interestingly, tissue-specific kinase preferences are not correlated with tissue-specific kinase expression. The tripartite graph connecting kinases, tissues and motifs reveals that some motifs are prominent in many tissues, but are only targeted by few kinases.

Similar to what we present in Chapter 2, in which we postulated that tissue-specific KATs and KDACs may control lysine acetylation, the findings presented here both at the sequence and structure level confirm the existence of tissue-specific protein kinases and phosphatases, as initially suggested by Huttlin *et al.* (Huttlin et al., 2010). Given the strong dependence of protein function on tissue context, it appears plausible that the intricate processes involved in kinase action, including the regulation of the catalytic domain by the hydrophobic spines (Schwartz and Murray, 2011), co-localization of the kinase and the substrate as a pre-requisite for substrate recruitment (Kobe et al., 2005), substrate sequestration or masking, which modulate availability for phosphorylation, may be tissue-specific. However, our analysis of the associations between kinases and amino acid sequence motifs failed to reveal any clear preferences of kinases for individual tissues. We therefore speculate that tissue specialization for kinase-substrate binding may be encoded at the 3D structure level.

Part IV

PREDICTION OF TISSUE-SPECIFIC PHOSPHORYLATION SITES BY INTEGRATING SEQUENCE- AND STRUCTURE-BASED FEATURES

4 PREDICTION OF TISSUE-SPECIFIC PHOSPHORYLATION SITES BY INTEGRATING SEQUENCE- AND STRUCTURE-BASED FEATURES

The recent advancements in high-throughput techniques enabled the identification of many novel phosphorylation sites; however, these techniques are still time-consuming and costly. Alternatively, many *in silico* studies have become very popular. Although successful, they still harbor some drawbacks as mentioned in Section 1.4, such as utilized features, generosity in redundancy elimination and so on. On the other hand, in Chapter 3 we have presented the evidence for tissue-specific kinases and proteases which regulate phosphorylation in a tissue-specific manner. To make use of this fact and address the above-mentioned drawbacks, here we present the first tissue-specific phosphorylation site prediction approach, TSPhosPred (Tissue-Specific Phosphorylation Prediction), based on the random forest (RF) algorithm. Our method uses sequence- and structure-based features as well as functional annotations. In order to obtain more accurate predictions, we also mapped the phosphorylated proteins onto Protein Data Bank (PDB) structures, and utilized experimental structures along with predicted structures. Both cross-validation and independent testing were applied, and prediction performance of each was presented. We also compared our method with an existing predictor on independent test data. Cross-tissues prediction was also performed in order to show the uniqueness of each tissue prediction model.

4.1 MATERIALS AND METHODS

4.1.1 *Data collection and preprocessing*

We used the same phosphorylation dataset, and performed similar preprocessing steps as defined in Chapter 3. In summary, we used the dataset from Lundby *et al.* where 31480 phosphorylation sites on 7280 proteins were established in 14 rat tissues, including brain (dissected into cerebellum, cortex and brainstem), blood, lung, heart, muscle, pancreas, kidney, liver, stomach, spleen, thymus, intestine, testis, and perirenal fat (Lundby et al., 2012b). In each tissue, the sequence position, intensity values, and the best-matching proteins of each phosphorylation peptide were collected from the dataset. We utilized the previously described method in Chapter 3 to identify the best-matching UniProtId for each phosphorylated peptide. Finally, we obtained 17917 phosphorylation sites on 5443 proteins Table 4.

In order to assess the influence of structural information into classification, we obtained 3D structures of phosphorylated proteins from the Protein Data Bank (PDB) using the procedure we previously de-

Datasets	Description	S	T	Y	Total
Initial dataset	Phosphorylation	14661	2832	424	17917
	Reference	348754	215735	100083	664572
PS1D-50	Phosphorylation	7065	1410	236	8711
	Reference	71201	43256	20090	134547
Structural dataset	Phosphorylation	729	232	69	1030
	Reference	7377	6564	4384	18325
Structural and solvent accessible dataset	Phosphorylation	489	181	53	723
	Reference	5941	5482	3981	15404
PS3D-70	Phosphorylation	379	126	36	541
	Reference	4162	3790	2804	10756

Table 4: Data summary of phosphorylation sites used in classification.

scribed in Chapter 2 and Chapter 3. With the resolution better than 3Å, we collected 1030 phosphorylation sites on 399 proteins Table 4.

4.1.2 *Training and independent test sets*

We generated prediction models for 17 tissues (including dissected parts of brain), 3 residue types and different subsets of features, each of them having their own training and independent test sets. All serine (S)/threonine (T)/tyrosine (Y) sites displaying non-zero intensity values in a tissue were considered in the positive set. The negative set consists of all S/T/Y residues not annotated as phosphorylated by Lundby *et al.*, and matched to those tissues on which the protein containing the negative site has at least one positive site. Even though all these sites are not necessarily true negatives, a large fraction of them is expected to be. We subsequently extracted the 21-mer sequences (from -10 to +10) surrounding each site in positive and negative sets.

- *Sequence-based data* - Homology reduction of 21-mers at the 50% identity level was performed separately on both positive and negative sets (PS1D-50) by using CD-HIT (Li and Godzik, 2006). Then, the second step homology reduction was done between positive and negative sets at the 45% identity level again using CD-HIT. If a 21-mer from positive set and a 21-mer from negative set had more than 45% sequence similarity, we kept the 21-mer in the positive set, and eliminated the corresponding similar 21-mer from the negative set. We called the resulting dataset, including positive and negative sets, PS1D.
- *Structure-based data* - Homology reduction of 21-mers was again performed separately on both positive and negative sets by using CD-HIT at the 70% identity level due to the lack of structurally known proteins (PS3D-70). Then, the second step homology reduction was done at the 45% sequence identity level as applied for the sequence-based data. We called the resulting dataset, including positive and negative sets, PS3D.

80% of each positive and negative set in final datasets was randomly selected to train models and evaluate the performance in 10-fold cross-validation tests. The remaining 20% was used as independent test data to evaluate if prediction models are over-fitting for

		PS1D			PS3D	
Tissue	Description	pS	pT	pY	pS	pT
Brain	Positive	1832	276	28	89	19
	Negative	19547	14345	6911	900	869
Brainstem	Positive	1052	147	17	61	9
	Negative	13941	9781	4711	685	668
Cortex	Positive	1320	185	16	68	14
	Negative	14801	10519	5112	771	759
Cerebellum	Positive	1287	165	18	61	11
	Negative	16629	11811	5617	733	705
Testis	Positive	985	102	12	33	4
	Negative	14737	9964	4555	480	473
Pancreas	Positive	239	25	3	7	2
	Negative	4460	2797	1299	182	184
Stomach	Positive	848	93	14	47	11
	Negative	13093	8740	4185	567	609
Liver	Positive	761	105	14	73	16
	Negative	12560	8167	3973	649	645
Perirenal fat	Positive	679	97	11	41	5
	Negative	10292	6773	3300	437	450
Intestine	Positive	943	103	13	50	8
	Negative	14325	9756	4597	598	634
Kidney	Positive	775	77	9	49	7
	Negative	11510	7803	3715	545	556
Spleen	Positive	1273	157	21	55	5
	Negative	17423	12052	5584	685	690
Thymus	Positive	1368	168	12	49	5
	Negative	17713	12444	5641	607	606
Lung	Positive	1276	153	23	52	5
	Negative	18580	12803	5822	625	656
Muscle	Positive	433	120	38	65	37
	Negative	7107	4560	2255	430	470
Heart	Positive	444	57	15	32	6
	Negative	7250	4745	2223	337	356
Blood	Positive	270	37	5	27	5
	Negative	4300	2841	1314	326	315

Table 5: Data summary of phosphorylation sites used in training depending on tissue and residue types in both PS1D and PS3D datasets.

the training data. The statistics of datasets depending on tissues and modification types is given in Table 5. In order to produce a valid machine learning model, we discarded the training sets having less than 20 positive samples. This also means that tyrosine phosphorylation sites were discarded from prediction models. Although the number threonine phosphorylation sites was not as sufficient as the number of serine phosphorylation sites, we still performed tissue-specific phosphorylation site prediction for threonine residues. However, note that the availability of more tissue-specific threonine phosphorylation sites would make the prediction model more robust.

4.1.3 *Feature extraction*

The extracted features were trained in 6 different categories: (i) sequence-based data (PS1D) with position-specific scoring matrix (PSSM) encoding, (ii) sequence-based data with binary encoding, (iii) sequence-based data with PSSM encoding plus functional features and predicted structural features, (iv) sequence-based data with binary encoding plus functional features and predicted structural features, (v) structure-based data (PS3D) with PSSM encoding plus functional features and real structural features, (vi) structure-based data with binary encoding plus functional features and real structural features (See Table 6).

4.1.3.1 *Sequence-based features*

Position-specific scoring matrix (PSSM): PSSM profile of a sequence shows the probability of each residue of the sequence at a specific position in the multiple sequence alignment. PSSM profiles were obtained from PSI-BLAST, using the default parameters and the filtered UniRef90 dataset from the UniProt Knowledgebase (UniProtKB) (UniProt, 2010). For each position of the peptide, PSSM scores of 20 different amino acids were considered. Therefore, PSSM encoding of each positive or negative site was represented in a feature vector with the dimension of 20 (amino acid type) x 21 (window size).

Binary encoding: In order to take into account the compositional characteristics of the amino acid sequences, we employed the simplest binary encoding algorithm. In this scheme we used 21 types of amino acids, which are given as ARNDCQEGHILKMFPSTWYVX.
Each amino acid is represented by a 21-dimensional binary vector that while A corresponds to (100000000000000000000), X corresponds to (000000000000000000001) and represents the artificial residues in the peptide sequences having less than 10 residues in the upstream or downstream region of the phosphorylated serine (S) / threonine (T) / tyrosine (Y) residues. In total, each positive or negative site is represented in a binary vector with the dimension of 21 (amino acid type) x 21 (window size).

K nearest neighbors (KNN) score: It has been known that close sequence

	seqPSSM	seqBinaryEnc	seqPSSM+predStr	seqBinaryEnc+predStr	seqPSSM+realStr	seqBinaryEnc+realStr
Position-specific scoring matrix	✓		✓		✓	
Binary encoding of linear amino acid content		✓		✓		✓
Binary encoding of spatial amino acid content					✓	✓
Spatial environment significance score					✓	✓
Accessibility score					✓	✓
Predicted accessibility score			✓	✓		
Secondary structure					✓	✓
Predicted secondary structure			✓	✓		
B-factor					✓	✓
Disordered/ordered region			✓	✓	✓	✓
SCOP class					✓	✓
Protein domain					✓	✓
Pathway			✓	✓	✓	✓
GO terms			✓	✓	✓	✓

Table 6: The summary of features used to train each prediction model.

clusters of phosphorylation sites plays important role in kinase and phosphatase targeting as substrates of same kinases share common sequences. In order to benefit from local sequence clusters of phosphorylated peptides, we followed the same approach defined in (Gao and Xu, 2010; Suo et al., 2012).

Briefly, (i) we calculated the distance between each query 21-mer to other 21-mers in both positive and negative sets. The distance function is defined as below:

$$\mathrm{Dist}(m_1, m_2) = 1 - \frac{\sum_{i=1}^{l} \mathrm{Sim}(m_1(i), m_2(i))}{l} \tag{1}$$

where l is the length of peptide sequence (l=21 in this study), and *Sim* is the amino acid similarity matrix based on a normalized amino acid substitution matrix defined as below:

$$\mathrm{Sim}(x, y) = \frac{M(x, y) - \min(M)}{\max(M) - \min(M)} \tag{2}$$

where x and y are two amino acids of the compared 21-mers, *M* is the substitution matrix (BLOSUM62 in this study), max(M) is the largest number in the matrix, and min(M) is the smallest number in the matrix.

(ii) After sorting distance scores of each query 21-mer, the percentage of k nearest neighbors from the positive set determines the KNN score of the corresponding 21-mer. Since sequences in the positive set are assumed to be similar to each other and distant from sequences in the negative set, positive sites are expected to have higher KNN scores, whereas negative sites are expected reversely.

Since the optimum k value depends on the ratio between positive set size and negative set size, we chose different k values for each prediction model (See Table 7). The details are given in Section 4.2.

		pS					pT				
		0.075%	0.1%	1%	2%	4%	0.25%	0.5%	1%	2%	4%
Blood	seqPSSM		✓						✓		
	seqBinaryEnc		✓						✓		
	seqPSSM+predStr		✓						✓		
	seqBinaryEnc+predStr		✓						✓		
	seqPSSM+realStr					✓				✓	✓
	seqBinaryEnc+realStr					✓				✓	✓
Brain	seqPSSM	✓	✓				✓	✓	✓		
	seqBinaryEnc	✓	✓				✓	✓	✓		
	seqPSSM+predStr	✓	✓				✓	✓	✓		
	seqBinaryEnc+predStr	✓	✓				✓	✓	✓		
	seqPSSM+realStr			✓	✓	✓				✓	✓
	seqBinaryEnc+realStr			✓	✓	✓				✓	✓
Brainstem	seqPSSM	✓	✓				✓	✓	✓		
	seqBinaryEnc	✓	✓				✓	✓	✓		
	seqPSSM+predStr	✓	✓				✓	✓	✓		
	seqBinaryEnc+predStr	✓	✓				✓	✓	✓		
	seqPSSM+realStr			✓	✓	✓				✓	✓
	seqBinaryEnc+realStr			✓	✓	✓				✓	✓
Cerebellum	seqPSSM	✓	✓				✓	✓	✓		
	seqBinaryEnc	✓	✓				✓	✓	✓		
	seqPSSM+predStr	✓	✓				✓	✓	✓		
	seqBinaryEnc+predStr	✓	✓				✓	✓	✓		
	seqPSSM+realStr			✓	✓	✓				✓	✓
	seqBinaryEnc+realStr			✓	✓	✓				✓	✓
Cortex	seqPSSM	✓	✓				✓	✓	✓		
	seqBinaryEnc	✓	✓				✓	✓	✓		
	seqPSSM+predStr	✓	✓				✓	✓	✓		
	seqBinaryEnc+predStr	✓	✓				✓	✓	✓		
	seqPSSM+realStr			✓	✓	✓				✓	✓
	seqBinaryEnc+realStr			✓	✓	✓				✓	✓
Heart	seqPSSM	✓	✓					✓	✓		
	seqBinaryEnc	✓	✓					✓	✓		
	seqPSSM+predStr	✓	✓					✓	✓		
	seqBinaryEnc+predStr	✓	✓					✓	✓		
	seqPSSM+realStr					✓					✓
	seqBinaryEnc+realStr					✓					✓
Intestine	seqPSSM	✓	✓				✓	✓	✓		
	seqBinaryEnc	✓	✓				✓	✓	✓		
	seqPSSM+predStr	✓	✓				✓	✓	✓		
	seqBinaryEnc+predStr	✓	✓				✓	✓	✓		
	seqPSSM+realStr			✓	✓	✓				✓	✓
	seqBinaryEnc+realStr			✓	✓	✓				✓	✓
Kidney	seqPSSM	✓	✓				✓	✓	✓		
	seqBinaryEnc	✓	✓				✓	✓	✓		
	seqPSSM+predStr	✓	✓				✓	✓	✓		
	seqBinaryEnc+predStr	✓	✓				✓	✓	✓		
	seqPSSM+realStr			✓	✓	✓				✓	✓
	seqBinaryEnc+realStr			✓	✓	✓				✓	✓
Liver	seqPSSM	✓	✓				✓	✓	✓		
	seqBinaryEnc	✓	✓				✓	✓	✓		
	seqPSSM+predStr	✓	✓				✓	✓	✓		
	seqBinaryEnc+predStr	✓	✓				✓	✓	✓		
	seqPSSM+realStr			✓	✓	✓				✓	✓
	seqBinaryEnc+realStr			✓	✓	✓				✓	✓
Lung	seqPSSM	✓	✓				✓	✓	✓		
	seqBinaryEnc	✓	✓				✓	✓	✓		
	seqPSSM+predStr	✓	✓				✓	✓	✓		
	seqBinaryEnc+predStr	✓	✓				✓	✓	✓		
	seqPSSM+realStr			✓	✓	✓				✓	✓

	seqBinaryEnc+realStr			✓	✓	✓				✓	✓
Muscle	seqPSSM	✓	✓					✓	✓		
	seqBinaryEnc	✓	✓					✓	✓		
	seqPSSM+predStr	✓	✓					✓	✓		
	seqBinaryEnc+predStr	✓	✓					✓	✓		
	seqPSSM+realStr					✓					✓
	seqBinaryEnc+realStr					✓					✓
Pancreas	seqPSSM		✓					✓	✓		
	seqBinaryEnc		✓					✓	✓		
	seqPSSM+predStr		✓					✓	✓		
	seqBinaryEnc+predStr		✓					✓	✓		
	seqPSSM+realStr					✓					✓
	seqBinaryEnc+realStr					✓					✓
Perirenal fat	seqPSSM	✓	✓				✓	✓	✓		
	seqBinaryEnc	✓	✓				✓	✓	✓		
	seqPSSM+predStr	✓	✓				✓	✓	✓		
	seqBinaryEnc+predStr	✓	✓				✓	✓	✓		
	seqPSSM+realStr			✓	✓	✓				✓	✓
	seqBinaryEnc+realStr			✓	✓	✓				✓	✓
Spleen	seqPSSM	✓	✓				✓	✓	✓		
	seqBinaryEnc	✓	✓				✓	✓	✓		
	seqPSSM+predStr	✓	✓				✓	✓	✓		
	seqBinaryEnc+predStr	✓	✓				✓	✓	✓		
	seqPSSM+realStr			✓	✓	✓				✓	✓
	seqBinaryEnc+realStr			✓	✓	✓				✓	✓
Stomach	seqPSSM	✓	✓				✓	✓	✓		
	seqBinaryEnc	✓	✓				✓	✓	✓		
	seqPSSM+predStr	✓	✓				✓	✓	✓		
	seqBinaryEnc+predStr	✓	✓				✓	✓	✓		
	seqPSSM+realStr			✓	✓	✓				✓	✓
	seqBinaryEnc+realStr			✓	✓	✓				✓	✓
Testis	seqPSSM	✓	✓				✓	✓	✓		
	seqBinaryEnc	✓	✓				✓	✓	✓		
	seqPSSM+predStr	✓	✓				✓	✓	✓		
	seqBinaryEnc+predStr	✓	✓				✓	✓	✓		
	seqPSSM+realStr			✓	✓	✓				✓	✓
	seqBinaryEnc+realStr			✓	✓	✓				✓	✓
Thymus	seqPSSM	✓	✓				✓	✓	✓		
	seqBinaryEnc	✓	✓				✓	✓	✓		
	seqPSSM+predStr	✓	✓				✓	✓	✓		
	seqBinaryEnc+predStr	✓	✓				✓	✓	✓		
	seqPSSM+realStr			✓	✓	✓				✓	✓
	seqBinaryEnc+realStr			✓	✓	✓				✓	✓

Table 7: The summary of k values chosen for each prediction model in different tissues.

4.1.3.2 *Structure-based features*

Spatial environment encoding: As we have mentioned in Chapter 3, spatial surroundings of phosphorylation sites follow certain specific patterns. To incorporate these patterns of spatial environments and benefit from them, we calculated the occurrence of 20 different amino acid types within the radial distances of 2 to 12 Å for both positive and negative sites. Then, each site was represented in a vector with the dimension of 10 (radial distance size) x 20 (amino acid type).

Spatial environment probability score: Similar to previous approach, we also represented the contribution of spatial surroundings with a probability score. First, we calculated the significance (p-value) of each amino acid type at a specific position (ranging from 2 to 12 Å in increments of 1 Å) using Fisher exact test. Then, we multiplied the p-values of all amino acids in the <12 Å radial distance of each site, which gives the spatial environment probability score of each positive or negative site.

Secondary structure: Secondary structure assignments for all positive and negative sites along with the residues surrounding them in the PS3D dataset were obtained from the DSSP database (Joosten et al., 2011). The obtained structures were categorized into 3: alpha, beta and loop.

We also utilized PSIPRED (McGuffin, Bryson, and Jones, 2000) to find predicted secondary structures for all sites in the PS1D dataset. The default parameters with the filtered UniRef90 dataset from the UniProt Knowledgebase (UniProtKB) were used (UniProt, 2010). The obtained structures were again categorized as alpha, beta and loop.

Solvent accessibility: The solvent accessibility of each positive and negative site along with their sequence surroundings in the PS3D dataset was calculated using NACCESS (Hubbard and Thornton, 1993).

The accessibility scores for phosphorylation and non-phosphorylation sites in the PS1D dataset were predicted using SPPIDER (Porollo and Meller, 2007).

B-factor: We obtained the B-factors of each site along with sequence surroundings in the PS3D dataset from Protein Data Bank.

Disordered region: We used DisEMBL (Linding et al., 2003) to predict disordered/unstructured regions within protein sequences. Each positive and negative site along with sequence surroundings in both PS1D and PS3D datasets were considered to reside in a disordered region if it was predicted by DisEMBL to be located in a region associated with either loops/coils, or hot loops, or missing coordinates. A positive or negative site was eventually represented with a 1-digit categorical feature that it can either be in a disordered, or an ordered region.

Structural folds and functional domains: We incorporated the structural folds and domains of each protein in the PS3D dataset from the SCOP database using the *class* and *protein domain* levels (Murzin et al., 1995). We represented structural folds in a categorical variable where it can be all alpha proteins, all beta proteins, alpha and beta proteins (a+b), alpha and beta proteins (a/b), multi-domain proteins (alpha and beta), small proteins, coiled coil proteins, and membrane and cell surface proteins and peptides.

4.1.3.3 *Functional features*

Gene ontology (GO) terms: GO terms of each protein in both PS1D and PS3D datasets were gathered from the downloadable rat-filtered dataset of Gene Ontology Consortium (Gene Ontology, 2015) as of May 26, 2015. We used 3-digit binary encoding where each digit represents the association of each site to biological processes, cellular components and molecular functions.

Pathways: We obtained the pathways each protein takes place from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa et al., 2006). KEGG pathways of phosphorylation and non-phosphorylation proteins from both PS1D and PS3D datasets were obtained.

4.1.4 *Machine learning*

Tissue-based phosphorylation site prediction was treated as a binary classification problem where each phosphorylated residue type can be classified as phosphorylated in a particular tissue or not phosphorylated at all. We used random forest (RF) machine learning technique (Breiman, 2001) implemented in the R (Team, 2009) Caret package (Kuhn, 2008), since there are many advantages of this method. RF ensembles the *ntree* decision trees, and trains each tree with a subset of training samples (sampling with replacement) that guarantees to use many of samples in the training data for each tree. Random forest gives an output value between 0 and 1 that is the fraction of decision trees voting for phosphorylation in a particular tissue. The decisions are not sensitive to outliers, and it does feature selection by building trees on better performing features.

Since random forest allows using a max of 32 categories for a feature, we separated features (domain and pathway features) having more than 32 different values into 2 categorical variables where each categorical variable may have a value between 0 and 31. Applying this approach, 1024 (32 x 32) different values can be represented.

We generated 2 (residue type) x 17 (tissue type) x 6 (feature category) prediction models, and each model was trained on 300 trees with a 10-fold cross-validation by splitting the data into 10 parts and using 9 parts for training and the remaining part for testing. As seen in Table 4, the number of negative samples surpasses that of positive samples. To overcome this class imbalance problem, we performed undersampling for the majority class with the ratio of 1:1 that positive and negative sample sizes are equal to each other when growing each tree. With this approach we guaranteed to use as many negative samples as possible in training, as the number of trees (300 in this study) used in random forest gets larger. The *mtry* value, which is the number of variables randomly sampled as candidates at each split, was tuned based on the best Area Under the Receiver-Operating Characteristic Curve (AUC) measure, and the best value of *mtry* was used to train each model.

4.1.5 *Prediction performance assessment*

In order to evaluate the predictions, we used Accuracy (ACC), Sensitivity (SEN), Specificity (SPE), and Area Under the Receiver-Operating Characteristic Curve (AUC) measures. AUC is the area under the Receiver-Operating Characteristic Curve plot where axes of the plot are true positive rate (SEN) and false positive rate (FPR). For each classifier, we adjusted the optimal cut-off threshold for class probabilities (the best of both specificity and sensitivity values), rather than choosing the conventional 50% cut-off. The overall qualities of the classifiers were mostly compared based on AUC values. The evaluation measures were formulized below:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$SEN = \frac{TP}{TP + FN} \tag{4}$$

$$SPE = \frac{TN}{TN + FP} \tag{5}$$

$$FPR = \frac{FP}{TN + FP} \tag{6}$$

TP, TN, FP and FN correspond to the numbers of true positives, true negatives, false positives and false negatives, respectively.

4.1.6 *Comparing with existing tools*

As we first implemented tissue-specific phosphorylation site prediction, there is no other tool performing prediction in a tissue-specific manner. We, therefore, compared the our performance to predictors developed for globally phosphorylated site prediction. Musite was the only tool that was still available, easily downloadable and used by the time we conducted this work ((Gao and Xu, 2010; Gao et al., 2010)). *M. musculus* pre-trained model with default parameters was used since no pre-trained model for rat is available. When calculating the sensitivity score for Musite, we only counted phosphorylated serines in a particular tissue as positives. The reason is that in our training sets we only included phosphorylated serine residues in the corresponding tissue as positive instances. Therefore, true positive rate, or sensitivity, was calculated for Musite results by dividing the number of predicted phosphorylated serine residues to the number of all phosphorylated residues in the corresponding tissue – the other predicted phosphorylated serine residues were just ignored.

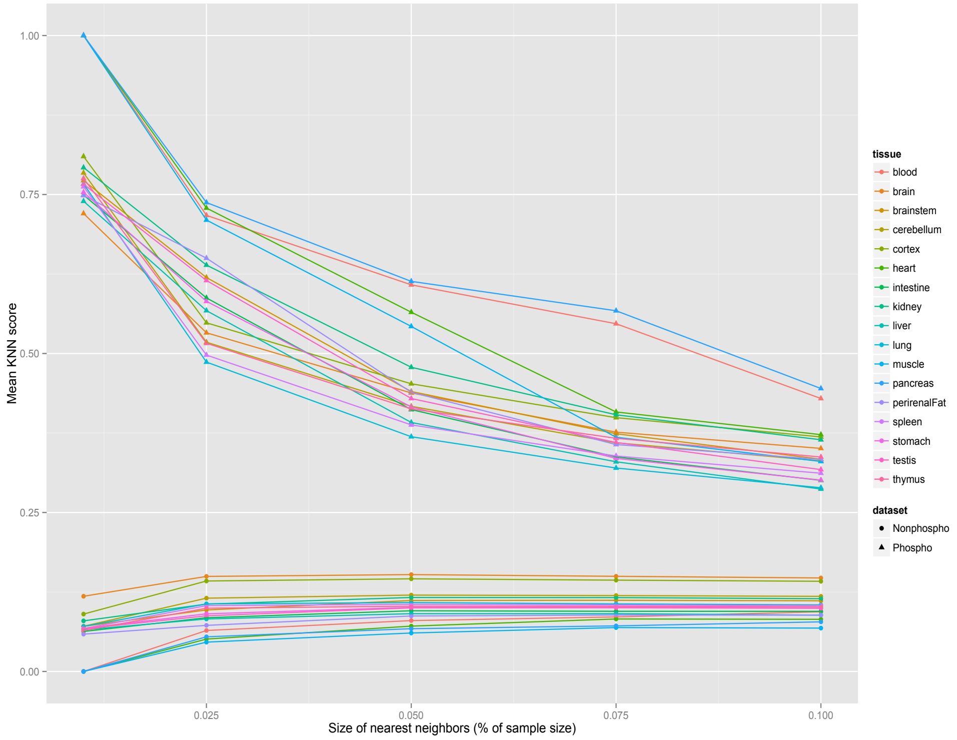


Figure 16: Comparison of KNN scores between serine phosphorylation and non-phosphorylation sites depending on tissues in the PS1D data.

4.2 RESULTS AND DISCUSSIONS

4.2.1 *Predictive performance of sequence environments surrounding phosphorylation sites*

In Chapter 3 we have shown that phosphorylation sites display tissue-specific preferences for certain residues in their linear amino acid sequence environment. For instance, only serine phosphorylation sites in perirenal fat have phenylalanine (F) residues at position +1, whereas only serine phosphorylation sites in pancreas and testis have a preference for glutamine (Q) residues at position -2. Histidine residues are only enriched at position +6 in blood, whereas phosphorylated serine residues in stomach and liver show preferences for asparagine (N) residues at position +3. All these results derived the significance of sequence environment surrounding phosphorylation sites used in tissue-specific phosphorylation prediction. The sequence environment content was represented as a feature in two ways: (i) PSSM profile encoding, which shows the probability of each residue of the query sequence at a specific position in the multiple sequence alignment, and (ii) Binary encoding, which resembles the amino acid content and position of residues surrounding phosphorylation sites and non-phosphorylation sites.

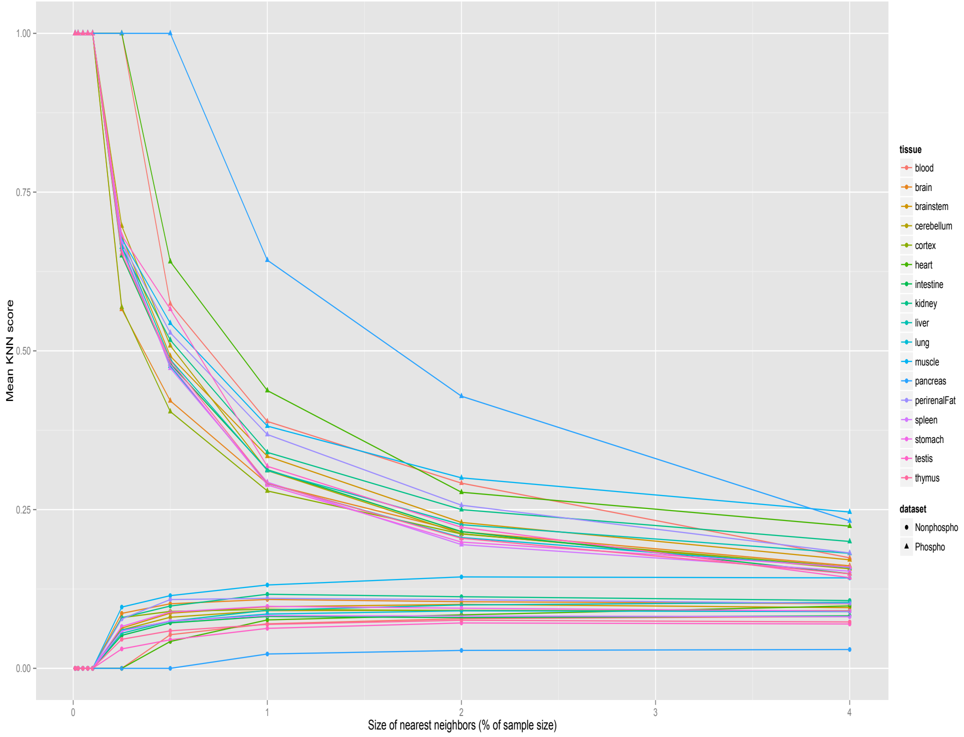


Figure 17: Comparison of KNN scores between serine phosphorylation and non-phosphorylation sites depending on tissues in the PS3D data.

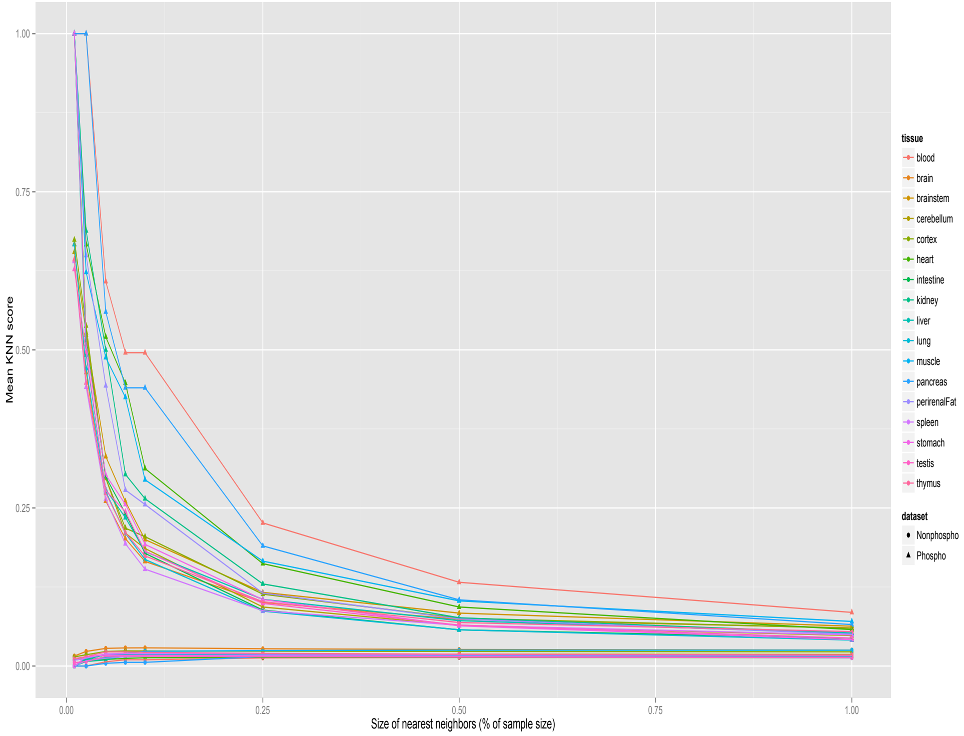


Figure 18: Comparison of KNN scores between threonine phosphorylation and non-phosphorylation sites depending on tissues in the PS1D data.

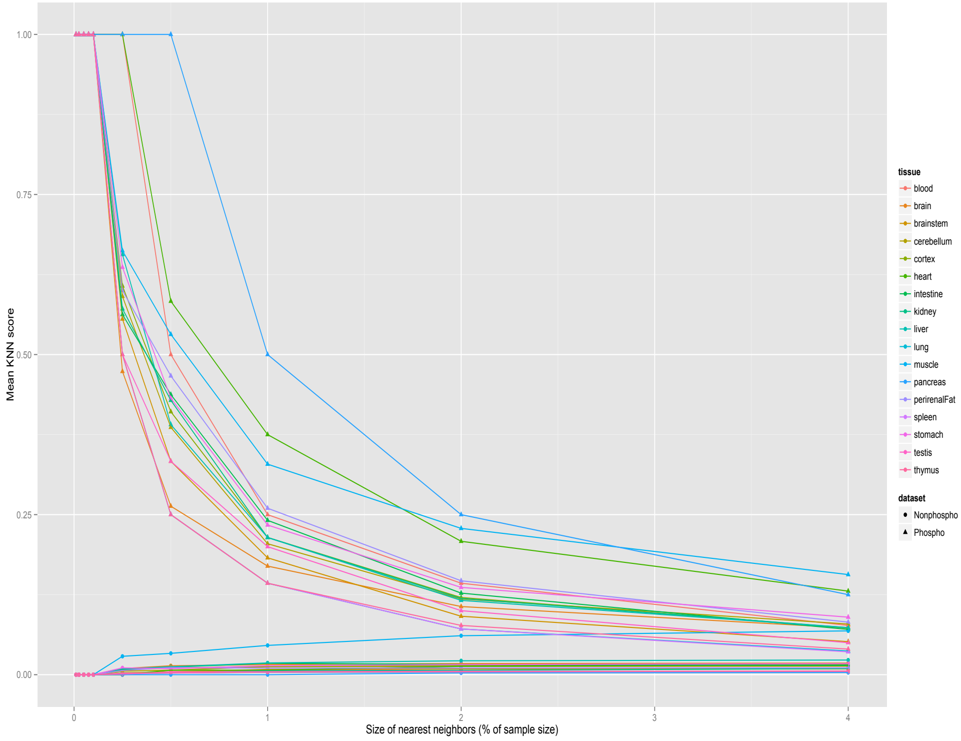


Figure 19: Comparison of KNN scores between threonine phosphorylation and non-phosphorylation sites depending on tissues in the PS3D data.

4.2.2 *KNN scores as features*

A KNN score is an indicator for similarity of local sequence surrounding a query site to local sequences surrounding sites in positive set and negative set. The higher score means that the query site is more similar to the positive set, whereas a lower score shows that it is more similar to the negative set. When the sizes of positive and negative sets are equal, the threshold can be set to 0.5, meaning that scores greater than 0.5 is more similar to the positive set and vice versa. However, we used all the negative samples in this study and handled the class imbalance problem by setting positive and negative sample sizes equal to each other while growing each tree of random forest. Therefore, as seen in Figure 16 - Figure 19, positive sites may also have scores smaller than 0.5, and the distributions are slightly different than the ones in previous studies (Gao and Xu, 2010; Suo et al., 2012).

The *k* nearest neighbors is set as the percentage of the size of the training set. Due to the class imbalance problem between positive and negative sets, this *k* value is highly sensitive to positive to negative ratio of each training set, and the size of dataset. This means that for a model, if the *k* value is too small, then all positive samples in the model might get a KNN score of 1 or very close to 1, and negative samples might get a KNN score of 0 or very close to 0 where the classifier using that *k* value as a parameter would give a very high accuracy even though this is not the case. On the other hand, if the *k* value is too large, then positive and negative samples might get similar KNN scores where those KNN scores used as feature in a predictor might not have any distinguishing power. Therefore, we chose the optimal k value/values for each model depending on tissue, modification type and used features Table 7.

Even though we set a very stringent threshold for sequence redundancy elimination and it is assumed that similarities found by KNN are not as a result of protein homology if the training set is non-redundant (Suo et al., 2012), we still observed that *k* value might cause bias on classification performance when the optimal *k* value is not chosen. Therefore, we trained each of 6 categories mentioned in Section 4.1 with and without KNN scores to better understand the contribution of KNN score, and to minimize the bias of protein homology on classification performance.

4.2.3 *Influence of spatial amino acid content and structural environment of phosphorylation sites on prediction*

We have shown in Chapter 3 that the amino acid composition in 3D varies in a tissue-dependent manner, which implies that both sequence and spatial environments of phosphorylation sites may play a role in determining the substrate-specificity of kinases and phosphatases across tissues. For instance, PSS in brain have a strong preference for arginine residues, and it is much more predominant than the enrichment in the 1D environment. Histidine and cysteine residues are strongly depleted in the 3D environment of PSS in kidney, whereas no preference for these residues is observed in the 1D environment. Therefore, we encoded spatial environment surrounding phosphorylation sites in an effort to increase the accuracy of tissue-specific phosphorylation site prediction. We also utilized secondary structure and accessibility scores of phosphorylation sites and the residues surrounding them obtained by mapping phosphorylated proteins on experimental protein structures. The dataset size (PS3D) is rather small in comparison to the dataset size obtained from sequence environment (PS1D). However, the performance of prediction using spatial content and experimental structures is quite comparable. To our knowledge this is the first study using experimental structures in the prediction of phosphorylation sites. In addition, we used predicted secondary structures, predicted disordered regions, and predicted accessibility scores in order to obtain a larger training dataset, and yield a better performance.

4.2.4 *The contribution of functional annotations on phosphorylation prediction*

Functional features have been incorporated into the prediction of posttranslational modification sites in previous studies, and shown that functional features are valuable features leading to a higher prediction performance (Fan et al., 2014; Li et al., 2014). In Chapter 3, we

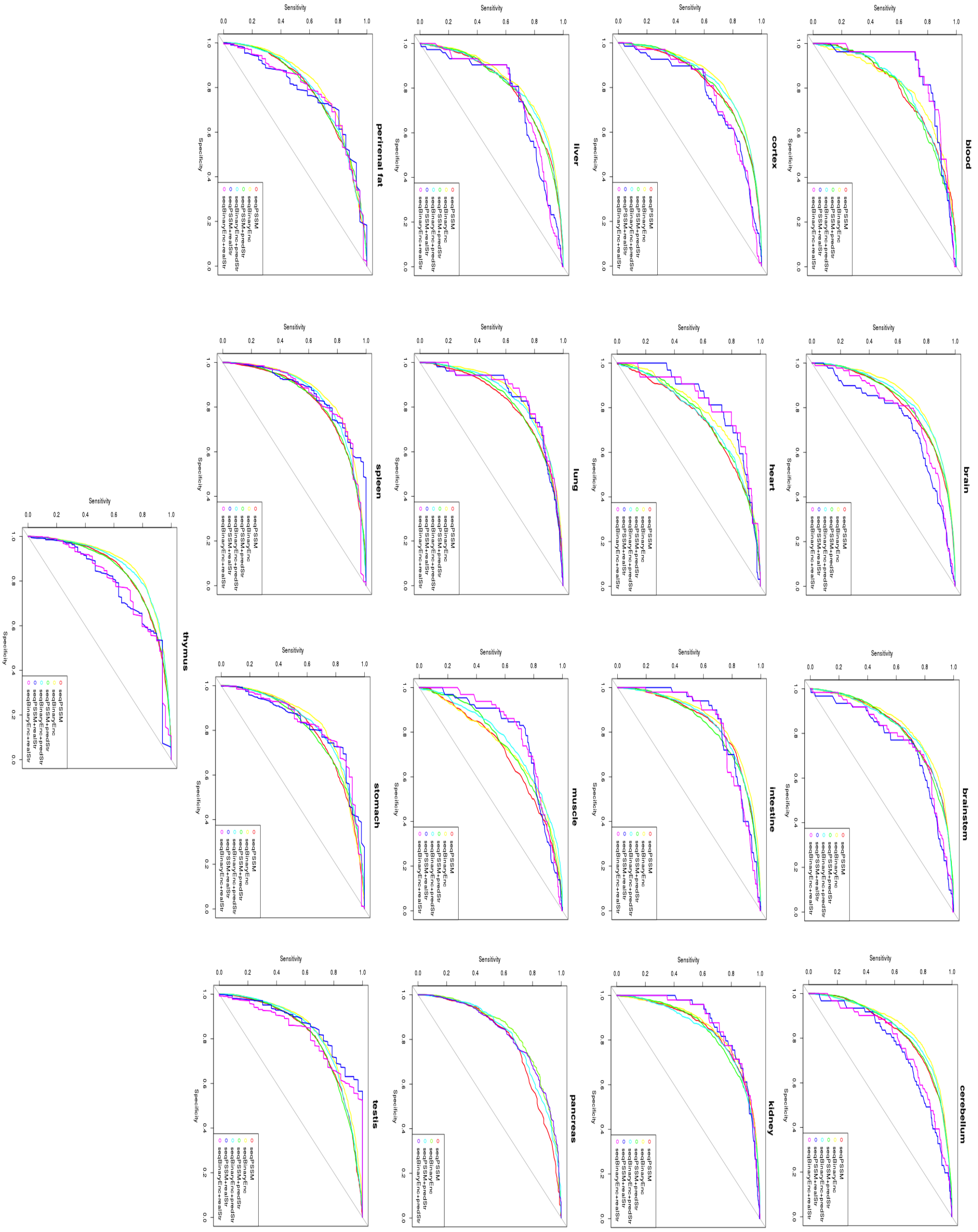


Figure 20: ROC curves of serine phosphorylation prediction models obtained from 10-fold cross-validation in different tissues without using the KNN score.

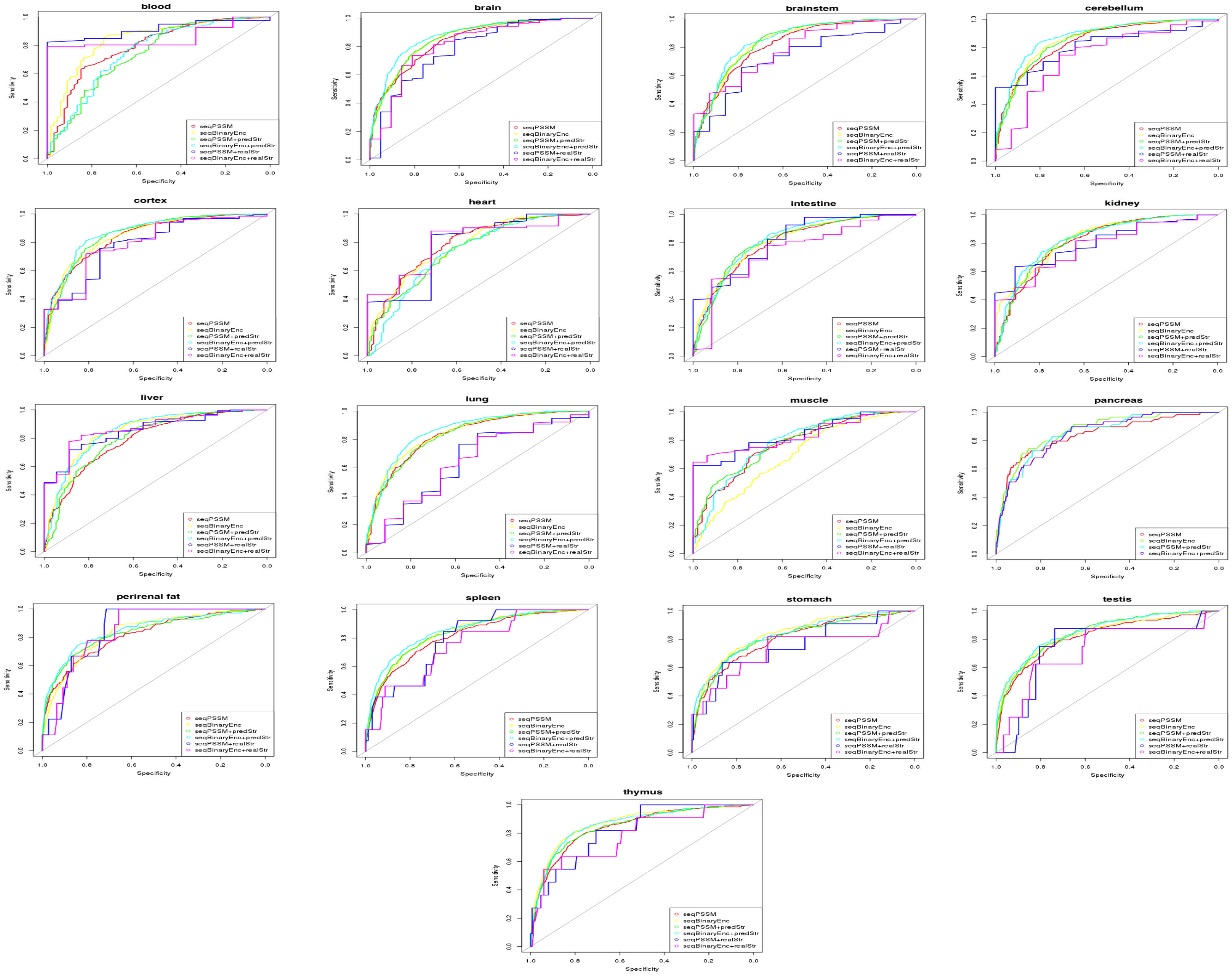


Figure 21: ROC curves of serine phosphorylation prediction models obtained from independent testing in different tissues without using the KNN score.

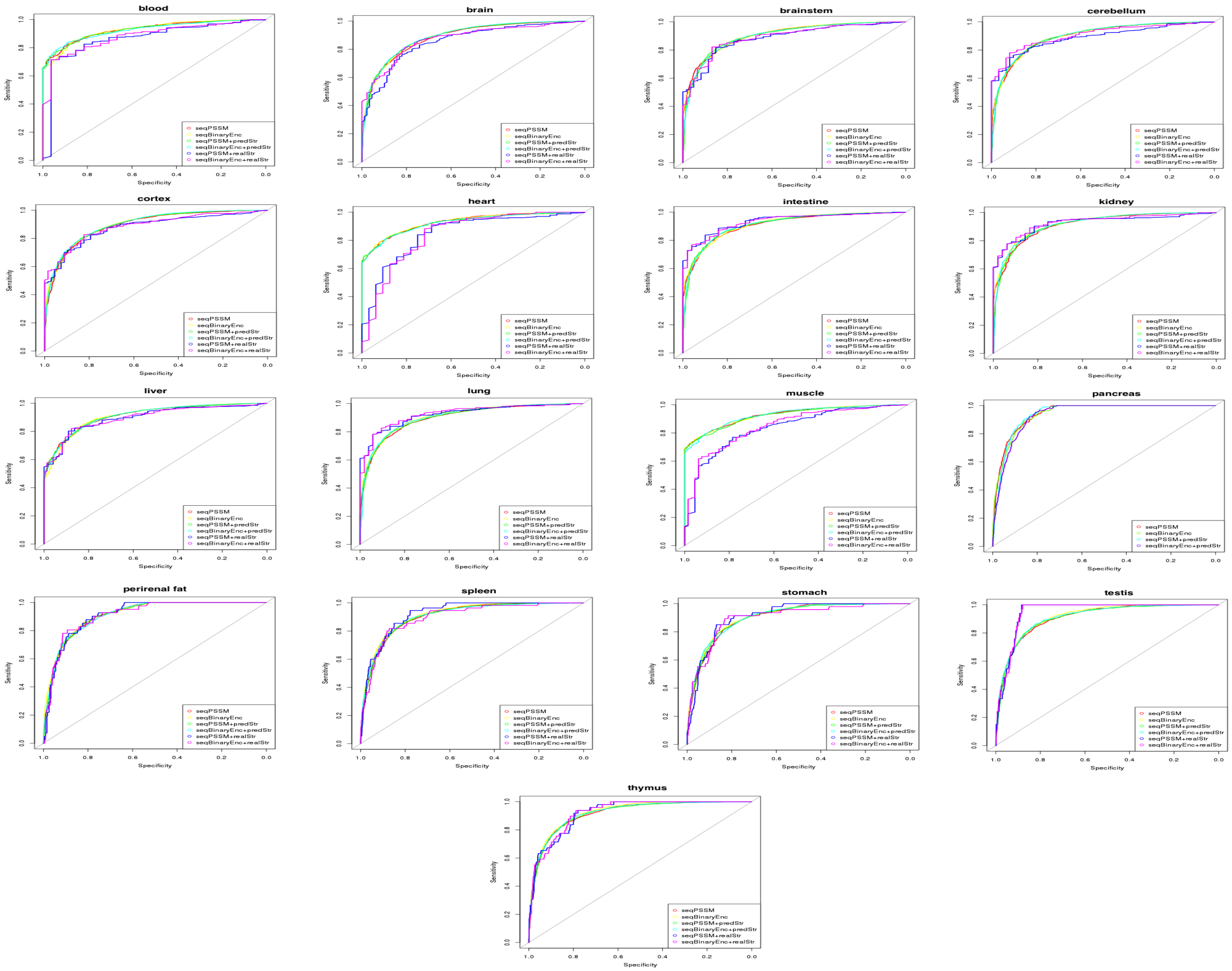


Figure 22: ROC curves of serine phosphorylation prediction models obtained from 10-fold cross validation in different tissues with using the KNN score.

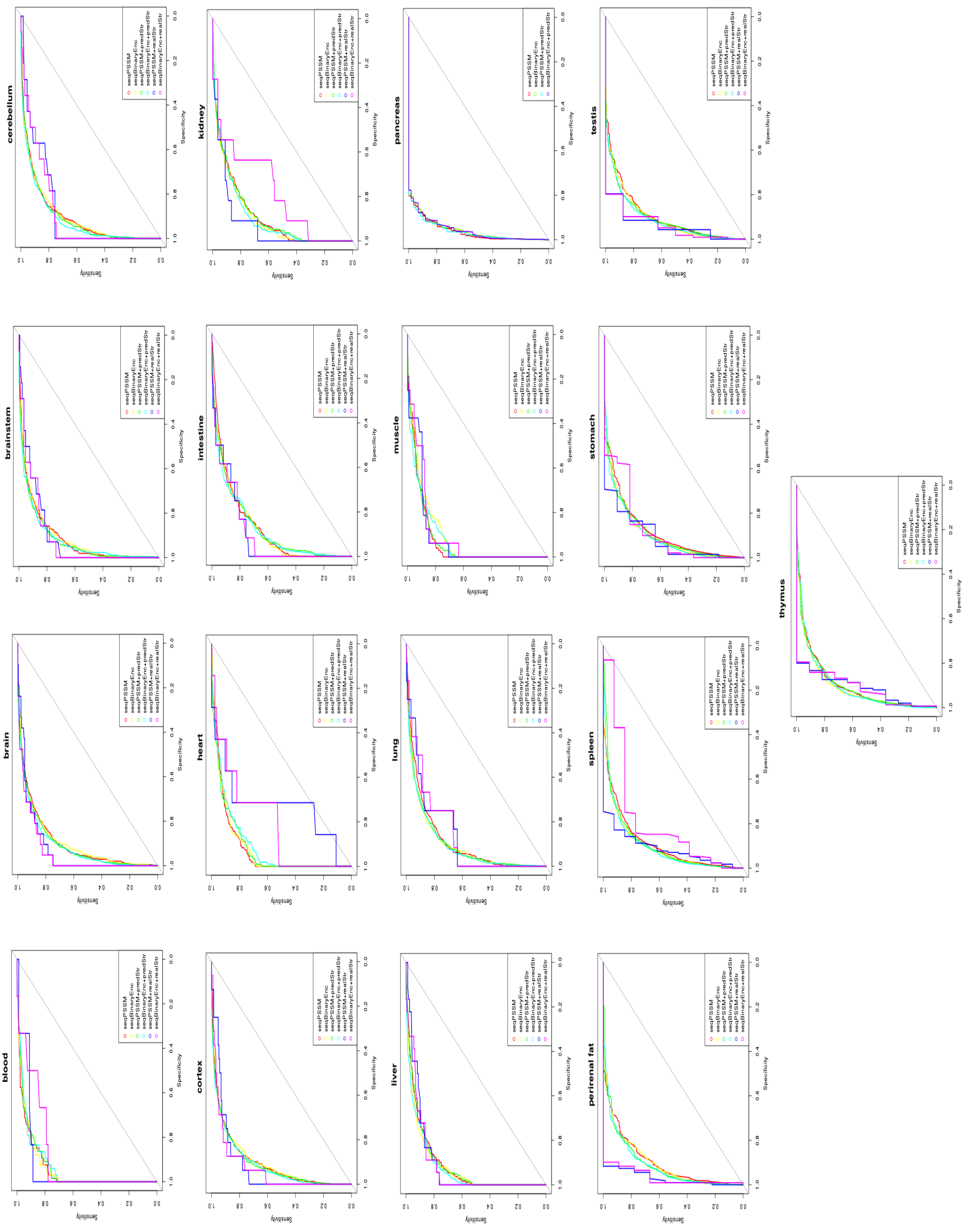


Figure 23: ROC curves of serine phosphorylation prediction models obtained from independent testing in different tissues with using the KNN score.

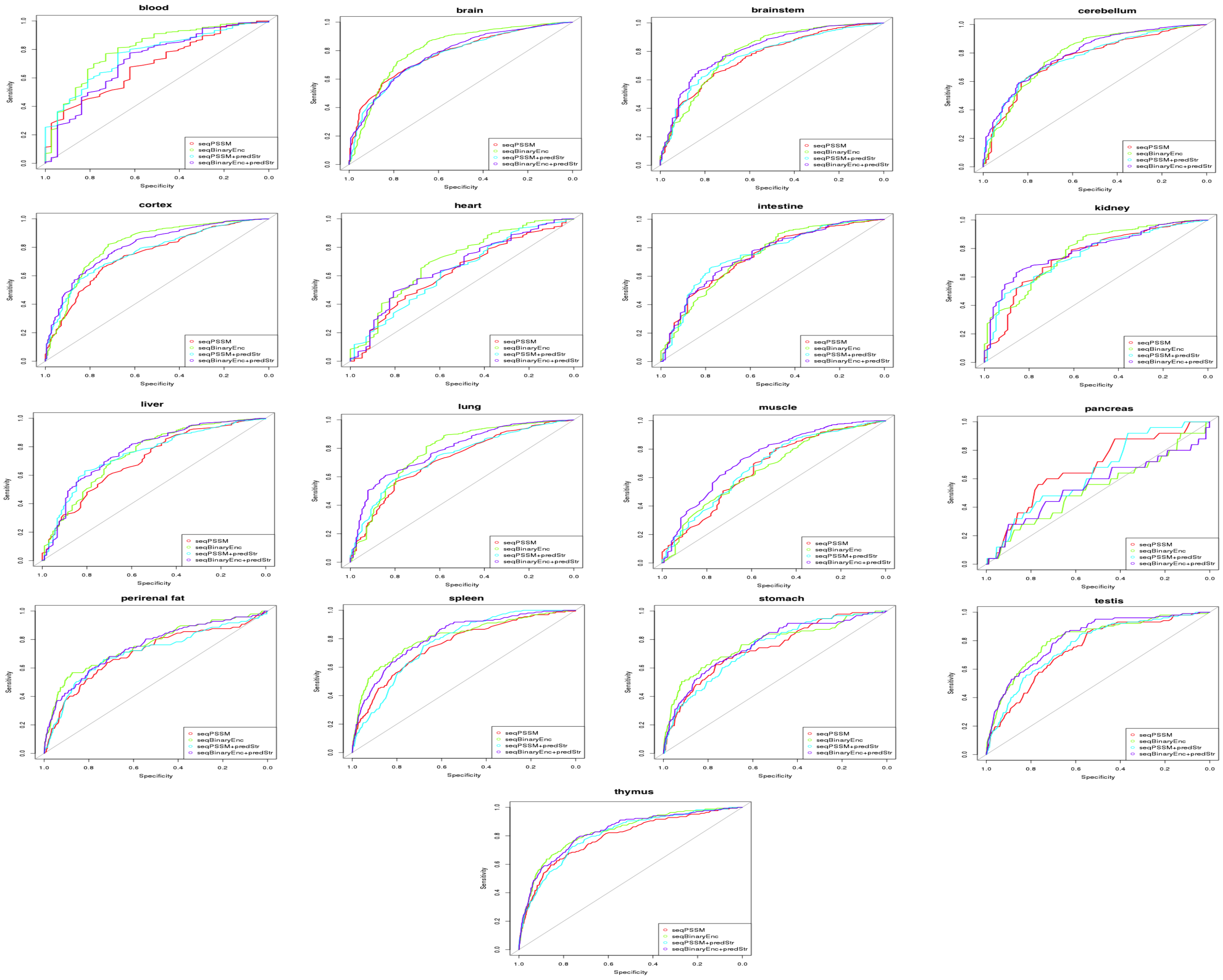


Figure 24: ROC curves of threonine phosphorylation prediction models obtained from 10-fold cross-validation in different tissues without using the KNN score.

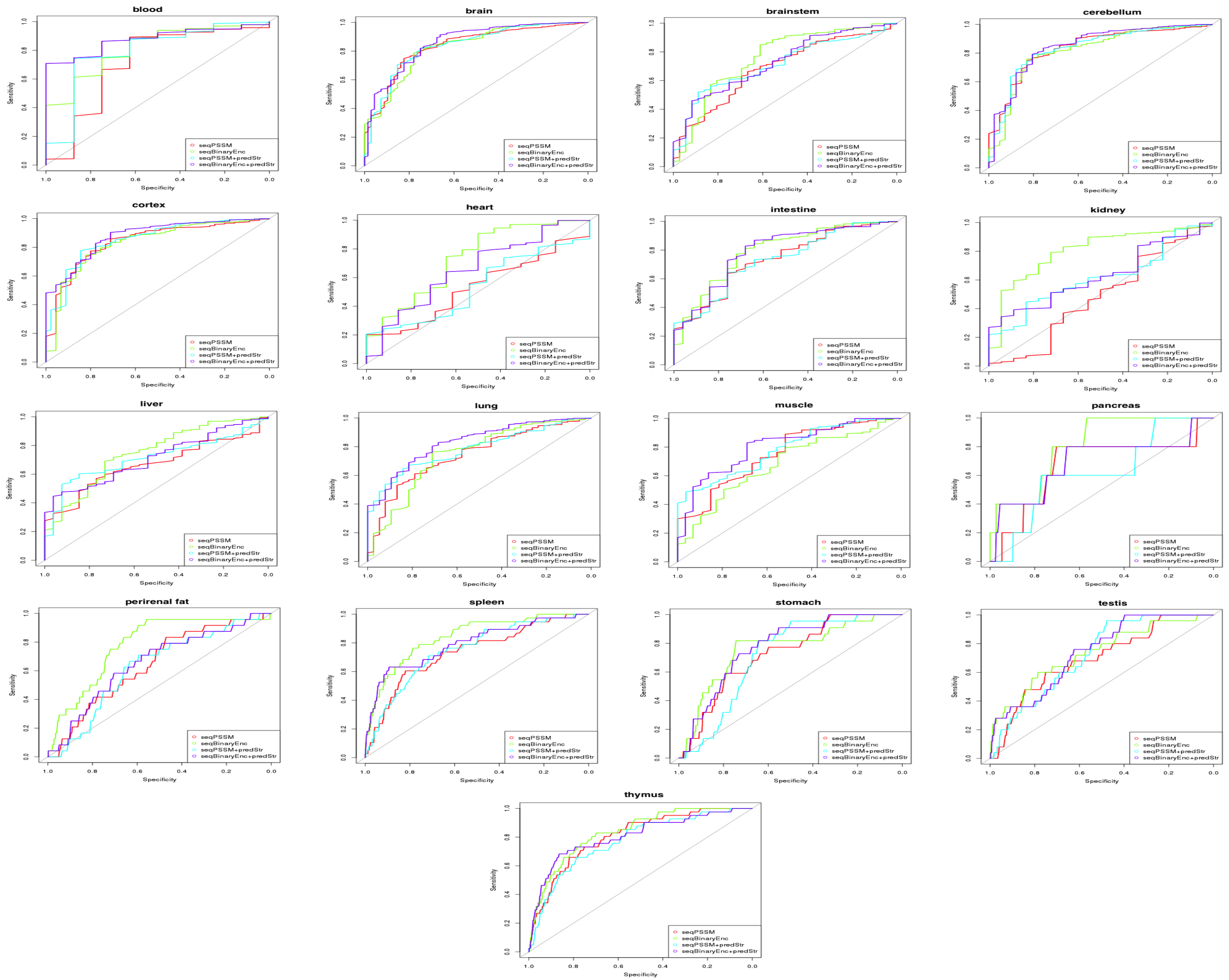


Figure 25: ROC curves of threonine phosphorylation prediction models obtained from independent testing in different tissues without using the KNN score.

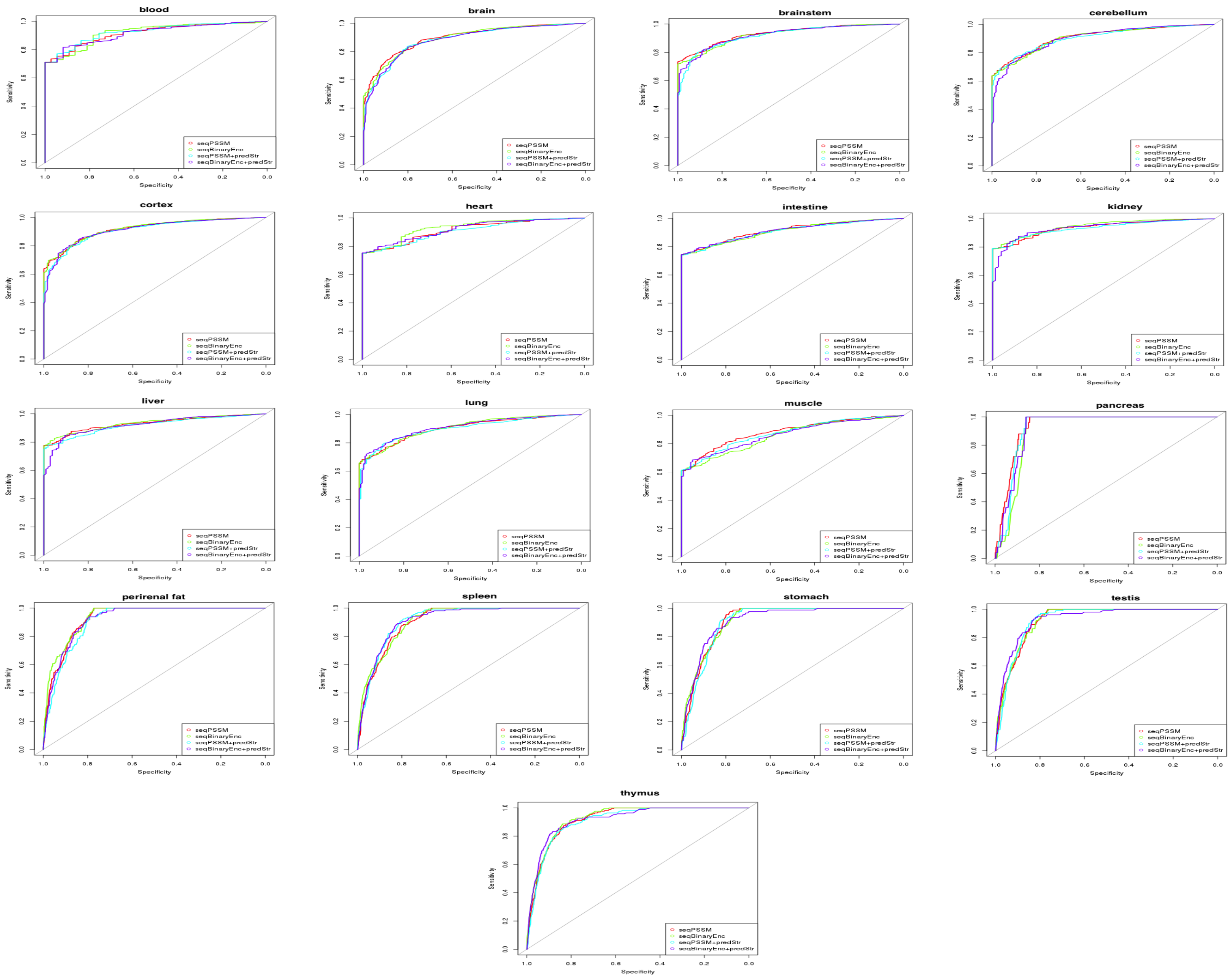


Figure 26: ROC curves of threonine phosphorylation prediction models obtained from 10-fold cross-validation in different tissues with using the KNN score.

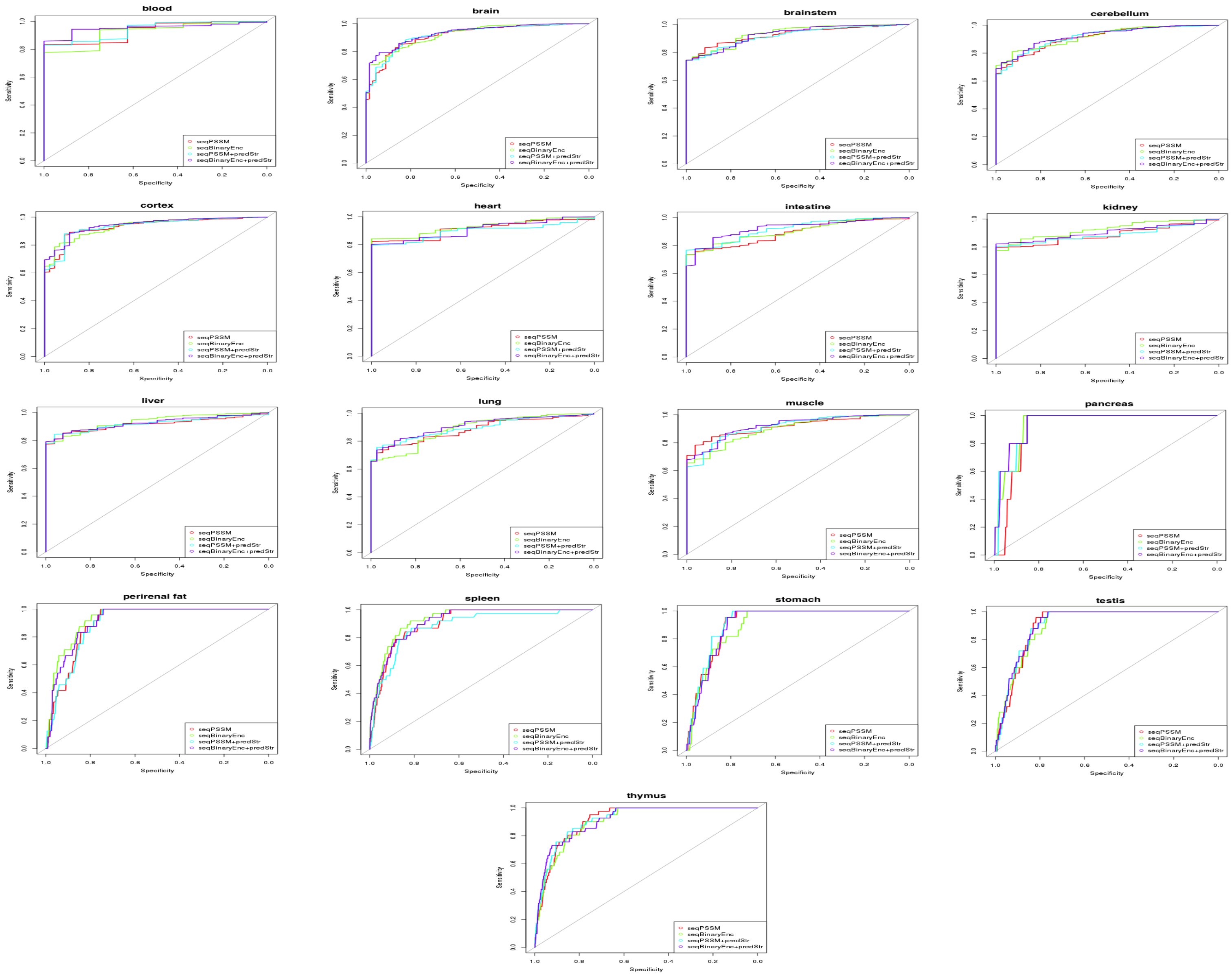


Figure 27: ROC curves of threonine phosphorylation prediction models obtained from independent testing in different tissues with using the KNN score.

have also observed that biological processes are regulated by phosphorylation in a tissue dependent manner. As a result, we utilized GO annotation and KEGG pathways to assess the contribution of functional annotations on tissue-specific phosphorylation site prediction. In this study, functional annotation did not contribute to prediction as significant as reported by previous studies. The possible reason may be the fact that we generated the negative set of non-phosphorylation sites from phosphorylated proteins. Since the entire rat proteome was not use as the negative set, the decision about the influence of functional annotation on tissue-specific phosphorylation prediction might not be accurate.

		pS				pT			
		AUC(%)	ACC(%)	SEN(%)	SPE(%)	AUC(%)	ACC(%)	SEN(%)	SPE(%)
Blood	seqPSSM	93.07/79.25	87.98/73.62	90.98/76.3	84.98/70.95	92.28/70.88	92.9/69.95	100/75.83	85.8/64.07
	seqBinaryEnc	92.98/80.46	87.48/76.5	90/76.3	84.95/76.7	93.18/81.51	92.75/79.83	97.5/81.67	87.99/77.99
	seqPSSM+predStr	93/79.74	87.55/75.9	90.26/79.39	84.84/72.42	93.05/75.95	93.31/77.23	100/82.5	86.61/71.96
	seqBinaryEnc+predStr	92.82/81.41	88.26/77.32	92.22/78.15	84.3/76.49	92.15/76.13	92.05/72.94	100/70	84.09/75.89
	seqPSSM+realStr	87/88.02	89.88/90.5	95/100	84.75/81	X	X	X	X
	seqBinaryEnc+realStr	89.23/88	91.38/90.76	100/100	82.77/81.52	X	X	X	X
Brain	seqPSSM	88.48/82.2	80.53/74.92	80.62/73.98	80.45/75.86	89.85/77.63	84.04/72.37	85.49/75.29	82.59/69.44
	seqBinaryEnc	88.88/85.4	81.23/78.16	80.68/78.99	81.78/77.34	89.5/81.03	83.24/77.82	85.16/76.85	81.32/78.78
	seqPSSM+predStr	88.92/82.92	81.26/75.36	81.11/74.42	81.42/76.31	88.9/76.87	83.12/73.38	82.59/77.88	83.65/68.88
	seqBinaryEnc+predStr	89.13/84.23	81.57/77.27	81.5/76.85	81.63/77.7	88.8/78.63	82.62/74	82.28/71.79	82.96/76.22
	seqPSSM+realStr	86.97/74.81	83.13/75.87	87.64/76.39	78.61/75.36	X	X	X	X
	seqBinaryEnc+realStr	87.93/78.49	84.49/78.19	85.42/78.61	83.56/77.78	X	X	X	X
Brainstem	seqPSSM	90.48/82.23	82.7175.31	83.02/75	82.4/75.62	93.24/76.23	87.76/73.33	89.14/75.62	86.38/71.04
	seqBinaryEnc	90.26/83.69	82.73/77.33	83.08/76.24	82.37/78.42	92.53/79.36	86.29/76.17	89.1/76.1	83.48/76.24
	seqPSSM+predStr	90.42/82.62	83.85/75.96	85.11/74.91	82.59/77.02	92.15/77.25	88.2/75.25	91.14/79.71	85.25/70.78
	seqBinaryEnc+predStr	90.18/82.73	83.51/77.18	83.36/75.18	83.66/79.18	92.39/80.97	87.29/77.58	87.76/80.14	86.82/75.02
	seqPSSM+realStr	89.71/77.09	87.52/76.96	91.67/76.9	83.37/77.01	X	X	X	X
	seqBinaryEnc+realStr	90.46/78.6	89.26/78.49	91.67/78.57	86.85/78.4	X	X	X	X
Cerebellum	seqPSSM	89.71/82.8	82.71/76.02	83.73/75.8	81.68/76.24	91.23/75.75	86.84/73.86	90.26/75.07	83.42/72.65
	seqBinaryEnc	89.77/84.94	82.16/78.58	81.9/77.86	82.42/79.3	90.76/78.74	84.26/75.88	85.4/73.31	83.12/78.46
	seqPSSM+predStr	89.94/83.4	82.93/76.46	82.65/75.95	83.21/76.96	91.11/77.33	85.32/73.97	86.65/76.32	83.99/71.61
	seqBinaryEnc+predStr	89.94/83.83	82.59/77.49	83.22/76.92	81.96/78.06	90.33/79.59	84.56/74.11	86.03/73.2	83.1/75.02
	seqPSSM+realStr	89.32/76.54	86.74/74.2	91.67/75.48	81.82/72.92	X	X	X	X
	seqBinaryEnc+realStr	90.85/78.6	87.78/78.29	91.67/78.81	83.9/77.77	X	X	X	X
Cortex	seqPSSM	89.32/82.16	81.96/75.36	82.81/74	81.11/76.73	92.04/74.34	85.99/71.82	86.96/72.92	85.03/70.71
	seqBinaryEnc	89.88/85.14	82.18/77.96	82.88/79.24	81.47/76.67	92.2/81.7	87.16/78.12	89.71/75.29	84.61/80.96
	seqPSSM+predStr	89.82/82.79	82.35/75.22	84.41/76.89	80.28/73.54	91.13/76.7	85.16/72.78	87.49/73.33	82.83/72.22
	seqBinaryEnc+predStr	90.09/84.15	82.12/77.21	82.58/76.36	81.66/78.06	91.6/81.29	86.28/76.5	88.71/77	83.85/76
	seqPSSM+realStr	87.88/78.21	85.14/75.74	86.67/80.95	83.61/70.52	X	X	X	X
	seqBinaryEnc+realStr	89.13/79.41	86.78/78.02	83.33/79.52	85.22/76.51	X	X	X	X
Heart	seqPSSM	92.98/74.71	86.22/70.67	89.18/71.53	83.25/69.81	91.69/69.14	88.99/65.02	95/66	82.98/64.03
	seqBinaryEnc	93.2/78.62	86.39/73.64	88.06/71.66	84.72/75.61	93.63/70.34	91.05/72.39	95/74.33	87.1/70.45
	seqPSSM+predStr	92.66/76.43	86.1/71.51	88.05/71.53	84.15/71.5	91.05/69.55	89.27/64.09	95/62.33	83.55/65.85
	seqBinaryEnc+predStr	92.77/77.16	86.2/73.02	88.07/72.54	84.33/73.5	92.33/69.11	90.46/68.41	95/74.33	85.92/62.49
	seqPSSM+realStr	85.7/83.88	84.98/84.07	88.33/85	81.63/83.15	X	X	X	X
	seqBinaryEnc+realStr	85.86/83.89	86.54/84.24	85.83/85	87.25/83.48	X	X	X	X
Intestine	seqPSSM	91.13/83.67	84.31/77.9	85.23/77.04	83.39/78.76	91.39/74.28	87.41/72.24	91/75.18	83.82/69.29
	seqBinaryEnc	91.58/85.42	83.98/78.74	84.73/79.33	83.24/78.16	90.93/74.88	87.11/72.52	94.18/70.82	80.04/74.22
	seqPSSM+predStr	91.22/83.08	84.49/76.44	85.33/76.62	83.66/76.26	90.93/76.7	86.52/75.86	91/79.09	82.04/72.64
	seqBinaryEnc+predStr	91.5/84.44	84.72/78.22	84.31/78.47	85.14/77.97	91.89/76.35	89.46/73.3	95/74.82	83.93/71.79
	seqPSSM+realStr	94.54/84.18	91.71/82.11	96/86	87.41/78.22	X	X	X	X
	seqBinaryEnc+realStr	93.9/83	90.89/81.88	94/84	87.79/79.75	X	X	X	X
Kidney	seqPSSM	91.45/84.63	84.12/78.12	84.6/77.35	83.63/78.88	92.92/75.09	90.51/75.05	95/78.93	86.02/71.18
	seqBinaryEnc	92.15/85.45	85.56/78.94	88.12/77.8	82.98/80.08	94.67/80.42	90.81/78.19	92.14/76.61	89.48/79.76
	seqPSSM+predStr	91.66/84.02	84.7/76.52	85/76.45	84.4/76.59	93.58/76.55	91.48/74.32	96.25/76.79	86.7/71.86
	seqBinaryEnc+predStr	91.84/84.04	85.17/77.71	87.09/76.64	83.24/78.78	94.19/81.4	90.97/78.03	92.32/84.29	89.62/71.78
	seqPSSM+realStr	92.93/89.17	91.7/86.82	96/85.5	87.41/88.15	X	X	X	X
	seqBinaryEnc+realStr	93.99/89.35	94.12/87.07	100/85.5	88.24/87.07	X	X	X	X
Liver	seqPSSM	91.16/80.45	83.73/73.81	84.34/76.45	83.12/71.18	93.21/70.92	90.79/69.44	94.18/70.55	87.4/68.33
	seqBinaryEnc	91.21/83.04	84.29/76.38	84.5/78.05	84.08/74.71	93.67/76.03	90.64/75.13	94.36/74.36	86.91/76
	seqPSSM+predStr	91.24/80.79	84.17/74.26	84.61/72.76	83.73/75.76	91.97/73.87	88.56/74.35	91.36/79.27	85.75/69.43
	seqBinaryEnc+predStr	91.28/81.85	83.57/76.16	83.84/74.77	83.3/77.56	92.2/77.28	89.26/75.82	91.64/74.64	86.87/77
	seqPSSM+realStr	90.75/77.48	87.08/78.17	90.18/83.57	83.99/72.77	X	X	X	X
	seqBinaryEnc+realStr	90.15/78.4	86.23/79.16	88.93/83.57	83.53/74.75	X	X	X	X
Lung	seqPSSM	89.95/81.53	82.73/74.72	81.87/73.94	83.59/75.5	90.22/72.54	84.15/70.69	86.92/71.17	81.39/70.2
	seqBinaryEnc	90.93/85.52	83.8/77.85	84.33/78.37	83.27/77.33	90.77/78.28	85.01/76.23	87.67/71.25	82.36/81.21
	seqPSSM+predStr	90.04/82.39	83.24/75.15	85.41/76.22	81.07/74.08	90.03/74.13	85.23/71.45	86.92/73.75	83.54/69.15
	seqBinaryEnc+predStr	90.72/84.52	83.59/77.46	83.93/76.88	83.25/78.04	90.98/79	85.72/73.65	89.46/76.42	81.98/70.87
	seqPSSM+realStr	92.73/85.44	90.47/86.11	94.33/89	86.61/83.23	X	X	X	X
	seqBinaryEnc+realStr	93.27/86.68	91.34/86.53	94.33/87	88.35/86.06	X	X	X	X
Muscle	seqPSSM	92.89/72.22	86.33/67.93	87.17/70.38	85.49/65.49	88.28/68.03	83.9/66.67	87.27/60.99	80.53/72.42
	seqBinaryEnc	92.97/74.47	86.3/70.43	89.39/67.88	83.2/72.98	85.57/66.47	79.48/65.68	85.83/67.5	73.14/63.86
	seqPSSM+predStr	92.59/74.85	86.02/69.77	87.91/70.36	84.14/69.17	87.39/68.04	82.63/67.43	88.18/64.55	77.08/70.32
	seqBinaryEnc+predStr	93.21/76.97	86.33/72.47	87.53/74.14	85.14/70.8	86.91/74.21	82.92/71	90/72.5	75.83/69.5
	seqPSSM+realStr	84.22/79.92	82.25/81.31	83.1/84.76	81.4/77.85	X	X	X	X

	seqBinaryEnc+realStr	86.44/81.24	83.89/79.88	84.52/80	83.26/79.77	X	X	X	X
Pancreas	seqPSSM	94.53/79.66	88.85/74.89	89.93/73.12	87.78/76.66	92.42/68.93	94.28/75.41	100/81.67	88.56/69.16
	seqBinaryEnc	94.5/83.58	90.04/78.3	93.73/77.41	86.35/79.19	90.66/68.12	93.98/56.15	100/50	87.95/62.3
	seqPSSM+predStr	94.66/81.6	90.49/76.03	94.57/76.11	86.41/75.96	91.99/68.83	94.71/65.97	100/63.33	89.42/68.62
	seqBinaryEnc+predStr	93.9/83.72	89.62/77.97	92.05/78.22	87.2/77.71	91.97/71.63	94.05/59.46	100/66.67	88.1/52.24
	seqPSSM+realStr	X	X	X	X	X	X	X	X
	seqBinaryEnc+realStr	X	X	X	X	X	X	X	X
Perirenal fat	seqPSSM	91.23/78.19	84.05/72.12	85.64/72.93	82.45/71.31	93/70.6	88.44/68.87	91.67/73.22	85.22/64.52
	seqBinaryEnc	91.46/80.56	84.52/74.88	85.13/74.67	83.91/75.1	93.81/76.63	89.04/74.8	93.89/73.11	84.19/76.5
	seqPSSM+predStr	91.19/79.21	84.32/72.74	85.51/72.04	83.13/73.45	91.89/71.38	87.45/71.9	91.56/72.33	83.34/71.47
	seqBinaryEnc+predStr	91.26/79.91	83.77/74.03	82.18/74.66	85.37/73.41	92.99/76.84	88.91/74.51	91.78/76.33	86.05/72.69
	seqPSSM+realStr	91.81/78.84	90.54/79.89	98/85	83.07/74.77	X	X	X	X
	seqBinaryEnc+realStr	91.81/80.87	90.29/78.3	95/82.5	85.59/74.1	X	X	X	X
Spleen	seqPSSM	90.51/82.71	83.5/75.92	84.66/75.37	82.34/76.47	90.67/74.2	87.25/71.97	87.25/71.96	82.42/71.98
	seqBinaryEnc	91.53/86.53	84.07/79.97	84.14/81.93	84.01/78.05	91.12/81	90.5/77.48	90.5/76.96	80.44/78.01
	seqPSSM+predStr	90.5/82.97	83.65/76.01	84.34/75.69	82.96/76.32	91.64/77.55	91.71/73.85	91.71/76.42	82.36/71.28
	seqBinaryEnc+predStr	90.86/84.78	84.44/77.85	85.94/77.37	82.94/78.32	91.52/81.75	89.79/76.86	89.7978.33	84.74/75.39
	seqPSSM+realStr	93.78/88.17	91.75/85.77	96/90.67	87.5/80.88	X	X	X	X
	seqBinaryEnc+realStr	90.92/85.62	89.91/85.54	90.33/89.33	89.49/81.74	X	X	X	X
Stomach	seqPSSM	90.14/80.34	83.5/73.92	85.55/71.8	81.45/76.03	91.5/74.17	89.05/72.93	95.89/69.67	82.21/76.2
	seqBinaryEnc	90.81/82.5	83.72/77.58	83.14/75.36	84.3/79.81	91.85/77.04	88.4/75.4	94.67/74.11	82.13/76.68
	seqPSSM+predStr	90.39/80.61	83.76/73.56	85.42/73.81	82.1/73.3	91.12/73.82	88.41/70.71	94.56/73	82.27/68.42
	seqBinaryEnc+predStr	90.74/82.51	83.11/76.39	82.89/76.41	83.33/76.38	91.62/77.15	88.65/74.87	92.33/78.44	84.97/71.29
	seqPSSM+realStr	91.39/83.83	89.42/83.25	91.5/87	87.33/79.5	X	X	X	X
	seqBinaryEnc+realStr	91.3/83.77	90.46/84.02	91.5/85	89.41/83.05	X	X	X	X
Testis	seqPSSM	91.14/80.82	84.03/74.89	84.97/73.79	83.09/75.98	92.77/74.89	90.03/73.9	95.09/76.73	84.98/71.08
	seqBinaryEnc	91.72/83.43	84.32/76.74	85.08/76.14	83.56/77.35	92.69/81.28	89.56/78.38	94.09/80.36	85.03/76.4
	seqPSSM+predStr	91.1/81.01	84.71/75.24	85.67/72.97	83.75/77.5	93.34/77.32	90.54/74.62	95.09/75.45	85.99/73.79
	seqBinaryEnc+predStr	91.44/82.63	84.82/76.34	86.7/74.92	82.93/77.77	93.17/82.43	89/78.58	90.09/78.36	87.91/78.79
	seqPSSM+realStr	95.21/87.05	95.78/86.96	100/96.67	91.56/77.26	X	X	X	X
	seqBinaryEnc+realStr	95.15/82.45	95.83/81.35	100/85	91.67/77.71	X	X	X	X
Thymus	seqPSSM	91.61/84.38	84.72/77.03	85.02/77.53	84.41/76.54	92.03/79.19	87.34/76.12	90.44/72.72	84.24/79.51
	seqBinaryEnc	92.54/87.61	85.1/80.94	86.18/81.07	84.01/80.81	92.68/83.19	87.93/78.58	90.48/78.6	85.38/78.56
	seqPSSM+predStr	91.63/84.58	84.84/77.31	85.02/76.72	84.67/77.88	91.73/81.61	87.57/77.1	89.26/77.43	85.87/76.77
	seqBinaryEnc+predStr	92/86.95	85.03/80.4	85.81/81.29	84.24/79.52	92.38/84.12	88.29/80.47	88.16/78.53	88.41/82.4
	seqPSSM+realStr	93.6/79.19	92.02/80.2	96/86	88.04/74.39	X	X	X	X
	seqBinaryEnc+realStr	92.99/80.14	90.23/81.31	89.5/88	90.95/74.61	X	X	X	X

Table 8: Classification performance analysis of 10-fold cross validation for each prediction model in different tissues with/without using the KNN score. For some tissues and modification types, the sizes of dataset containing real protein structures were not sufficient for the prediction. Therefore, the corresponding results are represented with "X".

4.2.5 *Predictive performance of the random forest model on tissue-specific phosphorylation sites*

To evaluate the performance of tissue-specific phosphorylation site prediction, 10-fold cross-validation was performed on 17 tissue-specific training sets for each serine and threonine residue types. We randomly divided each training dataset into 10 folds, and used each fold as the testing data to validate the trained model using the remaining 9 folds. Each fold was eventually used as a testing data. Note that the KNN score was calculated for each turn separately as the negative set changed in each turn. For each category defined in Section 4.1 10-fold cross-validation was applied, and the features contained in each category were summarized in Table 6.

The predictive performance of each model without using the KNN score was presented in Table 8, and corresponding AUC curves were plotted in Figure 20 and Figure 24 for serine and threonine residues, respectively. The prediction models for serine phosphorylation sites in all tissues concluded with considerably high performance – AUC scores ranging from 72.22 to 89.35, whereas prediction models for threonine phosphorylation sites yielded comparably high performance resulting in AUCs scores between 66.47 and 84.12. The findings show that the PSSM encoding, and binary encoding have the greatest impact on prediction performance. It is noteworthy to state here that we applied a very stringent sequence redundancy cut-off. In all tissues the binary encoding yielded a slightly higher performance than the PSSM encoding (AUCs of 0.822 and 0.851 for the PSSM encoding, and the binary encoding, respectively for serine phosphorylation sites in cortex). Incorporating the predicted structures and functional annotations along with the PSSM or binary encoding almost never improved the prediction performance except for muscle and blood. The model named as *seqBinary+predStr* has the highest AUCs of 0.814 and 0.77 for serine phosphorylation sites in blood and muscle, respectively. The features, in particular accessibility scores and secondary structures of phosphorylation sites and the residues surrounding them, had higher variable importance along with the PSSM or binary encoding scheme in prediction models. The models for phosphorylated serine residues using experimental protein structures along with the PSSM or binary encoding scheme, on the other hand, resulted in the highest performance in some tissues including perirenal fat, muscle, blood, heart, kidney, spleen and stomach. The feature spatial probability score contained in the categories *seqPSSM+realStr* and *seqBinary+realStr* outperformed among the other features, and leaded to a better prediction performance by having the highest variable importance in prediction models. These findings pointed out that predicted structures do not have the distinguishing power in tissue-specific phosphorylation site prediction as much as experimantal structures do.

With the addition of KNN score on top of six categories, a much larger increase in AUC scores was obtained for almost all prediction models in all tissues (See Table 8) – for serine phosphorylation sites ranging from 84.22 to 95.21 (See Figure 22), whereas for threonine phosphorylation sites ranging from 85.57 to 94.67 (See Figure 26). In particular, the prediction models for serine phosphorylation sites in blood and pancreas achieved very high performance, AUCs of 92.82 and 94.66, respectively even though the percentage of k nearest neighbors in those tissues was set to only 1% of the training dataset. This shows that the KNN score is a valuable feature in tissue-specific phosphorylation site prediction, but choosing the non-optimal k value would result in highly biased predictions. Our observations have shown that a very stringent sequence redundancy elimination is required prior to the KNN score calculation, and if the positive set to negative set size ratio is rather than 1:1, the k value becomes so sensitive to the size of dataset, and finding the optimal k value gets prior importance.

		pS				pT			
		AUC(%)	ACC(%)	SEN(%)	SPE(%)	AUC(%)	ACC(%)	SEN(%)	SPE(%)
Blood	seqPSSM	94.39/79.02	89.24/65.87	83.33/83.33	89.61/64.79	93.1/71.23	83.5/88.95	100/62.5	83.31/89.25
	seqBinaryEnc	94.39/84.07	86.85/84.57	88.06/74.63	86.78/85.2	92.17/81.48	77.99/74.65	100/75	77.75/74.65
	seqPSSM+predStr	93.75/73.98	86.36/65.43	86.36/69.7	85.3/65.17	93.97/79.27	83.22/74.41	100/87.5	83.03/74.26
	seqBinaryEnc+predStr	93.5/75.08	90.36/76.07	83.58/64.18	90.78/76.82	95.34/87.7	94.29/74.93	87.5/87.5	94.37/74.79
	seqPSSM+realStr	93.14/91.24	89.41/83.53	100/100	88.61/82.28	X	X	X	X
	seqBinaryEnc+realStr	88.89/85.39	79.31/80.46	100/100	77.78/79.01	X	X	X	X
Brain	seqPSSM	89.74/84.13	81.81/78.75	81.8/71.93	81.82/79.39	91.53/83.54	84.73/74.97	85.29/82.35	84.72/74.83
	seqBinaryEnc	89.59/85.64	80.72/79.32	83.81/76.59	80.43/79.57	92.17/82.91	82.37/78.68	85.29/77.94	82.32/78.69
	seqPSSM+predStr	90.49/85.9	82.21/76.79	83.11/79.83	82.13/76.5	91.96/83.49	86.35/81.87	83.82/75	86.4/82
	seqBinaryEnc+predStr	91.21/87.61	82.89/79.77	85.34/81.84	82.67/79.57	93.16/85.85	85.82/81.74	85.29/75	85.83/81.87
	seqPSSM+realStr	93.04/78.3	85.6/72.84	85.71/71.43	85.59/72.97	X	X	X	X
	seqBinaryEnc+realStr	94.35/81.12	83.67/74.29	95.24/80.95	82.59/73.66	X	X	X	X
Brainstem	seqPSSM	91.36/82.37	85/75.63	81.99/74.71	85.22/75.7	92.64/68.8	83.48/66.34	91.67/66.67	83.35/66.34
	seqBinaryEnc	91.19/84.11	80.76/79	86.26/75.57	80.34/79.25	93.49/75.84	81.22/84.68	88.89/61.11	81.1/85.03
	seqPSSM+predStr	92.08/84.51	86.53/83.44	81.99/71.26	86.87/84.36	91.91/71.35	83.27/61.08	86.11/75	83.23/60.87
	seqBinaryEnc+predStr	92.04/85.13	86.26/81.08	83.59/77.1	86.46/81.38	92.82/73.02	77.83/58.93	94.44/75	77.59/58.69
	seqPSSM+realStr	90.89/73.12	80.87/66.67	92.86/78.57	79.88/65.68	X	X	X	X
	seqBinaryEnc+realStr	91.7/78.87	84.78/68.48	85.71/71.43	84.71/68.24	X	X	X	X
Cerebellum	seqPSSM	90.04/84.17	82.8/76.65	83.44/74.38	82.75/76.83	91.27/83.54	77.27/74.92	90.24/82.93	77.09/74.81
	seqBinaryEnc	90.4/85.71	82.36/80.55	81.93/77.57	82.39/80.78	91.96/81.67	81.29/75.98	92.68/82.93	81.13/75.88
	seqPSSM+predStr	90.9/85.22	82.11/79.66	85/77.19	81.88/79.85	91.68/83.32	82.31/78.78	85.37/80.49	82.27/78.76
	seqBinaryEnc+predStr	91.76/87.66	81.67/82.72	85.67/81	81.36/82.85	92.37/84.7	86.94/79.28	82.93/80.49	86.99/79.27
	seqPSSM+realStr	88.36/81.02	77.44/70.77	100/78.57	75.69/70.17	X	X	X	X
	seqBinaryEnc+realStr	89.17/73.55	77.04/74.49	92.86/71.43	75.82/74.73	X	X	X	X
Cortex	seqPSSM	90.44/84.43	83.39/75.12	82.93/75.31	83.43/75.1	93.13/84.04	89.14/77.41	88.89/80	89.14/77.37
	seqBinaryEnc	90.18/84.9	86.18/78.31	79.94/76.29	86.73/78.49	93.4/83.02	81.64/74.68	93.33/82.22	81.44/74.55
	seqPSSM+predStr	90.88/85.7	83.76/76.17	82.01/81.4	83.92/75.7	93.41/85.33	88.09/77.72	91.11/84.44	88.04/77.6
	seqBinaryEnc+predStr	91.2/87.03	83.82/81.29	82.98/81.16	83.89/81.3	94.3/87.8	88.56/82.69	88.89/77.78	88.55/82.77
	seqPSSM+realStr	91.92/78.68	86.41/75.73	87.5/75	86.32/75.79	X	X	X	X
	seqBinaryEnc+realStr	92.79/79.38	88.94/72.6	87.5/81.25	89.06/71.88	X	X	X	X
Heart	seqPSSM	92.24/78.8	84.49/71.17	84.26/71.3	84.51/71.17	91.24/53.19	82.37/50.21	100/38.46	82.18/50.34
	seqBinaryEnc	91.95/77.73	85.17/68.31	85.46/71.82	85.16/68.1	92.78/73.13	84.4/74.56	100/64.29	84.22/74.68
	seqPSSM+predStr	92.03/74.88	79.06/75.4	87.96/62.04	78.53/76.2	89.58/53.76	79.87/66.58	100/46.15	79.65/66.81
	seqBinaryEnc+predStr	90.83/73.58	82.1/71.33	82.73/67.27	82.06/71.58	90.86/65.54	80.48/64.05	100/64.29	80.25/64.09
	seqPSSM+realStr	72.91/78.31	84.27/84.27	71.43/71.43	85.37/85.37	X	X	X	X
	seqBinaryEnc+realStr	81.41/80.12	81.11/86.67	71.43/71.43	81.93/87.95	X	X	X	X
Intestine	seqPSSM	87.49/80.51	79.76/71.23	79.92/75.64	79.75/70.94	89.02/73.53	77.44/63.88	92/76	77.29/63.76
	seqBinaryEnc	87.93/81.61	81.71/74.34	75.32/75.32	82.13/74.28	89.91/79.16	81.04/77.18	88/72	80.97/77.24
	seqPSSM+predStr	87.5/81.32	76.34/75.34	81.62/76.5	75.99/75.26	91.8/73.14	77.2/64.45	96/76	77.01/64.34
	seqBinaryEnc+predStr	88.12/82.46	76/77.44	81.28/73.19	75.65/77.72	92.19/79.02	85.67/72.96	88/76	85.64/72.93
	seqPSSM+realStr	90.48/82.07	76.88/81.25	91.67/66.67	75.68/82.43	X	X	X	X
	seqBinaryEnc+realStr	90.46/74.72	77.02/77.02	91.67/66.67	75.84/77.85	X	X	X	X
Kidney	seqPSSM	89.02/81.08	80.35/75	80.73/74.48	80.33/75.04	88.72/48.41	79.92/46.77	100/50	79.73/46.74
	seqBinaryEnc	89.85/83.39	82.35/80.55	79.79/70.98	82.52/81.2	92.5/80.12	85.82/79.12	88.89/72.22	85.79/79.18
	seqPSSM+predStr	88.9/81.83	80.32/75.16	80.21/75	80.33/75.17	88.97/61.48	80.63/51.3	100/72.22	80.45/51.1
	seqBinaryEnc+predStr	89.79/82.9	79.32/74.27	84.46/78.76	78.97/73.97	90.46/63.16	82.32/51.37	100/72.22	82.15/51.18
	seqPSSM+realStr	92.5/80.53	86.9/65.52	90.91/90.91	86.57/63.43	X	X	X	X
	seqBinaryEnc+realStr	78.61/77.87	82.99/80.27	63.64/63.64	84.56/81.62	X	X	X	X
Liver	seqPSSM	91.51/78.98	81.98/71.14	85.71/71.96	81.76/71.09	91.52/66.87	85.36/60.06	92.31/73.08	85.27/59.89
	seqBinaryEnc	92.39/84.59	83.12/79.96	87.37/78.42	82.86/80.06	93.42/74.71	83.12/69.13	92.31/73.08	82.99/69.08
	seqPSSM+predStr	91.4/80.06	83.16/74.78	85.71/72.49	83/74.92	92.05/71.09	84.49/60.59	96.15/84.62	84.34/60.29
	seqBinaryEnc+predStr	92.84/84.9	81.5/81.95	89.47/74.73	81.01/82.38	92.55/71.31	85.15/61.25	92.31/69.23	85.06/61.15
	seqPSSM+realStr	91.16/85.38	86.52/76.4	83.33/83.33	86.88/75.63	X	X	X	X
	seqBinaryEnc+realStr	92.06/86.92	86.03/82.12	88.89/83.33	85.71/81.99	X	X	X	X
Lung	seqPSSM	89.13/83.08	80.88/78.17	81.45/73.9	80.84/78.47	88.42/74.25	76.63/65.27	92.11/73.68	76.45/65.17
	seqBinaryEnc	90.17/84.05	82.83/73.94	82.39/79.56	82.86/73.56	88.87/74.74	80.51/76.28	78.95/71.05	80.53/76.34
	seqPSSM+predStr	89.97/83.08	78.78/73.73	86.48/77.67	78.25/73.46	89.51/78.36	77.47/67.32	94.74/81.58	77.26/67.15
	seqBinaryEnc+predStr	90.5/86.02	80.91/79.22	84.28/79.25	80.69/79.22	90.86/83.06	80.45/72.33	89.47/78.95	80.34/72.25
	seqPSSM+realStr	87.01/64.15	86.14/75.3	75/58.33	87.01/76.62	X	X	X	X
	seqBinaryEnc+realStr	86.7/64.02	82.14/60.12	75/66.67	82.69/59.62	X	X	X	X
Muscle	seqPSSM	93/77.34	82.82/71.01	92/72	82.3/70.96	91.79/75.31	84.37/68.65	88.89/66.67	84.26/68.7
	seqBinaryEnc	90.88/69.34	83.39/74.68	81.48/52.78	83.5/76.01	89.85/68.15	80.22/75.26	82.76/55.17	80.16/75.77
	seqPSSM+predStr	92.88/78.82	84.65/75.2	88/66	84.46/75.72	91.49/78.1	84.37/60.97	85.19/77.78	84.35/60.57
	seqBinaryEnc+predStr	92.33/78.5	82.22/78.5	84.26/65.74	82.1/79.28	92.18/80.53	86.56/82.62	82.76/68.97	86.66/82.97
	seqPSSM+realStr	91.83/84.11	86.07/77.87	93.75/75	84.91/78.3	X	X	X	X

	seqBinaryEnc+realStr	90.92/83.97	85.37/72.36	93.75/93.75	84.11/69.16	X	X	X	X
Pancreas	seqPSSM	96.31/83.98	91.13/83.63	88.14/72.88	91.29/84.2	90.93/66.67	85.23/70.31	100/80	85.12/70.24
	seqBinaryEnc	95.83/87.03	87.98/84.65	91.53/74.58	87.79/85.19	93.83/80.92	86.93/72.16	100/80	86.84/72.1
	seqPSSM+predStr	96.33/85.36	91.05/74.94	89.83/83.05	91.11/74.51	93.89/62.1	85.37/76.99	100/60	85.26/77.11
	seqBinaryEnc+predStr	95.82/84.87	87.98/74.77	93.22/79.66	87.7/74.51	94.65/69.03	85.37/65.77	100/80	85.26/65.67
	seqPSSM+realStr	X	X	X	X	X	X	X	X
	seqBinaryEnc+realStr	X	X	X	X	X	X	X	X
Perirenal fat	seqPSSM	90/81.4	79.08/78.64	86.31/68.45	78.61/79.31	90.12/64.23	82.33/51.67	87.5/75	82.26/51.34
	seqBinaryEnc	90.41/83.25	80.12/73.04	85.8/79.88	79.74/72.59	92.92/77.98	82.52/72.03	91.67/75	82.39/71.99
	seqPSSM+predStr	91.12/83.38	81.21/76.73	85.12/76.19	80.95/76.77	89.72/62.61	78.7/63.37	91.67/66.67	78.52/63.32
	seqBinaryEnc+predStr	91.35/85.03	84.13/82.09	82.25/75.74	84.25/82.5	91.54/65.47	85.61/58.28	83.33/70.83	85.64/58.1
	seqPSSM+realStr	97.07/85.91	92.31/73.5	100/100	91.67/71.3	X	X	X	X
	seqBinaryEnc+realStr	96.6/85.24	90.6/79.49	100/77.78	89.82/79.63	X	X	X	X
Spleen	seqPSSM	89.44/81.12	81.56/76.69	82.33/70.66	81.51/77.13	90.59/74.31	84.13/81.7	84.21/60.53	84.13/81.96
	seqBinaryEnc	90.54/83.48	83.63/79.41	82.7/71.7	83.7/79.98	93.05/84.09	86.1/78.56	86.84/76.32	86.09/78.59
	seqPSSM+predStr	89.65/83.23	81.71/75.64	83.28/76.34	81.6/75.59	88.32/74.93	81.43/71.51	86.84/71.05	81.36/71.51
	seqBinaryEnc+predStr	90.63/85.07	84.46/77.15	82.7/77.67	84.59/77.11	91.76/79.35	78.72/88.95	89.47/63.16	78.59/89.28
	seqPSSM+realStr	91.4/79.04	83.52/66.48	92.31/84.62	82.84/65.09	X	X	X	X
	seqBinaryEnc+realStr	79.51/74.79	83.7/64.67	76.92/76.92	84.21/63.74	X	X	X	X
Stomach	seqPSSM	88.77/80.46	80.14/78.15	82.86/67.62	79.97/78.83	91.27/72.32	82.48/63.64	95.46/72.73	82.35/63.55
	seqBinaryEnc	90.08/83.85	83.92/79.62	80.1/73.46	84.17/80.01	89.82/75.84	81.91/74.48	81.82/81.82	81.91/74.4
	seqPSSM+predStr	89.5/81.25	79.71/79.88	82.86/70.48	79.51/80.49	91.97/71.03	82.3/64.32	95.46/81.82	82.17/64.15
	seqBinaryEnc+predStr	90.47/83.17	82.8/74.59	81.99/76.78	82.86/74.45	90.81/77.25	81.69/73.03	95.45/72.73	81.55/73.03
	seqPSSM+realStr	91.53/75.26	80.13/84.77	90.91/63.64	79.29/86.43	X	X	X	X
	seqBinaryEnc+realStr	88.78/73.6	84.87/67.11	81.82/81.82	85.11/65.96	X	X	X	X
Testis	seqPSSM	89.96/81.56	79.72/72.12	85.31/79.18	79.35/71.65	90.82/69.43	81.88/74.98	96/60	81.74/75.13
	seqBinaryEnc	90.54/83.62	82.82/77.96	81.3/76.02	82.93/78.09	90.67/73.38	77.18/78.37	96/60	76.99/78.55
	seqPSSM+predStr	90.73/84.34	81.89/76.57	85.71/76.74	81.64/76.56	91.13/72.17	84.08/54.55	88/80	84.04/54.29
	seqBinaryEnc+predStr	91.28/85.24	83.03/80.23	84.55/75.61	82.93/80.54	91.31/73.99	83.14/62.54	88/76	83.09/62.41
	seqPSSM+realStr	93.75/74.26	91.27/74.6	87.5/87.5	91.53/73.73	X	X	X	X
	seqBinaryEnc+realStr	93.91/71.95	89.76/81.89	87.5/62.5	89.92/83.19	X	X	X	X
Thymus	seqPSSM	90.11/84.3	85.02/78.08	80/77.65	85.41/78.11	91.29/80.6	78.58/75.11	90.24/73.17	78.43/75.14
	seqBinaryEnc	91.03/87.07	82.7/79.49	84.75/81.82	82.54/79.31	90.31/83.36	77.34/78.48	90.24/75.61	77.17/76.5
	seqPSSM+predStr	90.22/84.78	81.59/81.61	83.24/74.41	81.46/82.17	91.45/77.21	82.97/78.49	85.37/65.85	82.94/78.65
	seqBinaryEnc+predStr	90.58/86.37	81.08/80.52	85.63/80.94	80.73/80.48	91.18/80.55	83.08/79.05	82.93/73.17	83.09/79.13
	seqPSSM+realStr	91.21/82	83.85/71.43	90.91/81.82	83.33/70.67	X	X	X	X
	seqBinaryEnc+realStr	91.45/78.18	84.57/84.57	90.91/63.64	84.11/86.09	X	X	X	X

Table 9: Classification performance analysis of independent testing for each prediction model in different tissues with/without using the KNN score. For some tissues and modification types, the sizes of datasets containing experimental protein structures were not sufficient for prediction. Therefore, the corresponding results are represented with "X".

4.2.6 *Prediction on independent test data*

In order to assess if prediction models are over-fitting for the training data, we also carried out the prediction models on independent test data. Note that proteins in independent test data were not used in training. In most tissues prediction models displayed similar performances on independent testing and 10-fold cross-validation, whereas the models obtained using the independent test data yielded higher performance in some tissues (See Table 9, Figure 21 and Figure 23 for serine residues with/without using the KNN score, Figure 25 and Figure 27 for threonine residues with/without using the KNN score). For instance, the model for phosphorylated serine residues using the category *seqBinaryEnc+predStr* without the KNN score in cortex achieved an AUC of 87.03, whereas the model using the same category in 10-fold cross-validation achieved an AUC of 84.15 in cortex. The same category in liver also yielded a higher performance using independent test set (AUC of 84.9) rather than 10-fold cross-validation (AUC of 81.85). These findings imply that the prediction models are not over-fitting for the training data.

4.2.7 *Cross-tissues performance evaluation on independent testing*

In order to assess the uniqueness performance of each tissue-specific model, we performed cross-tissues prediction against 13 tissues. The prediction was just conducted with the features defined in the *seqBinaryEnc* category to reduce the computational time and the complexity of representation of results. The independent test data of each tissue was created separately. We first split 70% of the sites in the PS1D dataset for training, and the remaining 30% for the testing. As there exist phosphorylation sites occurring in more than one tissue, note that we avoided the cases where a phosphorylation site was used in both training and testing. We subsequently conducted pairwise tests by assessing the prediction performance of independent test data of all other 12 tissues on a trained model of a particular tissue. The primary models, in almost all tissues, performed the highest phosphorylation site prediction performance on primary tissues as seen in Table 10 (based on AUCs). The performances did not vary substantially in non-primary tissues, because some phosphorylation sites are shared by more than one tissue, and as we do not have sufficient tissue-specific phosphorylation sites, we could not use mutually exclusive training sets of phosphorylation sites where each phosphorylation site just occur in one tissue. In other words, although there is not any overlap between training and testing sets, for instance training and testing sets of brain and kidney, respectively, this does not mean that a phosphorylation site in brain cannot be phosphorylated in kidney inherently, because the prediction model for brain was not trained with phosphorylation sites occurring only in brain. Given the fact that mutually exclusive sets of phosphorylation sites were not used for training the models, these findings still show that more accurate phosphorylation site predictions can be obtained when tissue-specific sites are taken into account. This fact supports the importance of tissue-specific phosphorylation site prediction.

4.2.8 *Performance comparison with the existing prediction tools*

We compared the performance of our prediction models in different tissues to that of Musite on independent test dataset (Gao and Xu, 2010). We used pre-trained M. Musculus prediction model of Musite with default parameters since training model on rat is not available. Our method outperformed in all tissues when sensitivity score is taken into account, whereas Musite performed slightly better in terms of specificity (Table 8). Note that when calculating sensitivity scores, we only counted phosphorylated serine residues in a particular tissue as true positives. As we mentioned before, the existing tools present

Training set/Independent set	Brain	Brainstem	Cerebellum	Cortex	Intestine	Kidney	Liver	Lung	Perirenal fat	Spleen	Stomach	Testis	Thymus
Brain	**85.32/78.28/78.43/78.26**	X	X	X	81/74.94/73.66/75.05	86.21/78.58/80.71/78.38	80.1/76.04/69.09/76.68	78.69/71.39/70.18/71.49	79.73/70.71/73.11/70.52	81.38/71.1/77.04/70.58	78.86/70.63/74.29/70.31	80.2/76.35/70.15/76.84	81.1/74.78/71.75/75.06
Brainstem	80.27/70.39/76.84/69.62	**83.48/79.61/73.6/80.07**	81.97/75.91/73.3/76.17	79.41/75.32/71.5/75.81	78.9/74.18/71.43/74.39	81.34/74.02/78.08/73.72	80.74/74.88/72.77/75.05	82.3/73.92/79.12/73.54	79.5/73.81/72.37/73.91	82.35/74.26/77.37/74	78.49/68.54/75/68.04	80.65/77.41/70.82/77.94	84.54/73.58/79.42/75.03
Cerebellum	81.68/75.97/73.47/76.39	80.87/77.45/71.8/78.29	**83.81/76.24/77.39/76.15**	78.51/70.28/76.15/69.36	78.72/74.18/69.71/74.54	82.16/74.81/75.47/74.75	80.72/77.29/69.78/77.98	81.7/78.07/74.09/78.37	77.48/71.55/72.87/71.46	81.97/77.57/75.09/77.79	77.31/70.35/73.47/70.08	80.05/78.79/67.58/79.67	83.98/77.2/77.34/77.18
Cortex	81.44/73.67/77.96/73.29	78.03/69.48/70/69.44	83.45/76.55/75.83/76.6	**85.13/77.48/77.33/77.5**	80.33/77.5/71.09/77.93	83.83/76.42/75.15/76.51	79.37/71.81/73.16/71.71	78.15/73.7/68.52/74.09	79.94/71.33/75.18/71.04	81.7/73.85/74.11/73.83	77.76/74.34/65.33/75.01	77.58/74.19/68.35/74.61	80.94/75.27/70.99/75.64
Intestine	81.25/73.76/74.46/73.66	77.28/68.56/72.29/68.18	81.88/75.08/73.21/75.28	81.71/71.37/78.67/70.49	**82.49/74.69/77.05/74.53**	84.56/76.41/74.78/76.59	76.92/71.58/71.8/71.57	82.84/78.35/72.95/78.84	81.3/77.3/73.59/77.58	82.76/74.4/75.4/74.3	77.82/74.61/66.45/75.58	77.27/66.9/73.66/66.27	81.19/74.17/75.09/74.08
Kidney	79.36/71.68/72.71/71.56	77.06/70.54/69.42/70.64	78.78/70.45/74.1/70.1	81.46/74.92/74.34/74.99	80.56/74.86/70.37/75.24	**84.21/73.58/78.97/73.22**	78.35/69.46/73.86/69.06	78.56/72.39/71.47/72.47	79.52/73.67/68.8/74.04	79.35/70.71/72.2/70.57	76.09/68.66/68.59/68.67	81.39/80.99/69.08/82.01	79.21/73.79/68.39/74.26
Liver	81.38/78.92/69.68/80.07	81.46/81.74/68.67/83.07	82.45/80.65/69.71/81.8	81.35/77.76/69.77/78.72	79.09/77.83/67.2/78.67	78.15/77.53/67.12/78.57	**82.69/75.41/76.49/75.35**	83.39/79.8/74.76/80.2	80.99/78.7/67.94/79.63	82.23/80.53/71.47/81.35	78.04/75.18/65.08/76.08	79.12/76.95/67.84/77.7	83.09/80.22/74.86/80.69
Lung	81.54/71.46/76.88/70.68	79.94/71.92/78.3/71.22	82.07/75.34/72.5/75.66	82.11/74/76.29/73.66	83.02/74.04/79.03/73.32	85.28/75.06/80.2/74.3	81.84/71.32/79.66/70.49	**83.71/76.35/77.2/76.29**	80.28/78.89/72.15/79.43	80.66/75.75/71.91/76.22	76.18/67.05/72.66/66.27	77.21/73.66/70.17/73.98	82.64/73.64/76.19/73.38
Perirenal fat	79.83/89.21/36.09/95.34	78.44/90.81/33.92/95.88	80.89/90.65/38.53/95.38	80.81/86.63/44.85/91.47	80.4/89.2/40.28/92.66	81.95/90.72/42.04/94.31	79.25/92.46/26.47/97.43	83.35/90.84/42.33/94.44	**82.34/88.62/47.64/91.32**	82.52/87.21/53.82/90.13	79.06/89.91/35.2/94.1	80.23/90.62/35.69/94.69	85.31/90.51/40.39/94.64
Spleen	81.97/87.49/37.01/94.42	79.8/84.5/49.42/88.52	82.44/87.43/48.1/91.83	83.34/87.15/43.73/92.88	82.75/88.24/53.97/91.49	86.9/89.87/43.59/94.89	79.01/90.7/28.57/95.94	81.58/90.11/34.18/95.14	79.13/89.3/35.56/94.46	**84.03/85.66/60.38/87.51**	76.09/88.01/32.68/94.41	76.3/87.72/40.37/91.72	85.48/86.46/56.32/89.63
Stomach	79.57/86.15/49.82/90.73	78.96/87.95/45.58/92.13	80.79/87.52/47.97/91.56	79.96/86.32/44.22/91.59	80.18/89.63/37.63/94.46	80.38/88.8/37.26/93.82	79.43/86.66/56.49/89.21	84.27/88.7/54.05/91.7	81.71/88.55/40.98/92.3	83.13/88.97/46.95/93.18	**81.17/89.69/45.11/92.57**	80.28/88.42/48.4/92.04	84.92/87.48/55.26/90.46
Testis	81.47/88.77/42.13/94.87	81.76/88.93/52.28/93.11	82.46/86.72/59.95/89.97	81.68/87.43/49.35/92.28	77.88/89.01/37.38/94.04	81.97/89.61/34.88/95.29	78.03/90.29/39.61/94.75	81.19/89.36/46.15/92.99	81.7/89.06/43.33/93.02	80.94/89.74/37.72/94.84	76.39/88.95/35.27/94.43	**83.94/90.74/50.95/93.4**	82.95/89.99/47.37/93.9
Thymus	81.94/85.91/53.52/90.66	81.96/86.84/55.76/90.44	81.79/88.5/32.04/95.37	81.61/84.81/54.85/89.09	77.12/86.21/33.82/91.82	82.98/88.28/40.16/93.67	78.75/88.46/42.24/92.41	82.46/88.37/40.68/92.85	78.44/85.51/40.2/89.29	81.36/85.56/43.9/91.75	77.64/88.81/33.55/95.3	78.58/88.09/44.44/91.89	**86.97/88.94/64.26/90.85**

Table 10: Cross-tissues prediction performance of serine phosphorylated sites based on independent testing in different tissues using the *seqBinaryEnc* category without the KNN score. As there is no mutually exclusive phosphorylation site between brain and its dissected components (brainstem, cortex and cerebellum), the corresponding cells are filled with "X". The numbers in each cell represent AUC, accuracy, sensitivity and specificity, respectively. The numbers in bold represent the results for test sets from the models trained using data in same tissue.

Tissue	TSPhosPred		Musite	
	Sn (%)	Sp (%)	Sn (%)	Sp (%)
Blood	83.33	89.61	37.88	91.69
Brain	81.8	91.53	47.59	89.12
Heart	84.26	84.51	40.74	88.21
Intestine	79.92	79.75	41.45	89.79
Kidney	80.73	80.33	45.31	90.27
Liver	85.71	81.76	45.5	91.25
Lung	81.45	80.84	48.74	89.57
Muscle	92	82.3	26	91.01
Pancreas	88.14	91.29	55.93	84.52
Perirenal fat	86.31	78.61	48.81	89.2
Spleen	82.33	81.51	49.84	89.06
Stomach	82.86	79.97	40.48	89.37
Testis	85.31	79.35	45.71	90.22
Thymus	80	85.41	53.53	89.99

Table 11: Performance comparison for phosphorylated serine residue prediction between Musite and TSPhosPred based on independent test sets in different tissues. The prediction model using the *seqPSSM* category with the KNN score was compared to Musite using *M. musculus* pre-trained model with default parameters.

prediction models with low sensitivity and high specificity; however, TSPhosPred achieved a great performance on both true positive and true negative rates.

4.3 CONCLUSIONS

In Chapter 2 and Chapter 3 we have underscored the importance of tissue-specific analysis of posttranslational modifications that the mechanism behind phosphorylation and acetylation follow a tissue-specific pattern. In particular for phosphorylation, we detected tissue-specific sequence and spatial motifs around phosphorylation sites along with tissue-specific structural preferences. These findings derived the existence of tissue-specific kinases and phosphatases that directed us to perform phosphorylation site prediction in a tissue-specific manner to obtain an improved performance in comparison to existing methods. In this work we have developed the novel comprehensive approach, TSPhoPred that performs a high-quality tissue-specific phosphorylation site prediction based on various informative features. In addition to sequence-based features, we also utilized structure-based features where both experimental structures and predicted structures were taken into account. We observed that experimental structures along with the encoding for linear amino acid content achieved the highest performance for some tissues. The cross-tissues prediction analysis with performance comparison with existing tools shows the originality of, and the necessity for tissue-specific phosphorylation site prediction. TSPhoPred provided not only high specificity, but also high sensitivity for almost all tissues, and it outperformed the existing tools developed for globally phosphorylated sites in terms of distinguishing power between phosphorylation sites and non-phosphorylation sites. However, given the fact that different training datasets might lead various performance results, this is the first study on the prediction of tissue-specific phosphorylation sites that there is not any existing tool available to make a fair comparison including only tissue-specific sites on training datasets.

This study has underscored that tissue-specific phosphoproteomics still harbors potential to reveal the regulatory mechanism of phosphorylation. Further experimental verifications are required following the findings here. As future work, a feature selection method can also be applied before prediction to avoid noisiness, and redundancy that heterogeneous features might lead (Li et al., 2014). Moreover, in order to better assess the influence of some features on tissue-specific phosphorylation site prediction, i.e. functional annotations, the prediction can be performed on a negative set generated from the entire rat proteome containing also non-phosphorylated proteins. However, this approach may also increase the false negative rate, because the phosphorylated proteins have already been well studied with respect to phosphorylation, as a result, extracted non-phosphorylation sites in those proteins are more likely to be true negatives (Trost and Kusalik, 2013). Last but not least, in this study we used the most comprehensive dataset for tissue-specific phosphorylation sites (containing 17 different tissues). However, experimentally identified phosphorylation sites in single tissues coming from separate studies can also be combined to the current training set to increase a training set size even though the aim in this study was to form the training data containing phosphorylation sites from the same experiment on the same species.

Part V

SUMMARY

5 SUMMARY

In this thesis we have conducted a comprehensive sequence- and structure-based analysis of tissue-specific acetylation and phosphorylation sites, and presented evidence that lysine acetylation sites and phosphorylation sites display tissue-specific preferences for certain residues both in their linear amino acid sequence and in spatial environments. We also showed tissue-specific characteristics for the structural organization around the binding sites of lysine acetylation sites and phosphorylation sites. We further demonstrated that LASs are generally more evolutionarily conserved than non-LASs, which is especially prominent in structurally regular regions and in proteins with known function. We also presented that both phosphorylated and acetylated proteins are specialized for various functions in different tissues, and this specialization is supported by tissue-specific key domain preferences.

The tripartite graph connecting kinases, tissues and motifs, on the other hand, reveals that some phosphorylation motifs are prominent in many tissues, but are only targeted by few kinases.

We were not able to detect significant differences in the abundance of paralogs of experimentally identified KATs across different tissues, which implies that tissue-specific preferences of lysine acetylation sites are not a result of tissue-specific KAT expression. Similarly, we indicated that while many kinases mediate phosphorylation in all tissues, there are also kinases that operate in a tissue-specific manner, and these tissue-specific kinase preferences are not correlated with tissue-specific kinase expression.

All these findings imply the existence of tissue-specific kinases and phosphatases and the existence of tissue-specific KATs and KDACs able to differentiate between various types of local structural environments beyond mere amino acid content in sequence and spatial environments.

These observations eventually directed us to perform phosphorylation site prediction in a tissue-specific manner to obtain an improved performance in comparison to existing prediction approaches. We have developed the novel comprehensive approach, TSPhoPred that performs a high-quality tissue-specific phosphorylation site prediction based on various informative features including structural characteristics of phosphorylation sites in addition to sequence-based properties. We also utilized experimental structures along with predicted structures to obtain more accurate predictions. Indeed, the results pointed out that experimental structures along with the encoding for linear amino acid content achieved the highest performance for certain tissues. Functional annotations were also used which did not make any significant contribution to the accuracy of predictions. To benefit the influence of functional annotations, the prediction can be performed on a negative set generated from the entire rat proteome containing also non-phosphorylated proteins. However, note that this approach would bring the overhead of high false negative rate. Comparison with an existing tool on independent testing indicated that TSPhoPred outperformed in all tissues in terms of true positive rates, and overcame the drawbacks of existing prediction models with low sensitivity and high specificity. Supportively, the cross-tissues prediction analysis demonstrated that the primary models performed the highest phosphorylation site prediction performance on primary tissues in almost all tissues. This finding proves originality of, and the necessity for tissue-specific phosphorylation site prediction.

Given the facts that acetylation and phosphorylation play roles in disease signaling pathways, and kinases and KATs are potential drugs against many diseases including cancer, we emphasize the importance of tissue-specific drug target designs. Altogether, this thesis provides a different aspect for the evolution of post-translational modifications.

The availability of tissue-specific glycosylation sites also enables a comprehensive analysis of tissue-specific glycosylation sites both in sequence and structural manners (Kaji et al., 2012). Further research is needed to perform the methods we defined and applied in this thesis on the dataset of tissue-specific glycosylation sites. As more tissue-specific post-translational modification sites are identified experimentally, more signals would be detected through our investigation approaches. In addition, our prediction approach can be applied on tissue-specific lysine acetylation sites using the analyzed features in Chapter 2, where we would expect to outperform the existing methods for lysine acetylation site prediction.

Part VI

APPENDIX

A

APPENDIX

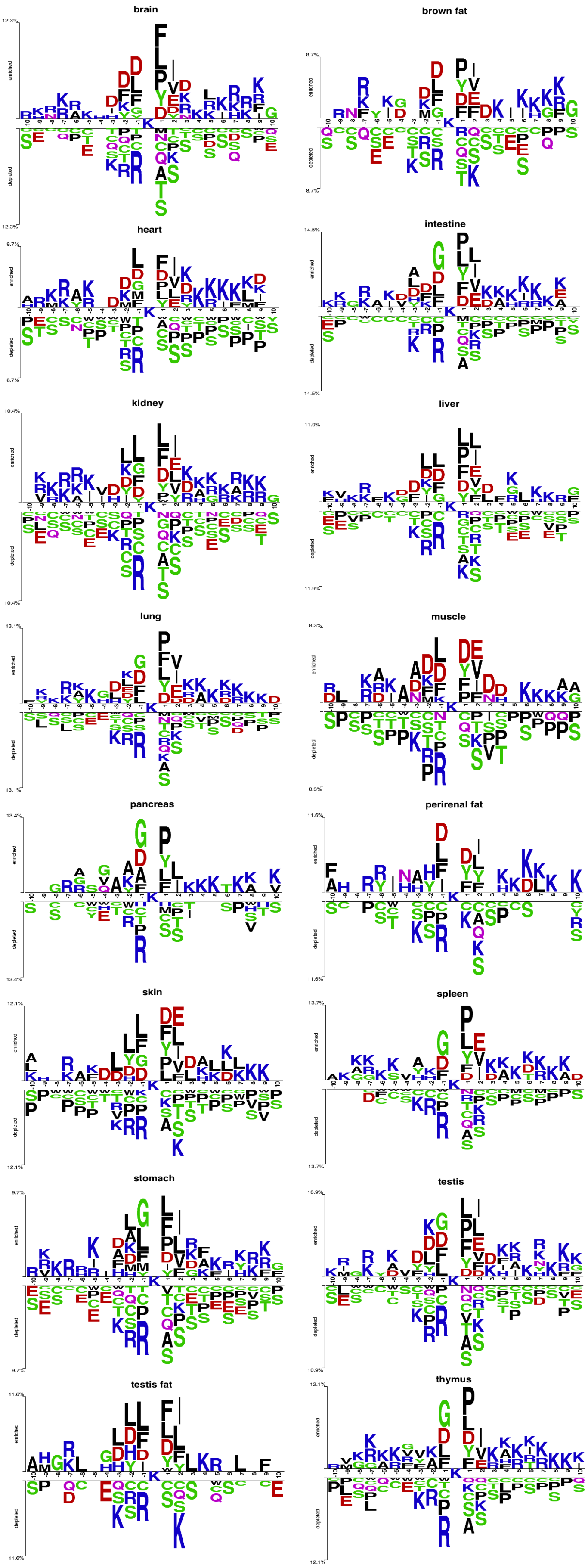


Figure 28: Two sample logo analysis of LASs in different tissues in the LAS1D dataset (See Figure 29 – Figure 44 for high resolution graphs).

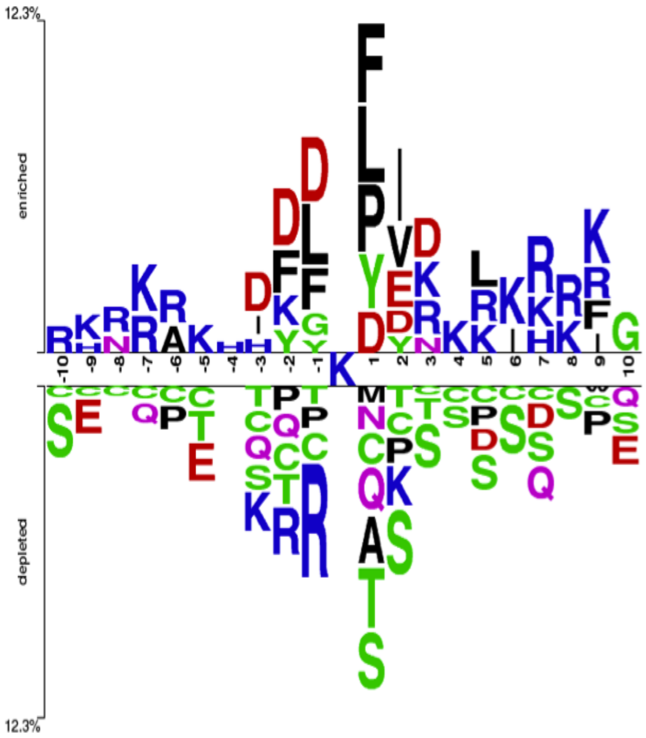


Figure 29: Two sample logo analysis of LASs from the LAS1D dataset in brain.

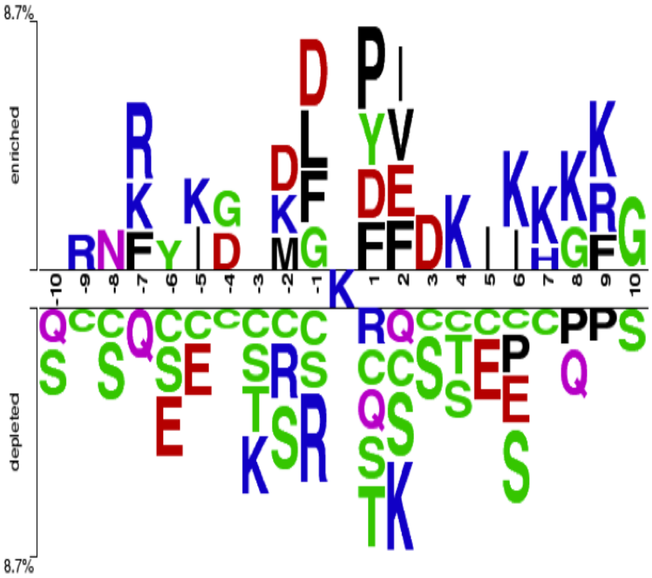


Figure 30: Two sample logo analysis of LASs from the LAS1D dataset in brown fat.

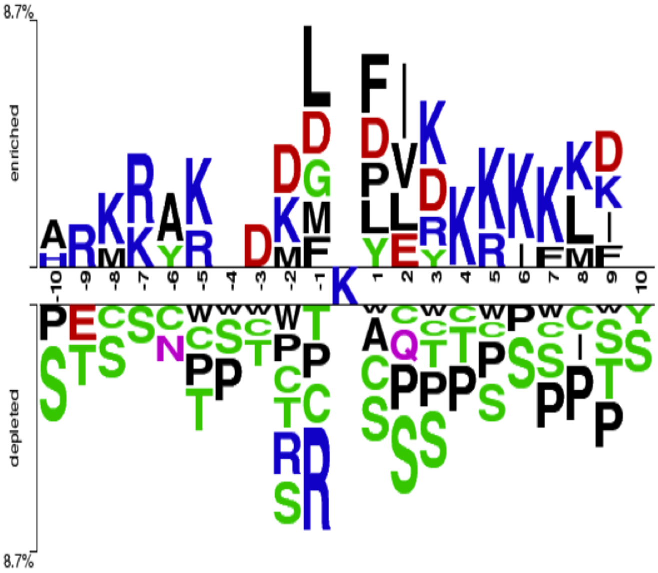


Figure 31: Two sample logo analysis of LASs from the LAS1D dataset in heart.

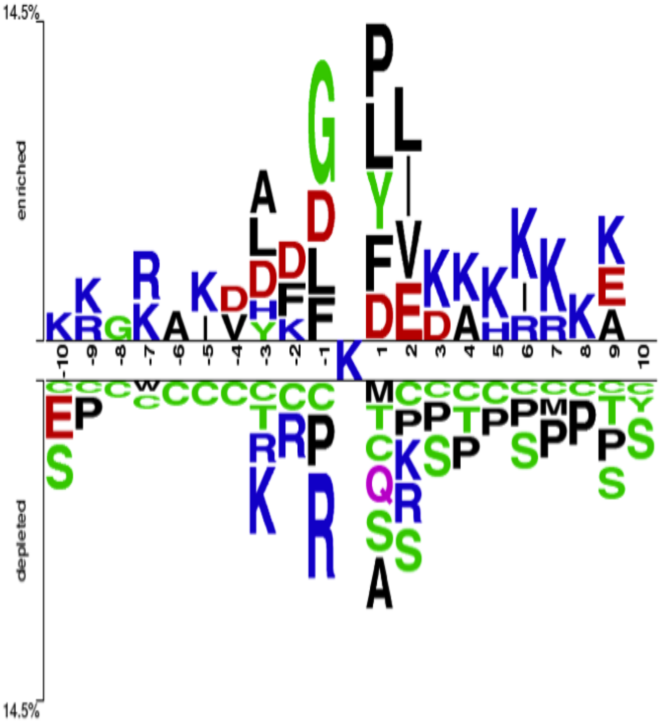


Figure 32: Two sample logo analysis of LASs from the LAS1D dataset in intestine.

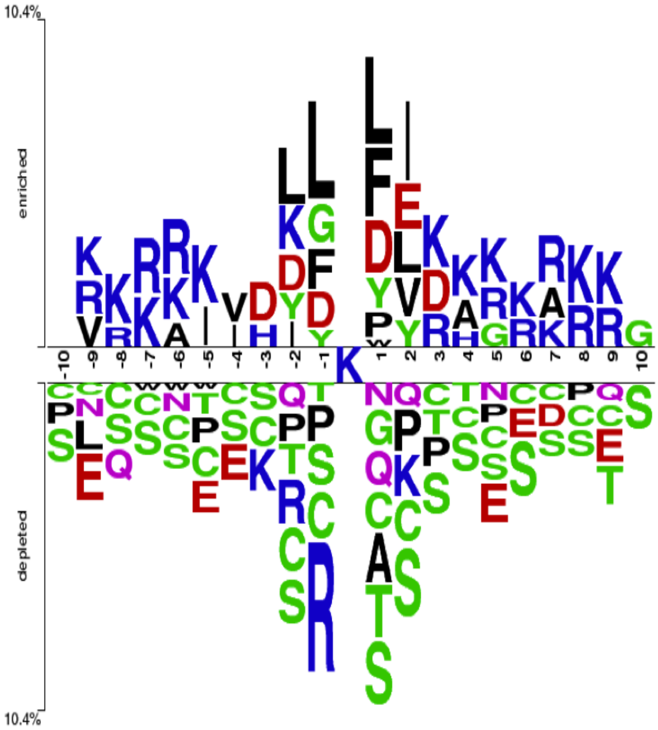


Figure 33: Two sample logo analysis of LASs from the LAS1D dataset in kidney.

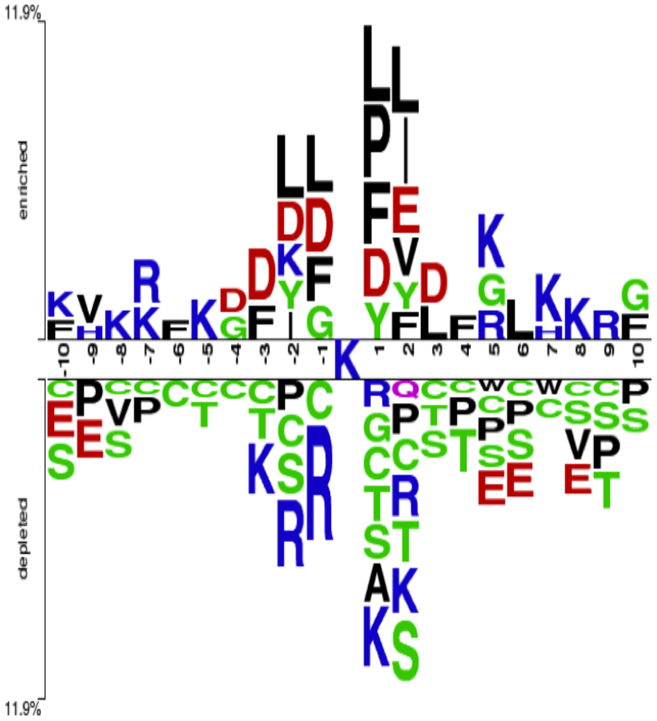


Figure 34: Two sample logo analysis of LASs from the LAS1D dataset in liver.

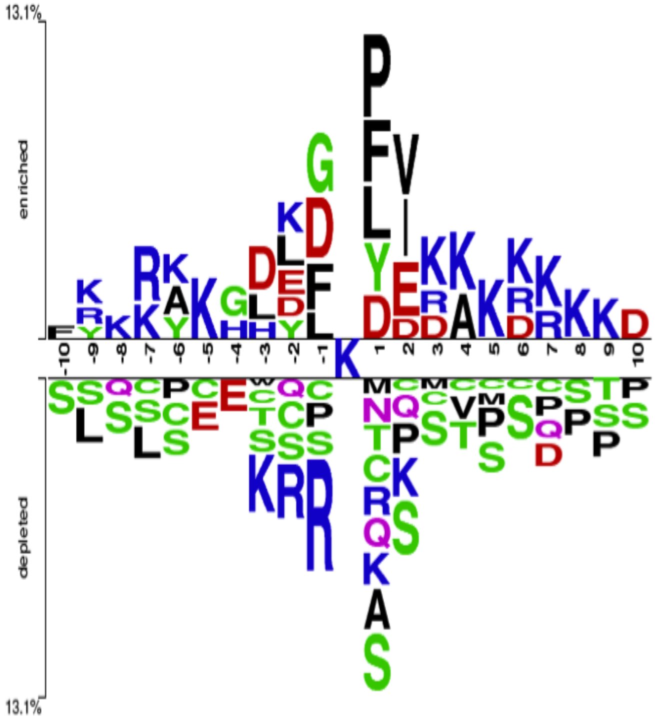


Figure 35: Two sample logo analysis of LASs from the LAS1D dataset in lung.

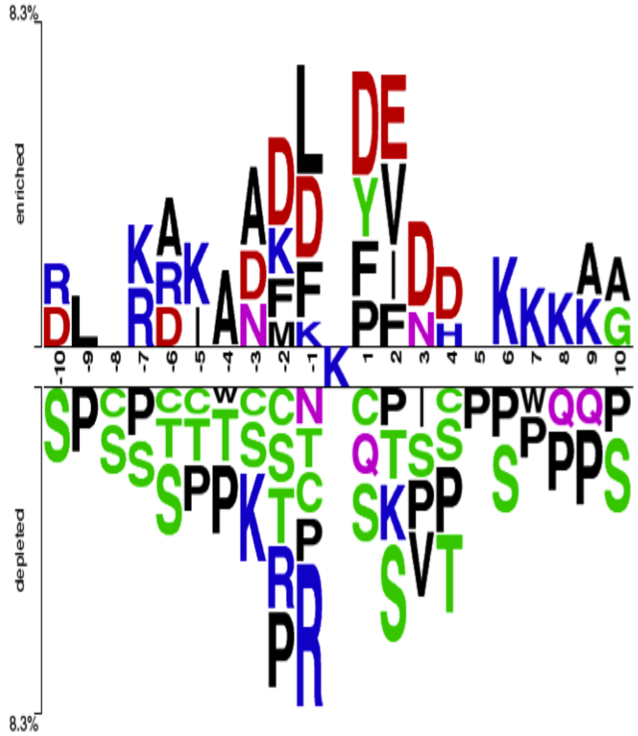


Figure 36: Two sample logo analysis of LASs from the LAS1D dataset in muscle.

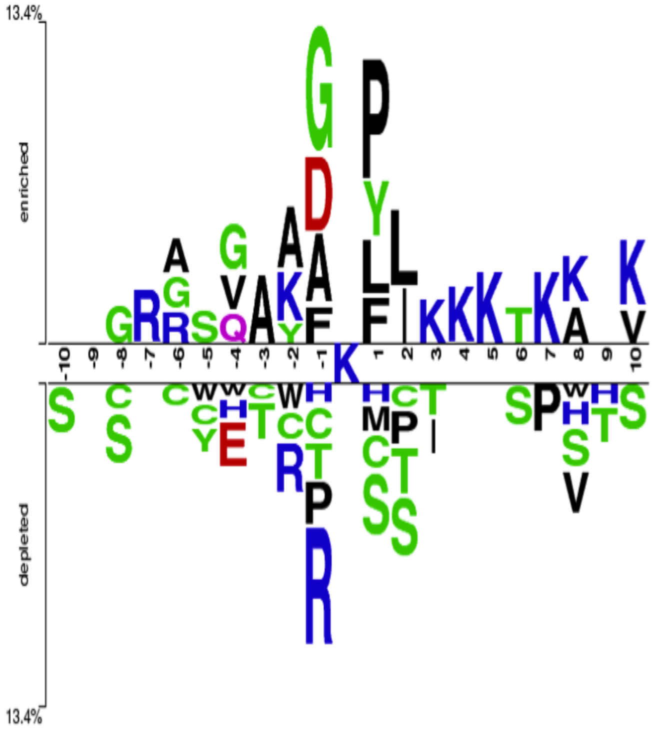


Figure 37: Two sample logo analysis of LASs from the LAS1D dataset in pancreas.

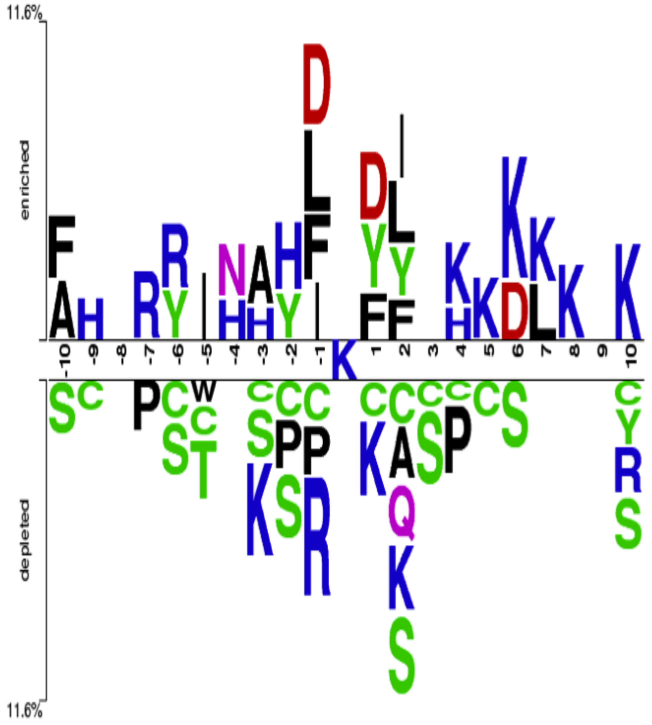


Figure 38: Two sample logo analysis of LASs from the LAS1D dataset in perirenal fat.

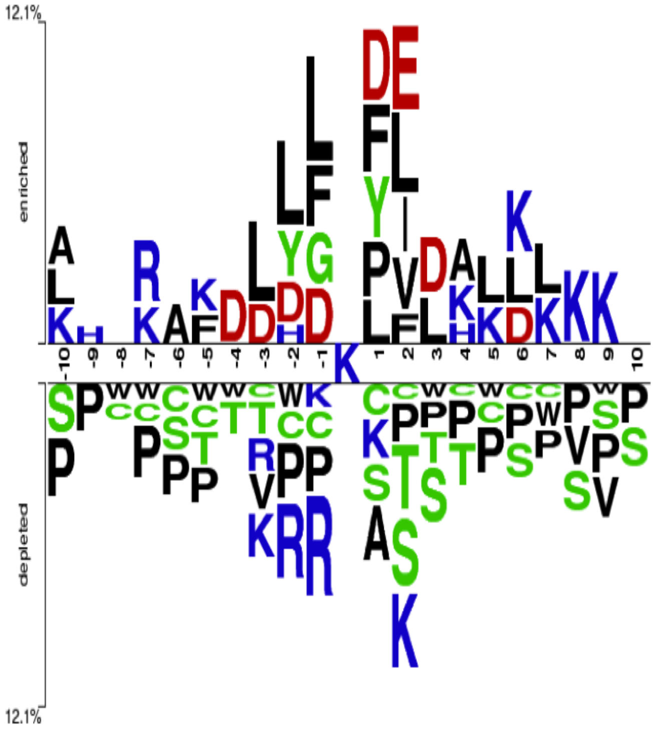


Figure 39: Two sample logo analysis of LASs from the LAS1D dataset in skin.

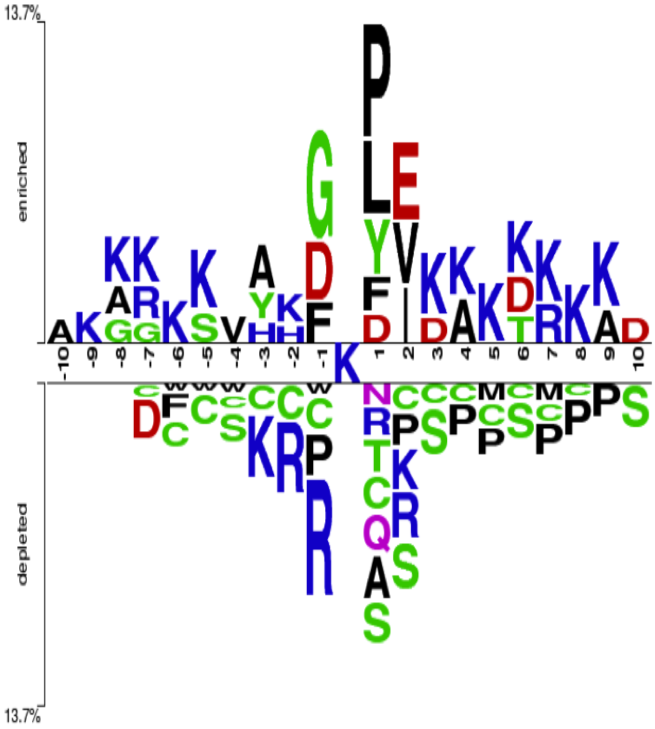


Figure 40: Two sample logo analysis of LASs from the LAS1D dataset in spleen.

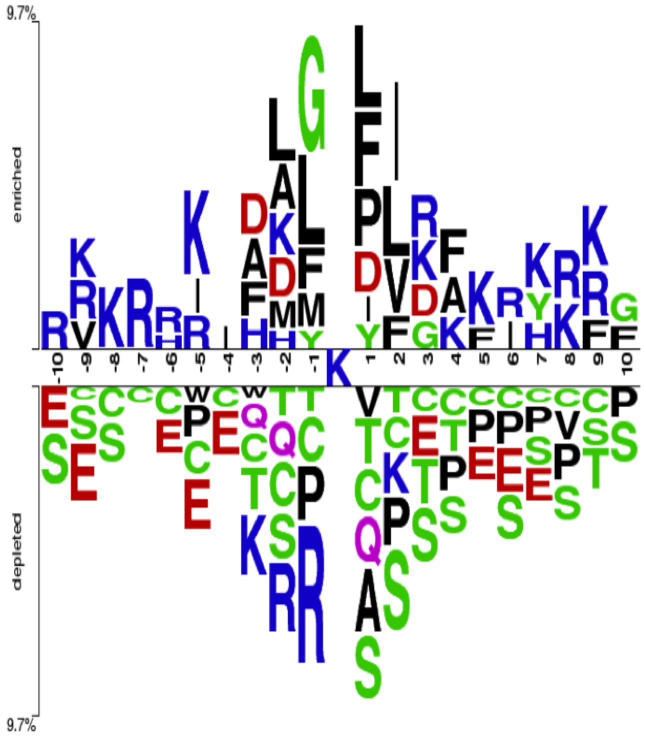


Figure 41: Two sample logo analysis of LASs from the LAS1D dataset in stomach.

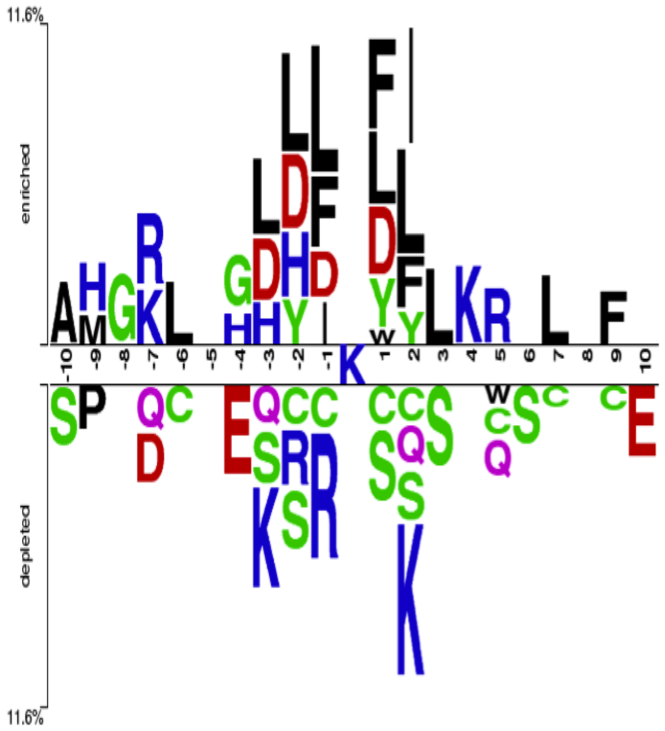


Figure 42: Two sample logo analysis of LASs from the LAS1D dataset in testis fat.

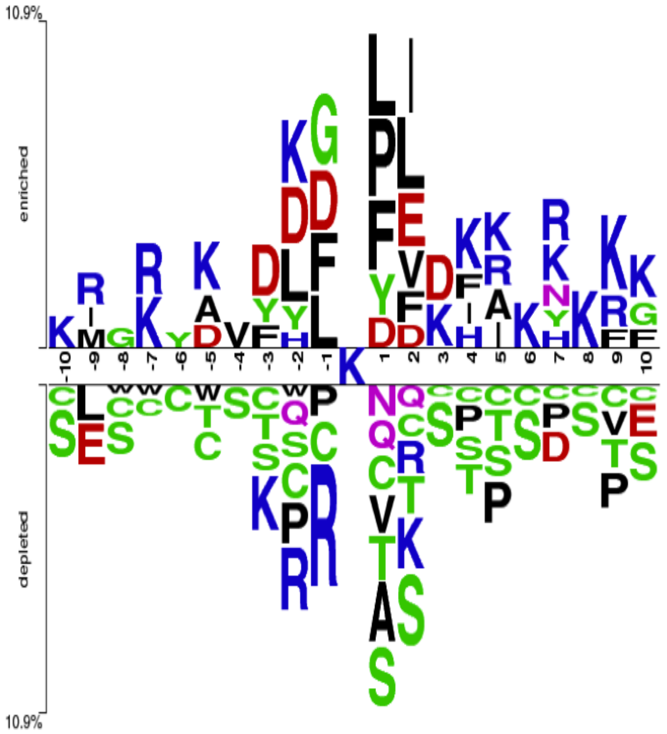


Figure 43: Two sample logo analysis of LASs from the LAS1D dataset in testis.

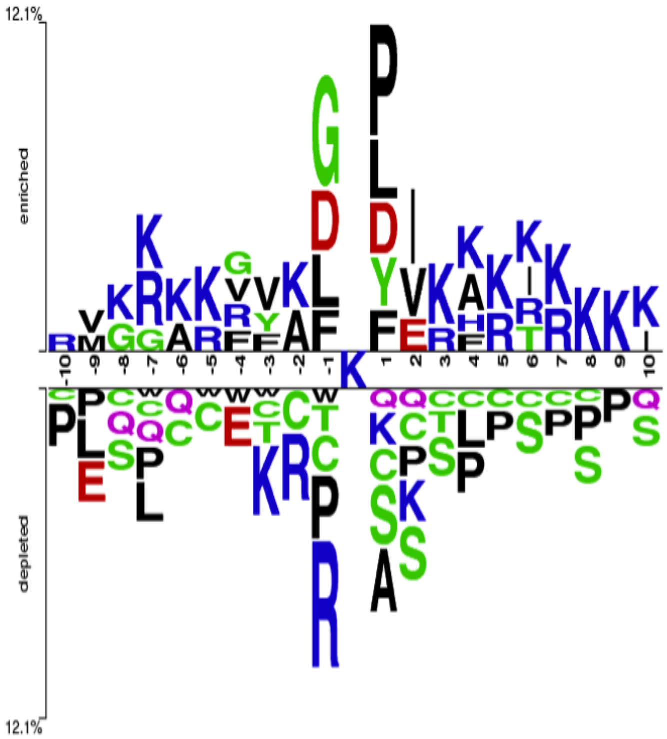


Figure 44: Two sample logo analysis of LASs from the LAS1D dataset in thymus.

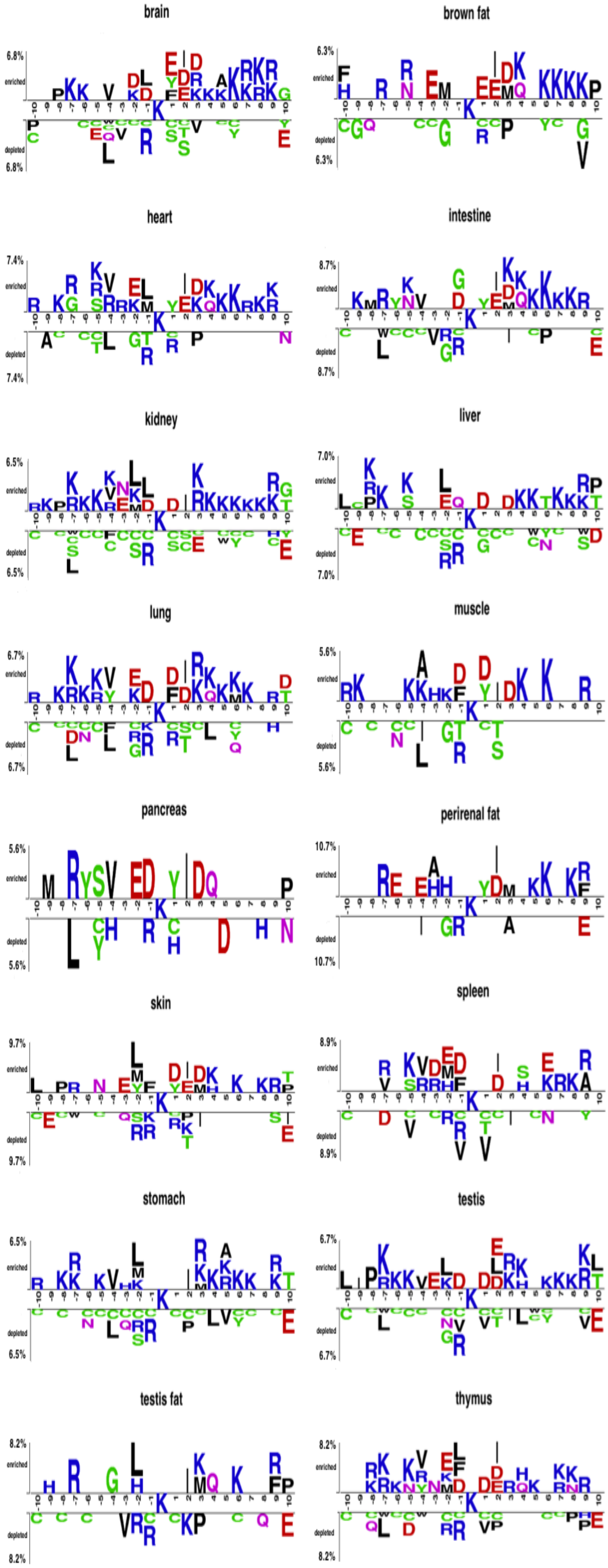


Figure 45: 1D environment of LASs from the LAS3D dataset in different tissues.

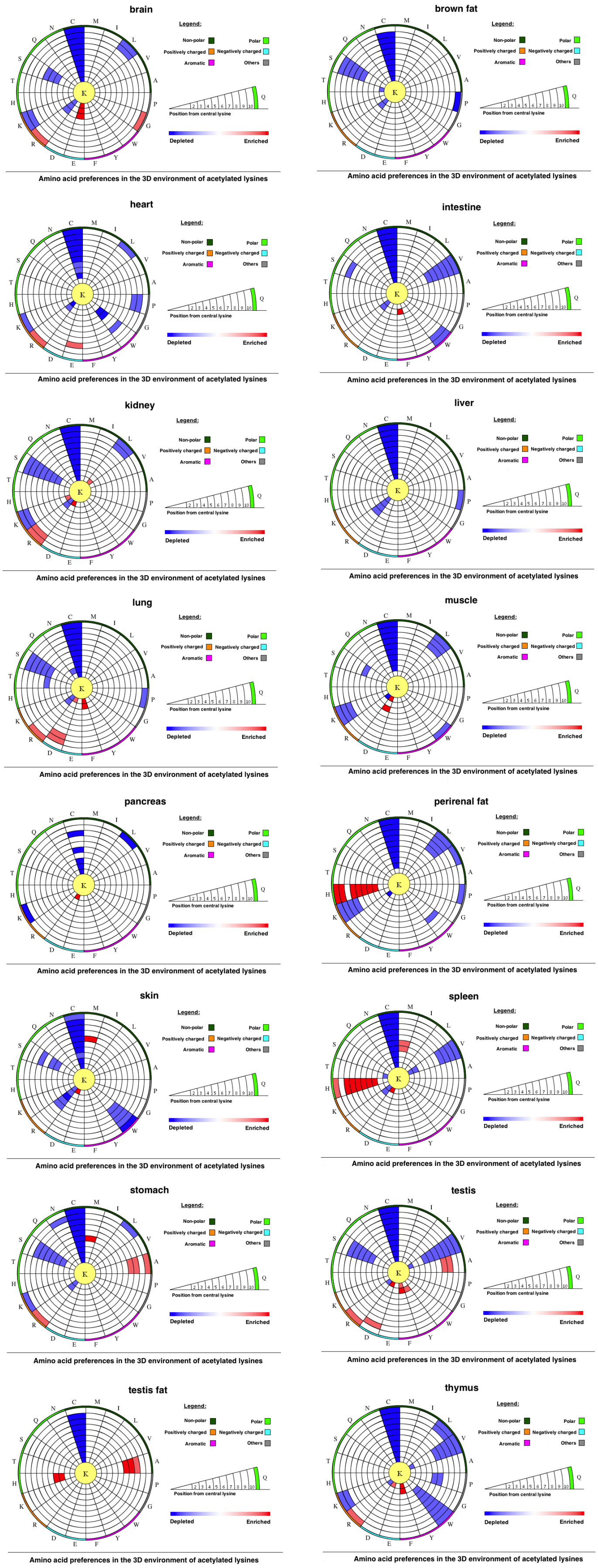


Figure 46: 3D environment of LASs from the LAS3D dataset in different tissues.

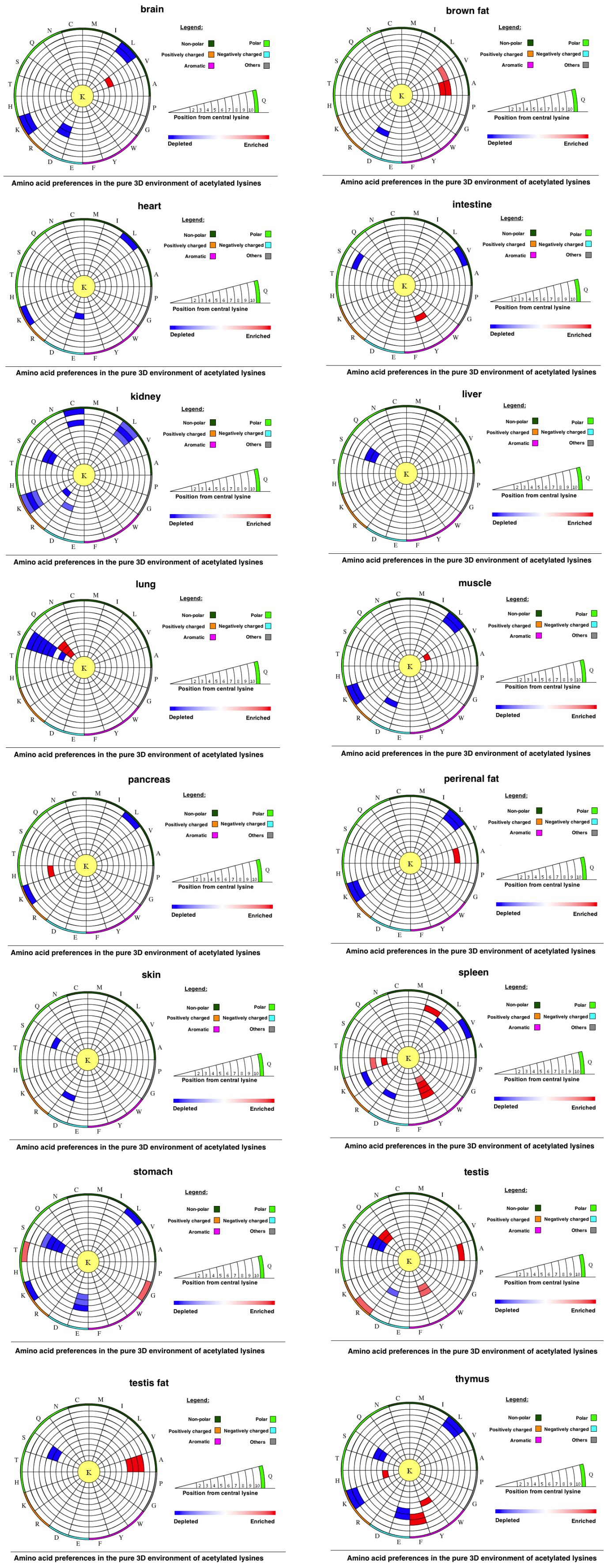


Figure 47: Pure 3D environment of LASs from the LAS3D dataset in different tissues.

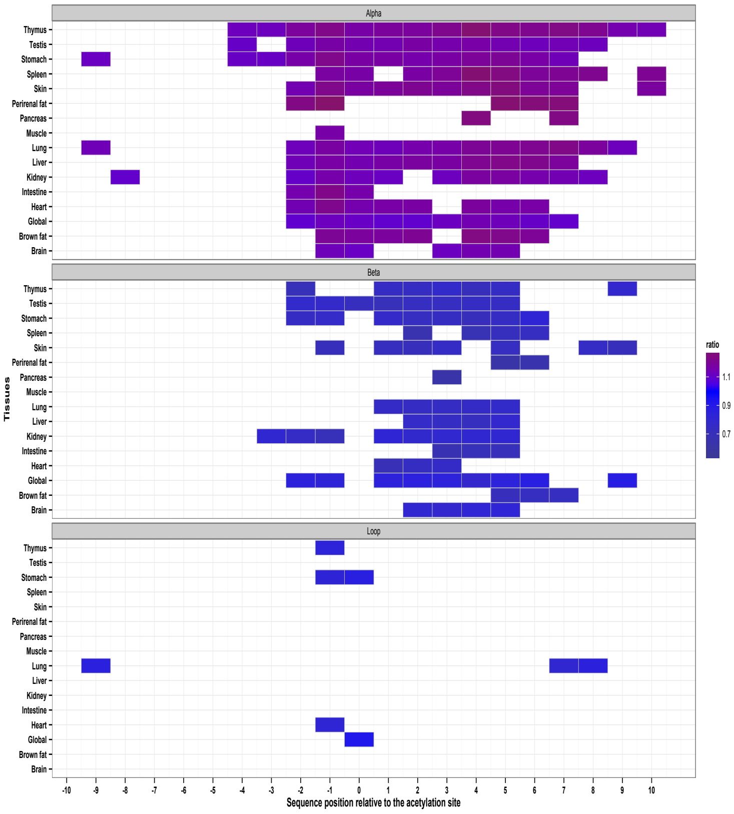


Figure 48: Secondary structure analysis of LASs and the residues surrounding them from the LAS3D dataset in different tissues. Ratios represented with shades of blue and red show the normalized number of LASs found in a particular secondary structure divided by the normalized number of non-LASs found in the corresponding structure. Non-significant ratios (p-value > 0.01) are represented with white cells. No structural preferences of LASs could be observed in testis fat.

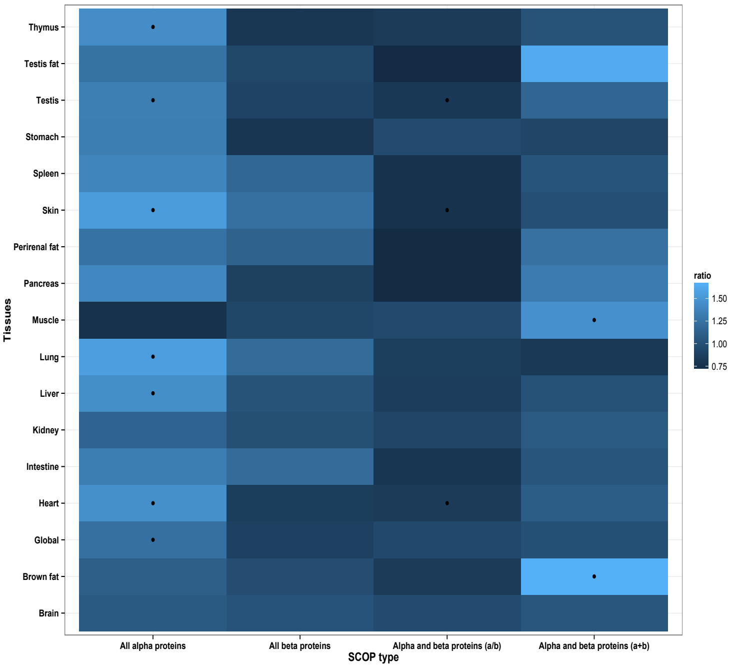


Figure 49: SCOP class analysis of LASs and the residues surrounding them in different tissues in the LAS3D dataset. Ratio represented with shades of blue show the normalized number of LASs found in a particular protein structural class divided by the normalized number of non-LASs found in the corresponding class. Black circles represent significant p-values (p < 0.01).

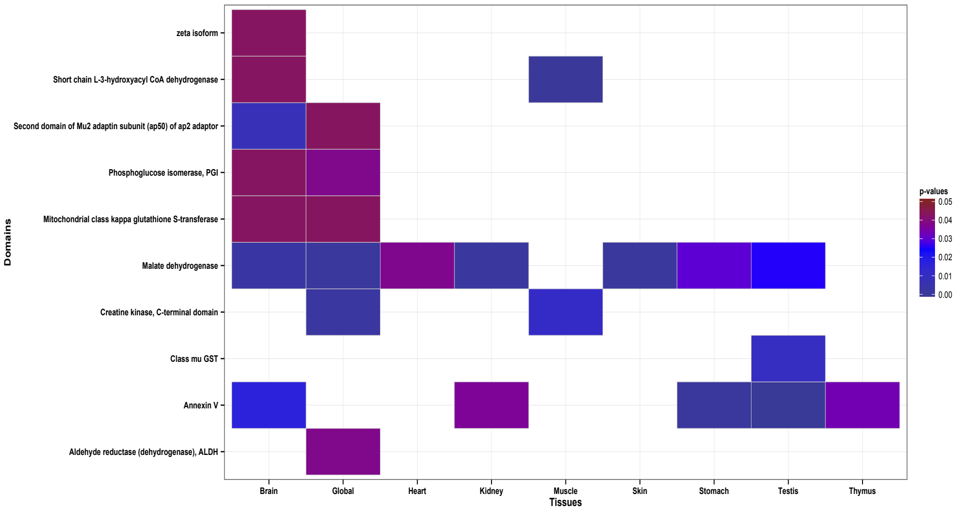
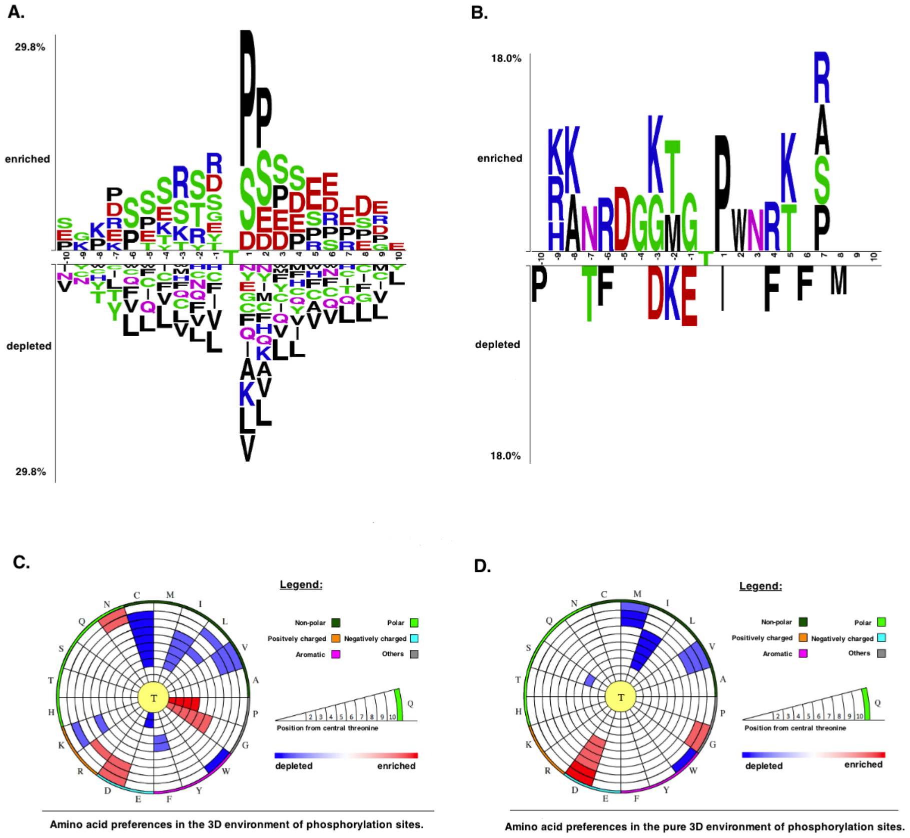


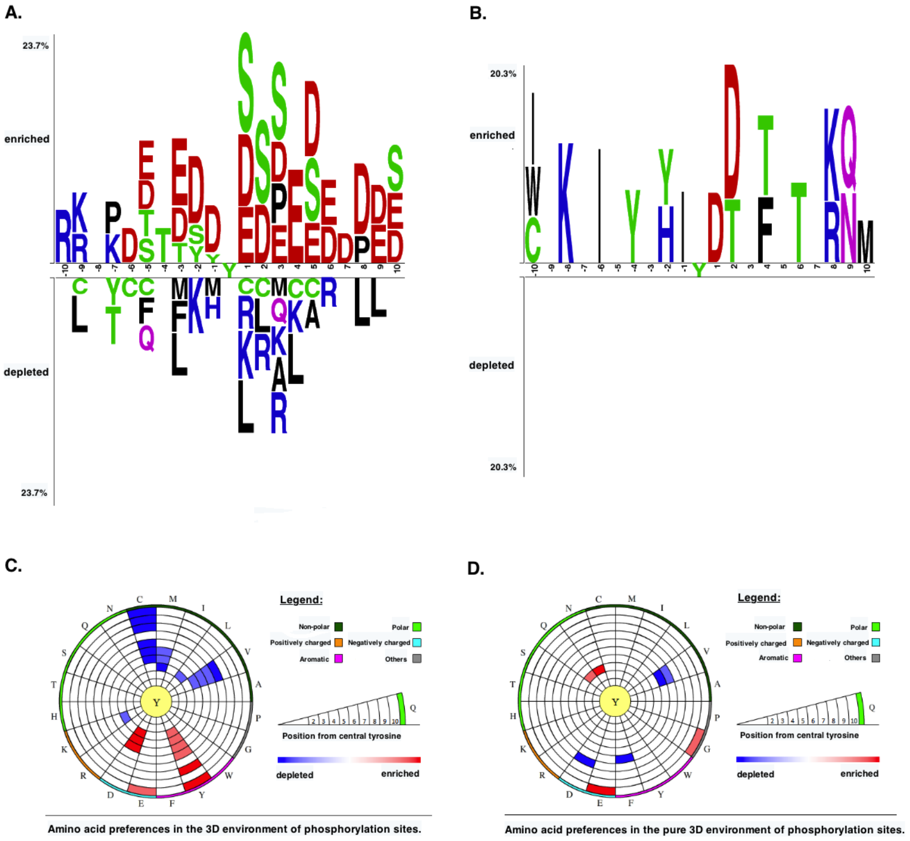
Figure 50: Global and tissue-specific occurrence of acetylated proteins from the LAS3D dataset in protein domains. Domains with a corrected p-value < 0.05 are considered significant.



Amino acid preferences in the 3D environment of phosphorylation sites.

Amino acid preferences in the pure 3D environment of phosphorylation sites.

Figure 51: Two sample logo analysis of global PTSs in the PS1D-70 dataset (A) and in the PS3D-90 dataset (B), 3D (C) and pure 3D (D) environments of PTSs in the PS3D-90 dataset.



Amino acid preferences in the 3D environment of phosphorylation sites.

Amino acid preferences in the pure 3D environment of phosphorylation sites.

Figure 52: Two sample logo analysis of global PYSs in the PS1D-70 dataset (A) and in the PS3D-90 dataset (B), 3D (C) and pure 3D (D) environments of PYSs in the PS3D-90 dataset.

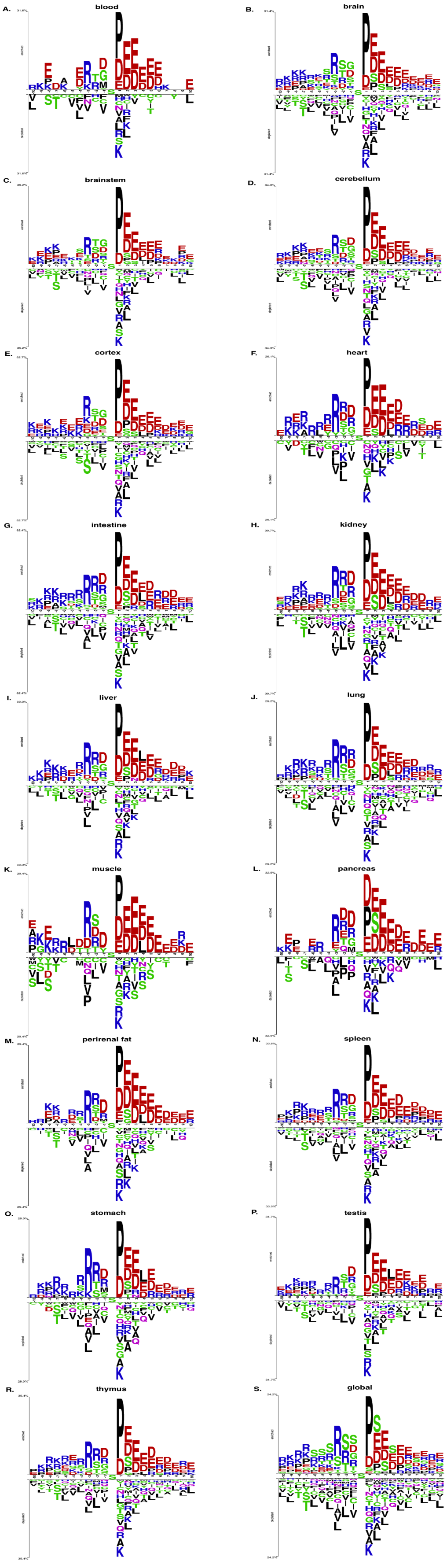


Figure 53: Two sample logo analysis of PSSs in different tissues in the PS1D-70 dataset (See Figure 54 - Figure 70 for high resolution graphs).

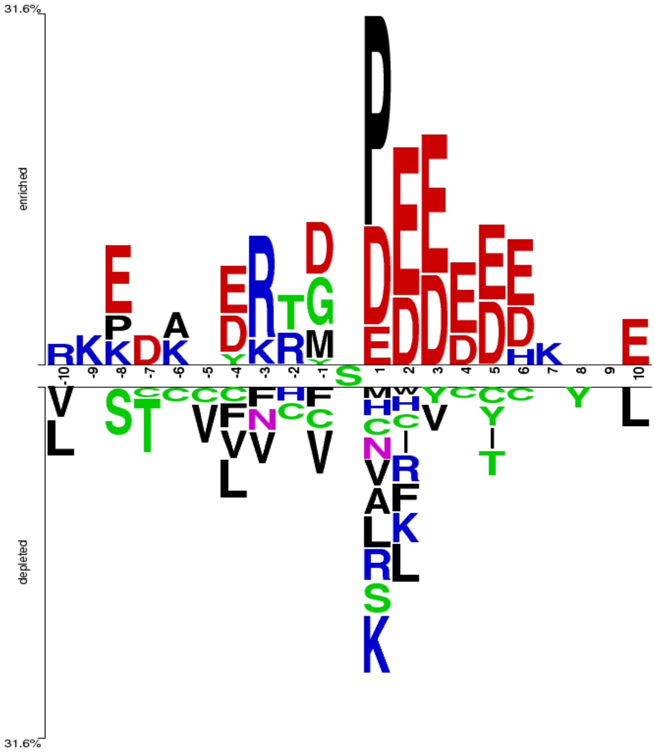


Figure 54: Two sample logo analysis of PSSs from the PS1D-70 dataset in blood.

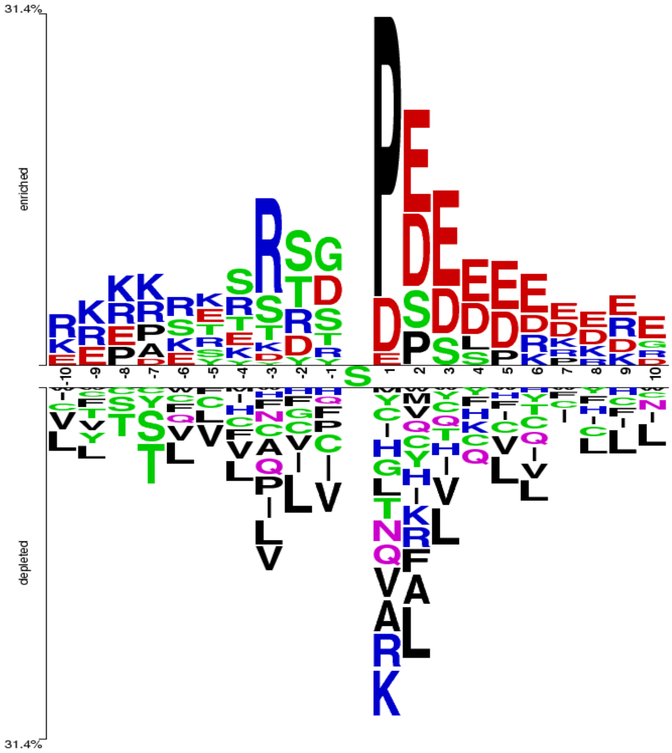


Figure 55: Two sample logo analysis of PSSs from the PS1D-70 dataset in brain.

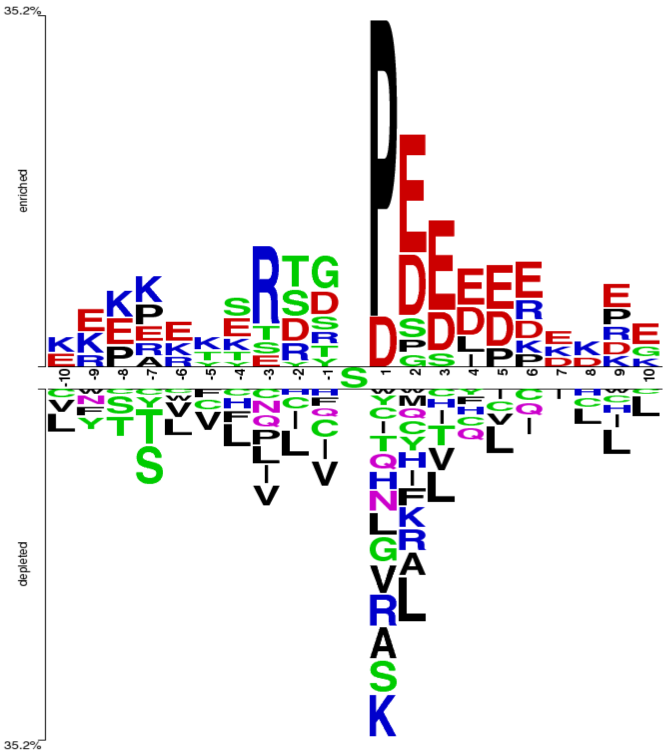


Figure 56: Two sample logo analysis of PSSs from the PS1D-70 dataset in brainstem.

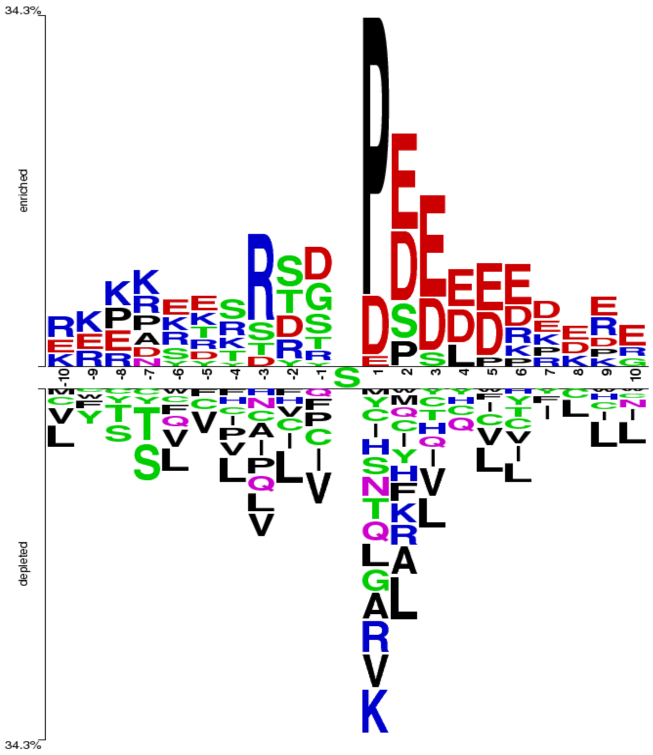


Figure 57: Two sample logo analysis of PSSs from the PS1D-70 dataset in cerebellum.

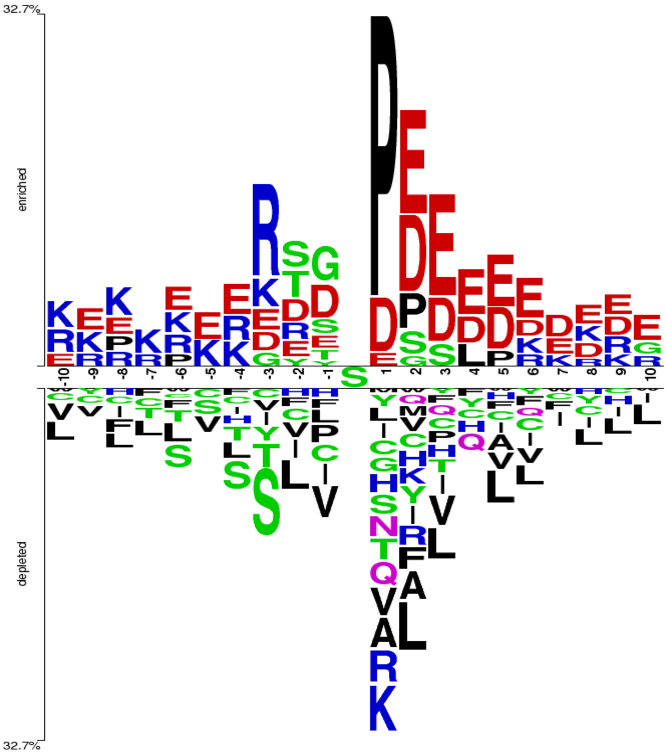


Figure 58: Two sample logo analysis of PSSs from the PS1D-70 dataset in cortex.

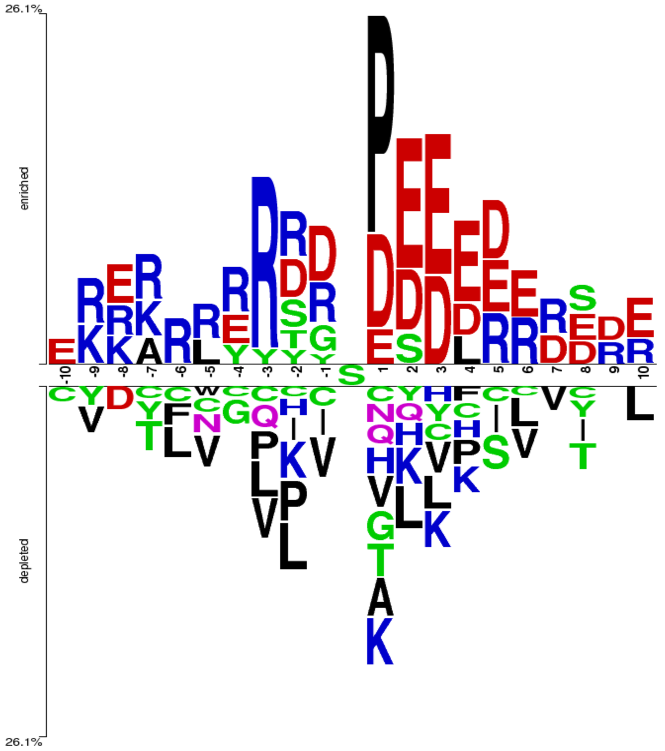


Figure 59: Two sample logo analysis of PSSs from the PS1D-70 dataset in heart.

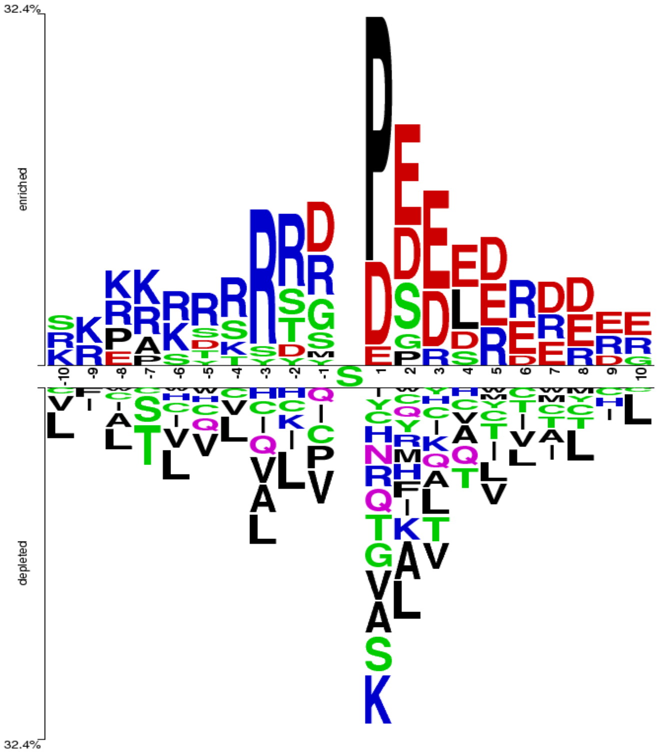


Figure 60: Two sample logo analysis of PSSs from the PS1D-70 dataset in intestine.

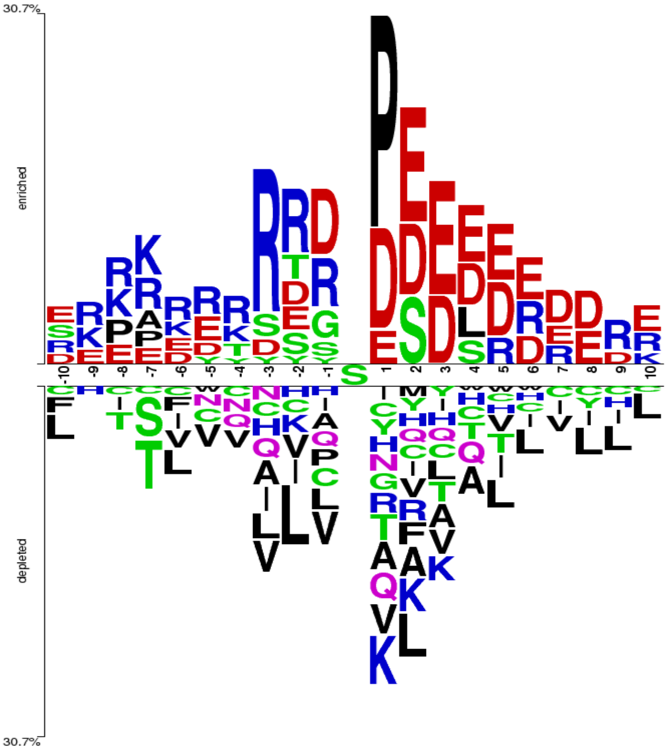


Figure 61: Two sample logo analysis of PSSs from the PS1D-70 dataset in kidney.

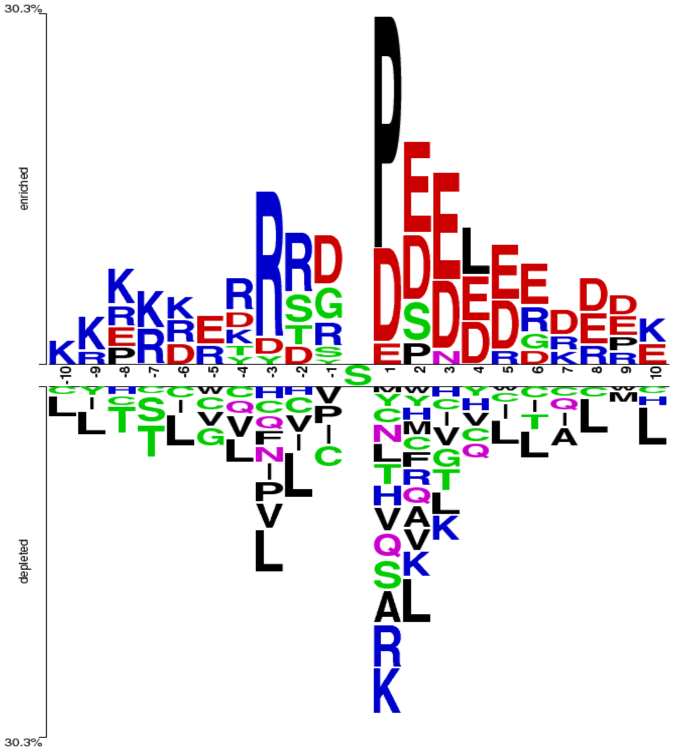


Figure 62: Two sample logo analysis of PSSs from the PS1D-70 dataset in liver.

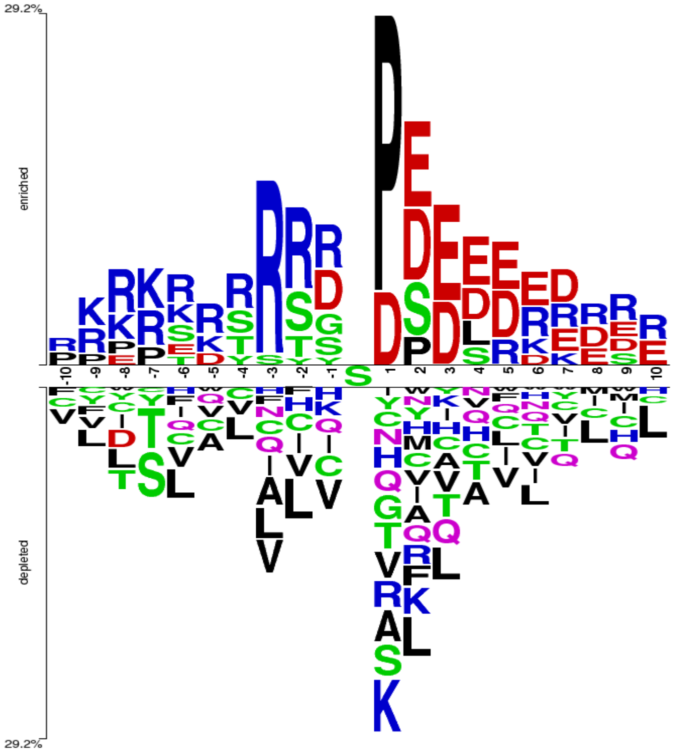


Figure 63: Two sample logo analysis of PSSs from the PS1D-70 dataset in lung.

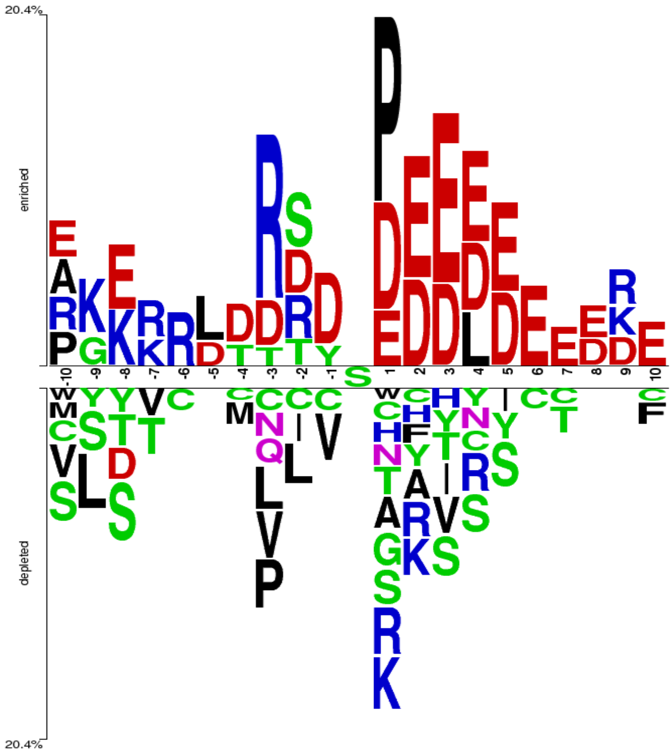


Figure 64: Two sample logo analysis of PSSs from the PS1D-70 dataset in muscle.

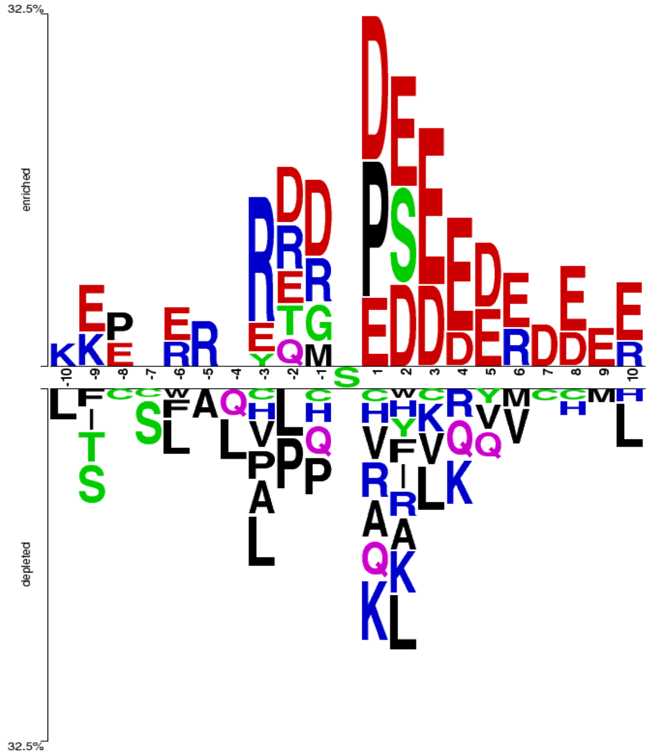


Figure 65: Two sample logo analysis of PSSs from the PS1D-70 dataset in pancreas.

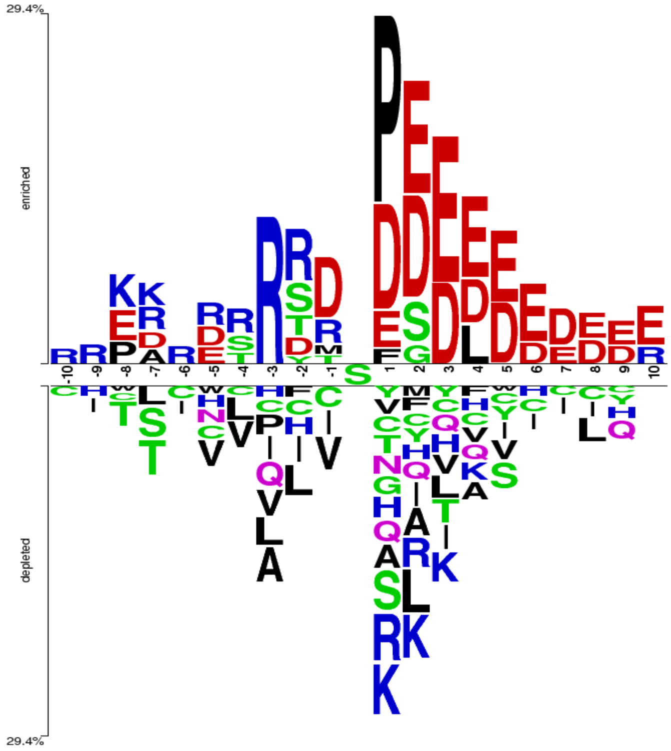


Figure 66: Two sample logo analysis of PSSs from the PS1D-70 dataset in perirenal fat.

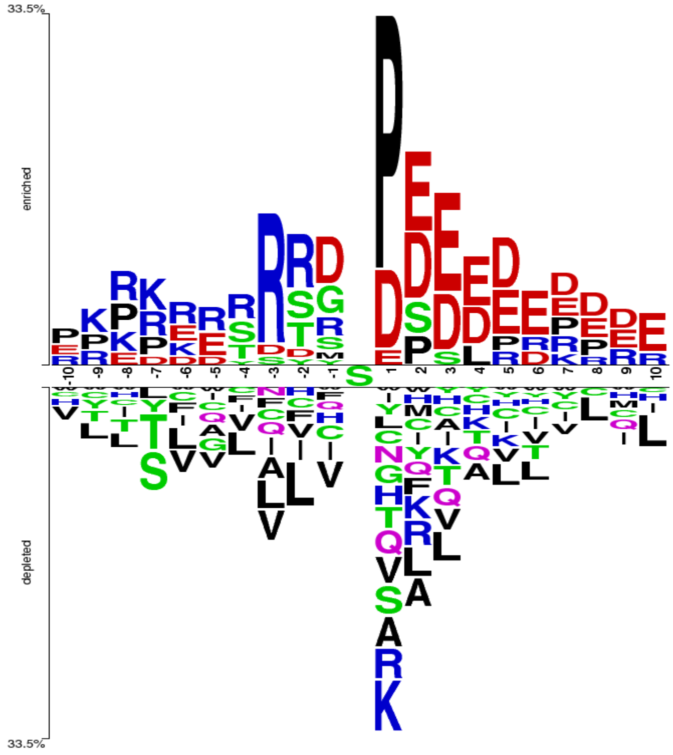


Figure 67: Two sample logo analysis of PSSs from the PS1D-70 dataset in spleen.

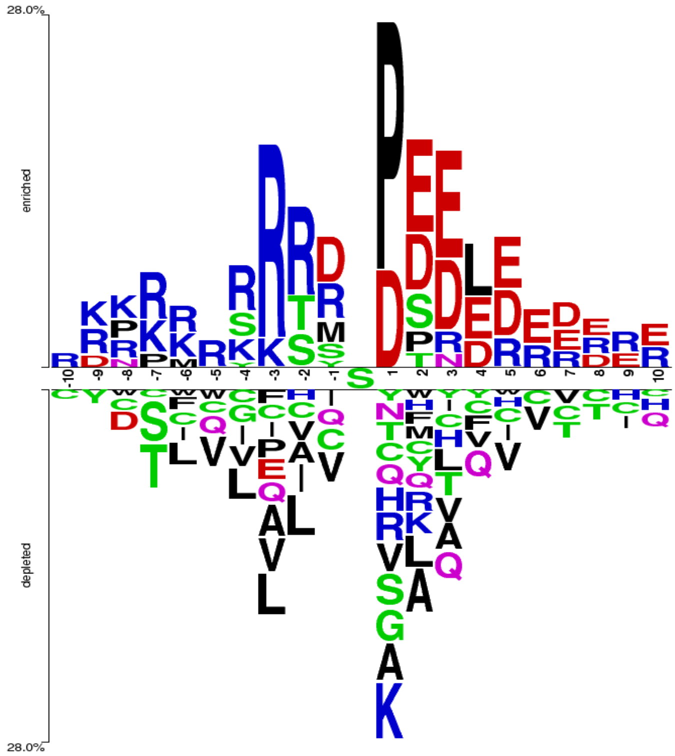


Figure 68: Two sample logo analysis of PSSs from the PS1D-70 dataset in stomach.

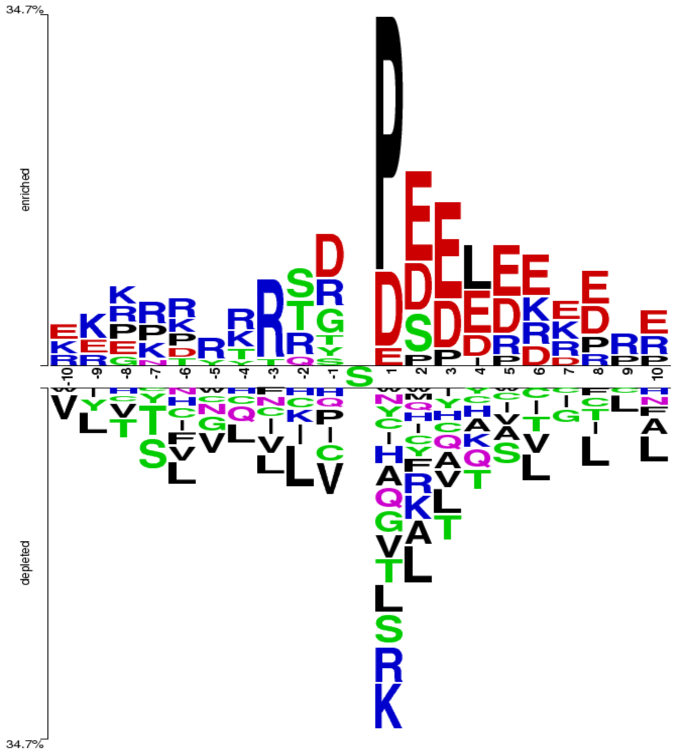


Figure 69: Two sample logo analysis of PSSs from the PS1D-70 dataset in testis.

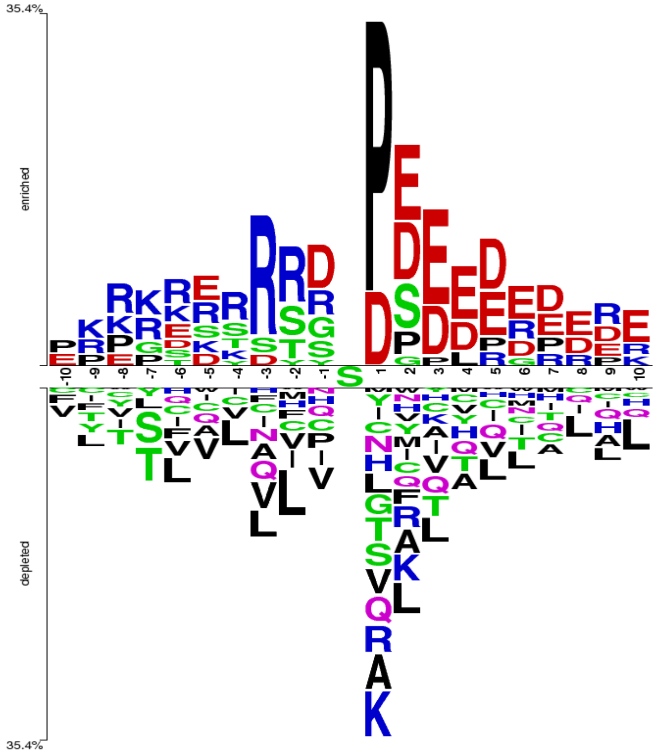


Figure 70: Two sample logo analysis of PSSs from the PS1D-70 dataset in thymus.

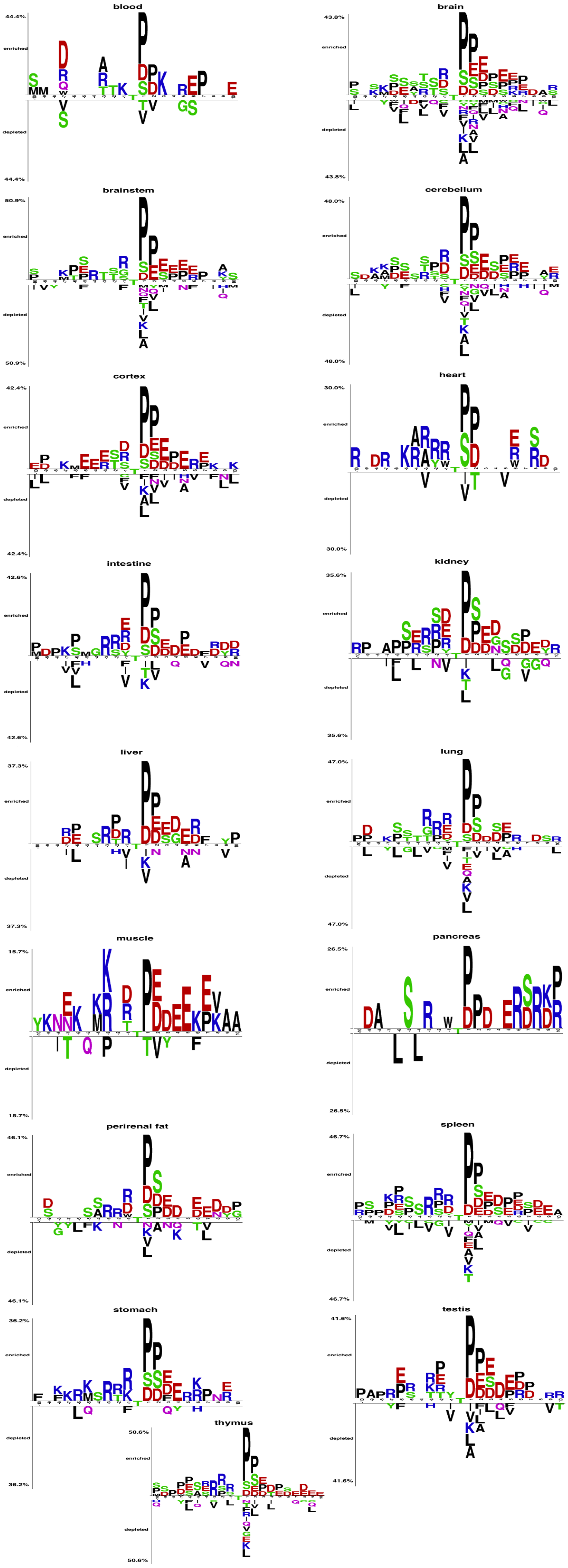


Figure 71: Two sample logo analysis of PTSs in different tissues in the PS1D-70 dataset.

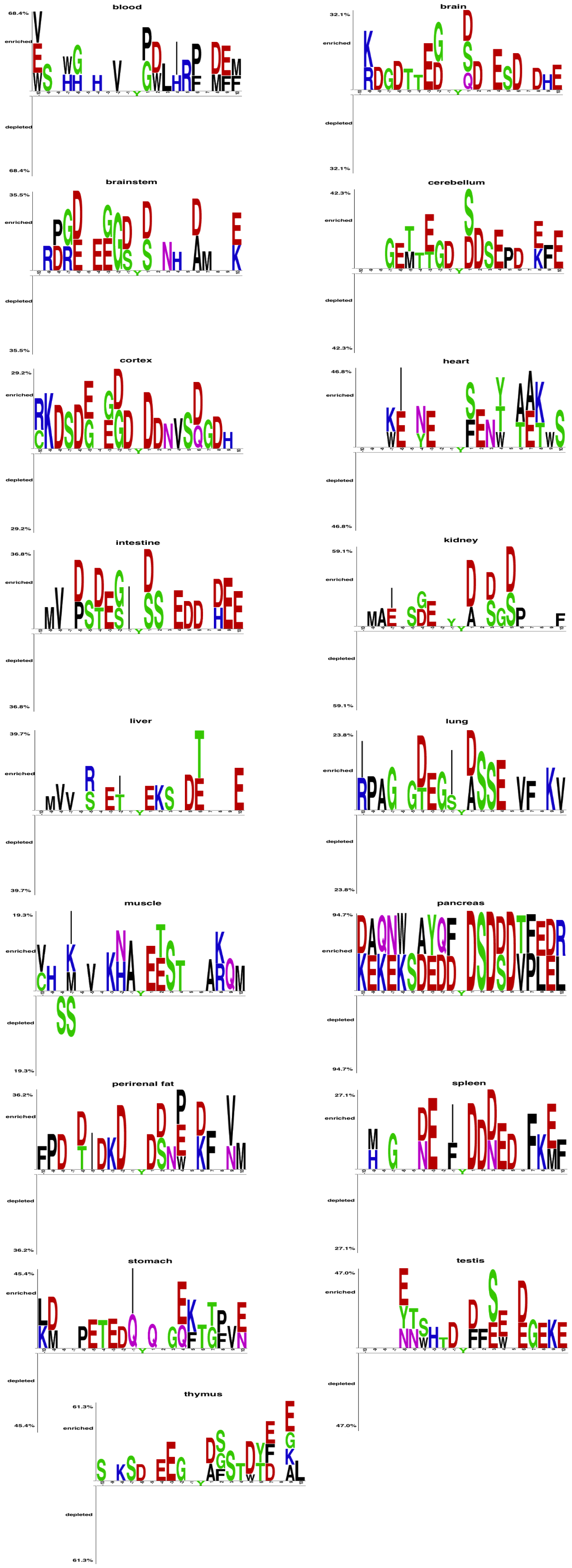


Figure 72: Two sample logo analysis of PYSs in different tissues in the PS1D-70 dataset.

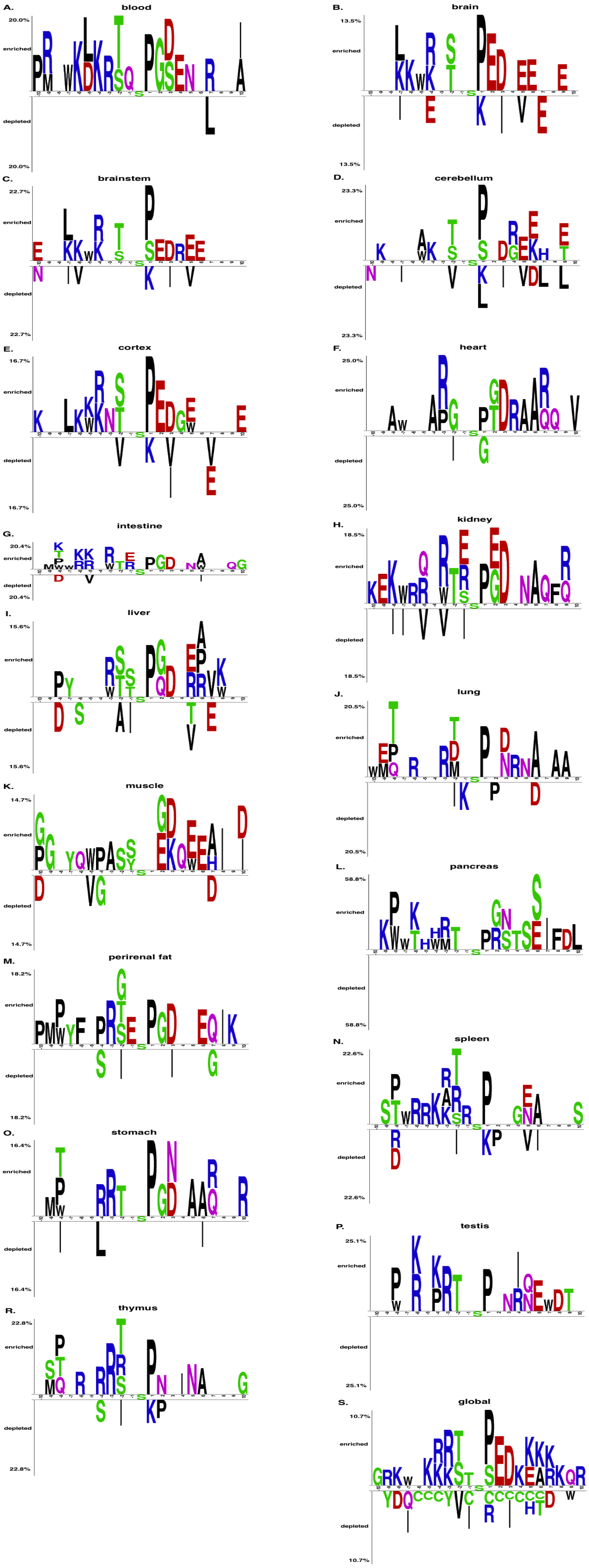


Figure 73: Two sample logo analysis of PSSs in different tissues in the PS3D-90 dataset.

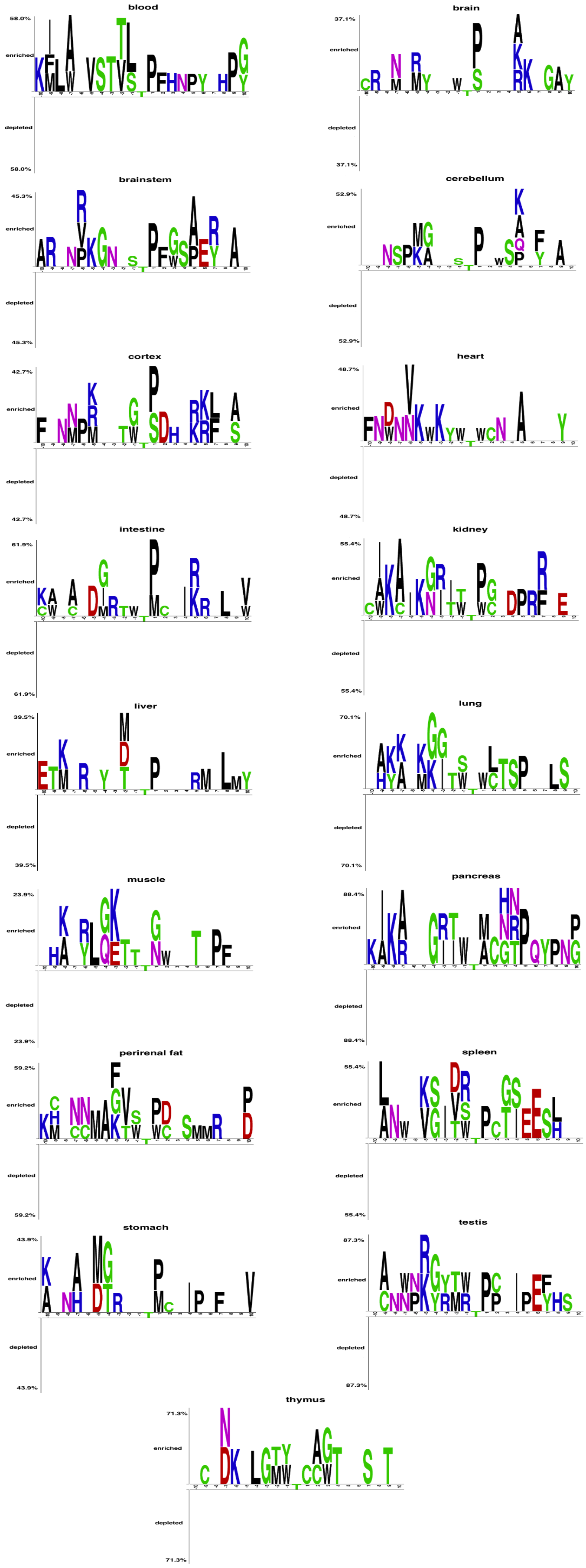


Figure 74: Two sample logo analysis of PTSs in different tissues in the PS3D-90 dataset.

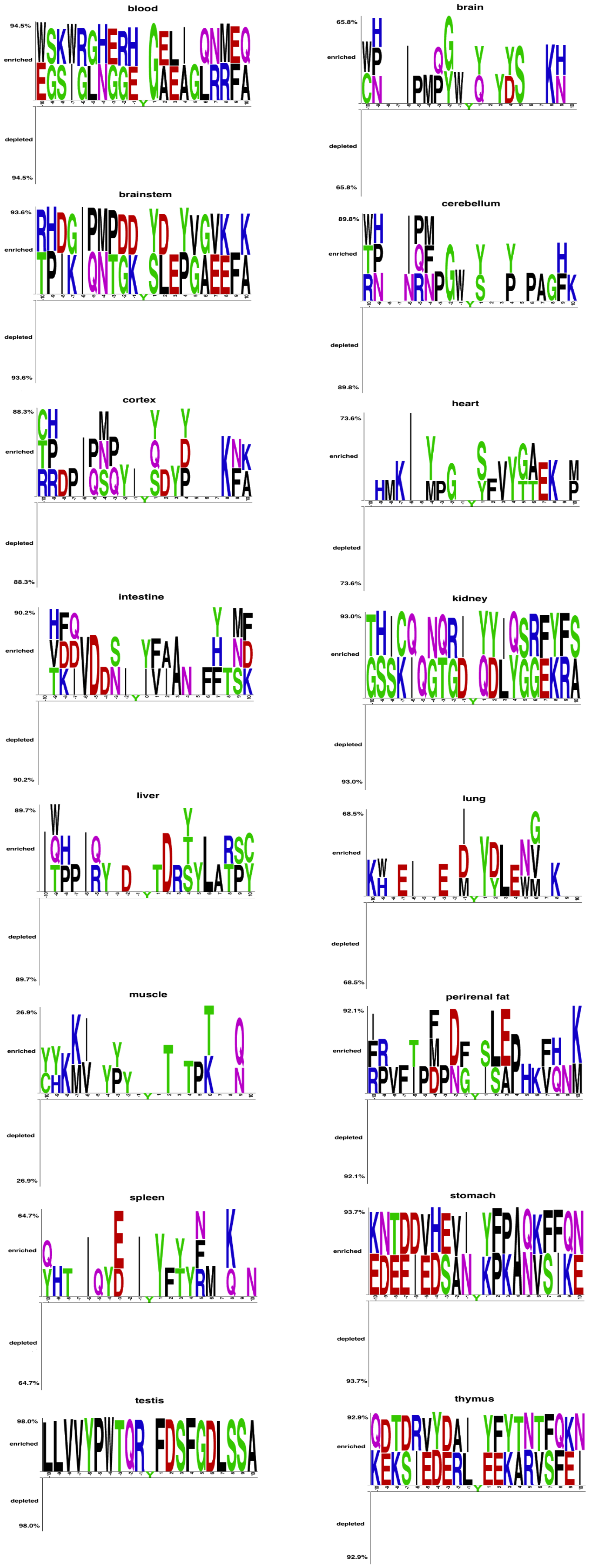


Figure 75: Two sample logo analysis of PYSs in different tissues in the PS3D-90 dataset.

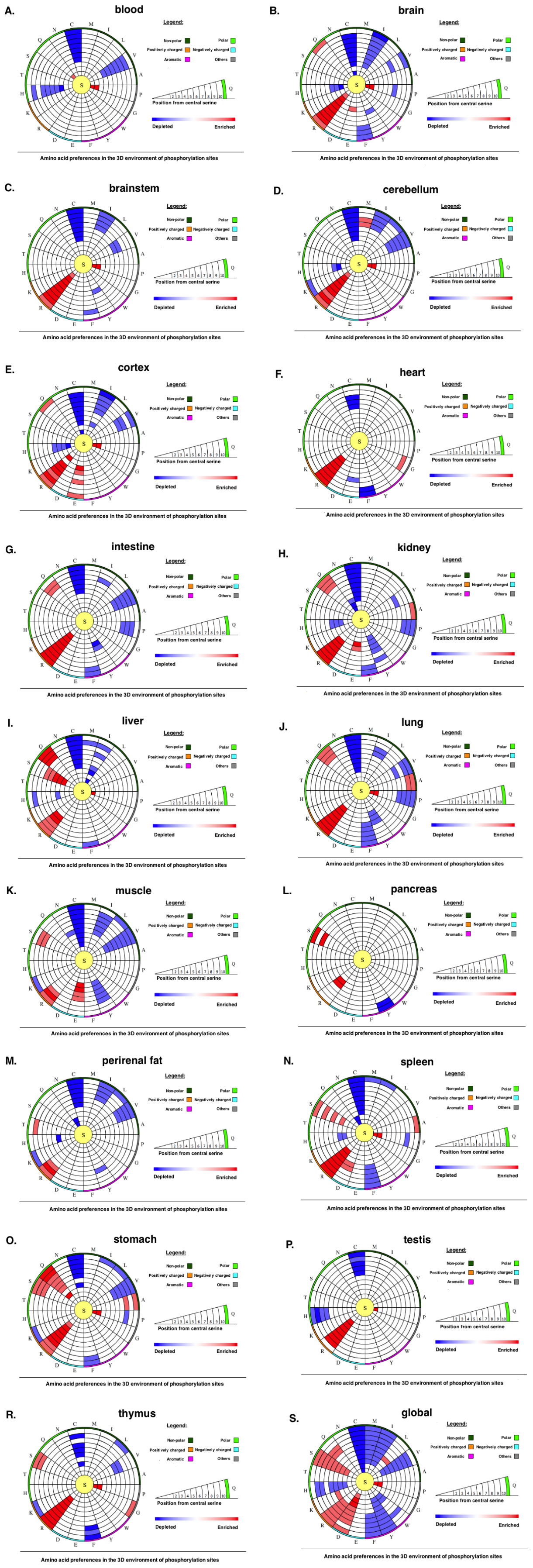


Figure 76: 3D environment of PSSs from the PS3D-90 dataset in different tissues.

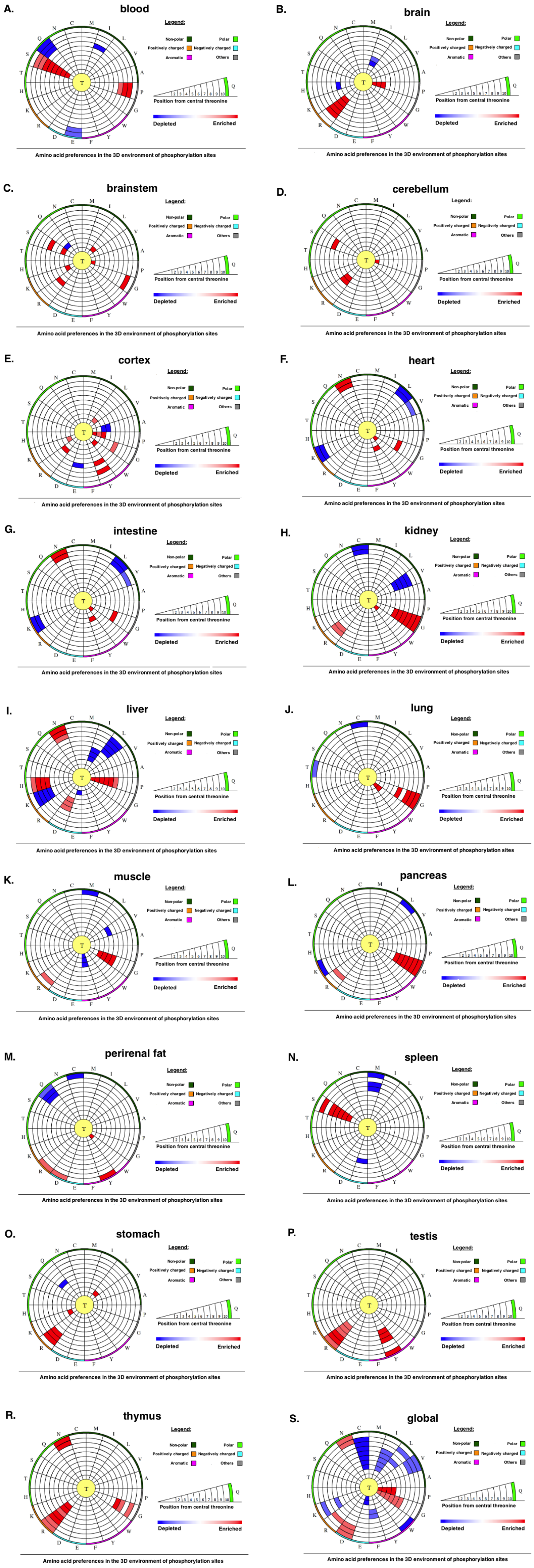


Figure 77: 3D environment of PTSs from the PS3D-90 dataset in different tissues.

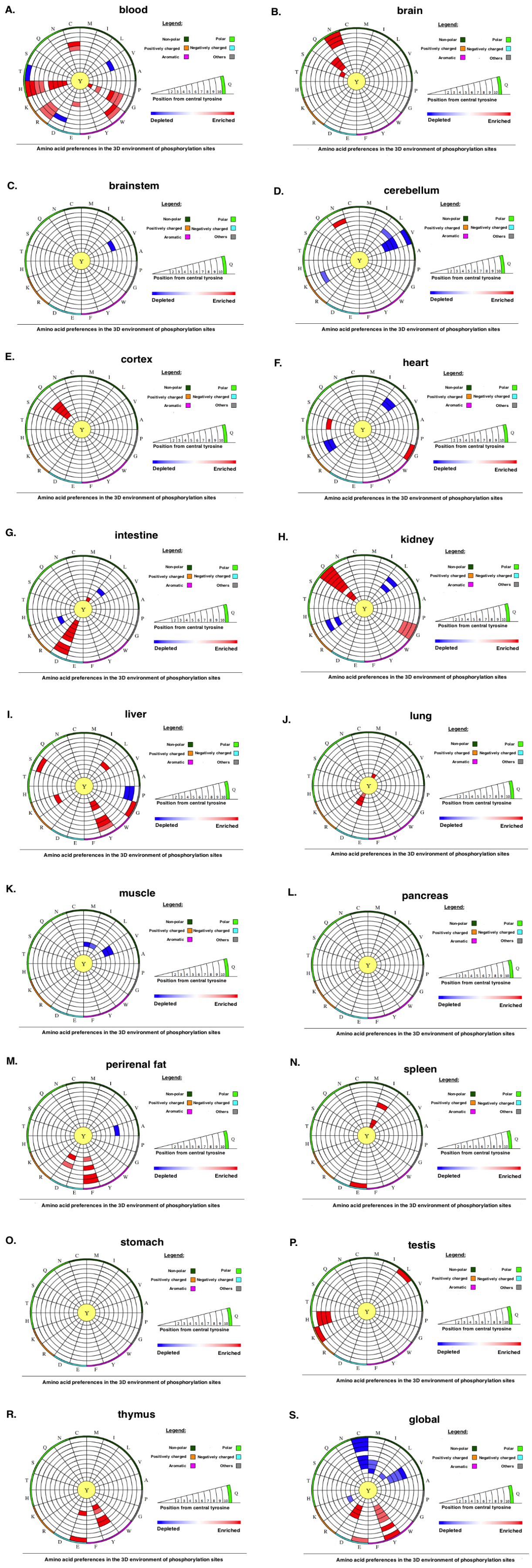


Figure 78: 3D environment of PYSs from the PS3D-90 dataset in different tissues.

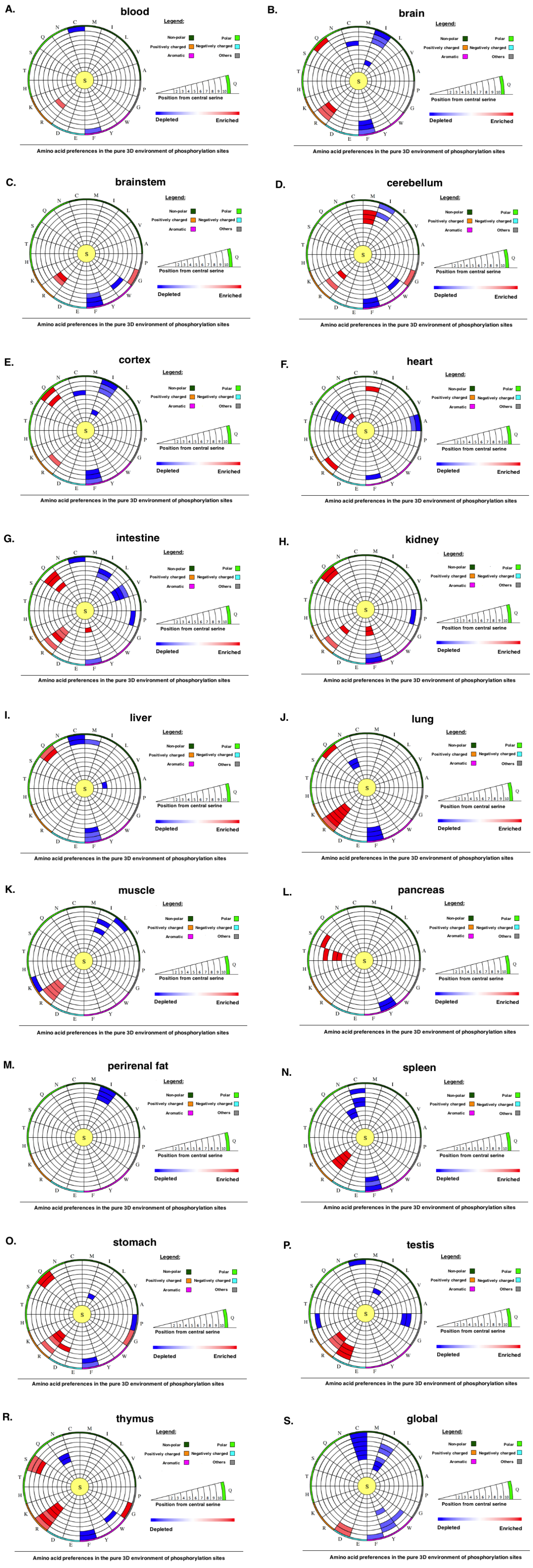


Figure 79: Pure 3D environment of PSSs from the PS3D-90 dataset in different tissues.

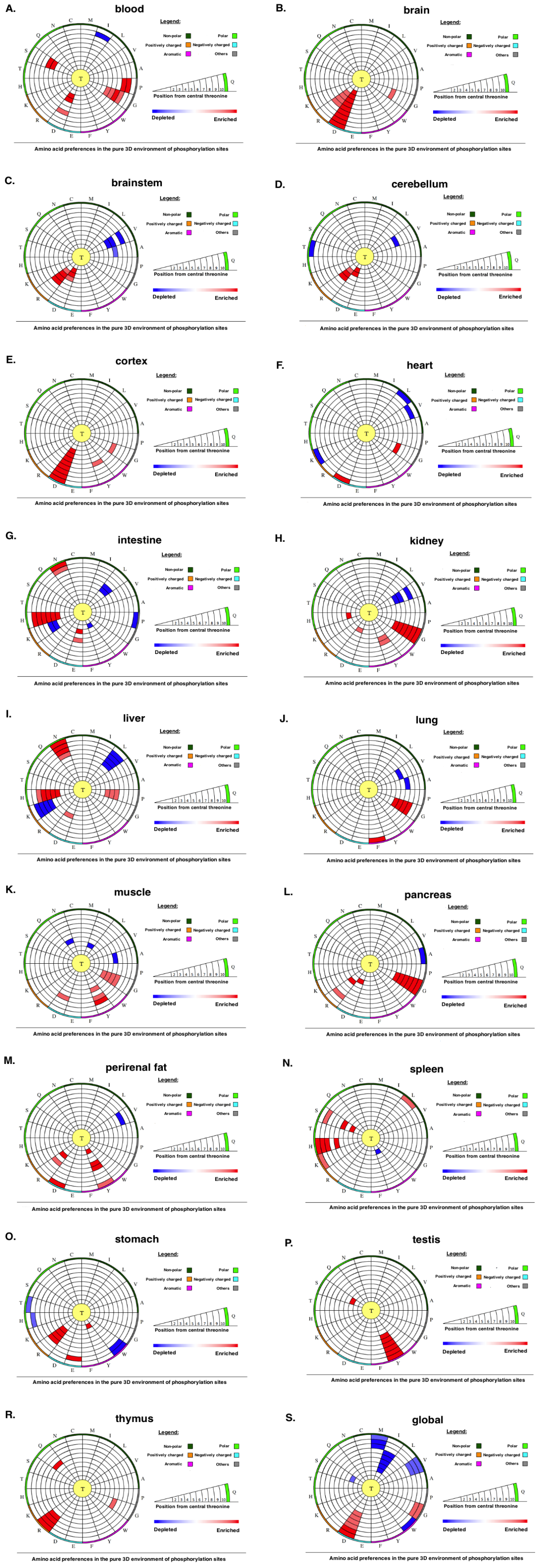


Figure 80: Pure 3D environment of PTSs from the PS3D-90 dataset in different tissues.

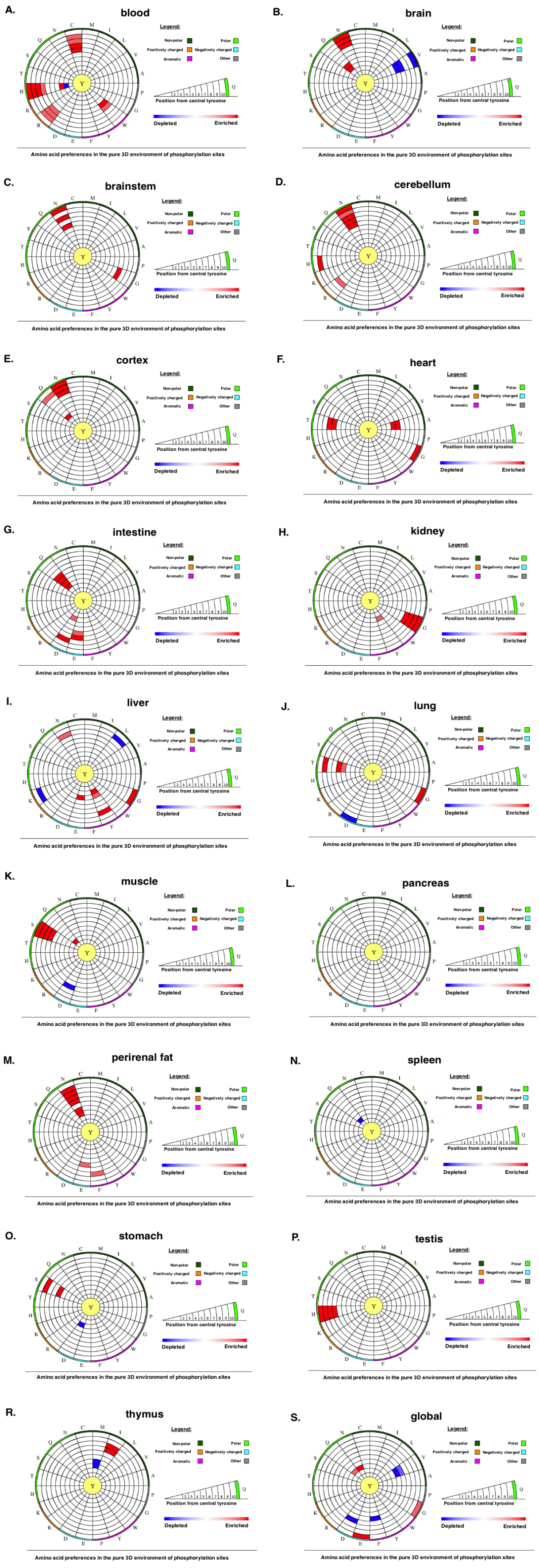


Figure 81: Pure 3D environment of PYSs from the PS3D-90 dataset in different tissues.

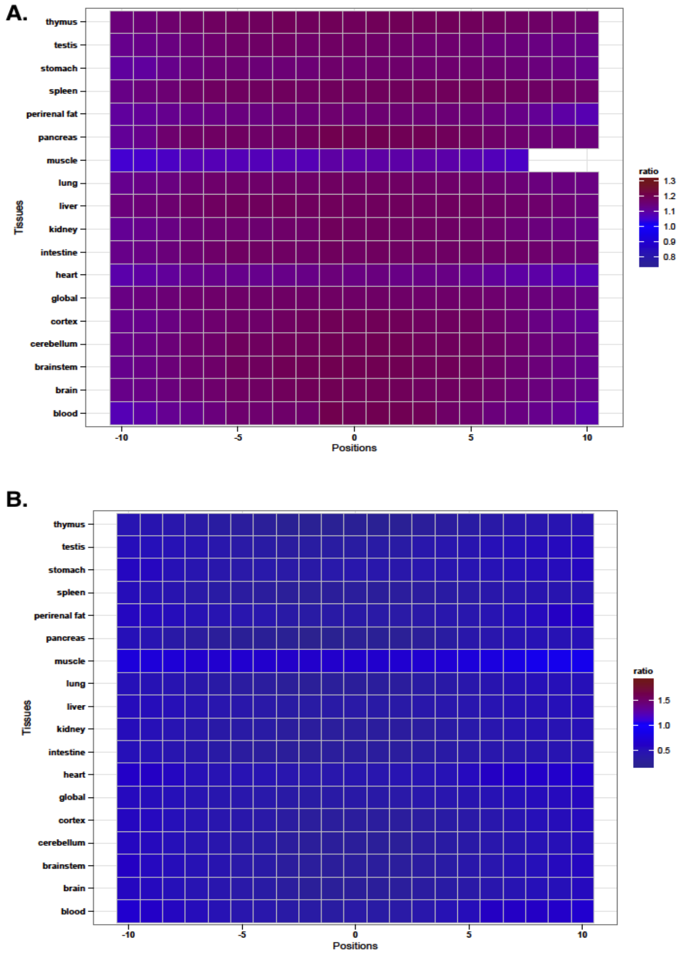


Figure 82: Disorder region analysis of PSSs and the residues surrounding them from the PS1D-70 dataset in different tissues. A. Ratio represented with shades of blue and red shows the normalized number of PSSs found in disorder regions divided by the normalized number of non-PSSs found in disorder regions. Non-significant ratios (p-value > 0.05) are represented with white cells. B. Ratio represented with shades of blue and red shows the normalized number of PSSs found in ordered regions divided by the normalized number of non-PSSs found in ordered regions. Non-significant ratios (p-value > 0.05) are represented with white cells. A and B are not complement of each other, because some phosphorylation sites with unknown regions also exist.

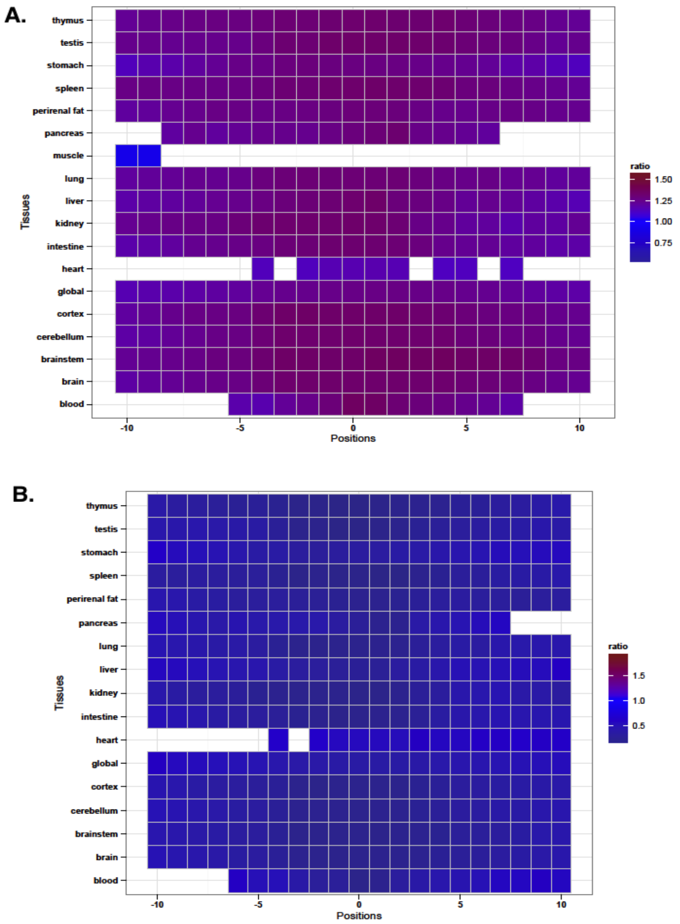


Figure 83: Disorder region analysis of PTSs and the residues surrounding them from the PS1D-70 dataset in different tissues. A. Ratio represented with shades of blue and red shows the normalized number of PTSs found in disorder regions divided by the normalized number of non-PTSs found in disorder regions. Non-significant ratios (p-value > 0.05) are represented with white cells. B. Ratio represented with shades of blue and red shows the normalized number of PTSs found in ordered regions divided by the normalized number of non-PTSs found in ordered regions. Non-significant ratios (p-value > 0.05) are represented with white cells. A and B are not complement of each other, because some phosphorylation sites with unknown regions also exist.

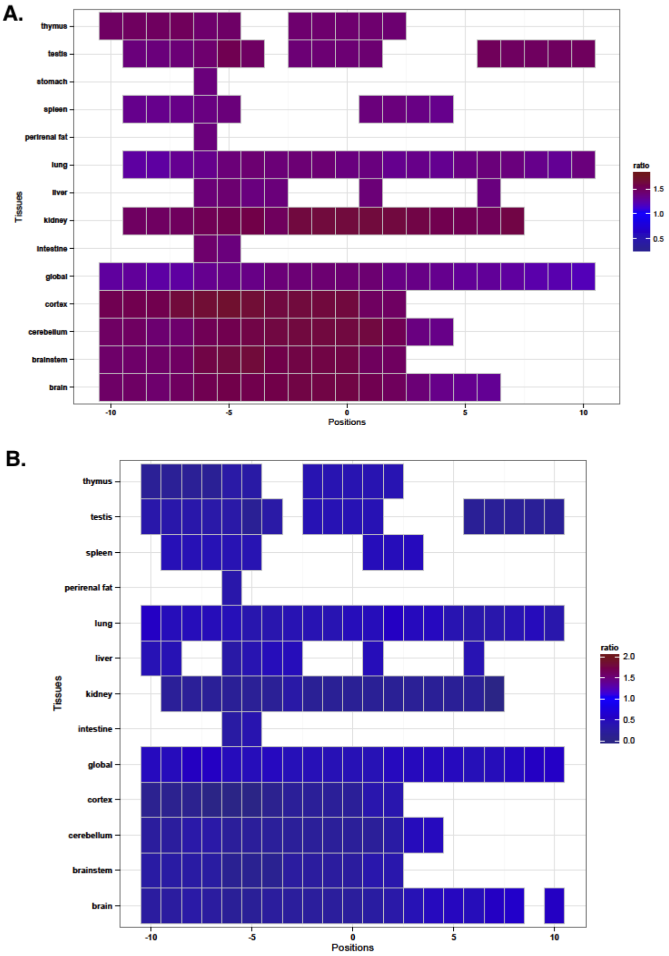


Figure 84: Disorder region analysis of PYSs and the residues surrounding them from the PS1D-70 dataset in different tissues. A. Ratio represented with shades of blue and red shows the normalized number of PYSs found in disorder regions divided by the normalized number of non-PYSs found in disorder regions. Non-significant ratios (p-value > 0.05) are represented with white cells. B. Ratio represented with shades of blue and red shows the normalized number of PYSs found in ordered regions divided by the normalized number of non-PYSs found in ordered regions. Non-significant ratios (p-value > 0.05) are represented with white cells. A and B are not complement of each other, because some phosphorylation sites with unknown regions also exist.

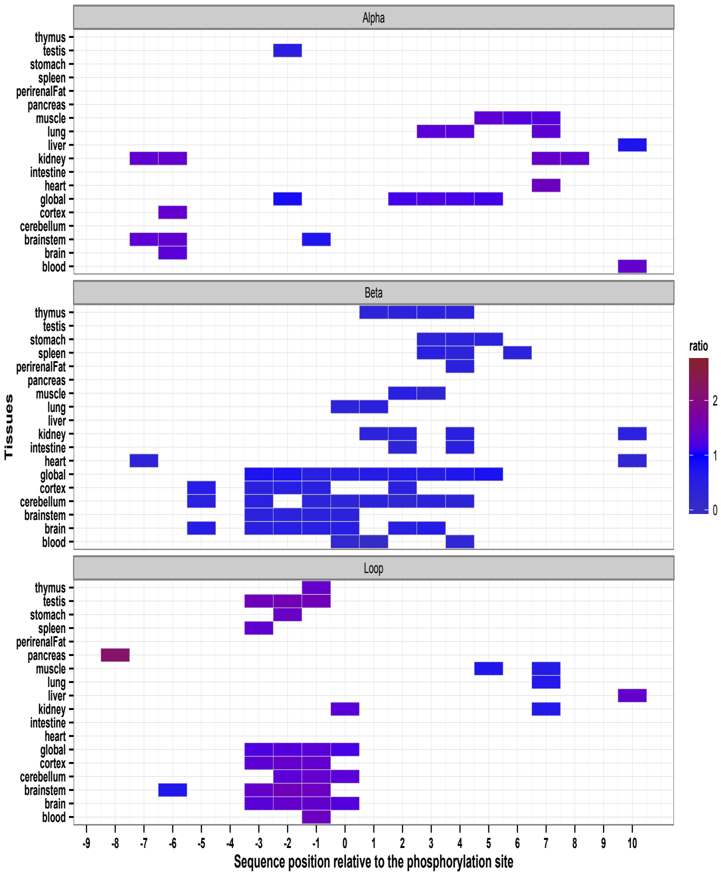


Figure 85: Secondary structure analysis of PSSs and the residues surrounding them from the PS3D-90 dataset in different tissues. Ratios represented with shades of blue and red show the normalized number of PSSs found in a particular secondary structure divided by the normalized number of non-PSSs found in the corresponding structure. Non-significant ratios (p-value > 0.05) are represented with white cells.

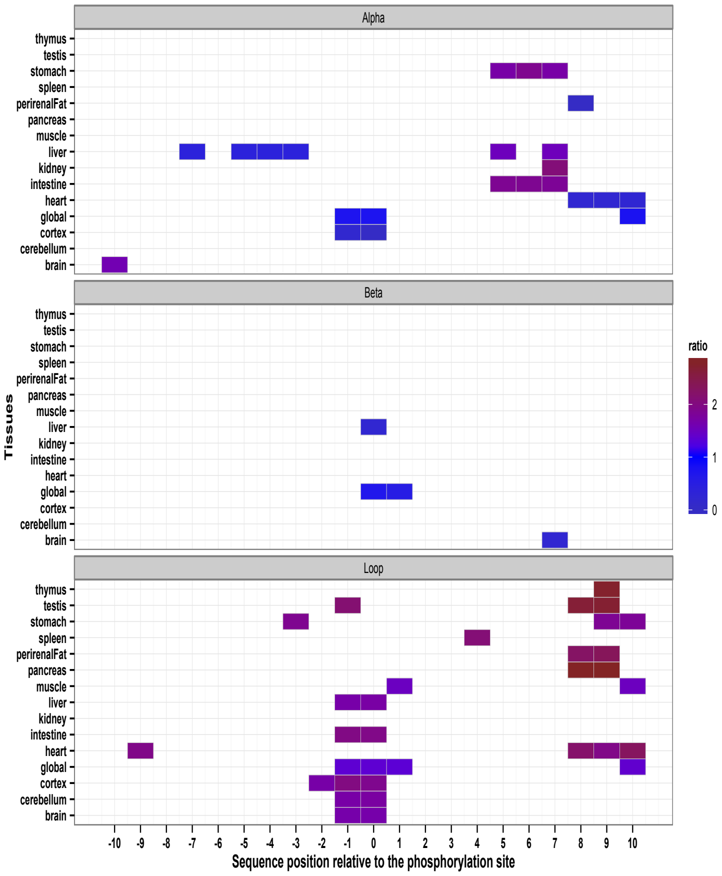


Figure 86: Secondary structure analysis of PTSs and the residues surrounding them from the PS3D-90 dataset in different tissues. Ratios represented with shades of blue and red show the normalized number of PTSs found in a particular secondary structure divided by the normalized number of non-PTSs found in the corresponding structure. Non-significant ratios (p-value > 0.05) are represented with white cells.

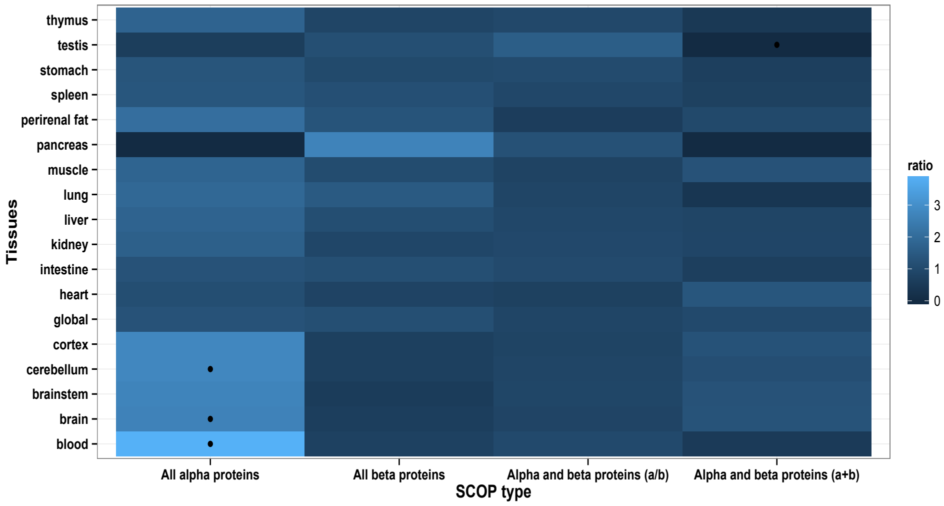


Figure 87: SCOP class analysis of serine phosphorylation sites and the residues surrounding them in different tissues in the PS3D-90 dataset. Ratio represented with shades of blue shows the normalized number of PSSs found in a particular protein structural class divided by the normalized number of non-PSSs found in the corresponding class. Black circles represent significant p-values ($p < 0.05$).

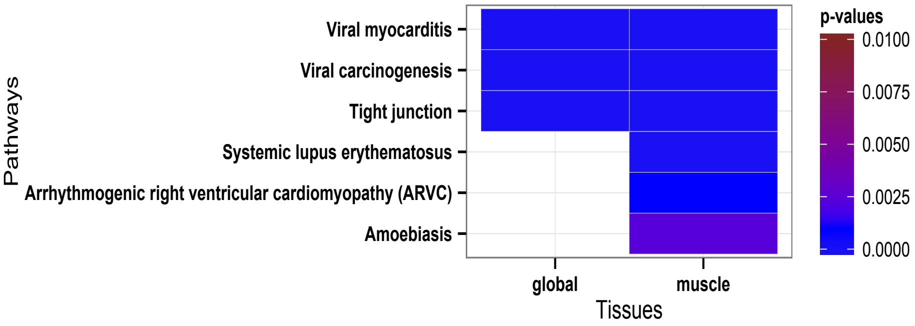


Figure 88: KEGG pathway analysis of the threonine phosphorylated proteins from the PS1D-70 dataset. Pathways with a corrected p-value < 0.01 in each tissue are considered significant.

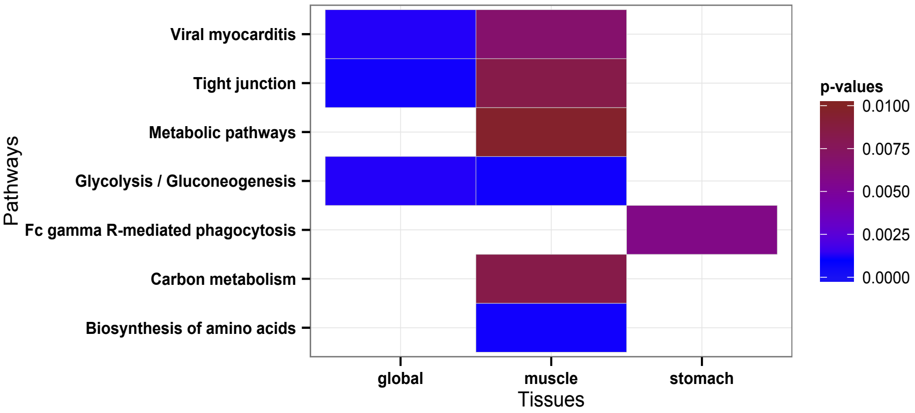


Figure 89: KEGG pathway analysis of the tyrosine phosphorylated proteins from the PS1D-70 dataset. Pathways with a corrected p-value < 0.01 in each tissue are considered significant.

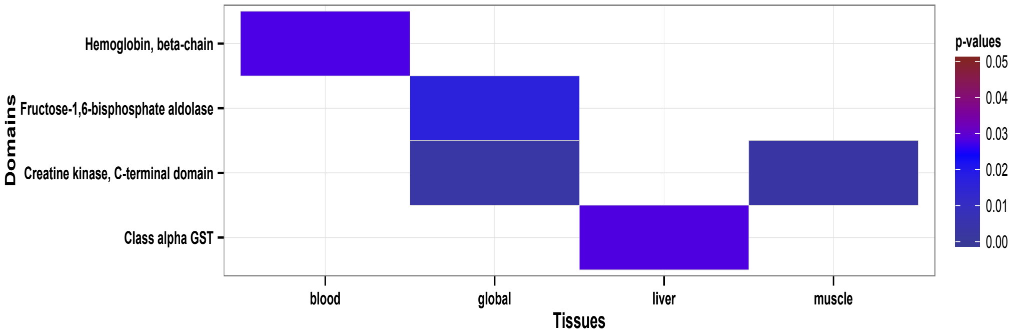


Figure 90: Global and tissue-specific occurrence of serine phosphorylated proteins from the PS3D-90 dataset in protein domains. Domains with a corrected p-value < 0.05 are considered significant.

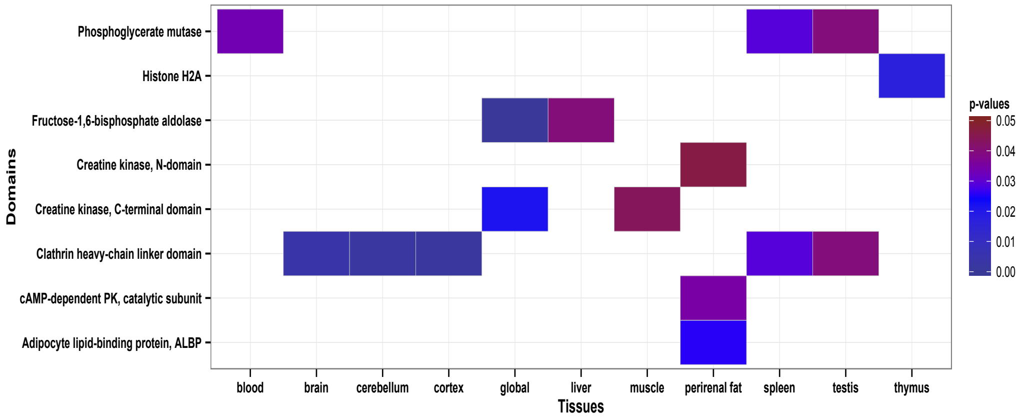


Figure 91: Global and tissue-specific occurrence of threonine phosphorylated proteins from the PS3D-90 dataset in protein domains. Domains with a corrected p-value < 0.05 are considered significant.

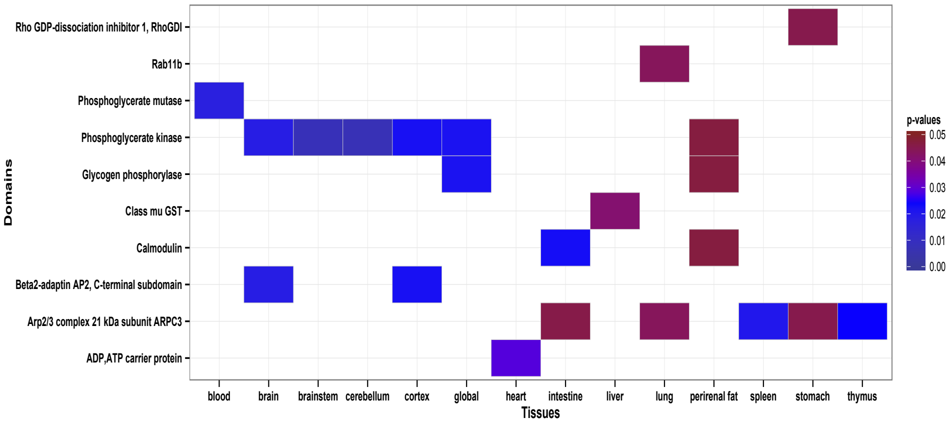


Figure 92: Global and tissue-specific occurrence of tyrosine phosphorylated proteins from the PS3D-90 dataset in protein domains. Domains with a corrected p-value < 0.05 are considered significant.

Tissue	Sequence motif	Motif score
Global	AcK-X-L-X-X-X-X-X-K	25.60
	L- AcK	16
	AcK-L	16
	G-AcK	13.91
	AcK-X-I	16
	AcK-I	16
	AcK-F	16
	AcK-X-F	15.65
	AcK-Y	16
	F-AcK	14.57
	I-AcK	12.07
	Y-AcK	15.11
	F-X-AcK	10.61
	F-X-X-X-AcK	10.27
	A-AcK	10.22
	F-X-X-X-X-AcK	10.39
	L-X-AcK	9.89
	I-X-AcK	9.86
	V-AcK	9.74
	AcK-V	9.35
	I-X-X-AcK	9.24
	L-X-X-AcK	8.97
	AcK-X-X-X-X-X-X-X-X-L	9.34
	Y-X-X-AcK	8.84
	AcK-X-X-F	7.83
	AcK-X-L	7.45
	AcK-X-X-X-X-F	8
	L-X-X-X-X-X-X-X-X-AcK	7.72
	L-X-X-X-X-X-X-X-AcK	6.89
	F-X-X-AcK	7.28
	AcK-X-X-X-X-X-X-X-L	7.31
	L-X-X-X-X-X-X-X-X-X-AcK	7.25
	AcK-X-X-X-X-X-X-L	8.17
	F-X-X-X-X-X-X-X-X-AcK	7
	AcK-X-V	6.70
	F-X-X-X-X-X-X-X-X-X-AcK	7.35
	AcK-X-X-I	7
	AcK-X-X-X-X-X-X-X-X-F	6.99
	AcK-X-X-X-X-X-F	6.26
	Y-X-X-X-X-X-AcK	6.41
Brain	AcK-X-I	16
	AcK-F	16
	AcK-Y	12.40
	AcK-L	11.72
	F-AcK	11.11
	L-AcK	10.09
	F-X-AcK	11.04
	AcK-X-F	9.04
	F-X-X-X-X-AcK	9.41
	AcK-X-V	9.10
	Y-X-AcK	8.77
	L-X-AcK	7.03
	F-X-X-X-AcK	9.01
	I-AcK	7.16
	AcK-I	8.48

	F-X-X-X-X-X-AcK	7.56
	L-X-X-AcK	7.87
	I-X-X-AcK	7.90
	AcK-X-L	8.27
Brown fat	F-AcK	7.42
	D-X-X-AcK	6.43
	AcK-X-F	6.18
	G-AcK	6.58
Heart	AcK-F	10.80
	AcK-X-I	9.91
	L-AcK	9.86
	L-X-AcK	9.69
	AcK-L	6.72
	AcK-I	8.54
	L-X-X-X-X-AcK	6.55
	F-X-X-X-X- AcK	6.65
	M-X-X-X-X-X-X-X-AcK*	7.44
	AcK-Y-X-X-X-X-X-X-L*	14.31
Intestine	G-AcK	11.20
	E-X-AcK-Y*	17.69
	AcK-L	8.40
	AcK-X-I	7.21
	AcK-X-V	7.56
	L-AcK	8.08
Kidney	L-X-AcK	11.21
	AcK-X-I	16
	AcK-F	16
	AcK-X-L-X-X-X-X-X-K	21.62
	L-AcK	15.95
	AcK-L	16
	AcK-X-F	11.73
	AcK-Y	9.76
	I-X-AcK	9.00
	V-X-AcK*	9.81
	F-X-AcK	9.46
	Y-X-AcK	9.74
	F-AcK	7.46
	AcK-I	9.74
	AcK-X-X-L*	7.22
	L-X-X-AcK	7.38
	F-X-X-X-AcK	7.94
	AcK-X-X-X-X-X-X-X-X-X-L	8.20
	AcK-X-X-F	7.70
	AcK-X-X-X-X-I	7.52
	AcK-X-X-I*	7.62
	AcK-X-X-X-X-L*	7.82
Liver	AcK-F	13.15
	L-X-AcK	10.82
	AcK-L	14.37
	AcK-X-I	8.97
	L-X-X-X-X-AcK	7.06
	AcK-X-F	8.82
	AcK-X-L	6.91
	AcK-Y	7.22
	F-X-X-X-X-AcK	6.77
	AcK-X-X-X-X-F	7.41
	AcK-X-X-X-F	8.37

	F-X-X-X-AcK	8.67
	L-AcK	10.06
	F-X-X-X-X-X-X-AcK	7.33
	L-X-X-X-X-X-X-X-X-X-AcK	7.32
	Y-X-AcK	8.14
	F-X-AcK	7.92
	AcK-X-X-X-L	6.83
	AcK-X-X-X-X-X-L	6.45
Lung	AcK-Y	11.14
	F-AcK	11.35
	G-AcK-X-X-X-D*	19.03
	AcK-F	9.80
	AcK-X-I	9.03
	L-AcK	8.80
	AcK-L	8.94
	G-AcK	8.99
Muscle	L-X-AcK	8.42
	AcK-Y	7.72
	L-X-X-X-X-AcK	7.75
	D-X-AcK*	7.16
	AcK-X-I	7.55
	D-X-X-X-X-X-X-AcK-X-L*	14.10
	D-X-X-AcK	7.65
Pancreas	AcK-L	7.97
	AcK-X-A	6.20
Perirenal fat	AcK-Y	6.05
	L-X-X-X-X-AcK	6.13
	F-AcK	7.21
Skin	AcK-Y	11.54
	L-X-AcK	11.94
	AcK-F	10.73
	L-AcK	9.71
	L-D-X-AcK*	13.42
	AcK-X-I	8.37
	AcK-X-L	12.81
	AcK-L	7.06
	AcK-X-X-Y*	6.03
Spleen	AcK-Y	7.36
	G-AcK	7.88
	F-AcK	6.32
	AcK-X-A	6.05
	AcK-X-I	6.32
Stomach	L-X-AcK	15.35
	AcK-X-I	16
	AcK-F	11.64
	AcK-L	10.04
	AcK-X-F	7.39
	F-X-X-X-X-AcK	8.48
	L-AcK	9.98
	AcK-Y	7.52
	G-AcK	16
	AcK-X-X-X-X-X-X-X-F	7.49
	AcK-X-X-X-X-X-L	8.70
	AcK-I	8.55
	V-X-AcK-X-L*	15.81
	Y-AcK	7.76
	F-X-X-X-X-X-X-X-X-X-AcK	7.43

	F-X-X-X-AcK	8.46
	F-AcK	7.06
	F-X-X-X-X-X-X-AcK	7.48
	AcK-X-X-X-X-X-X-X-X-X-F*	7.79
	L-X-X-X-X-X-AcK	7.40
	L-X-X-X-X-AcK	7.06
	I-X-X-X-X-X-X-X-X-X-AcK*	7.73
	L-X-X-X-X-X-X-X-X-X-AcK	7.97
	L-X-X-X-X-X-X-X- AcK	9.57
	AcK-X-X-X-X-X-X-X-L	8.39
	AcK-X-M*	7.38
	AcK-X-X-X-X-X-X-F	6.97
	I-X-X-X-X-X-X-X-AcK*	6.80
	I-X-X-X-X-X-AcK*	7.01
	I-X-X-X-X-AcK*	6.03
	AcK-X-X-X-X-X-X-I*	7.01
	AcK-X-X-X-X-X-X-X-X-F	6.19
Testis	AcK-L	11.97
	AcK-F	11.86
	L-AcK	7.63
	AcK-X-F	11.83
	F-X-X-X-X-AcK	9.71
	L-X-AcK	7.95
	AcK-Y	8.28
	F-X-AcK	8.46
	AcK-X-L	10.71
	AcK-X-X-F	8.54
	AcK-X-X-X-X-X-X-X-F	8.34
	F-X-X-X-X-X-X-X-X-X-AcK	7.30
	Y-X-X-AcK*	7.92
	F-X-X-X-AcK	8.41
	L-X-X-X-X-AcK	10.48
	AcK-X-I	13.09
	AcK-X-V	7.4
	L-X-X-X-X-X-AcK	9.5
	L-X-X-X-X-X-X-X-AcK	6.79
	AcK-X-X-X-X-X-X-X-L	8.93
	AcK-X-X-X-X-X-X-X-X-L	8.86
	L-X-X-X-X-X-X-AcK*	8.11
	AcK-X-X-X-X-X-X-X-X-X-X-L*	9.39
	AcK-X-X-X-X-X-X-X-Y*	7.47
	F-AcK	6.79
	AcK-X-X-X-X-X-X-L	6.18
	L-X-X-X-AcK*	6.73
	AcK-X-X-X-X-X-X-X-X-Y*	6.7
Testis fat	AcK-F	7.95
	AcK-X-I	6.92
	AcK-L	6.97
	L-X-AcK	8.25
	L-AcK	7.39
Thymus	G-AcK	10.38
	AcK-X-I	14.04
	L-AcK	8.89
	F-AcK	8.52
	AcK-Y	7.42
	AcK-L	6.59
	AcK-X-F	6.74

	AcK-I	6.56
	AcK-F	7.05
	AcK-X-X-X-X-X-I*	6.02

Table 12: Summary of sequence motifs associated with LASs in the LAS1D dataset. Motifs in bold correspond to tissue-specific motifs. Motifs in bold with stars (*) are not observed in global LASs. AcK and X represent acetylated lysine residues and wildcard residues, respectively. Only statistically significant motifs are shown ($p < 0.000001$).

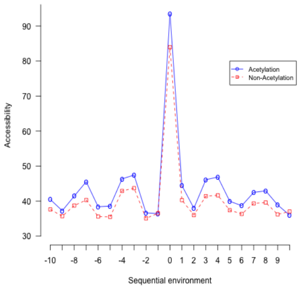
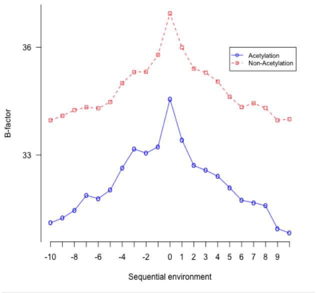
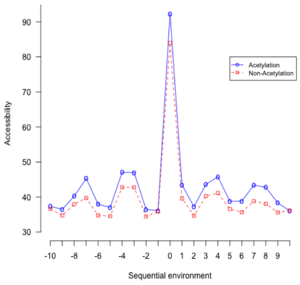
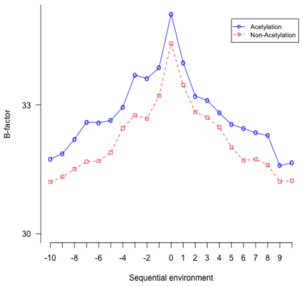
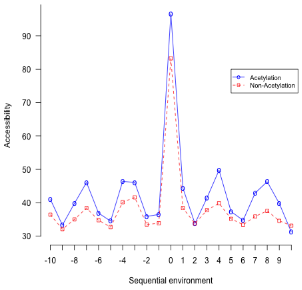
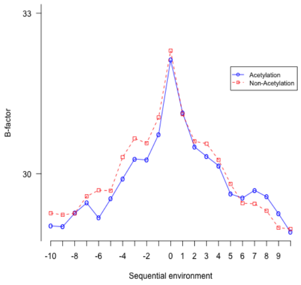
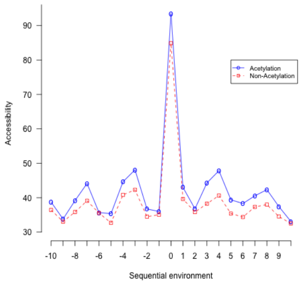
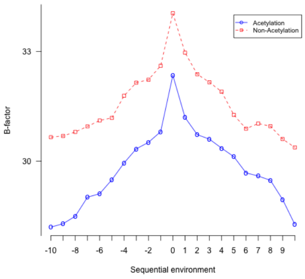
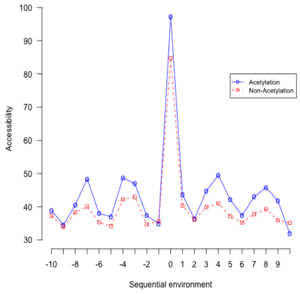
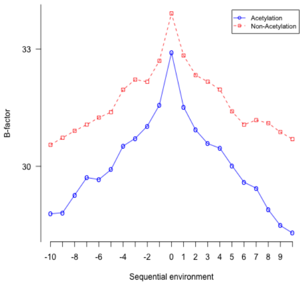
Tissue	Sequence motif	Motif score
Global	K-X-X-X-X-X-AcK-X-X-X-X-X-X-X-X-D	8.40
	AcK-X-D-X-X-X-X-X-K	9.76
	AcK-F-X-X-X-X-K	8.19
	K-X-D-X-X-X-X-X-AcK	9.67
	K-X-X-X-X-X-AcK-X-X-X-T	7.44
	D-X-X-AcK	4.07
	AcK-X-X-X-X-X-X-X-K	3.77
	K-X-X-X-X-AcK-X-X-X-X-X-X-X-X-X-F	10.94
	K-X-X-X-X-AcK	6.05
	R-X-X-X-X-X-X-X-X-X-AcK	3.80
	K-X-X-X-X-X-X-X-AcK	3.93
	AcK-X-X-X-X-K	4.74
	G-X-X-X-AcK	4.53
	K-X-X-X-T-X-AcK	7.25
	K-X-AcK	3.39
	AcK-X-K	4.28
Brain	**K-F-X-X-X-X-AcK***	8.02
	AcK-F-X-X-X-X-K	7.95
	AcK-X-X-F-X-X-X-X-K*	7.02
	K-X-X-X-X-X-X-X-AcK	3.59
	K-X-X-X-X-X-AcK	3.46
	E-X-AcK-V*	8.81
	AcK-X-X-X-X-X-X-X-A*	3.83
	AcK-X-X-X-X-X-K*	3.67
	K-X-X-X-X-X-X-X-X-AcK	4.29
	AcK-X-X-X-X-X-X-R*	4.11
Brown fat	**F-X-X-X-X-X-X-AcK***	3.30
	F-AcK	3.33
	K-X-X-X-X-X-X-X-AcK	3.54
Heart	**M-AcK***	4.12
	E-X-AcK*	4.38
	K-X-X-X-X-X-AcK-X-X-X-X-E*	6.90
	K-X-X-X-X-X-X-X-AcK	3.25
	G-X-X-X-X-X-X-X-X-X-AcK-X-X-X-K*	7.80
Intestine	No motif	
Kidney	R-X-X-X-X-X-X-X-X-X-AcK	4.49
	K-X-X-X-X-X-AcK	3.70
	AcK-X-X-X-X-X-K	3.37
	K-X-X-X-X-X-X-X-AcK-X-X-R*	7.15
	L-X-AcK-X-X-X-K*	7.04
Liver	**D-X-X-X-X-AcK***	3.33
	D-X-X-X-X-X-AcK*	3.09
Lung	**AcK-X-X-R***	3.29
	AcK-X-D-X-X-X-X-X-K	6.37
Muscle	K-X-X-X-X-X-AcK	3.79
	AcK-X-X-X-X-X-K	4.53
	K-X-X-X-X-X-X-X-X-AcK	4.64
	AcK-X-X-X-X-X-X-X-X-K*	4.01
Pancreas	S-X-X-X-X-AcK	3.05
Perirenal fat	No motif	
Skin	F-AcK	3.34
	L-X-X-X-L-X-AcK*	6.39
	AcK-X-X-Y*	3.25
Spleen	S-X-X-X-X-AcK	3.27
	AcK-X-X-X-X-X-K	3.52
	K-X-X-X-X-X-X-X-AcK	3.22

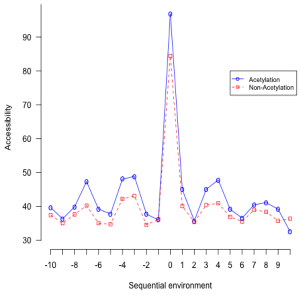
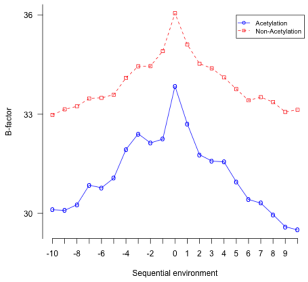
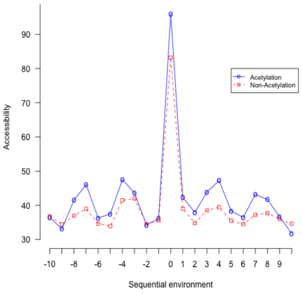
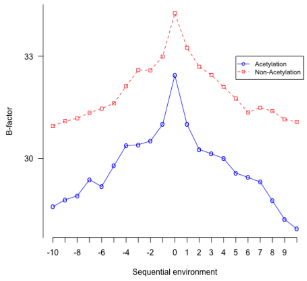
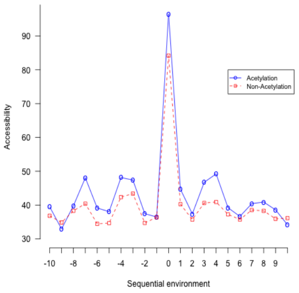
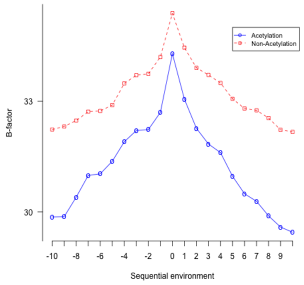
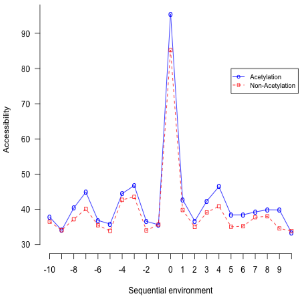
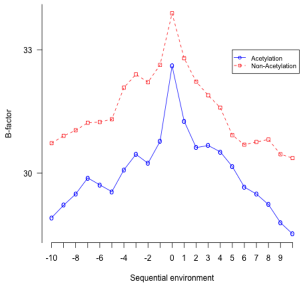
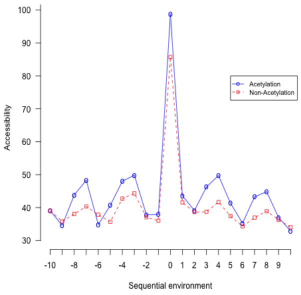
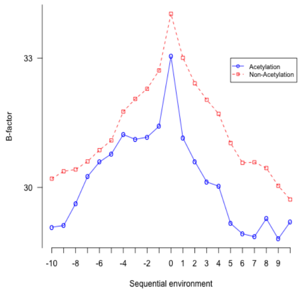
	AcK-X-X-X-X-K	3.24
	K-X-X-X-X-X-AcK	3.45
	AcK-X-X-X-X-X-X-X-X-R	3.07
Stomach	No motif	
Testis	**D-X-X-AcK**	3.66
	P-X-X-X-X-X-X-X-AcK*	3.70
	R-X-X-X-X-X-X-X-X-X-AcK	3.53
Testis fat	No motif	
Thymus	F-AcK	3.82
	AcK-X-X-X-X-X-X-X-X-R	3.76
	K-X-X-X-X-X-X-X-AcK	3.27
	R-X-X-X-X-X-X-X-X-X-AcK	3.30
	AcK-X-I*	3.42
	AcK-G*	3.05

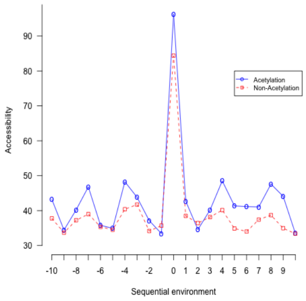
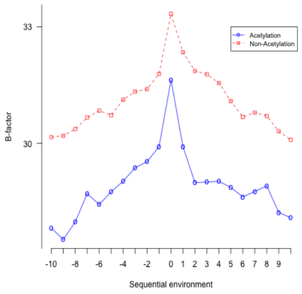
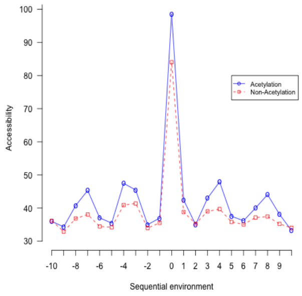
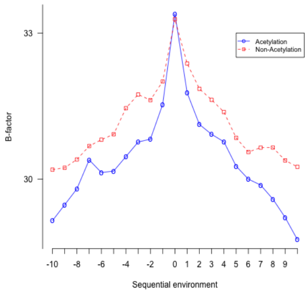
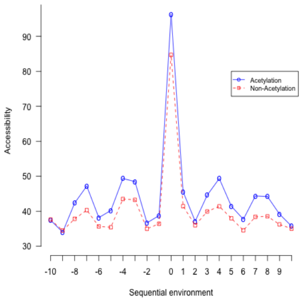
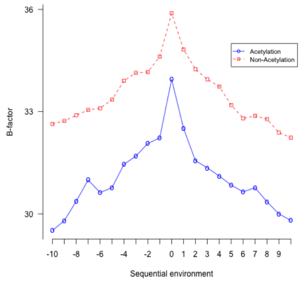
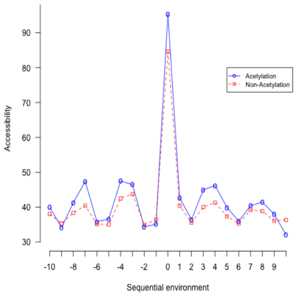
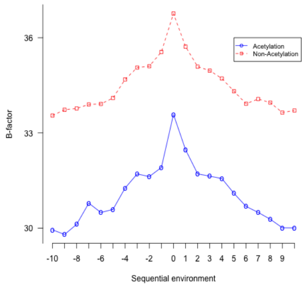
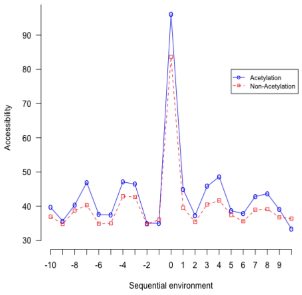
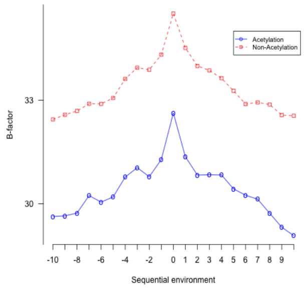
Table 13: Summary of sequence motifs associated with LASs in the LAS3D dataset. Motifs in bold correspond to tissue-specific motifs. Motifs in bold with stars (*) are not observed in global LASs. AcK and X represent acetylated lysine residues and wildcard residues, respectively. Only statistically significant motifs are shown ($p < 0.001$).

Tissue	% of conserved LASs	% of conserved non-LASs	p-value	% of conserved LASs in functional proteins	% of conserved non-LASs in functional proteins	p-value
All tissues	45.07	41.3	6.17 x 10-7	59.85	50.16	9.35 x 10-14
Brain	48.12	40	0.003	67.39	52.61	3.08 x 10-13
Brown fat	47.34	46.74	0.786	62.25	50.69	4.793 x 10-4
Heart	46.96	47.73	0.704	66.76	50.31	2.391 x 10-9
Intestine	47.22	46.07	0.642	56.52	44.94	6.71 x 10-4
Kidney	46.04	44.82	0.342	61.15	53.5	4.396 x 10-5
Liver	48.12	46.44	0.356	59.76	55.52	0.099
Lung	49.03	44.54	0.006	49.03	44.54	2.104 x 10-10
Muscle	47.41	46.73	0.719	70.15	47.75	4.323 x 10-18
Pancreas	45.69	44.92	0.783	48.43	41.21	0.083
Perirenal fat	51.28	48.75	0.51	58	44.1	0.007
Skin	51.21	49.3	0.398	62.41	46.24	1.987 x 10-7
Spleen	46.42	44.24	0.332	46.42	44.24	0.024
Stomach	46.87	46.61	0.861	59.73	53.22	0.002
Testis	47.72	44.96	0.063	58.46	48.34	2.092 x 10-6
Testis fat	50	52.28	0.54	57.9	43.11	0.003
Thymus	50.87	45.31	0.001	56.81	48.19	3.464 x 10-4

Table 14: Tissue-specific evolutionary conservation analysis of LASs and non-LASs.

Tissue	# of LASs	# of non-LASs	Solvent accessibility	B-factor scores
Global	2218	8777		
Brain	959	4827		
Brown fat	416	1877		
Heart	584	2177		
Intestine	365	2215		

Kidney	1116	5361		
Liver	680	3241		
Lung	741	4129		
Muscle	581	2213		
Pancreas	227	1313		

Perirenal fat	190	981		
Skin	473	2718		
Spleen	377	2327		
Stomach	862	4621		
Testis	882	4811		

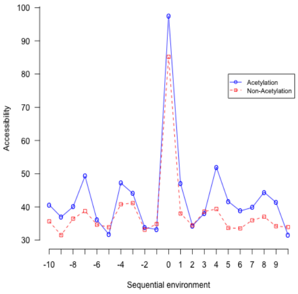
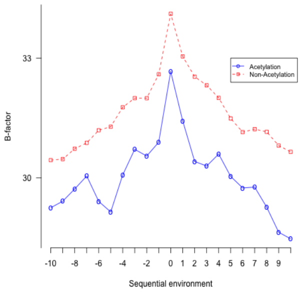
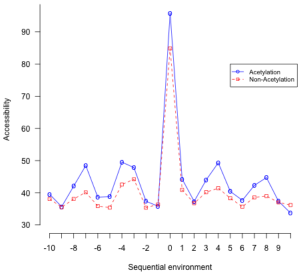
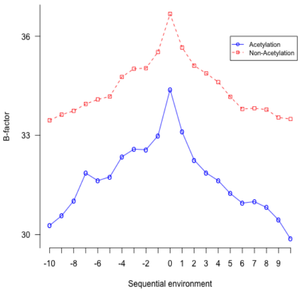
Testis fat	232	1329		
Thymus	604	3594		

Table 15: Accessibility and B-factor analysis of LASs in different tissues.

Tissue	AA	Motif	Motif Score
Blood	S	pS-P	16.00
		pS-X-E	14.09
		R-X-X-pS	11.21
		pS-D-X-E	15.39
	T	pT-P	8.34
	Y	No motif	
Global	S	R-pS-X-S-P	42.91
		R-S-X-pS-P	42.66
		pS-P-X-X-S-P	47.48
		R-pS-X-S	32.00
		S-P-pS	32.00
		S-P-X-X-pS-P	44.63
		S-X-X-X-pS-P	32.00
		pS-X-X-X-S-P	32.00
		pS-X-S-P	32.00
		R-X-X-S-X-pS	32.00
		R-X-X-pS-P	29.14
		R-S-X-pS	32.00
		pS-P-X-X-X-R	29.66
		pS-D-E-E	42.17
		pS-P-X-X-X-K	29.59
		pS-E-X-E-X-D	41.24
		pS-X-D-E-X-E	41.09
		pS-D-D-E	42.37
		pS-P-X-R	27.24
		pS-D-X-E-D	38.52
		pS-E-E-E	40.94
		pS-D-X-E	32.00
		pS-X-X-S-P	32.00
		R-X-X-X-pS-P	23.53
		R-X-X-pS-X-E	32.00
		pS-S-P	32.00
		R-X-pS-X-S	32.00
		pS-P-X-X-X-X-X-X-X-R	24.71
		pS-X-D-E	32.00
		R-X-G-pS	27.54
		S-P-X-pS	32.00
		pS-D-E-D	36.13
		pS-P-X-X-X-X-X-X-X-K	24.11
		pS-P-X-X-X-X-X-X-X-E	30.68
		pS-X-E-E	25.67
		pS-X-X-D-X-X-E	25.24
		pS-D-X-D-X-E	32.83
		pS-X-X-X-X-D-X-E	28.13
		R-X-X-pS-X-X-X-X-E	26.21
		R-X-X-X-X-X-X-X-pS-P	23.89
		E-E-X-X-X-X-X-X-X-pS	26.96
		R-R-X-pS	28.19
		S-P-X-X-pS-X-X-X-X-R	38.18
		R-X-X-S-X-X-pS	32.00
		pS-P-R	25.82
		pS-X-X-X-X-E-E	23.71
		pS-D-D-D	37.75
		R-X-X-S-P-X-X-X-X-X-pS	31.03
		R-X-X-S-X-X-X-X-X-pS	29.73
		pS-P-X-X-X-X-X-X-X-R	25.89

pS-R-X-X-S	32.00
pS-X-D-D	25.14
pS-X-E-D	23.76
pS-P-X-X-R	29.37
D-X-D-pS	26.27
R-X-X-S-X-X-X-pS	27.96
pS-X-X-R-X-X-S	30.42
pS-X-X-X-X-X-R-X-X-S	32.00
R-R-X-X-pS	22.50
pS-X-D-X-D	24.57
R-X-X-X-X-X-X-X-X-X-pS-P	23.31
R-X-pS-S	25.08
S-P-X-X-X-X-X-X-X-pS-X-X-X-X-X-X-X-X-R	36.27
R-X-X-S-X-X-X-X-pS	25.99
D-D-X-X-X-X-X-pS	23.79
S-P-X-X-X-X-X-X-X-X-pS-X-X-X-X-X-X-X-R	36.22
pS-X-S-X-X-X-X-X-X-R	22.55
D-X-X-X-X-X-X-X-X-pS	16.00
R-X-X-S-X-X-X-X-X-X-pS-X-X-X-L	29.67
R-X-X-X-X-X-X-X-X-pS-P	22.04
pS-X-X-X-X-R-X-X-S	32.00
R-X-pS-P	22.31
pS-X-S-X-D	22.76
pS-P	16.00
pS-X-X-X-X-X-X-X-X-X-D	16.00
pS-X-X-X-X-X-X-X-D-X-E	23.40
pS-X-X-X-X-X-X-X-X-X-R	16.00
R-X-X-X-pS-X-X-X-X-X-X-S	23.60
P-X-X-X-X-X-X-X-X-X-pS-X-X-X-X-X-X-R-X-X-S	35.57
R-X-X-pS	16.00
S-X-X-X-X-X-X-X-X-X-pS-X-X-X-X-X-X-X-R	23.09
pS-X-D-X-X-D	26.06
R-X-pS	16.00
pS-X-X-D	16.00
pS-X-X-E-E	25.76
pS-X-X-X-X-X-X-X-X-R	15.65
R-X-S-X-pS	22.64
D-X-X-X-X-X-X-X-X-X-pS	13.87
S-X-X-X-X-X-X-X-pS-X-X-X-D	21.07
R-X-X-X-X-X-X-X-pS	13.28
pS-X-X-X-X-X-X-X-E-X-E	19.05
R-X-X-X-X-X-X-X-X-X-pS	12.49
pS-X-X-R	14.65
D-X-X-X-X-X-X-pS	11.28
pS-X-X-X-X-X-X-D	12.54
R-X-X-X-X-X-X-pS	11.29
pS-X-X-X-X-X-R	10.74
R-X-X-X-X-pS	10.75
K-X-pS	10.65
D-X-X-X-pS	9.84
pS-X-X-X-X-X-X-R	9.66
pS-X-X-K	9.74
K-X-X-X-X-X-X-X-pS	9.77
pS-R	8.13
pS-D	8.62
pS-X-X-X-X-X-X-X-R	7.88
pS-X-X-E	8.37

		pS-X-R	8.27
		pS-X-X-X-X-K	8.36
		pS-X-X-S	8.02
		P-X-X-X-X-X-X-pS	6.02
		pS-X-X-X-R	6.10
		pS-X-X-G	6.32
	T	pT-P-P	32.00
		pT-P-E	23.30
		pT-X-S-P	24.71
		pT-S-P	32.00
		pT-P	16.00
		S-X-pT	15.65
		pT-X-X-X-X-E	13.33
		pT-X-S	8.42
		pT-X-D	9.06
		S-X-X-pT	7.56
		pT-D-X-E	12.85
	Y	pY-X-X-X-E	6.76
		pY-D	6.78
		pY-X-X-S	6.03
Brain	S	**pS-P-E***	24.35
		R-X-X-pS-P	23.65
		E-X-X-X-X-X-X-X-pS-P*	23.69
		pS-P-X-X-E*	22.98
		pS-P	16.00
		pS-D-D-E	44.19
		pS-D-X-E	30.32
		R-X-X-pS	16.00
		pS-X-E-D	22.36
		pS-E-X-E	30.18
		pS-X-E	16.00
		pS-X-D-D	25.74
		R-X-pS	15.95
		pS-X-D	13.97
		pS-X-S-P	28.00
		E-X-X-X-X-pS*	13.64
		D-X-X-X-X-pS	11.67
		R-X-X-X-X-X-X-pS	11.57
		pS-X-X-X-X-X-X-X-X-D*	11.61
		pS-X-X-X-X-X-X-X-X-R	10.42
		pS-X-X-X-X-X-X-X-X-E*	10.41
		R-X-X-X-X-X-X-X-X-pS	9.81
		E-X-X-X-X-X-X-X-X-pS*	8.35
		pS-X-X-X-X-E	11.02
		pS-X-X-X-X-R	8.34
		R-X-X-X-X-X-X-X-pS	7.64
		pS-X-X-X-X-X-R	8.94
		R-X-X-X-X-pS	7.36
		R-X-X-X-X-X-X-X-X-X-pS	6.26
		pS-X-X-X-X-X-X-X-X-K*	6.05
		pS-X-X-X-R	7.50
		K-X-pS	6.40
		E-X-X-X-X-X-pS	6.41
		pS-X-X-E	6.12
		E-X-X-X-X-X-X-X-pS	6.06
	T	pT-P-P	25.98
		pT-P	16.00

		pT-S-P	20.41
		pT-X-X-X-X-E	6.94
	Y	No motif	
Heart	S	R-X-X-pS-P	26.15
		pS-P-X-X-X-R	22.71
		pS-D-E-E	38.96
		R-X-X-pS	16.00
		pS-P	16.00
		pS-X-D-E	22.41
		R-pS	11.21
		pS-D-X-D	17.52
		pS-E-X-E	14.14
		pS-X-X-X-X-X-X-X-X-R	6.95
		R-X-X-X-X-X-X-X-X-pS	6.29
		E-pS	6.68
	T	No motif	
	Y	No motif	
Intestine	S	R-X-X-pS-P	31.26
		pS-P-X-X-X-X-R	30.31
		R-X-X-X-X-X-X-X-X-pS-P	24.63
		pS-D-D-E	41.11
		R-R-X-X-pS	22.15
		pS-D-E-E	40.90
		pS-P-X-X-X-R	23.47
		pS-D-X-E	29.48
		R-X-X-X-pS-P	22.17
		R-X-X-pS	16.00
		pS-P	16.00
		R-pS-X-S	22.02
		pS-D-X-D	30.72
		pS-E-X-E	32.00
		R-X-X-X-X-X-X-X-X-X-pS	16.00
		pS-X-X-R	16.00
		R-X-pS	16.00
		pS-X-X-X-X-X-X-X-X-X-R	14.29
		pS-X-X-X-X-X-X-X-R	12.08
		pS-R	12.38
		pS-X-D	11.79
		pS-X-X-X-R	10.46
		R-X-X-X-X-X-X-X-pS	11.63
		pS-X-X-X-X-R	10.95
		E-X-X-X-X-X-X-X-pS	9.48
		pS-X-X-X-X-X-X-E	8.77
		R-X-X-X-X-X-X-pS	8.31
		R-X-X-X-pS	8.50
		R-pS	6.89
		R-X-X-X-X-X-pS	6.17
		pS-X-X-X-X-X-X-R	6.35
	T	pT-P-P	22.82
		pT-P	14.91
	Y	No motif	
Kidney	S	R-X-X-pS-P	30.61
		pS-P-X-X-X-X-R	29.10
		pS-D-D-E	41.79
		pS-D-E-E	40.00
		pS-P	16.00
		R-R-X-pS	24.51

		pS-E-X-E	29.30
		R-X-R-X-X-pS*	22.03
		R-pS-X-S	22.06
		pS-X-E-D	25.75
		D-pS-D*	30.11
		pS-X-D-D	24.74
		R-X-X-pS	13.99
		pS-X-X-X-X-X-X-X-D	13.67
		R-X-pS	11.95
		pS-X-X-X-X-X-X-X-X-X-R	11.49
		D-X-X-X-X-pS	9.93
		R-X-X-X-pS	8.80
		R-X-X-X-X-X-X-pS	9.88
		D-X-X-X-X-X-X-pS	7.51
		R-X-X-X-X-X-X-X-pS	8.24
		pS-X-X-X-R	9.50
		E-X-X-X-X-X-X-X-pS	6.72
		pS-X-X-X-X-E	7.22
		R-X-X-X-X-X-X-X-X-X-pS	7.45
		pS-X-X-X-X-R-X-X-S	13.06
		D-X-pS*	6.06
		pS-X-X-X-X-X-X-X-X-R	6.53
	T	pT-P	16.00
	Y	No motif	
Liver	S	pS-P-X-X-X-X-R	22.35
		pS-P	16.00
		pS-D-D-E	40.42
		R-X-X-pS	16.00
		pS-D-X-E	32.00
		pS-X-X-X-X-D*	16.00
		pS-X-X-X-X-R	10.45
		R-X-pS	9.72
		pS-X-X-X-X-X-X-X-D	10.04
		R-pS	8.52
		E-pS	7.30
		pS-X-X-X-X-X-X-X-R	7.12
		pS-X-X-X-X-X-X-X-X-X-R	6.63
		R-X-X-X-X-X-X-X-X-pS	6.30
	T	pT-P	16.00
	Y	No motif	
Muscle	S	R-X-X-pS	16.00
		pS-P	12.55
		pS-D-X-E	29.26
		E-X-X-X-X-X-X-X-pS	8.56
		pS-X-X-D	7.53
		pS-X-X-X-X-E	6.61
	T	pT-P	7.78
	Y	No motif	
Lung	S	pS-P-X-X-X-X-R	32.00
		R-X-X-pS-P	28.55
		pS-P-X-X-X-R	24.74
		pS-P-X-X-X-X-X-X-X-X-R	22.97
		R-R-X-pS	25.01
		pS-D-D-E	40.92
		R-X-X-X-pS-P	24.26
		pS-D-E-E	38.79
		R-X-X-pS-X-E	22.52

		pS-P-X-X-K	23.01
		pS-D-X-E	29.48
		R-pS-X-S	24.63
		pS-P	16.00
		R-X-X-pS	16.00
		R-X-pS	16.00
		pS-D-X-D	32.00
		pS-X-X-X-X-X-X-X-X-R	16.00
		R-X-X-S-X-pS	23.45
		R-X-X-X-X-X-X-X-pS	16.00
		pS-X-X-X-X-R-X-X-S	22.71
		pS-X-X-X-X-X-X-X-X-X-R	15.48
		R-X-X-X-pS	16.00
		pS-E-X-E	30.53
		pS-X-X-X-X-X-R	13.73
		R-X-X-X-X-X-X-pS	14.75
		R-pS	10.78
		pS-X-X-R	11.79
		pS-X-X-X-X-X-X-E	10.92
		R-X-X-X-X-X-X-X-X-pS	11.27
		pS-X-X-X-R	10.98
		R-X-X-X-X-X-X-X-X-X-pS	12.40
		pS-R	10.51
		pS-X-X-X-X-X-X-X-R	10.52
		D-pS*	9.36
		K-X-X-pS*	10.49
		R-X-X-X-X-X-pS	9.23
		E-X-X-pS	8.28
		pS-X-X-X-X-R	8.48
		pS-X-X-X-X-X-X-D	7.13
		pS-X-X-X-X-X-X-X-K*	6.44
	T	pT-P-P	22.72
		pT-P	16.00
	Y	No motif	
Pancreas	S	pS-D-X-E	29.53
		R-X-X-pS	13.67
		pS-E-E*	20.99
		pS-X-D-D	15.45
		pS-P	8.98
		pS-X-X-E	8.18
		D-X-X-X-X-pS	7.00
	T	No motif	
	Y	No motif	
Perirenal fat	S	pS-P	16.00
		pS-D-D-E	40.21
		pS-D-X-E	31.05
		R-X-X-pS	16.00
		pS-X-E	13.36
		pS-X-D-D	21.17
		D-X-X-X-X-X-X-pS	10.15
		R-pS	10.86
		pS-X-D	8.77
		pS-X-X-R	7.30
		E-X-X-X-X-X-X-X-pS	7.08
		R-X-X-X-X-X-X-X-X-pS	8.04
		pS-X-X-E	6.03
	T	pT-P	16.00

	Y	No motif	
Spleen	S	R-X-X-pS-P	26.33
		pS-P-X-X-X-X-X-R*	25.31
		P-X-pS-P*	23.39
		pS-P	16.00
		pS-D-D-E	45.40
		pS-D-E-E	38.28
		R-R-X-pS	25.72
		pS-E-X-E-X-D	38.08
		pS-D-X-E	30.72
		R-X-X-pS	16.00
		pS-E-X-E	31.65
		pS-D-X-D	32.00
		R-pS	16.00
		pS-X-X-X-X-X-X-R	14.65
		pS-X-D	13.74
		pS-X-X-R	13.28
		R-X-X-S-X-X-pS	17.85
		R-X-pS	10.67
		E-X-E-X-X-X-X-X-pS*	18.65
		pS-X-X-X-X-X-E*	9.81
		pS-X-X-X-R	10.24
		pS-R	10.12
		R-X-X-X-X-X-X-X-pS	10.28
		pS-X-X-X-X-X-R	10.02
		pS-X-X-X-X-X-X-X-X-X-R	9.79
		E-X-X-pS	8.19
		pS-X-X-X-X-X-X-X-X-X-E*	8.51
		pS-X-X-X-X-X-X-X-R	7.41
		R-X-X-X-pS	9.24
		E-X-X-X-X-X-pS	6.65
		E-X-X-X-X-X-X-X-pS	7.24
	T	pT-P-P	25.31
		pT-P	16.00
		pT-D	6.28
	Y	No motif	
Stomach	S	R-X-X-pS-P	26.03
		R-X-X-X-pS-P	24.24
		R-X-X-pS-X-X-D*	22.47
		pS-P-X-X-X-R	23.53
		R-X-X-pS	16.00
		K-X-X-X-X-X-pS-P*	22.31
		pS-D-E-E	38.25
		pS-P	16.00
		pS-D-D-E	39.91
		R-X-pS	16.00
		pS-X-X-X-X-X-X-X-X-R	14.95
		R-pS	12.52
		pS-D-X-D	21.24
		pS-X-X-X-X-R	11.93
		R-X-X-S-X-pS	18.87
		R-X-X-X-X-X-X-pS	10.69
		pS-X-X-R	9.79
		pS-D-X-E	18.71
		pS-X-X-X-X-X-X-X-X-X-R	8.20
		R-X-X-X-X-X-X-X-X-pS	8.60
		pS-X-X-X-X-X-R	7.32

		R-X-X-X-X-X-X-X-pS	7.93
		pS-X-X-X-R	7.87
		R-X-X-X-X-X-X-X-X-X-pS	6.57
	T	pT-P	16.00
	Y	No motif	
Testis	S	R-X-X-pS-P	26.22
		pS-P-X-X-X-R	24.33
		pS-P-X-X-X-K*	22.92
		pS-D-D-E	38.49
		pS-P	16.00
		pS-D-X-E	32.00
		R-X-X-pS	16.00
		pS-E-X-E	32.00
		pS-D-X-D	26.98
		pS-X-E-X-L*	19.10
		R-pS	11.41
		pS-X-X-X-X-X-X-X-X-X-R	10.30
		pS-X-X-X-X-X-X-X-X-X-D	9.05
		pS-X-X-X-X-X-X-R	9.06
		pS-X-X-R	8.16
		R-X-X-X-X-X-X-pS	8.70
		K-X-X-pS-X-X-X-X-X-X-X-X-X-E*	15.87
		pS-X-X-X-X-E	7.29
		R-X-X-X-X-pS	6.42
	T	pT-P-P	26.54
		pT-P	16.00
		pT-D	7.64
	Y	No motif	
Thymus	S	R-X-X-pS-P	28.88
		pS-P-X-X-X-R	25.51
		pS-P-X-R	23.29
		pS-P-X-X-K	22.13
		pS-D-D-E	43.35
		pS-P-X-X-X-X-X-X-R	22.36
		pS-D-E-E	46.21
		pS-P	16.00
		R-R-X-pS	26.74
		pS-E-X-E-X-D	39.88
		pS-E-X-E	30.91
		R-X-X-pS	16.00
		pS-D-X-E	32.00
		pS-R-S*	22.56
		pS-D-X-D	32.00
		R-X-X-S-X-pS	23.74
		pS-X-X-X-X-D-E*	22.91
		R-X-pS	16.00
		D-E-X-X-X-X-X-X-X-pS*	22.00
		D-D-X-X-X-X-X-pS	26.94
		R-X-X-X-X-X-X-X-pS	14.68
		pS-X-X-X-X-R-X-X-S	22.40
		pS-X-D	14.00
		pS-X-X-X-X-X-X-X-X-X-R	12.85
		E-E-X-X-X-X-X-X-X-pS	17.79
		pS-X-X-X-X-R-X-X-S	17.79
		pS-X-E	10.30
		pS-X-X-R	10.84
		R-pS	11.15

	R-X-X-X-X-X-X-X-X-pS	10.51
	pS-X-X-X-R	10.90
	R-X-X-X-X-X-X-X-X-X-pS	10.57
	D-X-X-X-X-X-X-pS	9.88
	R-X-X-X-X-X-X-pS	9.75
	R-X-X-X-pS	8.87
	pS-X-X-X-X-X-X-X-R	8.34
	E-X-pS*	7.61
	R-X-X-X-X-X-pS	7.24
	pS-X-X-X-X-R	8.03
	pS-X-R	8.44
	K-X-X-X-X-pS*	8.08
	pS-X-P*	7.68
	pS-X-X-X-X-E	7.84
T	pT-P-P	26.36
	pT-P	16.00
Y	No motif	

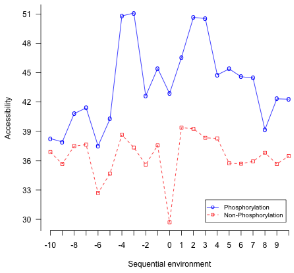
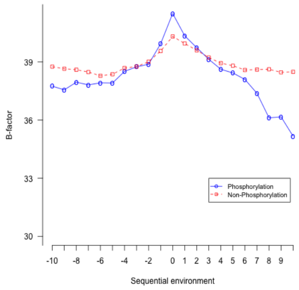
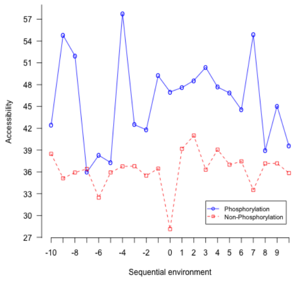
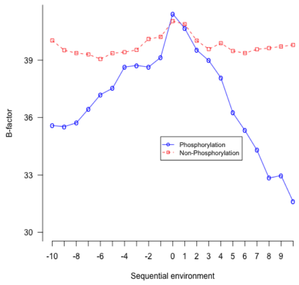
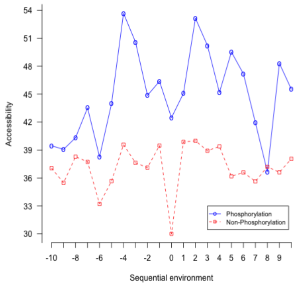
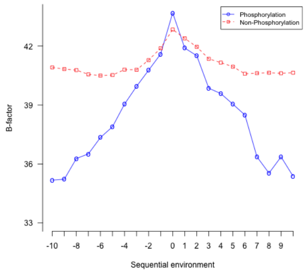
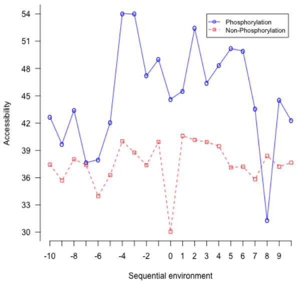
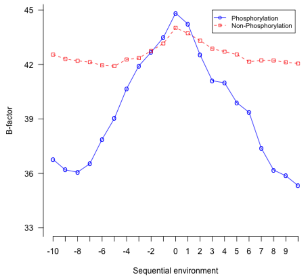
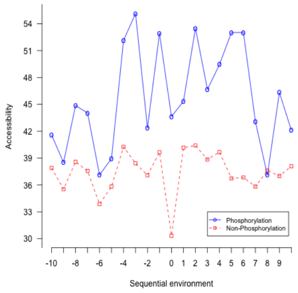
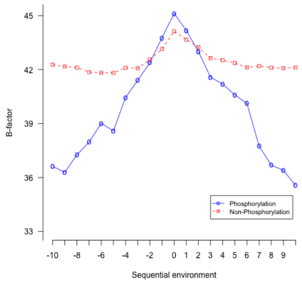
Table 16: Sequence motifs of all phosphorylation sites at 70% sequence similarity with the $p < 0.000001$ significance threshold. The motif score represents the sum of the negative log probabilities used to fix each position of the motif. The higher the motif score, the more statistically significant the corresponding motif is.

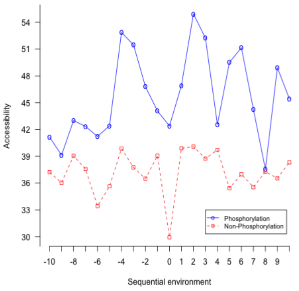
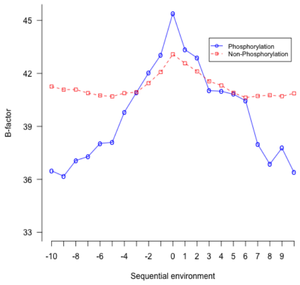
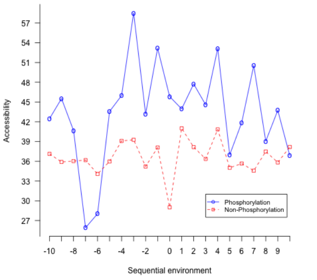
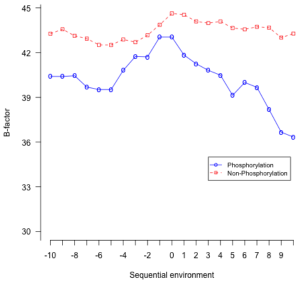
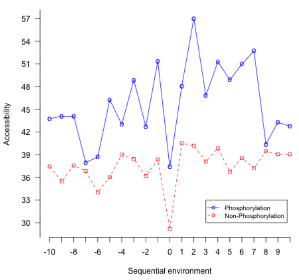
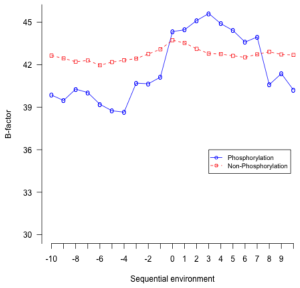
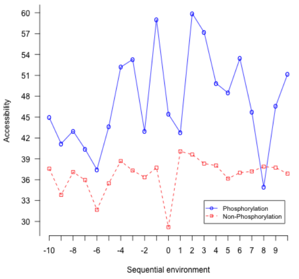
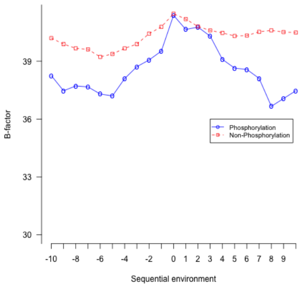
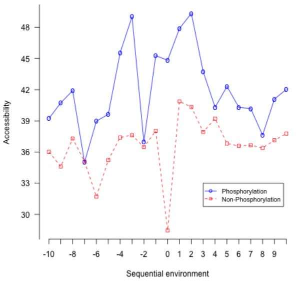
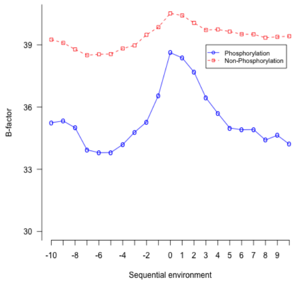
AA/Pos	-6	-5	-4	-3	-2	-1	1	2	3	4	5	6
G	n/n	n/n	n/n	n/+	-/n	+/+	-/n	n/n	n/n	n/n	n/n	n/n
S	+/+	+/n	+/+	+/+	+/+	+/+	n/+	+/+	+/+	+/+	n/+	n/n
T	n/n	+/n	+/n	+/+	+/n	+/n	-/n	n/n	-/n	n/n	n/n	-/n
Y	n/+	+/+	+/n	+/+	+/n	+/n	-/n	-/-	-/n	-/n	n/-	-/n
C	-/n	-/n	-/-	-/-	-/n	-/n	-/n	-/n	-/n	-/n	-/-	-/-
N	n/n	n/n	n/n	-/-	n/-	n/-	-/-	n/-	n/n	n/n	n/n	n/n
Q	-/n	n/n	n/n	-/n	n/n	-/n	-/-	-/-	-/-	-/n	n/-	-/-
K	+/n	+/n	+/-	+/n	n/-	n/n	-/-	-/-	n/-	-/-	n/n	+/n
R	+/n	+/n	+/n	+/-	+/-	+/n	-/-	-/-	n/-	n/-	n/-	+/+
H	n/n	n/n	-/-	-/-	-/n	-/-	-/n	-/-	-/n	-/n	-/-	-/-
D	n/+	n/+	n/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+
E	+/+	+/n	+/n	n/+	n/+	n/+	+/n	+/+	+/+	+/+	+/+	+/+
P	n/n	n/n	n/-	-/n	n/n	-/n	+/+	+/+	n/-	n/n	+/+	n/n
A	n/n	n/+	n/+	-/n	n/n	n/n	-/-	-/n	n/n	n/n	n/n	n/n
W	-/-	n/n	n/n	-/n	n/n	n/n	n/n	-/n	-/n	n/n	-/n	n/n
F	-/n	-/n	-/n	-/n	-/n	-/n	n/-	-/-	n/-	-/-	-/n	n/n
L	-/n	-/-	-/n	-/-	-/n	n/-	-/-	-/-	-/-	+/-	-/-	-/-
I	n/-	n/n	-/-	-/-	-/n	-/-	-/-	-/-	-/-	n/n	-/n	-/-
M	n/n	n/n	-/-	n/n	n/n	n/n	-/-	-/n	n/n	n/n	n/n	n/n
V	-/-	-/-	-/-	-/-	-/n	-/-	-/-	-/-	-/-	n/-	-/-	-/-

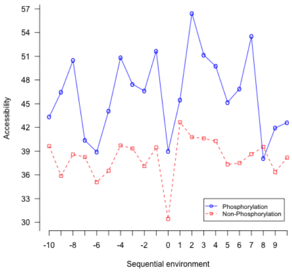
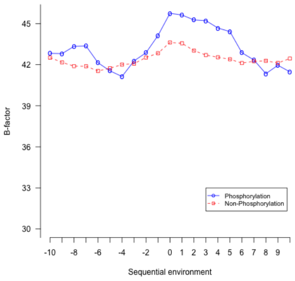
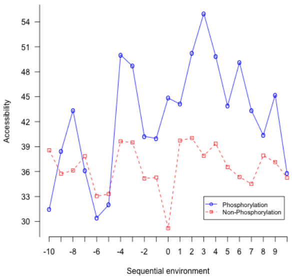
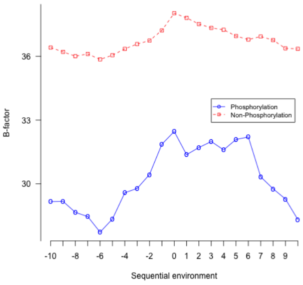
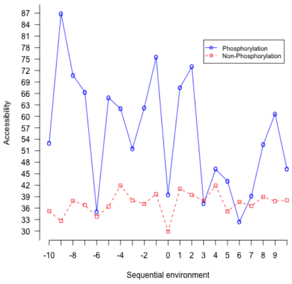
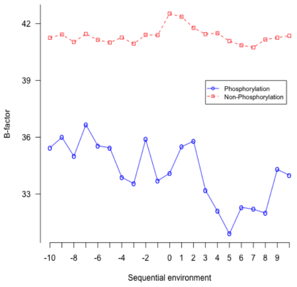
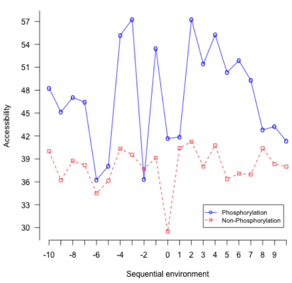
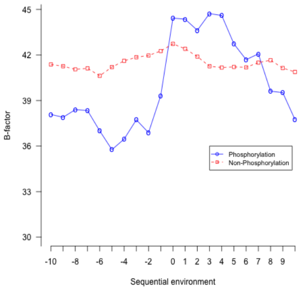
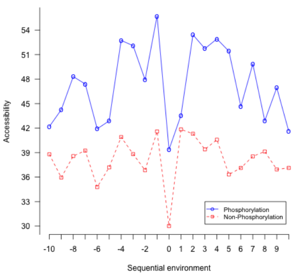
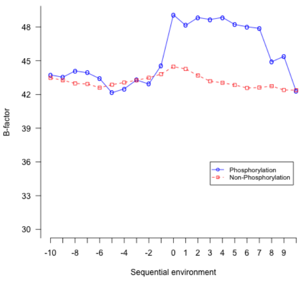
Table 17: Two sample logo comparison of phosphorylation sites between our study and the study of Lundby *et al.* in brain. "+" and "-" represent enrichment and depletion of amino acids at a particular position, respectively, whereas "n" represents the lack of depletion or enrichment. Numerator of each fraction in each cell corresponds to the finding in our study, whereas denominator corresponds to the observation in the study of Lundby et al. 58.8% of the cases overlap in both studies (n/n, +/+, -/-). In 27.5% of the cases we found a particular amino acid enriched/depleted in the corresponding position, and the study by Lundby et al. found no signal, whereas in 12.1% of the cases it is vice versa.

AA/Pos	-6	-5	-4	-3	-2	-1	1	2	3	4	5	6
G	n/n	-/-	n/n	n/+	n/n	+/n	-/n	n/n	n/-	n/n	n/-	n/n
S	n/n	n/n	n/n	n/n	+/+	+/n	-/+	+/+	n/n	n/n	-/n	n/-
T	+/n	n/n	+/n	+/n	+/n	+/+	-/n	n/n	-/-	-/-	n/n	-/-
Y	n/n	+/n	+/n	n/-	n/n	+/n	-/-	-/n	-/n	-/n	n/n	n/n
C	-/-	-/n	-/n	-/n	-/n	-/n	-/n	-/n	-/n	-/n	-/n	-/n
N	-/n	-/n	n/n	-/n	n/n	n/n	-/n	n/n	n/n	n/n	n/n	n/n
Q	n/n	n/n	-/-	n/n	+/n	-/-	-/n	-/-	-/n	-/-	n/n	n/n
K	+/+	n/n	+/n	n/-	-/-	n/-	-/-	-/n	n/-	-/n	n/n	+/+
R	+/n	+/+	+/+	+/-	+/-	+/n	-/-	-/-	n/-	n/-	+/n	+/n
H	-/n	n/n	-/n	n/n	-/n	-/n	-/n	-/n	-/-	-/n	n/n	n/n
D	+/+	n/+	n/n	n/+	n/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+
E	n/+	n/+	n/+	n/+	n/+	n/+	+/+	+/+	+/+	+/+	+/+	+/+
P	+/n	n/n	n/n	n/n	n/n	-/n	+/n	+/n	+/+	n/n	+/n	n/n
A	n/n	n/n	n/n	n/n	n/-	n/n	-/n	-/-	-/-	-/-	-/n	n/n
W	n/n	-/n	n/n	n/n	n/n	n/n	-/-	-/n	n/n	n/n	-/n	n/n
F	-/-	n/-	n/n	-/n	n/n	n/-	n/-	-/-	n/n	n/n	n/n	n/n
L	-/-	n/n	-/n	-/n	-/-	n/n	-/-	-/-	-/n	+/+	n/-	-/-
I	-/-	n/n	n/n	-/n	-/-	-/-	-/-	-/-	-/n	+/n	-/n	-/n
M	n/n	n/-	n/n	n/n	n/-	n/n	n/-	-/n	n/-	n/n	n/n	n/n
V	-/n	-/-	n/n	-/n	n/n	-/-	-/-	n/n	-/-	n/-	-/n	-/-

Table 18: Two sample logo comparison of phosphorylation sites between our study and the study of Lundby *et al.* in testis. "+" and "-" represent enrichment and depletion of amino acids at a particular position, respectively, whereas "n" represents the lack of depletion or enrichment. Numerator of each fraction in each cell corresponds to the finding in our study, whereas denominator corresponds to the observation in the study of Lundby et al. 57.5% of the cases overlap in both studies (n/n, +/+, -/-). In 29.2% of the cases we found a particular amino acid enriched/depleted in the corresponding position, and the study by Lundby et al. found no signal, whereas in 12.1% of the cases it is vice versa.

Tissue	# of PSS	# of non-PSS	Solvent accessibility	B-factor scores
Global	423	4162		
Blood	38	625		
Brain	122	1758		
Brainstem	81	1318		
Cerebellum	82	1411		

Cortex	93	1483		
Heart	44	668		
Intestine	69	1139		
Kidney	66	1062		
Liver	97	1305		

Lung	73	1216		
Muscle	95	889		
Pancreas	10	349		
Perirenal fat	56	850		
Spleen	73	1330		

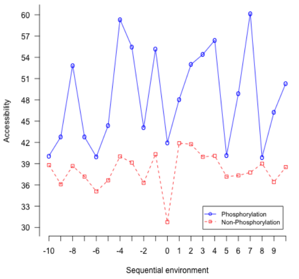
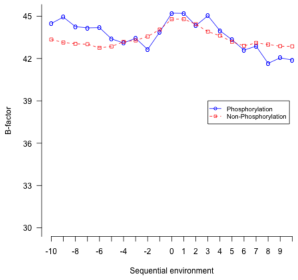
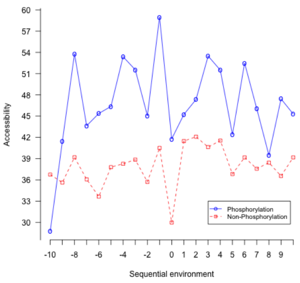
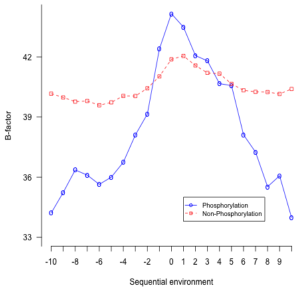
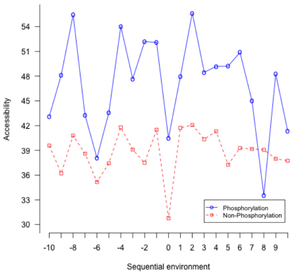
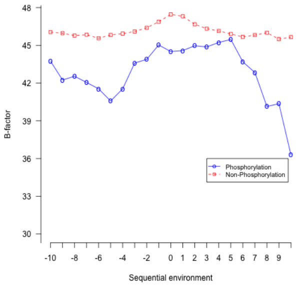
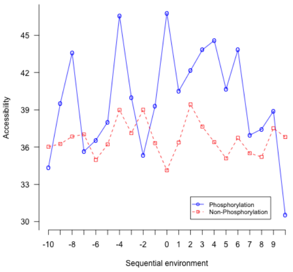
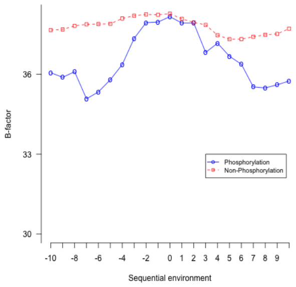
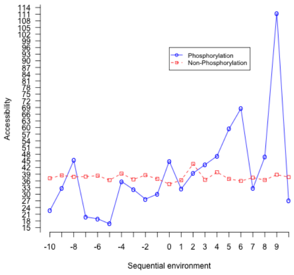
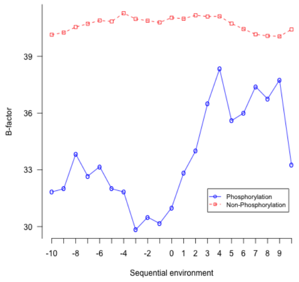
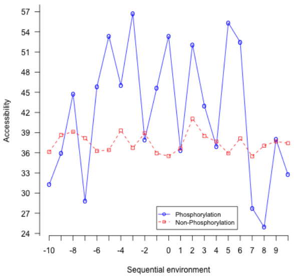
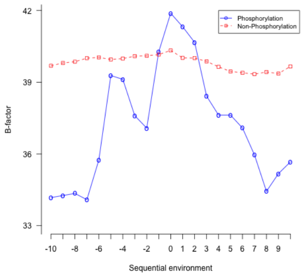
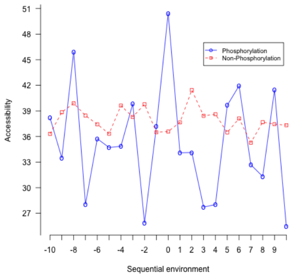
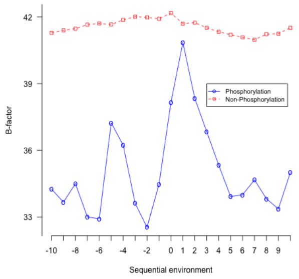
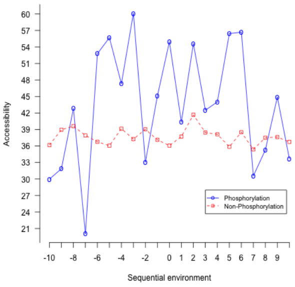
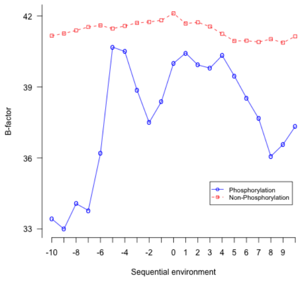
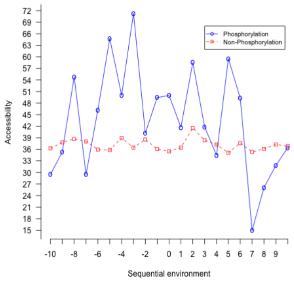
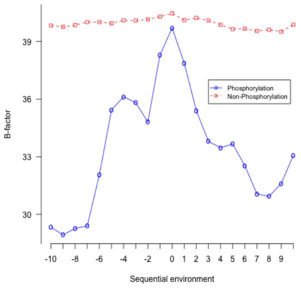
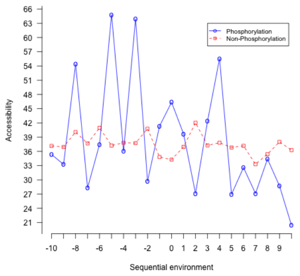
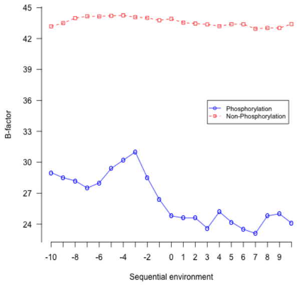
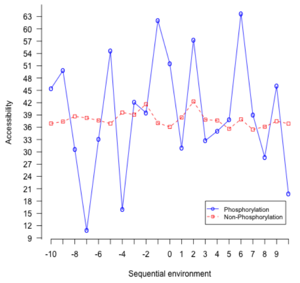
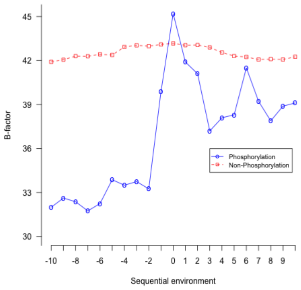
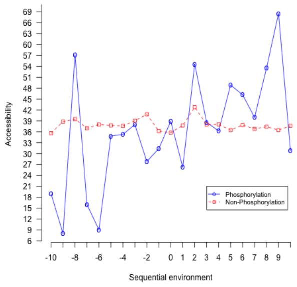
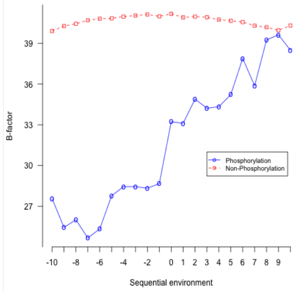
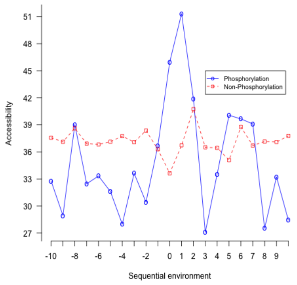
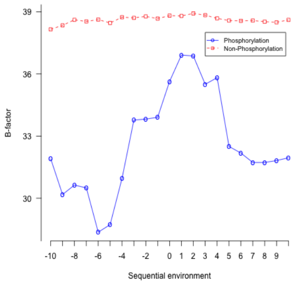
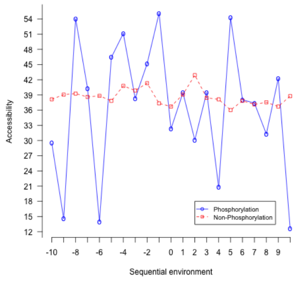
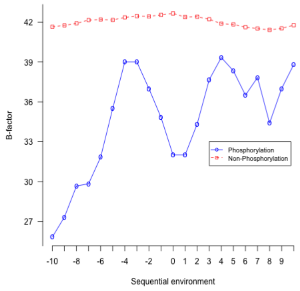
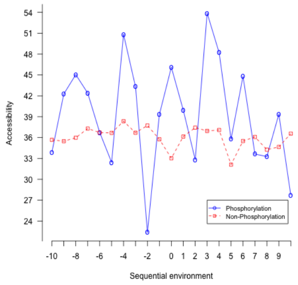
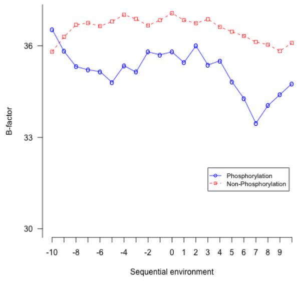
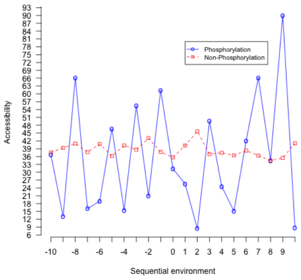
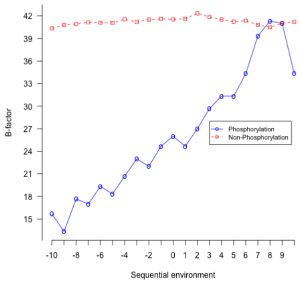
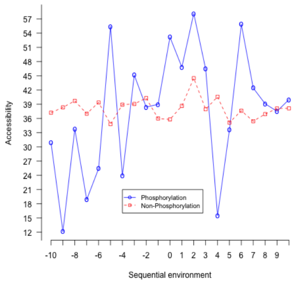
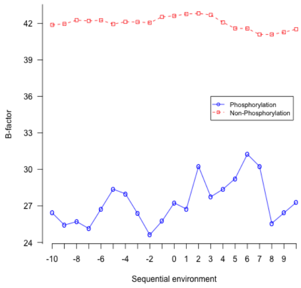
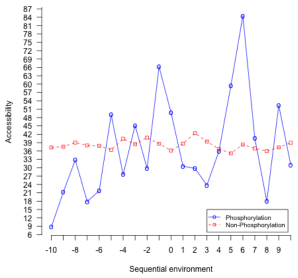
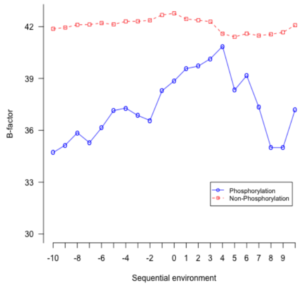
Stomach	66	1107		
Testis	45	922		
Thymus	65	1154		

Table 19: Accessibility and B-factor analysis of PSS in different tissues in the PS3D-90 dataset.

Tissue	# of PTS	# of non-PTS	Solvent accessibility	B-factor scores
Global	140	3790		
Blood	6	567		
Brain	26	1597		
Brainstem	13	1223		
Cerebellum	16	1281		

Cortex	17	1399		
Heart	10	652		
Intestine	10	1146		
Kidney	9	1002		
Liver	22	1179		

Lung	6	1176		
Muscle	51	847		
Pancreas	3	319		
Perirenal fat	8	796		
Spleen	7	1280		

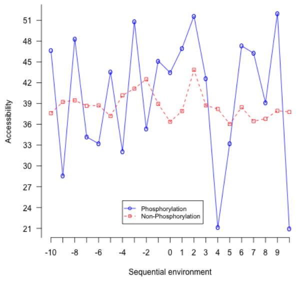
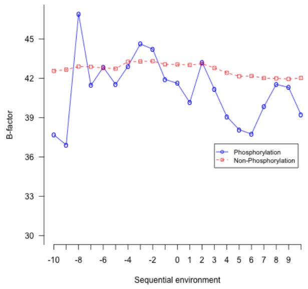
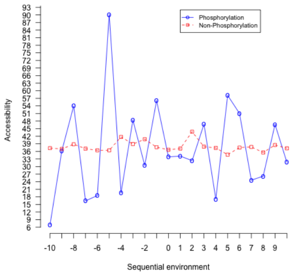
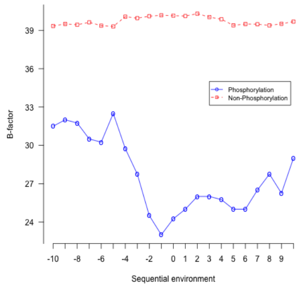
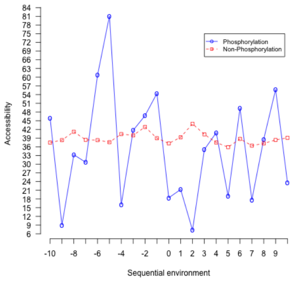
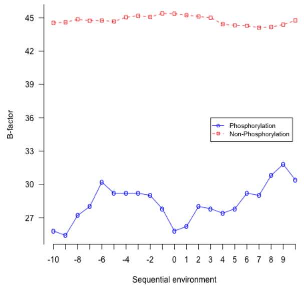
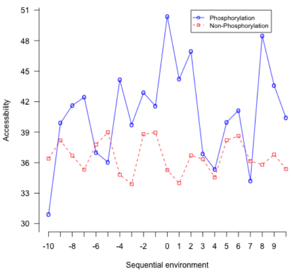
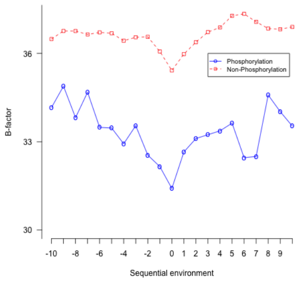
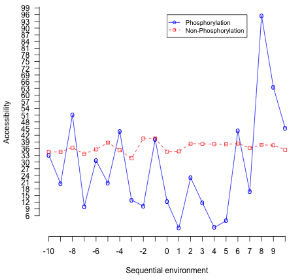
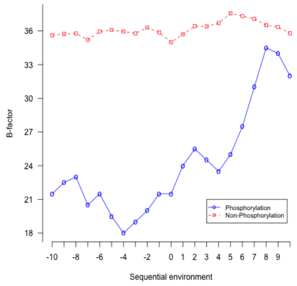
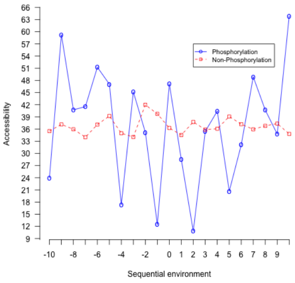
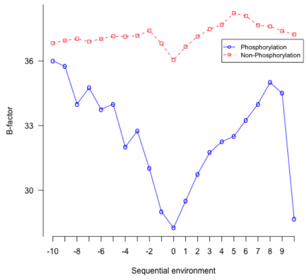
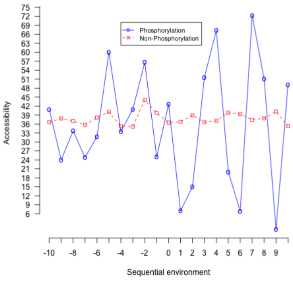
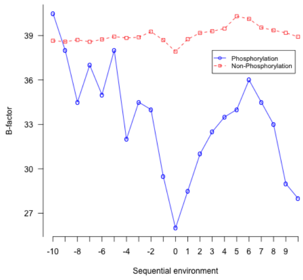
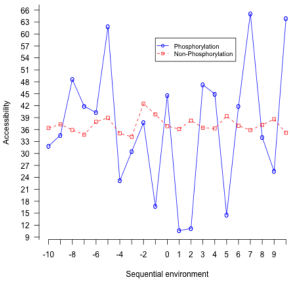
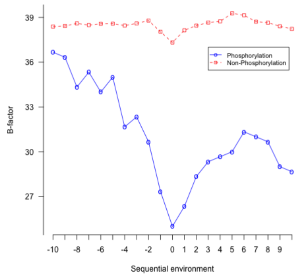
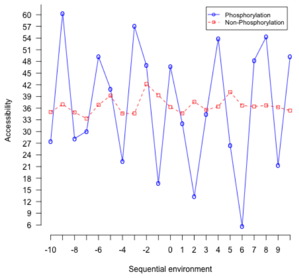
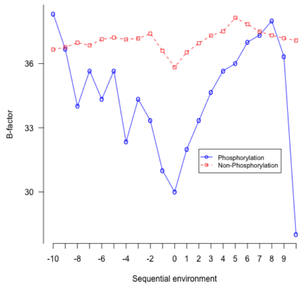
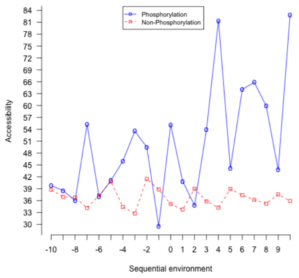
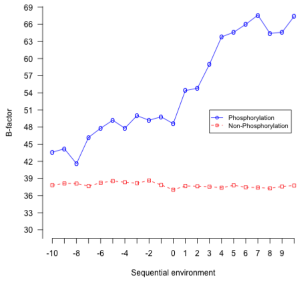
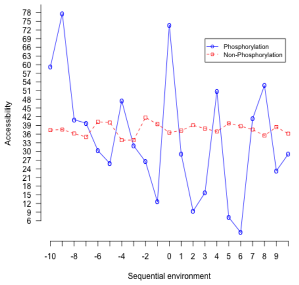
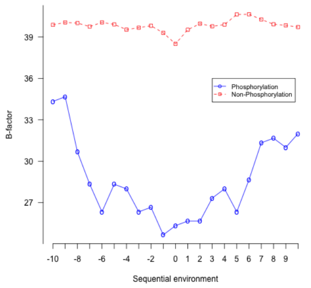
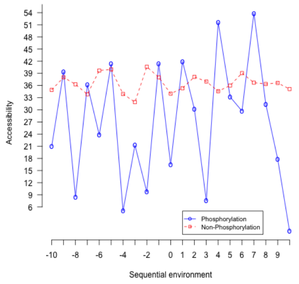
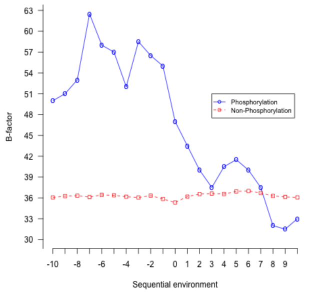
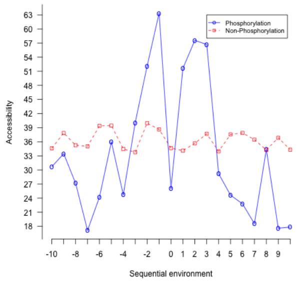
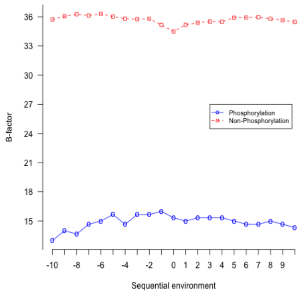
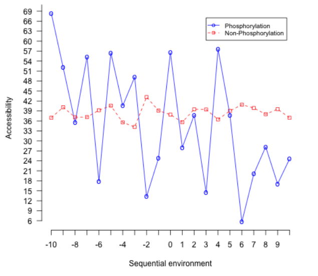
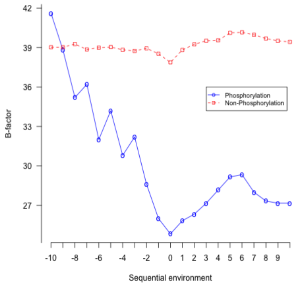
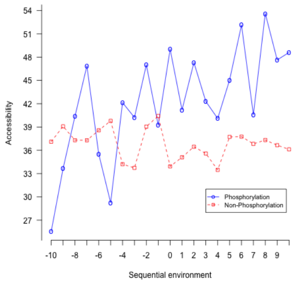
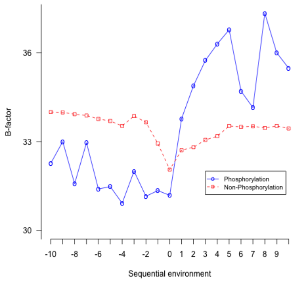
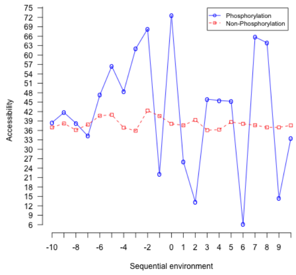
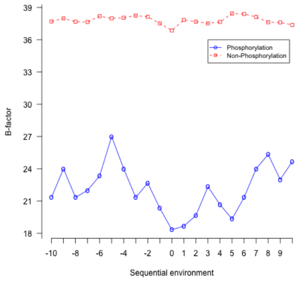
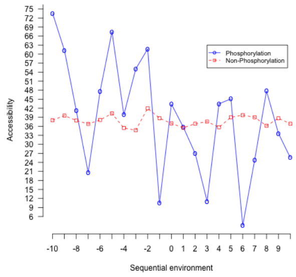
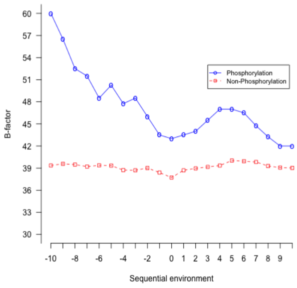
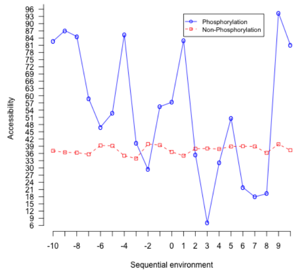
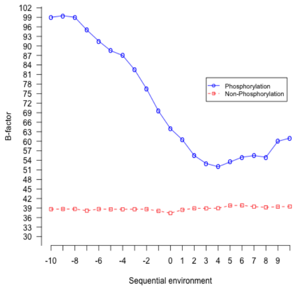
Stomach	13	1128		
Testis	4	848		
Thymus	5	1110		

Table 20: Accessibility and B-factor analysis of PTS in different tissues in the PS3D-90 dataset.

Tissue	# of PYS	# of non-PYS	Solvent accessibility	B-factor scores
Global	46	2804		
Blood	2	397		
Brain	4	1159		
Brainstem	2	829		
Cerebellum	3	906		

Cortex	3	974		
Heart	5	452		
Intestine	3	788		
Kidney	2	731		
Liver	3	918		

Lung	6	834		
Muscle	27	602		
Pancreas	0	276	Not applicable	Not applicable
Perirenal fat	3	557		
Spleen	4	949		
Stomach	2	727		

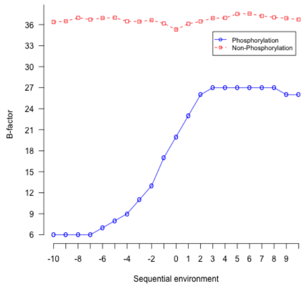
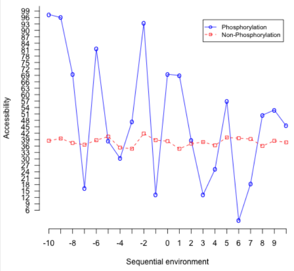
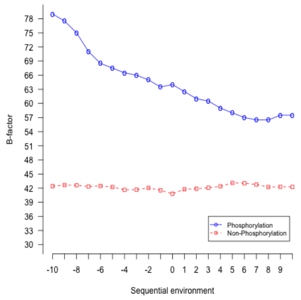
Testis	1	661	Not applicable	
Thymus	2	826		

Table 21: Accessibility and B-factor analysis of PYS in different tissues in the PS3D-90 dataset.

Basu, A. et al. (2009). "Proteome-wide prediction of acetylation substrates." In: *Proc Natl Acad Sci U S A* 106.33, pp. 13785–90.

Baxa, C. A., R. S. Sha, M. K. Buelt, A. J. Smith, V. Matarese, L. L. Chinander, K. L. Boundy, and D. A. Bernlohr (1989). "Human adipocyte lipid-binding protein: purification of the protein and cloning of its complementary DNA." In: *Biochemistry* 28.22, pp. 8683–90.

Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne (2000). "The Protein Data Bank." In: *Nucleic Acids Res* 28.1, pp. 235–42.

Berndsen, C. E., T. Tsubota, S. E. Lindner, S. Lee, J. M. Holton, P. D. Kaufman, J. L. Keck, and J. M. Denu (2008). "Molecular functions of the histone acetyltransferase chaperone complex Rtt109-Vps75." In: *Nat Struct Mol Biol* 15.9, pp. 948–56.

Blom, N., S. Gammeltoft, and S. Brunak (1999). "Sequence and structure-based prediction of eukaryotic protein phosphorylation sites." In: *J Mol Biol* 294.5, pp. 1351–62.

Blom, N., T. Sicheritz-Ponten, R. Gupta, S. Gammeltoft, and S. Brunak (2004). "Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence." In: *Proteomics* 4.6, pp. 1633–49.

Breiman, R. F. (2001). "Vaccines as tools for advancing more than public health: perspectives of a former director of the National Vaccine Program office." In: *Clin Infect Dis* 32.2, pp. 283–8.

Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden (2009). "BLAST+: architecture and applications." In: *BMC Bioinformatics* 10, p. 421.

Chen, X., S. P. Shi, S. B. Suo, H. D. Xu, and J. D. Qiu (2014). "Proteomic analysis and prediction of human phosphorylation sites in subcellular level reveal subcellular specificity." In: *Bioinformatics*.

Chen, Y. C., K. Aguan, C. W. Yang, Y. T. Wang, N. R. Pal, and I. F. Chung (2011). "Discovery of protein phosphorylation motifs through exploratory data analysis." In: *PLoS One* 6.5, e20025.

Chen, Z., X. J. Chen, M. Xia, H. W. He, S. Wang, H. Liu, H. Gong, and Y. B. Yan (2012). "Chaperone-like effect of the linker on the isolated C-terminal domain of rabbit muscle creatine kinase." In: *Biophys J* 103.3, pp. 558–66.

Chou, M. F. and D. Schwartz (2011). "Biological sequence motif discovery using motif-x." In: *Curr Protoc Bioinformatics* Chapter 13, Unit 13 15–24.

Choudhary, C., C. Kumar, F. Gnad, M. L. Nielsen, M. Rehman, T. C. Walther, J. V. Olsen, and M. Mann (2009). "Lysine acetylation targets protein complexes and co-regulates major cellular functions." In: *Science* 325.5942, pp. 834–40.

Choudhary, C., B. T. Weinert, Y. Nishida, E. Verdin, and M. Mann (2014). “The growing landscape of lysine acetylation links metabolism and cell signalling.” In: *Nat Rev Mol Cell Biol* 15.8, pp. 536–50.

Clements, A., A. N. Poux, W. S. Lo, L. Pillus, S. L. Berger, and R. Marmorstein (2003). “Structural basis for histone and phosphohistone binding by the GCN5 histone acetyltransferase.” In: *Mol Cell* 12.2, pp. 461–73.

Consortium, The UniProt (2014). “Activities at the Universal Protein Resource (UniProt).” In: *Nucleic Acids Research* 42.D1, pp. D191–D198.

Damle, N. P. and D. Mohanty (2014). “Deciphering kinase-substrate relationships by analysis of domain-specific phosphorylation network.” In: *Bioinformatics* 30.12, pp. 1730–8.

Dou, Y., B. Yao, and C. Zhang (2014). “PhosphoSVM: prediction of phosphorylation sites by integrating various protein sequence attributes with a support vector machine.” In: *Amino Acids* 46.6, pp. 1459–69.

Duan, Guangyou and Dirk Walther (2015). “The Roles of Post-translational Modifications in the Context of Protein Interaction Networks.” In: *PLoS Comput Biol* 11.2, pp. 1–23.

Durek, P., C. Schudoma, W. Weckwerth, J. Selbig, and D. Walther (2009). “Detection and characterization of 3D-signature phosphorylation site motifs and their contribution towards improved phosphorylation site prediction in proteins.” In: *BMC Bioinformatics* 10, p. 117.

Durek, P., R. Schmidt, J. L. Heazlewood, A. Jones, D. MacLean, A. Nagel, B. Kersten, and W. X. Schulze (2010). “PhosPhAt: the Arabidopsis thaliana phosphorylation site database. An update.” In: *Nucleic Acids Res* 38.Database issue, pp. D828–34.

Fan, W., X. Xu, Y. Shen, H. Feng, A. Li, and M. Wang (2014). “Prediction of protein kinase-specific phosphorylation sites in hierarchical structure using functional information and random forest.” In: *Amino Acids* 46.4, pp. 1069–78.

Fillingham, J., J. Recht, A. C. Silva, B. Suter, A. Emili, I. Stagljar, N. J. Krogan, C. D. Allis, M. C. Keogh, and J. F. Greenblatt (2008). “Chaperone control of the activity and specificity of the histone H3 acetyltransferase Rtt109.” In: *Mol Cell Biol* 28.13, pp. 4342–53.

Gao, J. and D. Xu (2010). “The Musite open-source framework for phosphorylation-site prediction.” In: *BMC Bioinformatics* 11 Suppl 12, S9.

Gao, J., J. J. Thelen, A. K. Dunker, and D. Xu (2010). “Musite, a tool for global prediction of general and kinase-specific phosphorylation sites.” In: *Mol Cell Proteomics* 9.12, pp. 2586–600.

Geiss-Friedlander, R. and F. Melchior (2007). “Concepts in sumoylation: a decade on.” In: *Nat Rev Mol Cell Biol* 8.12, pp. 947–56.

Gene Ontology, Consortium (2015). “Gene Ontology Consortium: going forward.” In: *Nucleic Acids Res* 43.Database issue, pp. D1049–56.

Gnad, F., J. Gunawardena, and M. Mann (2011). "PHOSIDA 2011: the posttranslational modification database." In: *Nucleic Acids Res* 39.Database issue, pp. D253–60.

Grotenbreg, G. and H. Ploegh (2007). "Chemical biology: dressed-up proteins." In: *Nature* 446.7139, pp. 993–5.

Henriksen, P., S. A. Wagner, B. T. Weinert, S. Sharma, G. Bacinskaja, M. Rehman, A. H. Juffer, T. C. Walther, M. Lisby, and C. Choudhary (2012). "Proteome-wide analysis of lysine acetylation suggests its broad regulatory scope in Saccharomyces cerevisiae." In: *Mol Cell Proteomics* 11.11, pp. 1510–22.

Hjerrild, M., A. Stensballe, T. E. Rasmussen, C. B. Kofoed, N. Blom, T. Sicheritz-Ponten, M. R. Larsen, S. Brunak, O. N. Jensen, and S. Gammeltoft (2004). "Identification of phosphorylation sites in protein kinase A substrates using artificial neural networks and mass spectrometry." In: *J Proteome Res* 3.3, pp. 426–33.

Hornbeck, P. V., J. M. Kornhauser, S. Tkachev, B. Zhang, E. Skrzypek, B. Murray, V. Latham, and M. Sullivan (2012). "PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse." In: *Nucleic Acids Res* 40.Database issue, pp. D261–70.

Hou, T., G. Zheng, P. Zhang, J. Jia, J. Li, L. Xie, C. Wei, and Y. Li (2014). "LAceP: lysine acetylation site prediction using logistic regression classifiers." In: *PLoS One* 9.2, e89575.

Hubbard, S. J. and J. M. Thornton (1993). "'NACCESS', computer program." In:

Huttlin, E. L., M. P. Jedrychowski, J. E. Elias, T. Goswami, R. Rad, S. A. Beausoleil, J. Villen, W. Haas, M. E. Sowa, and S. P. Gygi (2010). "A tissue-specific atlas of mouse protein phosphorylation and expression." In: *Cell* 143.7, pp. 1174–89.

Iakoucheva, L. M., P. Radivojac, C. J. Brown, T. R. O'Connor, J. G. Sikes, Z. Obradovic, and A. K. Dunker (2004). "The importance of intrinsic disorder for protein phosphorylation." In: *Nucleic Acids Res* 32.3, pp. 1037–49.

Imamura, H., N. Sugiyama, M. Wakabayashi, and Y. Ishihama (2014). "Large-scale identification of phosphorylation sites for profiling protein kinase selectivity." In: *J Proteome Res* 13.7, pp. 3410–9.

Jakob, B., J. Splinter, S. Conrad, K. O. Voss, D. Zink, M. Durante, M. Lobrich, and G. Taucher-Scholz (2011). "DNA double-strand breaks in heterochromatin elicit fast repair protein recruitment, histone H2AX phosphorylation and relocation to euchromatin." In: *Nucleic Acids Res* 39.15, pp. 6489–99.

Joosten, R. P., T. A. te Beek, E. Krieger, M. L. Hekkelman, R. W. Hooft, R. Schneider, C. Sander, and G. Vriend (2011). "A series of PDB related databases for everyday needs." In: *Nucleic Acids Res* 39.Database issue, pp. D411–9.

Kaji, H. et al. (2012). "Large-scale identification of N-glycosylated proteins of mouse tissues and construction of a glycoprotein database, GlycoProtDB." In: *J Proteome Res* 11.9, pp. 4553–66.

Kanehisa, M., S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa (2006). "From genomics to chemical genomics: new developments in KEGG." In: *Nucleic Acids Res* 34.Database issue, pp. D354–7.

Khoury, G. A., R. C. Baliban, and C. A. Floudas (2011). "Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database." In: *Sci Rep* 1.

Kim, S. C. et al. (2006). "Substrate and functional diversity of lysine acetylation revealed by a proteomics survey." In: *Mol Cell* 23.4, pp. 607–18.

Kobe, B., T. Kampmann, J. K. Forwood, P. Listwan, and R. I. Brinkworth (2005). "Substrate specificity of protein kinases and computational prediction of substrates." In: *Biochim Biophys Acta* 1754.1-2, pp. 200–9.

Kreegipuu, A., N. Blom, and S. Brunak (1999). "PhosphoBase, a database of phosphorylation sites: release 2.0." In: *Nucleic Acids Res* 27.1, pp. 237–9.

Kuhn, Max (2008). "Building predictive models in R using the caret package." In: *Journal of Statistical Software* 28.5, pp. 1–26.

Laborde, E. (2010). "Glutathione transferases as mediators of signaling pathways involved in cell proliferation and cell death." In: *Cell Death Differ* 17.9, pp. 1373–1380.

Levy, E. D., S. W. Michnick, and C. R. Landry (2012). "Protein abundance is key to distinguish promiscuous from functional phosphorylation based on evolutionary information." In: *Philos Trans R Soc Lond B Biol Sci* 367.1602, pp. 2594–606.

Li, T., P. Du, and N. Xu (2010). "Identifying human kinase-specific protein phosphorylation sites by integrating heterogeneous information from various sources." In: *PLoS One* 5.11, e15411.

Li, W. and A. Godzik (2006). "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences." In: *Bioinformatics* 22.13, pp. 1658–9.

Li, X. D., Y. J. Yang, Y. J. Geng, J. L. Zhao, H. T. Zhang, Y. T. Cheng, and Y. L. Wu (2012). "Phosphorylation of endothelial NOS contributes to simvastatin protection against myocardial no-reflow and infarction in reperfused swine hearts: partially via the PKA signaling pathway." In: *Acta Pharmacol Sin* 33.7, pp. 879–87.

Li, Y., M. Wang, H. Wang, H. Tan, Z. Zhang, G. I. Webb, and J. Song (2014). "Accurate in silico identification of species-specific acetylation sites by integrating protein sequence-derived and functional features." In: *Sci Rep* 4, p. 5765.

Linding, R., L. J. Jensen, F. Diella, P. Bork, T. J. Gibson, and R. B. Russell (2003). "Protein disorder prediction: implications for structural proteomics." In: *Structure* 11.11, pp. 1453–9.

Linding, R., L. J. Jensen, A. Pasculescu, M. Olhovsky, K. Colwill, P. Bork, M. B. Yaffe, and T. Pawson (2008). "NetworKIN: a resource for exploring cellular phosphorylation networks." In: *Nucleic Acids Res* 36.Database issue, pp. D695–9.

Lu, Z., Z. Cheng, Y. Zhao, and S. L. Volchenboum (2011). "Bioinformatic analysis and post-translational modification crosstalk prediction of lysine acetylation." In: *PLoS One* 6.12, e28228.

Lundby, A. et al. (2012a). "Proteomic analysis of lysine acetylation sites in rat tissues reveals organ specificity and subcellular patterns." In: *Cell Rep* 2.2, pp. 419–31.

Lundby, A., A. Secher, K. Lage, N. B. Nordsborg, A. Dmytriyev, C. Lundby, and J. V. Olsen (2012b). "Quantitative maps of protein phosphorylation sites across 14 different rat organs and tissues." In: *Nat Commun* 3, p. 876.

Maksimoska, J., D. Segura-Pena, P. A. Cole, and R. Marmorstein (2014). "Structure of the p300 histone acetyltransferase bound to acetyl-coenzyme A and its analogues." In: *Biochemistry* 53.21, pp. 3415–22.

Manni, S., J. H. Mauban, C. W. Ward, and M. Bond (2008). "Phosphorylation of the cAMP-dependent protein kinase (PKA) regulatory subunit modulates PKA-AKAP interaction, substrate phosphorylation, and calcium signaling in cardiac cells." In: *J Biol Chem* 283.35, pp. 24145–54.

McGuffin, L. J., K. Bryson, and D. T. Jones (2000). "The PSIPRED protein structure prediction server." In: *Bioinformatics* 16.4, pp. 404–5.

Miller, M. L. et al. (2008). "Linear motif atlas for phosphorylation-dependent signaling." In: *Sci Signal* 1.35, ra2.

Morrison, R. S., Y. Kinoshita, M. D. Johnson, T. Uo, J. T. Ho, J. K. McBee, T. P. Conrads, and T. D. Veenstra (2002). "Proteomic analysis in the neurosciences." In: *Mol Cell Proteomics* 1.8, pp. 553–60.

Murzin, A. G., S. E. Brenner, T. Hubbard, and C. Chothia (1995). "SCOP: a structural classification of proteins database for the investigation of sequences and structures." In: *J Mol Biol* 247.4, pp. 536–40.

Needleman, S. B. and C. D. Wunsch (1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins." In: *J Mol Biol* 48.3, pp. 443–53.

Obenauer, J. C., L. C. Cantley, and M. B. Yaffe (2003). "Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs." In: *Nucleic Acids Res* 31.13, pp. 3635–41.

Okanishi, H., K. Kim, R. Masui, and S. Kuramitsu (2013). "Acetylome with structural mapping reveals the significance of lysine acetylation in Thermus thermophilus." In: *J Proteome Res* 12.9, pp. 3952–68.

Olsen, J. V. and M. Mann (2013). "Status of Large-scale Analysis of Post-translational Modifications by Mass Spectrometry." In: *Mol Cell Proteomics* 12.12, pp. 3444–3452.

Ostlund, G., T. Schmitt, K. Forslund, T. Kostler, D. N. Messina, S. Roopra, O. Frings, and E. L. Sonnhammer (2010). "InParanoid 7: new algorithms and tools for eukaryotic orthology analysis." In: *Nucleic Acids Res* 38.Database issue, pp. D196–203.

Patel, J., R. R. Pathak, and S. Mujtaba (2011). “The biology of lysine acetylation integrates transcriptional programming and metabolism.” In: *Nutr Metab (Lond)* 8, p. 12.

Porollo, A. and J. Meller (2007). “Prediction-based fingerprints of protein-protein interactions.” In: *Proteins* 66.3, pp. 630–45.

Poux, A. N. and R. Marmorstein (2003). “Molecular basis for Gcn5/PCAF histone acetyltransferase selectivity for histone and nonhistone substrates.” In: *Biochemistry* 42.49, pp. 14366–74.

Recht, J. et al. (2006). “Histone chaperone Asf1 is required for histone H3 lysine 56 acetylation, a modification associated with S phase in mitosis and meiosis.” In: *Proceedings of the National Academy of Sciences* 103.18, pp. 6988–6993.

Rojas, J. R., R. C. Trievel, J. Zhou, Y. Mo, X. Li, S. L. Berger, C. D. Allis, and R. Marmorstein (1999). “Structure of Tetrahymena GCN5 bound to coenzyme A and a histone H3 peptide.” In: *Nature* 401.6748, pp. 93–8.

Roskoski R., Jr. (2015). “A historical overview of protein kinases and their targeted small molecule inhibitors.” In: *Pharmacol Res* 100, pp. 1–23.

Sadoul, K., J. Wang, B. Diagouraga, and S. Khochbin (2011). “The tale of protein lysine acetylation in the cytoplasm.” In: *J Biomed Biotechnol* 2011, p. 970382.

Saunders, N. F., R. I. Brinkworth, T. Huber, B. E. Kemp, and B. Kobe (2008). “Predikin and PredikinDB: a computational framework for the prediction of protein kinase peptide specificity and an associated database of phosphorylation sites.” In: *BMC Bioinformatics* 9, p. 245.

Schwartz, D. and S. P. Gygi (2005). “An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets.” In: *Nat Biotechnol* 23.11, pp. 1391–8.

Schwartz, P. A. and B. W. Murray (2011). “Protein kinase biochemistry and drug discovery.” In: *Bioorg Chem* 39.5-6, pp. 192–210.

Shannon, P., A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker (2003). “Cytoscape: a software environment for integrated models of biomolecular interaction networks.” In: *Genome Res* 13.11, pp. 2498–504.

Shao, J., D. Xu, L. Hu, Y. W. Kwan, Y. Wang, X. Kong, and S. M. Ngai (2012). “Systematic analysis of human lysine acetylation proteins and accurate prediction of human lysine acetylation through bi-relative adapted binomial score Bayes feature representation.” In: *Mol Biosyst* 8.11, pp. 2964–73.

Sirover, M. A. (2012). “Subcellular dynamics of multifunctional protein regulation: mechanisms of GAPDH intracellular translocation.” In: *J Cell Biochem* 113.7, pp. 2193–200.

Su, M. G. and T. Y. Lee (2013). “Incorporating substrate sequence motifs and spatial amino acid composition to identify kinase-specific phosphorylation sites on protein three-dimensional structures.” In: *BMC Bioinformatics* 14 Suppl 16, S2.

Suo, S. B., J. D. Qiu, S. P. Shi, X. Y. Sun, S. Y. Huang, X. Chen, and R. P. Liang (2012). “Position-specific analysis and prediction for protein lysine acetylation based on multiple features." In: *PLoS One* 7.11, e49108.

Suo, S. B., J. D. Qiu, S. P. Shi, X. Chen, S. Y. Huang, and R. P. Liang (2013). "Proteome-wide analysis of amino acid variations that influence protein lysine acetylation." In: *J Proteome Res* 12.2, pp. 949–58.

Suo, S. B., J. D. Qiu, S. P. Shi, X. Chen, and R. P. Liang (2014). "PSEA: Kinase-specific prediction and analysis of human phosphorylation substrates." In: *Sci Rep* 4, p. 4524.

Team, R. Development Core (2009). *R: A Language and Environment for Statistical Computing*.

ThermoFisher Scientific. https://www.thermofisher.com/de/de/home/life-science/protein-biology/protein-biology-learning-center/protein-biology-resource-library/pierce-protein-methods/overview-post-translational-modification.html.

Trost, B. and A. Kusalik (2011). "Computational prediction of eukaryotic phosphorylation sites." In: *Bioinformatics* 27.21, pp. 2927–35.

— (2013). "Computational phosphorylation site prediction in plants using random forests and organism-specific instance weights." In: *Bioinformatics* 29.6, pp. 686–94.

Tyanova, S., J. Cox, J. Olsen, M. Mann, and D. Frishman (2013). "Phosphorylation variation during the cell cycle scales with structural propensities of proteins." In: *PLoS Comput Biol* 9.1, e1002842.

UniProt, Consortium (2010). "The Universal Protein Resource (UniProt) in 2010." In: *Nucleic Acids Res* 38.Database issue, pp. D142–8.

Vacic, V., L. M. Iakoucheva, and P. Radivojac (2006). "Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments." In: *Bioinformatics* 22.12, pp. 1536–7.

Villen, J., S. A. Beausoleil, S. A. Gerber, and S. P. Gygi (2007). "Large-scale phosphorylation analysis of mouse liver." In: *Proc Natl Acad Sci U S A* 104.5, pp. 1488–93.

Wang, Benlian, Han Wei, Lakshmi Prabhu, Wei Zhao, Matthew Martin, Antja-Voy Hartley, and Tao Lu (2015). "Role of Novel Serine 316 Phosphorylation of the p65 Subunit of NF-?B in Differential Gene Regulation." In: *Journal of Biological Chemistry* 290.33, pp. 20336–20347. eprint: http://www.jbc.org/content/290/33/20336.full.pdf+html.

Wang, M., M. Weiss, M. Simonovic, G. Haertinger, S. P. Schrimpf, M. O. Hengartner, and C. von Mering (2012). "PaxDb, a database of protein abundance averages across all three domains of life." In: *Mol Cell Proteomics* 11.8, pp. 492–500.

Weinert, B. T., S. A. Wagner, H. Horn, P. Henriksen, W. R. Liu, J. V. Olsen, L. J. Jensen, and C. Choudhary (2011). "Proteome-wide mapping of the Drosophila acetylome demonstrates a high degree of conservation of lysine acetylation." In: *Sci Signal* 4.183, ra48.

Wijk, K. J. van, G. Friso, D. Walther, and W. X. Schulze (2014). "Meta-Analysis of Arabidopsis thaliana Phospho-Proteomics Data Reveals Compartmentalization of Phosphorylation Motifs." In: *Plant Cell* 26.6, pp. 2367–2389.

Xue, Y., X. Gao, J. Cao, Z. Liu, C. Jin, L. Wen, X. Yao, and J. Ren (2010). "A summary of computational resources for protein phosphorylation." In: *Curr Protein Pept Sci* 11.6, pp. 485–96.

Yao, C. et al. (2013). "Role of Fas-associated death domain-containing protein (FADD) phosphorylation in regulating glucose homeostasis: from proteomic discovery to physiological validation." In: *Mol Cell Proteomics* 12.10, pp. 2689–700.

Zhao, Y. and O. N. Jensen (2009). "Modification-specific proteomics: strategies for characterization of post-translational modifications using enrichment techniques." In: *Proteomics* 9.20, pp. 4632–41.

Zhao, Y., X. Li, X. Sun, Y. Zhang, and H. Ren (2012). "EMT phenotype is induced by increased Src kinase activity via Src-mediated caspase-8 phosphorylation." In: *Cell Physiol Biochem* 29.3-4, pp. 341–52.