DOCTOR OF PHILOSOPHY

Computational structure analysis and prediction of Ser/Thr modified by O-GlcNAc in human proteins

Britto Borges, Thiago

Award date:
2016

COMPUTATIONAL STRUCTURE ANALYSIS AND PREDICTION OF Ser/Thr MODIFIED BY *O*-GlcNAc IN HUMAN PROTEINS

By

Thiago Britto Borges

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
AT
UNIVERSITY OF DUNDEE
DUNDEE, UNITED KINGDOM
NOVEMBER 2016

Contents

List of Figures

List of Tables

Acknowledgements

First and foremost, I would like to express my gratitude to my supervisor, Geoff Barton, for the opportunity to work in his group, for his support and patience. I am also grateful to Nicola Stanley-Wall and Bill Hunter for their advice during the thesis committee meetings. I wish to thank the University of Dundee IT department for computing support. Thanks to Daan van Aalten and his group for fruitful discussion on the *O*-GlcNAc topic. Thanks to the Python Programming language community, specially to the Numpy/Scipy, BioPython, Pandas and SciKit-Learn developers and contributors. Many thanks to current and former members of the GJB group Alexey Drozdetskiy, Charles Ofoegbu, Chris Cole, David Martin, Fábio Madeira, Jim Procter, Joe Ward, Kimon Froussios, Kira Mourão, Kostas Moutselos, Marek Gierlinski, Mungo Carstairs, Nancy Giang, Nick Schurch, Pietà Schofield, Reema Singh, Stuart MacGowan and Tara Eckenrode, for technical assistance, enjoyment and scientific discussion. Special thanks to Marialuisa, Stuart, Kira and Nadja who helped me overcome my writing struggles by revising this material. My warmest thanks to my family and friends, close or overseas, who are too many to fit on this page. Love to my partner, Nadja, for sharing and supporting me throughout the hardest of moments and also the good ones. Finally, My acknowledgements would

not be complete without thanking the joint programme between the Coordenação de Aperfeiçamento de Pessoal de Nível Superior (CAPES, process 1529/12-9) and the University of Dundee, which founded this PhD project.

To my parents, to my brothers, and to Nadja.

STATEMENT

School of Life Sciences, University of Dundee

I certify that Thiago Britto Borges has satisfied all the terms and conditions of the relevant Ordinance and Regulations to qualify in submitting this thesis, entitled ‘Computational structure analysis and prediction of Ser/Thr modified by *O*-GlcNAc in human proteins’, in application for the degree of Doctor of Philosophy.

Date: November 2016

Supervisor: ____________________________
Prof Geoffrey J. Barton

DECLARATION

School of Life Sciences, University of Dundee

I hereby declare that the work described in this thesis, entitled 'Computational structure analysis and prediction of Ser/Thr modified by *O*-GlcNAc in human proteins', is my own; that I am the author of this thesis; that it has not previously been put forward in submission for any other degree or qualification; and that I have consulted references herein.

Date: November 2016

Applicant: ______________________________
Thiago Britto Borges

Abstract

This thesis studies the post-translational modifications of proteins by *O*-GlcNAcylation with a computational biology approach. The *O*-GlcNAc transferase (OGT), the enzyme that catalyses the protein *O*-GlcNAcylation, targets specific Serines and Threonines (S/T) of intracellular proteins. However, while other post-translationally modified residues, including phosphorylated ones, occur within sites distinguished by their amino acid sequences, less than 25% of known *O*-GlcNAc sites match to a sequence pattern. The small signal on the sequence patterns of multiple sites leads to the question whether the sites' structure defines the pattern recognised by OGT. The thesis then focuses on the structural features of the modified sites that could help distinguish potential sites from non-modifiable ones.

1622 *O*-GlcNAc sites were collected from the scientific literature. Next, 143 sites were mapped to protein 3D structure in the PDB. Modified S/T were 1.7 times more likely than unmodified S/T in the same protein to be annotated in the REMARK465 field of the PDB file, which defines missing regions in the protein structure, suggesting that these sites may be in structurally disordered regions. Clustering the structure of *O*-GlcNAc sites leads to ten distinct groups indicating the sites' structural diversity. The study was extended by the analysis of features predicted from the sequence of

O-GlcNAcylated proteins with Jpred4 and 3 disorder predictors, DisEMBL, IUpred and JRonn. Overall, disorder scores and proportion of S/T in coils confirmed that *O*-GlcNAc sites tend to be disordered.

A new classifier for *O*-GlcNAc-site (POGSPSF) was developed and trained with sequence, predicted secondary structure and disordered from 1 283 non-redundant sites. The POGSPSF Random Forest model achieved 71% area under the ROC curve in a blind test. Predictions were applied to around 2.5 million S/T in the human proteome. Nuclear and cytoplasmic protein were over-represented among the top ranking proteins. Top scoring sites were also more likely to be phosphorylated. Also, novel and potential proteins were identified within the predictions.

Abbreviations

AAS Amino Acid Substitution.

ANN Artificial Neural Network.

API Application Programming Interface.

AUC Area Under the Curve.

CDS Coding DNA Sequence.

CI Confidence Intervals.

dbOGAP Database of *O*-GlcNAcylated Proteins and Sites.

DBSCAN Density-Based Spatial Clustering of Applications with Noise.

DSSP Define Secondary Structure of Proteins.

EBI European Bioinformatics Institute.

eOGT EGF domain-specific *O*-GlcNAc-Transferase.

ExAC The Exome Aggregation Consortium.

FE Fold Enrichment.

FN False Negatives.

FP False Positives.

FPR False Positive Rate.

GO Gene Ontology.

GRCh37 Genome Reference Consortium human (version 37 from March 2009).

GS Globular Set.

HGVS Human Genome Variation Society.

JSON JavaScript Object Notation.

MCC Matthews Correlation Coefficient.

mmCIF macromolecular Crystallographic Information File.

mOGT mitochondrial *O*-GlcNAc Transferase (OGT).

MSS Modification Sequence Sites.

ncOGT nuclear and cytoplasmic OGT.

NGS Next-Generation Sequencing.

nsSNV Non-Synonymous Single Nucleotide Variant.

OGA O-GlcNAcase.

OGT *O*-GlcNAc Transferase.

PDB Protein Data Bank.

PDBe PDB in Europe.

POGSPSF Predict *O*-GlcNAc Sites from Protein Sequences and Features.

PolyPhen-2 Polymorphism Phenotyping (version 2).

ProteoFAV Protein Feature Aggregator and Variants.

PSSM Position-Specific Scoring Matrix.

PTM Post-Translational Modification.

RBF Radial Basis Function.

RMSD Root-Mean-Square Deviation.

ROC Receiver Operating Characteristic.

RSA Relative Solvent Accessibility.

RVIS Residual Variation Intolerance Score.

S/T Serine/Threonine.

SIFT Sorting Intolerant from Tolerant.

SIFTS the Structure Integration with Function, Taxonomy and Sequence.

SNNS Stuttgart Neural Network Simulator.

SNV Single Nucleotide Variant.

sOGT short OGT.

SQL Structured Query Language.

SS131 Structure Sites with backbone.

SS143 Structure Sites.

sSNV synonymous Single Nucleotide Variant.

SVM Support Vector Machines.

TCGA The Cancer Genome Atlas.

TN True Negatives.

TP True Positives.

TPR Tetratricopeptide Repeats.

TPR True Positive Rate.

UDP Uridine Diphosphate.

UniProt the Universal Protein Resource.

UniProtKB UniProt KnowledgeBase.

USS Unmodification Sequence Sites.

VEP Variant Effect Predictor.

XML Extensible Markup Language.

Z-DOPE Discrete Optimized Protein Energy.

Chapter 1

Introduction

1.1 The central dogma of molecular biology

The central dogma of molecular biology defines how the information flows from the DNA molecule to proteins, via RNA molecules. Crick (1970) sets out the principles of information transfer between the three macromolecules. The central concept is the direction in which information flows: DNA → RNA → protein as illustrated in Figure 1.1. This dogma is common to all living beings, from the simplest unicellular to complex multicellular organisms. Crick also highlighted and discussed possible exceptions.

Figure 1.1 emphasizes three processes that are also basic to living organisms. The molecular basis of these processes is better understood today than when the central dogma was first described; moreover, the molecular basis varies among studied organisms. The DNA molecule stores the genetic information that is replicated, during cell duplication. So the two daughter cells inherit the genetic information from the two copies of DNA produced during replication.

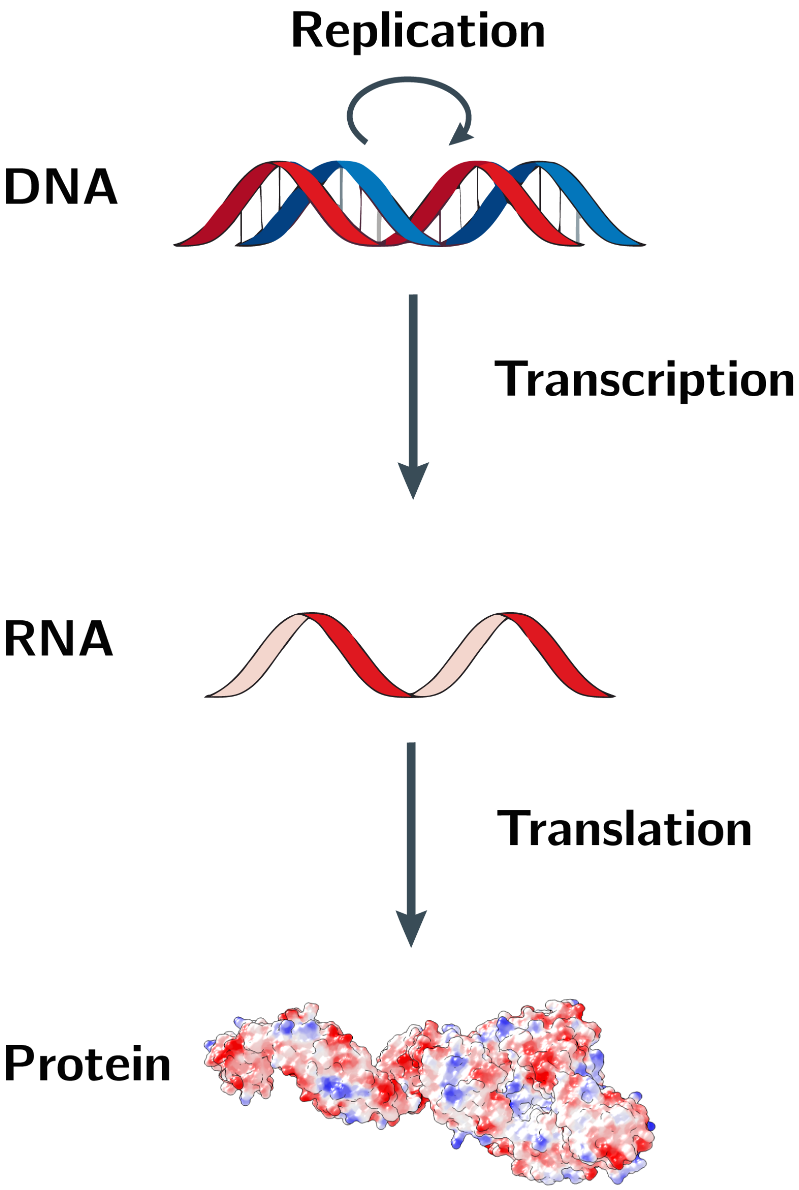


Figure 1.1: A simplified representation of the central dogma of molecular biology. The central dogma establishes the flow of genetic information or the sequential transfer of information from the DNA to the proteins. The DNA conserves genetic information among generations. The RNA carries the information from the DNA and subsequently, is decoded into proteins. The protein image refers to the proteins' three-dimensional atomic model built and discussed in Chapter 5. As proteins are central in this work, their representation was not simplified. Modified from Fu et al. (2014)

The DNA → RNA information transfer occurs during the transcription process. The transcription process uses DNA as a template and generates a transcript, an RNA molecule containing the information encoded by the DNA. There are various types of RNA molecules with different functions within the cell and the regulation of transcript levels is a major step in cell physiology. However, the study of RNA molecules is not in the scope of this thesis. Ribosomes are molecular machines that translate the information encoded in the transcript, building the proteins.

All macromolecules in biological systems are built of monomers. 4 nucleotides make the DNA and RNA: adenine, guanine, thymine (substituted by uracil in RNA) and cytosine. The complexity for protein is higher, since proteins are built up from 20 amino acids. Table 1.1 lists the 20 amino acids, which differ in physico-chemical properties, like size, shape and charge. Proteins will vary in length and amino acid composition, differences that yield proteins with unique properties.

Ribosomes are molecular machines that synthesise proteins by reading information from transcripts. During translation, the ribosomes extend the nascent protein chain by linking the next amino acid to the chain, which has its carboxyl covalently bound to the amino group of the nascent chain. After that, a peptide bond is formed between the two amino acids. The first amino acid of the proteins' chains contains a free amino group and is called the N-terminus. The other end, which holds a free carboxyl group, is called C-terminus.

From the evolutionary perspective, one can define a gene as a unit of hereditary characteristics (Alberts et al., 2010). Genes are segments of the DNA containing the information for proteins, RNAs and elements that regulate the transcription process. The Coding DNA Sequence (CDS) is the region of the gene that contains

Table 1.1: List of the 20 standard amino acids, sorted by 1-letter name. The Type column classifies each amino acid in 6 physico-chemical groups. The Polarity shows the side chain polarity. The Charge column shows the side chain charge at neutral pH.

Amino Acid	**3-Letter**	**1-Letter**	**Type**	**Polarity**	**Charge**
Alanine	Ala	A	aliphatic	nonpolar	neutral
Cysteine	Cys	C	sulfur-containing	nonpolar	neutral
Aspartic acid	Asp	D	acid/amide	acidic polar	negative
Glutamic acid	Glu	E	acid/amide	acidic polar	negative
Phenylalanine	Phe	F	aromatic	nonpolar	neutral
Glycine	Gly	G	aliphatic	nonpolar	neutral
Histidine	His	H	basic	basic polar	positive
Isoleucine	Ile	I	aliphatic	nonpolar	neutral
Lysine	Lys	K	basic	basic polar	positive
Leucine	Leu	L	aliphatic	nonpolar	neutral
Methionine	Met	M	sulfur-containing	nonpolar	neutral
Asparagine	Asn	N	acid/amide	polar	neutral
Proline	Pro	P	cyclic	nonpolar	neutral
Glutamine	Gln	Q	acid/amide	polar	neutral
Arginine	Arg	R	basic	basic polar	positive
Serine	Ser	S	hydroxyl-containing	polar	neutral
Threonine	Thr	T	hydroxyl-containing	polar	neutral
Valine	Val	V	aliphatic	nonpolar	neutral
Tryptophan	Trp	W	aromatic	nonpolar	neutral
Tyrosine	Tyr	Y	aromatic	polar	neutral

the information for the protein sequence. The start and end codons in a CDS indicate the protein start and end, or the N- and the C-terminus, respectively.

Genes and other biological sequences can be conserved. Conserved sequences are identical or similar, when the substitution of a few amino acids is allowed. Conserved sequences inherited from a common ancestor of two species are named homologous. Homologous sequences that diverge in sequence composition but not in function are named orthologous. Two sequences are paralogous if they originate from a gene duplication event, and resulting sequences may have different sequence composition or function. These three terms, orthologous, paralogous and homologous, organise the comparison of biological sequences, which can be made via a multiple sequence alignment. Biological sequences include not only DNA, RNA and protein sequences but also the order of polysaccharide, another class of macromolecules, discussed in Section 1.3.2.

However, DNA, RNA and proteins are not a simple string of monomers. They occupy the three-dimensional space, forming complex structures that are related to their function. The Section 1.2 discusses the relationship between the structure and function of proteins.

1.2 Proteins

"Proteins ... are the molecules that put cells' genetic information in action." (Alberts et al., 2010). Proteins drive every dynamic process in cells. Small chains (< 30) of amino acids are called peptides, and therefore proteins are also named polypeptides. Enzymes are proteins that catalyse chemical reactions. However, enzymes are not the only type of proteins, and proteins have a large functional repertoire, like structuring the cell shape, sensing the extracellular environment, moving and regulating other macromolecules within the cell. A fact often ignored is that a single protein may have multiple unrelated functions (Jeffery, 1999; Henderson and Martin, 2014).

After translation, each protein folds into a unique three-dimensional structure. Experiments with the ribonuclease enzyme indicated that the protein sequence determines the enzyme biological activity and, therefore, its three-dimensional structure (Anfinsen, 1973). In the protein native three-dimensional structure, the protein achieves a state of minimum energy. The experiments also demonstrated that outside the cell's conditions proteins would fold into an alternative conformation without biological activity, implying that the minimum energy state is not unique. However, under certain conditions, the protein would fold into the native - or biological active - state. Thus protein sequence, structure and function are associated.

Given the possible combinations of the 20 amino acids and their interactions, protein folding is hard to study. However, two main physical components drive the process. Firstly, the protein core forms from hydrophobic amino acids, which tend to avoid contact with the solvent (water) and other hydrophilic side chains. Secondly, contacts between amino acids, mainly local hydrogen bonds, stabilise the protein structure. Ongoing research aims to understand better and simulate the protein folding process. It is important to note that proteins are not static, but rather dynamic macromolecules that fluctuate among conformations. The collection of conformation and their transitions can also be associated with the protein function (Karplus and Kuriyan, 2005; Henzler-Wildman and Kern, 2007).

Structural biology is a sub-area of biological sciences that studies the structure of proteins and how changes in a protein structure influence its function in the cell.

1.2.1 Structural biological concepts

A hierarchy of 4 levels organises proteins' structure. The primary structure is the linear sequence of amino acids in the protein chain, encoded by the CDS in a gene. The secondary structure accounts for the local structure of the backbone atoms of the polypeptide chain. The backbone of a residue comprises four atoms: the Cα, an O, an N and a C. The Cβ atom links the residue backbone to its side chain, which confers the physico-chemical properties of amino acids (See Table 1.1). Hydrogen bonds stabilise the local structure and form regular structures, called secondary structure elements. There are several patterns that, in general, can be simplified into two main groups by the local structure of proteins: α-helix (H), or helices, and β-strands (E) (Pauling and Corey, 1951; Pauling et al., 1951). A third type incorporates the lack of a regular structure, coils (C). For H, the amino groups of a residue i forms a hydrogen bond with the carboxyl group if residue $i - 4$ and regularity comprise several residues, yielding a helix. In E the backbone atoms of three or more residues are connected to the backbone atoms of residues not immediately adjacent to the structure, producing an **E**xtended planar conformation.

Interactions among secondary structure elements of a protein produce a higher-level structure called a fold. The tertiary structure comprises the three-dimensional position of every amino acid and also the contacts between non-local amino acids and arrangement of amino acid side chains. The quaternary structure includes the protein associations, the protein-protein and protein-ligand interactions. Proteins can comprise multiple subunits that assemble to promote the protein function. Homo-oligomers are protein complexes with multiple copies of the same protein; hetero-oligomers are protein complexes with more than one polypeptide chain. Proteins bind to a vast array of ligands from ions, like zinc, to molecules, such as ATP or water. These interactions can vary in time scale, while some proteins will always participate in complexes, and others will only interact temporarily. In fact, many proteins work in groups, called protein complexes, rather than individually.

X-ray crystallography is the most popular method to determine protein structure at an atomic level. The method has several steps and is labour intensive. First, the target protein needs to be pure and homogeneous for crystallisation, a process in which the protein molecules pack together in a ordered three-dimensional array. Subsequently, the crystal is exposed to X-rays generating a diffraction pattern. The atoms' electron density can be deduced from the diffraction pattern, but not before a series of processing steps that aim to understand the nature of the crystal, like its symmetry, the unit cell parameters, orientation and resolution limit, and to solve the phase problem. The atomic model is build from the election density and undergoes several iterations of refinement. Finally, it will be validated and submitted to the Protein Data Bank (PDB). More recently, the number of protein structures determined by nuclear magnetic resonance and cryo-electron microscopy is increasing rapidly.

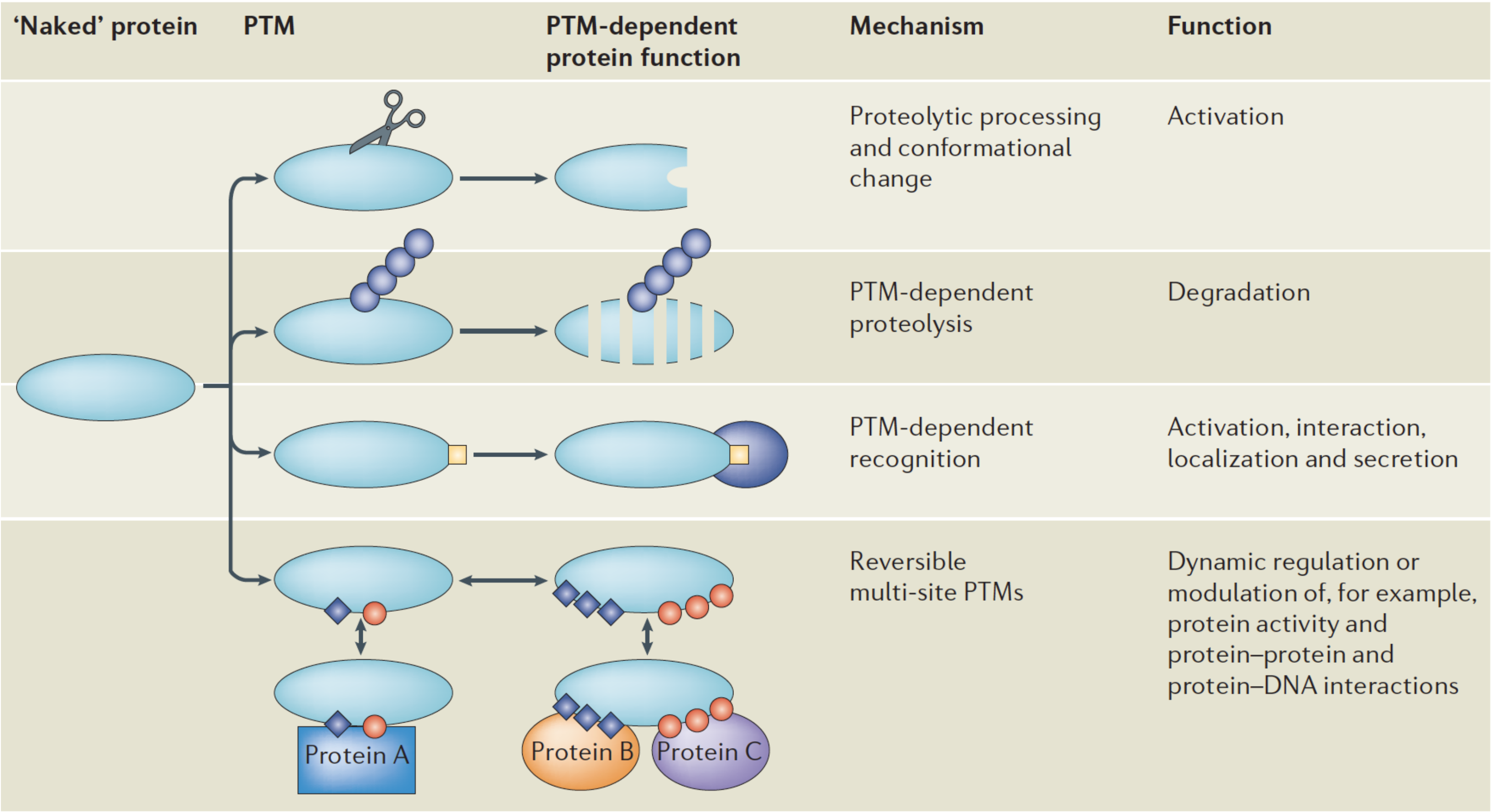
Although the definition of protein domains may change from field to field, domains are typically considered compact regions of the protein three-dimensional structure with some degree of functional independence (Siddiqui and Barton, 1995). An example of a domain found in proteins with different functions is the Tetratricopeptide

Repeats (TPR) domain (PFAM website, 2016b), known for mediating protein-protein interactions (Hirano et al., 1990; Das et al., 1998). Proteins can have more than one domain, so domains work like protein modules. Motifs are short regions of the primary structure that also have some independent function, such ligand binding. Both domains and motifs set patterns or amino acid signatures that can be studied to understand their function. Also, motifs and domains are often conserved among proteins with similar function. The protein function can be regulated, and the next section describes one fundamental regulatory process.

1.3 Protein post-translational modifications

The collection of proteins in a cell is named the proteome, and two major components control the proteome diversity. The first mechanism is at the RNA level, involving a series of processes that include RNA splicing. The second mechanism occurs at the protein level.

After translation and folding, a protein can have its backbone cleaved at a specific point or one of its residues attached to chemical groups. These modifications are called Post-Translational Modification (PTM). Proteases are enzymes that irreversibly cleave proteins at unspecific or specific points. Conversely, the reversible attachment of chemical groups to polypeptide chains is the second type of PTMs. The modifications occur in sites, specific positions within the protein chain. Overall, these modifications participate in the cell signalling process that coordinates the actions within a cell in response to extra- or intracellular messengers. Figure 1.2 describes 4 types of PTM by functional roles. This thesis focuses on the last group:



'Naked' protein	PTM	PTM-dependent protein function	Mechanism	Function
			Proteolytic processing and conformational change	Activation
			PTM-dependent proteolysis	Degradation
			PTM-dependent recognition	Activation, interaction, localization and secretion
			Reversible multi-site PTMs	Dynamic regulation or modulation of, for example, protein activity and protein–protein and protein–DNA interactions

Figure 1.2: PTM can be organised by function into four groups. The first is proteolytic processing that activates the protein or peptide function. Proteolytic processing is also related to the controlled cell death process, named apoptosis and the second represents PTM-dependent proteolysis when a protein is directed to specialised organelles where they are degraded. In the third group, the modification enables protein-protein interactions. The fourth group is probably the most common one, and contains multiple reversible PTM. Reprinted with permission from Macmillan Publishers Ltd: Nature Reviews Molecular Cell Biology, (Jensen, 2006), copyright 2006.

reversible multi-site PTM.

Regularly protein functional models obtained from experiments show PTM as on and off switches, which describe proteins' gain or loss functions upon modification. A more accurate classification of how reversible PTM regulates protein functions considers a few points. Upon modification, the protein can change cellular localisation (Fushimi et al., 1997) or be degraded (Poizat et al., 2005). The PTM may also: modulate enzyme activity and change its affinity to ligands (Hilário-Souza et al., 2011); either disrupt or direct protein-protein interactions; affect protein structure and dynamics (Xin and Radivojac, 2012). Accordingly, PTM adds a layer of control over protein function that is more complex than a switch. Over 235 chemical modifications have been experimentally catalogued (Khoury et al., 2011). The modification types vary in size and physicochemically from a charged phosphate group to a 70-residue long protein, ubiquitin. Although some PTM types are very well studied, others were barely confirmed *in vivo*. Multiple enzymes can target the same protein so that the protein can be modified in different sites by different PTMs types (Cohen, 2000). The evidence of multi-site PTM increases the protein's complexity, because of all the possible combinations of sites and modification types. Of course, not all combinations are observed in experiments. Moreover, enzymes that catalyse PTM may compete for a particular residue. Also, a modified site may disrupt the subsequent modification of a nearby site, for mutually exclusive sites; some sites are only available for modification after the modification of an adjacent site. Examples of sites that do not affect the modifications are also known.

Not all PTM sites have an impact on protein function (Beltrao et al., 2012). It is experimentally challenging to track down a PTM site function. Moreover, even if the functional impact is unobserved in an experiment, it should not be entirely discarded. PTM can only slightly change the protein structure (Xin and Radivojac, 2012) and a particular functional impact may be unobserved. A site which is conserved among homologous proteins is more likely to be functional. However, the phylogenetic comparison of sites is not simple. For example, Beltrao et al. (2012) describes a phosphorylation at Ser40 of the fungal protein Skp1. The modification promotes an interaction that, in the human ortholog, is instead mediated by phosphorylation at Tyr20. So, although the sites are not conserved between the fungal and human proteins, the function is. Overall, the conservation analysis of PTM is hard because of the lack of information on the sites' functions in different organisms; therefore the function of modification should be studied on a site by site basis.

Most of the enzymes that catalyse PTM recognise a particular segment within the protein sequence, called a sequon. Another important characteristic of PTM sites is their sub-stoichiometry. Not all protein copies will be modified at the same point in time, and the ratio of unmodified / modified may be related to its function. Because kinases are very popular drug targets, their structure and function have been extensively studied (Dumas, 2001). Section 1.3.1 describes this modification in more detail.

1.3.1 Reversible protein phosphorylation

Reversible protein phosphorylation is the most studied PTM and is present in the 6 kingdoms of life. Hubbard and Cohen (1993) estimated that, in eukaryote cells, one-third of the proteins will undergo phosphorylation. In humans, the protein kinase superfamily, the family of enzymes that catalyse protein phosphorylation, is one of the biggest protein families, encoded by $\approx 2\%$ of the genome (Manning, 2002). "It is difficult to find a physiological reaction that is not directly or indirectly affected by protein phosphorylation" (Fischer, 2016). A subset of known cells' physiological processes are regulated by reversible protein phosphorylation:

1. Cell death

2. Cell cycle

3. Cytoskeletal rearrangement

4. Cell differentiation

5. Immune response

6. Transcription

7. Translation

8. Metabolism

Malfunction of the phosphorylation regulatory system has been linked to a series of hereditary diseases and cancer, as reviewed by Cohen (2001).

The reversible phosphorylation cycles the attachment and removal of a phosphate to Ser/Thr/Tyr. The reaction consists of the transfer of ATP's γ-phosphate to the side chain of residues within specific sites of the target protein. Phosphorylation in bacteria and fungi can also modify His and Asp residues. Phosphatases remove the attachment; this family of enzymes varies from promiscuous to very specific to a site. Both kinases and phosphatases are regulated some times by other kinases, in processes called kinase cascades. The process is explained by the kinase mediated kinase-activation, where one kinase regulates the following kinase in the cascade. Malfunction of the phosphorylation regulatory system has been linked to a series of hereditary diseases and cancer, as reviewed by Cohen (2001).

The side chains of Ser and Thr contains a hydroxyl group (-OH) that is uncharged in cellular conditions. Upon modification at physiological pH, the O-phosphate has 2 negative charges and can pair with cationic residues, such as Arg (Walsh et al., 2005). Protein conformational change often accompanies phosphorylation. The change may occur in the secondary, tertiary or quaternary structures to accommodate the phosphoryl group and its charges. Conformation transition differs from structure to structure, as reviewed by Johnson and Lewis (2001). For instance, the Tau protein - a protein involved in neurodegenerative diseases - is globally disordered in solution (Schweers et al., 1994). Upon phosphorylation, Tau's transient α-helix stabilises, and this stabilisation changes the way it interacts with tubulin (Sibille et al., 2012) and, hence, its function - which is to induce tubulin polymerization into microtubules. A general rule is that phosphorylation induces small structural changes, which affects the local structure and reduces conformational heterogeneity (Xin and Radivojac, 2012).

1.3.2 Protein glycosylation

Polysaccharides are one of the four macromolecules that constitute living cells, the other 3 been nucleic acids (DNA and RNA), proteins, and lipids. The polysaccharide chains are formed by monosaccharides, which are also named sugars. Glycosylation is the enzymatic linkage of saccharides to proteins, lipids and other saccharides. The attachment of mono or polysaccharide to a protein forms glycoproteins, the most widespread PTM.

Varki et al. (2009) classify mammalian glycosylation products into 7 classes. An N-acetylglucosamine attached to an asparagine characterises the N-glycan; these polysaccharides are very diverse in composition and branched. Hyaluronan is a group of extracellular and linear polysaccharides formed by N-acetylglucosamine and glucuronic acid disaccharides in tandem. Mucins O-glycans have an N-acetylgalactosamine α-linked to Serine/Threonine (S/T) in the target protein; non-mucin O-glycans have different sugar core, such as mannose, fucose or glucose. The Glycosylphosphatidylinositols (GPI) are covalently bound to a protein and a phosphoethanolamine group (Et-P), which works as a membrane anchor to proteins. Another form of glycolipids also occurs, specially glycosphingolipids in vertebrates. Glycosaminoglycans are composed of linear disaccharide tandem repeats and have a very specific protein binding activity. The only class found to happen in intracellular proteins is the β-*O*-GlcNAcylation of S/T.

Structural characteristics determine the glycoproteins' diversity. The glycosidic bond joins the carbohydrate to another chemical group in the α and β forms, depending on the saccharide and group stereochemistry. Also, a polysaccharide chain can have multiple branches, increasing the polymer complexity. Figure 1.3 illustrates the diversity of saccharides in mammalian organisms.

Since protein glycosylation builds up such diversity, it is hard to pinpoint a function for this PTM. In mammalian organisms, most groups of protein glycosylation happen in the lumen of the endoplasmic reticulum and the Golgi apparatus, two specialised compartments in the cell. The glycoproteins transit in the secretory pathway and stay on the cell-surface, facing the extracellular side of the membrane or

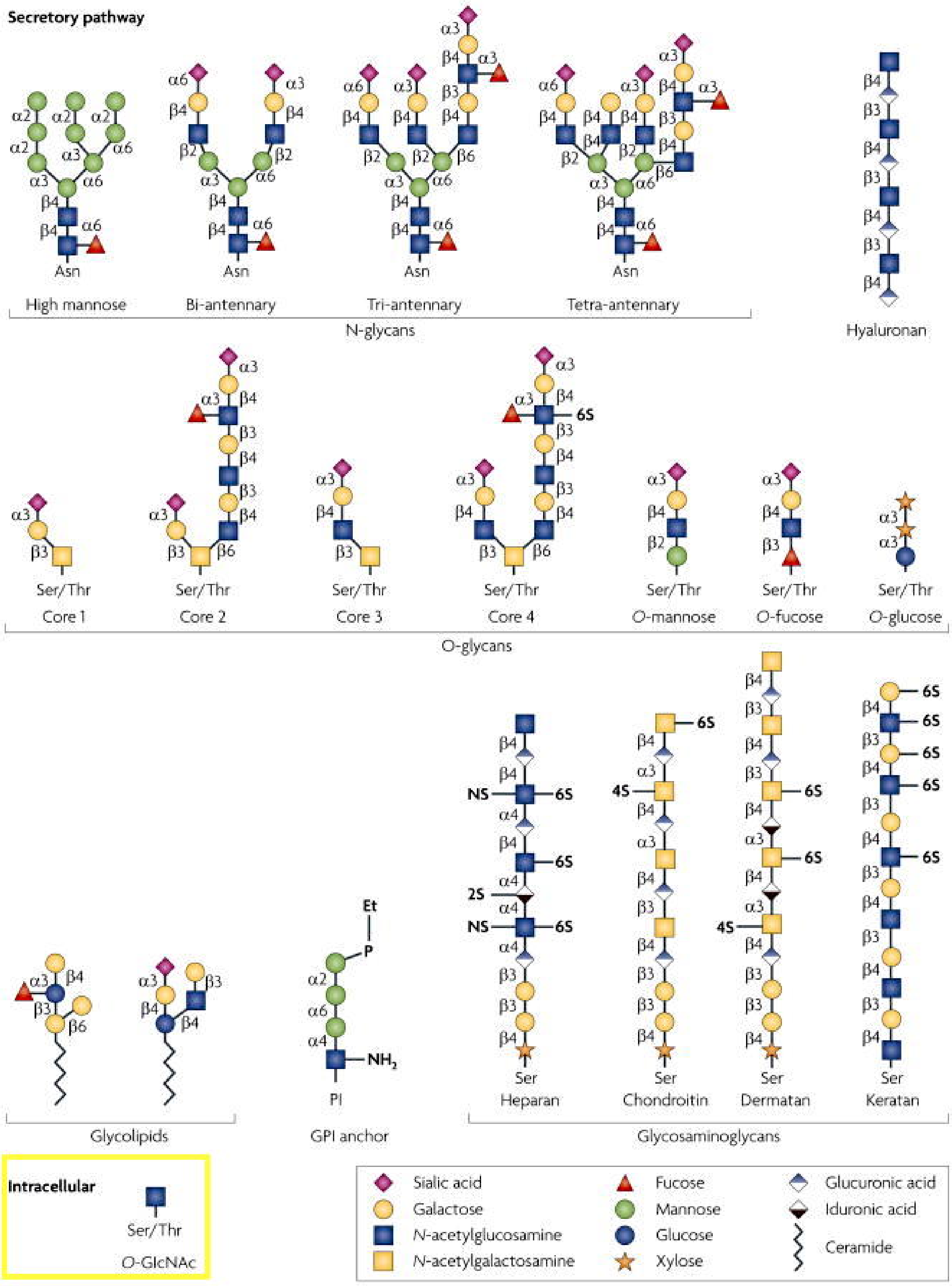


Figure 1.3: The different types of mammalian saccharides. Glycoproteins are complex molecules, part due to the diversity of saccharides and the polymer structure. *O*-GlcNAc, highlighted in yellow, is a monosaccharide intracellular modification of proteins and the theme of this thesis. Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Immunology, (Marth and Grewal, 2008), copyright 2008.

secreted into the extracellular medium. These glycoproteins that are exposed to the extracellular medium are involved in cell-to-cell recognition, cell death and receptors activation. In addition, most of the secreted protein are attached to saccharides, and part of these glycoproteins will form the extracellular matrix.

The collection of glycoproteins changes over time and in response to signals. Differently from RNA and protein sequences, DNA does not directly encode the polysaccharide sequence. While replication, transcription and translation are template-dependent processes, the polysaccharide sequence depends on the order that different enzymes will attach a saccharide to the chain in the secretory pathway.

Glycosyltransferase and glycosidase are enzymes that catalyse the attachment and the removal of saccharides from proteins. Almost every type of protein glycosylation is irreversible, so once the saccharide is attached to the protein, it won't be removed until the protein degradation.

Glycobiology, the science that studies protein glycosylation, has a community of its own. Protein glycosylation is an abundant and diverse in composition and structure of the modifiers. This work focuses on the protein *O*-GlcNAcylation modification, and its interplay with other PTMs.

1.4 Protein *O*-GlcNAcylation

An important, but yet overlooked, form of PTM is *O*-GlcNAcylation, also called *O*-GlcNAc. Torres and Hart (1984) first described the glycosylation of intracellular proteins, in particular, proteins of the nuclear envelope. Later *O*-GlcNAc modification was found in nuclear, cytoplasmic and mitochondrial proteins (Holts and Hart, 1986).

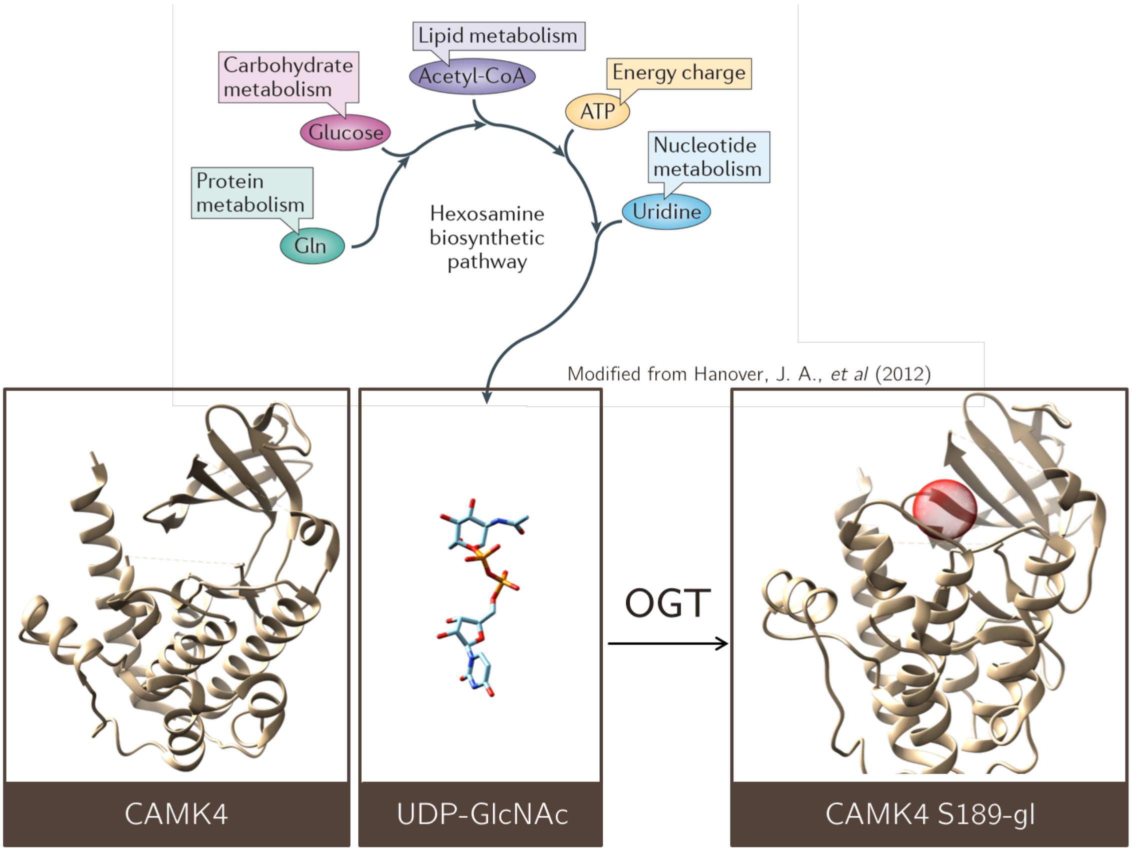


Figure 1.4: OGT has two substrates: the UDP-GlcNAc and a target protein. In the top panel, hexosamine biosynthetic pathway produces UDP-GlcNAc. 4 metabolic pathways plus the ATP molecule converges in the hexosamine biosynthetic pathway. The pathway uses 2%-5% of the glucose that enters the cell; since glutamine, ATP and glucose are precursors of the UDP-GlcNAc, its intracellular level is associated with the cell's nutritional state, hence UDP-GlcNAc is a nutrient sensor. The molecule is also reactant for the synthesis of other glycolipids and glycoproteins. In the bottom panel, OGT modifies the Calcium/calmodulin-dependent protein kinase type IV (CaMK IV) in several residues. The red sphere represents the approximate location of the Ser189 modification, which occurs in an unobserved segment of the protein structure 2w4o (unpublished).

The modification of intracellular proteins by O-β-glycosylation is common to eukaryotes. The modification has been demonstrated experimentally in *Trichoplax adhaerens* (Selvan et al., 2015), *Caenorhabditis elegans* (Lubas et al., 1997), *Drosophila melanogaster* (Sinclair et al., 2009), plants (Chen et al., 2005) and mammals, but not in yeast. This thesis focuses on mammalian organisms since the vast majority of known sites were obtained from those organisms.

Several differences separate *O*-GlcNAc from other classes of protein glycosylation. As already mentioned, *O*-GlcNAc occurs outside the secretory pathway. Also, glycoproteins have complex structures and composition, as highlighted in Figure 1.3, in contrast to *O*-GlcNAc, which is a monosaccharide. Furthermore, while most mature glycoproteins are stable, the *O*-GlcNAc modification is actively attached and removed in a controlled manner (Kearse and Hart, 1991; Chou et al., 1992). Thus *O*-GlcNAc is a dynamic PTM, as is regulatory protein phosphorylation. The catalytic cycle for protein *O*-GlcNacylation and protein phosphorylation is also similar. Two enzymes control *O*-GlcNAc cycling: the glycosyltransferase OGT and the glycosidase O-GlcNAcase (OGA). The first enzyme catalyses the β-linked attachment of a single GlcNAc from the UDP-GlcNAc molecules to the side chain of a S/T in the acceptor protein. For the opposite reaction, OGA cleaves the GlcNAc from the protein.

Protein phosphorylation and *O*-GlcNAc have two major differences. First, a vast range of enzymes, more than 500 kinases and 200 phosphatases, drives protein phosphorylation, while only OGT and OGA control *O*-GlcNAc cycling. Figure 1.4 shows the modification processes and describes the role of the UDP-GlcNAc. The second difference regards the modification site. While phosphorylation has a clear sequon, a consensus motif for modification, the same is not true for protein *O*-GlcNAcylation. In the early days (Gupta et al., 1999), when around 50 *O*-GlcNAc sites were mapped to their protein sequences, 50% of sites held the pattern `PV[ST]` for residues -2, -1 and 0, where the residues zero is the modified S/T. Chapters 2 and 3 further discuss this point, but overall the scientific literature agrees on the lack of clear sequon for sites modified by *O*-GlcNAc, in contrast to kinases, which have clear sequons.

Interestingly, some S/T known to be *O*-GlcNAc are also phosphorylated. The Yin-O-Yang hypothesis describes the interplay between the two modifications (Hart et al., 1995). OGT and kinase targeted sites were observed in various proteins. So some kinases compete with OGT for the site's modification. In fact, *O*-GlcNAc and phosphosites occur not only in the same residues but also in neighbouring residues and Chapter 4 discusses this event. The structural effects of protein phosphorylation have been extensively studied (Johnson and Barford, 1993; Johnson and Lewis, 2001; Xin and Radivojac, 2012). However, the effects of *O*-GlcNAc modification on substrate's structure is unknown, due to the lack of full-length OGT-modified protein structures. Overall, there are many more reports and data on protein phosphorylation than on protein *O*-GlcNAcylation.

The number of scientific reports on protein phosphorylation outnumbers the reports on *O*-GlcNAc by 1 to 2 orders of magnitude. However, the role of protein *O*-GlcNAcylation in health and disease has also been partially established. *O*-GlcNAcylation of Akt1 has been associated with pancreatic cell death (Kang et al., 2008). Proper *O*-GlcNAc cycling is also required for oocyte maturation (Lefebvre et al., 2004), indicating that the modification participates in the cell cycle. OGT targets various cytoskeletal proteins (Ramirez-Correa et al., 2008; Kakade et al., 2016) and the disruption of the modification has deleterious implications for the cell. Although the PTM was discovered in lymphocytes, its role in the T cell-mediated (Swamy et al., 2016) and innate (Ryu and Do, 2011) response to infection was only recently described. Processes like transcription (Kelly et al., 1993), translation (Datta et al., 1989; Zeidan et al., 2010) and macromolecules metabolism (Patti et al., 1999) have been linked to *O*-GlcNAcylation and have been extensively reviewed in the scientific literature (Hardivillé and Hart, 2014; Hanover et al., 2012). More recently the modification was found involved in stress resistance in mammalian cells (Zachara and Hart, 2004), heat resistance in *Drosophila melanogaster* (Radermacher et al., 2014) and interplaying with phosphorylation in the control of the circadian clock (Kaasik et al., 2013). OGT and OGA play a role in several diseases such as diabetes, neurodegeneration and cancer, as extensively reviewed elsewhere (de Queiroz et al., 2014; Bond and Hanover, 2013; Ngoh et al., 2010; Hart et al., 2011).

The two following sections detail the structure and function of the enzymes that control *O*-GlcNAc cycling.

1.4.1 *O*-GlcNAc transferase

OGT (Enzyme Commission number 2.4.1.255) was first isolated in 1993, but the human gene was only cloned later (Lubas et al., 1997)(Ensembl identifier `ENSG00000147162`). The human gene is located on chromosome X and encodes 3 proteins produced by alternative splicing. The longest product is a protein with 110 kDa and 1046 residues expressed constitutively in all human tissues. The gene has a highly conserved primary sequence among mammalians (Kreppel et al., 1997) and is essential for

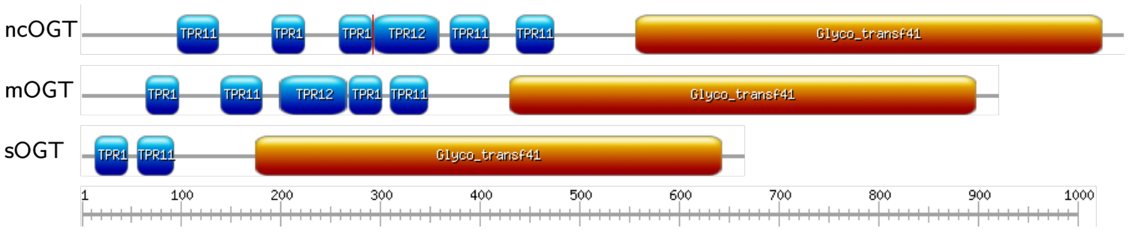


Figure 1.5: Domain architecture for OGT 3 isoforms. Designed with `http://prosite.expasy.org/mydomains/`. nc, nucleocytoplasmic; m, mithocondrial; s, short.

embryogenesis and viability in mammalians (Shafi et al., 2000; O'Donnell et al., 2004).

Figure 1.5 shows the three OGT isoforms. The nucleocytoplasmic (1 042 residues), mitochondrial (920 residues) and short (665 residues) OGT differ in the number of TPRs in the enzyme N-terminus. Also, the nucleocytoplasmic and the mitochondrial version have a signal peptide directing the protein to the nucleus and the mitochondria, respectively. The biological role of the short and mitochondrial isoforms are unknown and the mitochondrial OGT is not necessary to the modification of known mitochondrial targets (Love, 2002; Trapannone et al., 2016). In this work, OGT refers to the nucleocytoplasmic isoform.

The enzyme comprises two distinct regions, one in each terminus. The N-terminus contains several repeats of the TPR domain, which was present in 766 human proteins in the InterPro, a database of protein domains and motifs (EMBL-EBI website, 2016). The domain has 34 residues and a degenerated sequence, with key conserved positions, as revealed in the domain multiple sequence alignment (Sikorski et al., 1990). The conserved positions form a pattern of hydrophobic segments that folds in a pair of anti-parallel α-helices. The domains occur in a wide range of proteins in 3 to 16 tandem repeats and different arrangements, reviewed by D'Andrea and

Regan (2003). Despite its widespread presence in unrelated proteins, the domain has a well-established function of mediating protein-protein interactions.

OGT's N-terminus contains 12.5 TPRs that oligomerizes into an extended superhelix (Jínek et al., 2004). OGT constructs without the N-terminus domains can target peptide substrate *in vitro* (Lubas and Hanover, 2000), but cannot modify full-length proteins (Kreppel and Hart, 1999). Hence the enzyme function requires its N-terminal (Iyer and Hart, 2003). Furthermore, OGT works in a functional complex *in vivo* (Wells et al., 2004; Cheung et al., 2008; Perez-Cervera et al., 2013) and databases of binary protein-protein interactions inform the vast number of reported and inferred OGT interactors. The inner surface of the superhelix resembles the peptide-binding site of importin α protein, with potential functional implications (Jínek et al., 2004). Based on the structure similarity, Jínek et al. (2004) speculated whether the enzyme's N-terminus works only as a molecular scaffold for protein-protein interactions or have a more active role in substrate recognitions by OGT.

The C-terminus of OGT harbours the enzyme's active site and is also called the catalytic domain. The domain is classified into the family glycosyltransferase 41, exclusive to β-N-acetylglucosaminyltransferase. The region is subdivided into two lobules and an intervening domain. The two lobules yield a pocket where the UDP-GlcNAc binds before the protein recognition processes when the target protein interacts with the enzyme. Three protein segments are responsible for nucleotide binding, from residue 905 to residue 931, which is essential to the catalytic cycle (Lazarus et al., 2011). The C-terminus is also involved in the protein translocation to the plasma membrane mediated by a phosphatidylinositol binding motif (Yang et al., 2008).

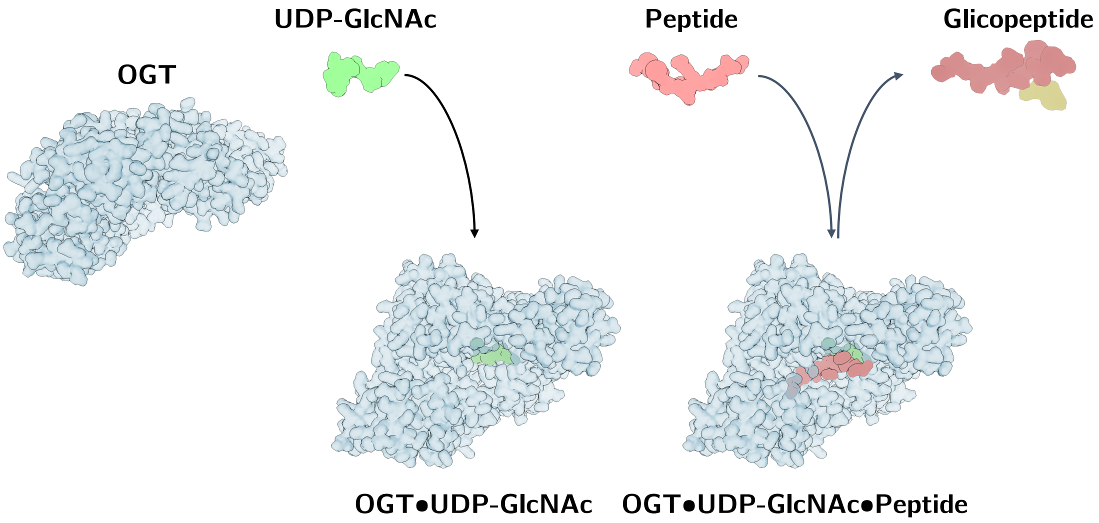


Figure 1.6: Schematic representation of OGT mechanism. Lazarus et al. (2012) obtained the three-dimensional structure of OGT and the two substrates. Instead of using the full-length OGT, the enzyme construct had 4.5 TPR. Also, the enzyme has as substrate an incompetent peptide (Ser mutated to Ala) instead of a full-length protein and UDP-5SGlcNAc, a very slow substrate to OGT. These conditions are needed to capture the reaction, which occurs very rapidly otherwise. The mechanism of the reaction catalysed by OGT is sequential. Figure based on Lazarus et al. (2012).

The details of the OGT catalytic mechanism were obtained by independent work of two groups (Schimpl et al., 2012; Lazarus et al., 2012). The proposed mechanism is named an ordered sequential bi-bi reaction and Figure 1.6 illustrates it. The cycle starts with UDP-GlcNAc, the sugar donor, binding to the OGT at its binding site in the active site. Next, the protein substrate binds to the OGT-UDP-GlcNAc complex. The transfer reaction follows the complex formation, with OGT transferring the sugar moiety from the UDP-GlcNAc to the hydroxyl group of the target S/T. Subsequently, the enzymes liberate the glycoprotein and later the UDP, finalising the reaction cycle.

OGT activity is regulated by the intracellular level of UDP-GlcNAc (Taylor et al., 2009). The enzyme has high affinity for UDP-GlcNAc, in the μmolar range. Although the affinity for several peptides has been calculated (Iyer and Hart, 2003), the better enzyme kinetics for different proteins is lacking. The enzyme activity is inversely proportional to the levels of the glucose in a pathway that increases OGT gene expression and directs the enzyme to specific targets, probably by different protein-protein interactions (Cheung and Hart, 2008).

More recently, another *O*-GlcNAc transferase was identified in the secretory pathway, the EGF domain-specific *O*-GlcNAc-Transferase (eOGT) (Enzyme Commission number 2.4.1.255). The enzyme belongs to the glycosyltransferase family 61 and target S/T of extracellular-targeted protein that contains the eukaryotic growth factor-like domains. OGT and eOGT have different substrates (Müller et al., 2013).

1.4.2 O-GlcNAcase

The MGEA5 gene encodes the OGA protein (Enzyme Commission number 3.2.1.169) with 916 residues. The human organism ubiquitously expresses the enzyme, with elevated transcript levels in the brain and pancreas. The O-GlcNAcase activity was detected in the cell nucleus and cytoplasm (Comtesse et al., 2001; Gao et al., 2001).

OGA is a neutral β-N-acetylglucosaminidase that catalyses the O-GlcNAc hydrolysis, but not GalNAc. The enzyme was first isolated from spleen (Dong and Hart, 1994). The protein belongs to the glycoside hydrolase family 84 in the CAZy Database (Cazy website, 2016), which is part the glycoside hydrolase superfamily observed from single cellular organisms to humans. The domain with NAGidase activity sits in the N-terminus (PFAM website, 2016a). The enzyme also contains other domains, including a histone acetyltransferase domain in the C-terminus, but the biological role of the domain in the C-terminus is unknown since the enzyme does not have histone acetyltransferase activity (Rao et al., 2013).

Specific inhibition of the β-N-acetylglucosaminidase activity in the human cell led to increased global levels of protein *O*-GlcNAcylation (Dorfmueller et al., 2010, 2009), suggesting the importance of the enzyme in the *O*-GlcNAc process. However, OGA site specificity remains uncertain. The bacterial homolog of OGA showed β-N-acetylglucosaminidase activity for human substrates. The crystal structure of the homolog with three glycopeptides demonstrated that the substrate recognition occurs in a sequence-independent fashion (Schimpl et al., 2012). Additional molecular dynamics experiments suggested that the interactions between glycopeptides and OGA are not structure specific (Martin et al., 2014). The structural data show that the glycopeptides' backbone atoms are interacting with the enzyme. If, however, experiments confirm that OGA is site specific, it may prove to be a valuable tool in the study of the modification.

1.5 Detection of *O*-GlcNAc

Protein *O*-GlcNAcylation can be detected at the protein, site or residue level. Most of the studies aim to map unambiguously the site position for a given protein, and determine its biological role. 'Gold-standard' or *bona fide* sites are defined as genuine sites unambiguously mapped to the protein sequence. The functional significance of most known *O*-GlcNAc sites is uncertain, and mutagenesis experiments are the best available to determine a site's functions. The following sections briefly comment on experimental methods to detect the sites. Computational methods that predict the modification sites are examined in Chapter 3.

1.5.1 Radio-labelled sugar donor

The first studies that discovered and characterised the subcellular location of *O*-GlcNAc used radiolabelled UDP-Galactose to detected the modification. The glycoproteins extracted from cells can be treated with β-1,4-galactosyltransferase and to incorporate the radiolabelled sugar, which can be detected by autoradiography (Torres and Hart, 1984; Holts and Hart, 1986). However, the method is not convenient because of the use of radioisotopes and the cross-reaction with terminal *O*-GlcNAc in glycoproteins in the secretory pathway and extracellular medium.

1.5.2 Antibodies and lectins

Several antibodies have been developed to detect protein *O*-GlcNAcylation. Antibodies that recognise terminal GlcNAc, *O*-GlcNAcylated S/T and site-specific modifications are available (Comer et al., 2001; Holt, 1987). Ma and Hart (2014) consider the use of antibodies to be more sensitive and convenient than radiolabelling. However, the antibodies can also cross-react with other glycoproteins and approaches that minimise these cross-reactions have been proposed, as reviewed by Banerjee et al. (2013), including the use of multiple antibodies instead of one.

Lectins are plant proteins that bind to specific sugar moieties. Succinylated wheat germ agglutinin has been primarily employed in the enrichment of β-*O*-GlcNAcylated proteins or peptides in chromatography columns and can achieve up to 50 000 fold enrichment of nucleocytoplasmic glycoproteins. The enrichment techniques combined with mass spectrometry helped to detect many of the known sites.

1.5.3 Mass spectrometry

Although radiolabelling and antibodies are efficient to probe the modification at a protein or peptide level, they cannot determine modification position without peptide sequencing. Technical advances in mass spectrometry have led to an increase in the number of experimentally determined *O*-GlcNAc sites from 50 in the year 2000 to more than 1 000 sites today (Hornbeck et al., 2012).

Mass spectrometry has replaced Edman degradation as the protein sequencing method of choice. Edman degradation works as an iterative and controlled cleavage of the protein N-terminus followed by amino acid identification. The identification of around 20 amino acids could determine the protein; however, the whole experiment is slow, cannot detect protein modifications and depends on a unique N-terminus. Mass spectrometry uses a different approach that enables it to identify and even quantify a mixture of proteins in hours.

Figure 1.7 illustrates a mass spectrometry experiment. Mass spectrometry does not require protein purification, and the first step is to isolate the protein sample by fractionation. Next, the proteins are separated with SDS-page electrophoresis. The proteins are excised from the gel and digested with trypsin and, subsequently, the peptides are separated by high-performance liquid chromatography. The peptides are ionised, accelerated, and the mass/charge ratio is detected. The profile of peptide fragments forms a fingerprint that can match to an entry in a peptide fragment database. Also, a shift of 210 Da from the expected fragment mass indicates the *O*-GlcNAc. The peptide fragments can undergo a second round of ionisation that breaks the peptide bonds and generates a series of mass-charge peaks. The analysis

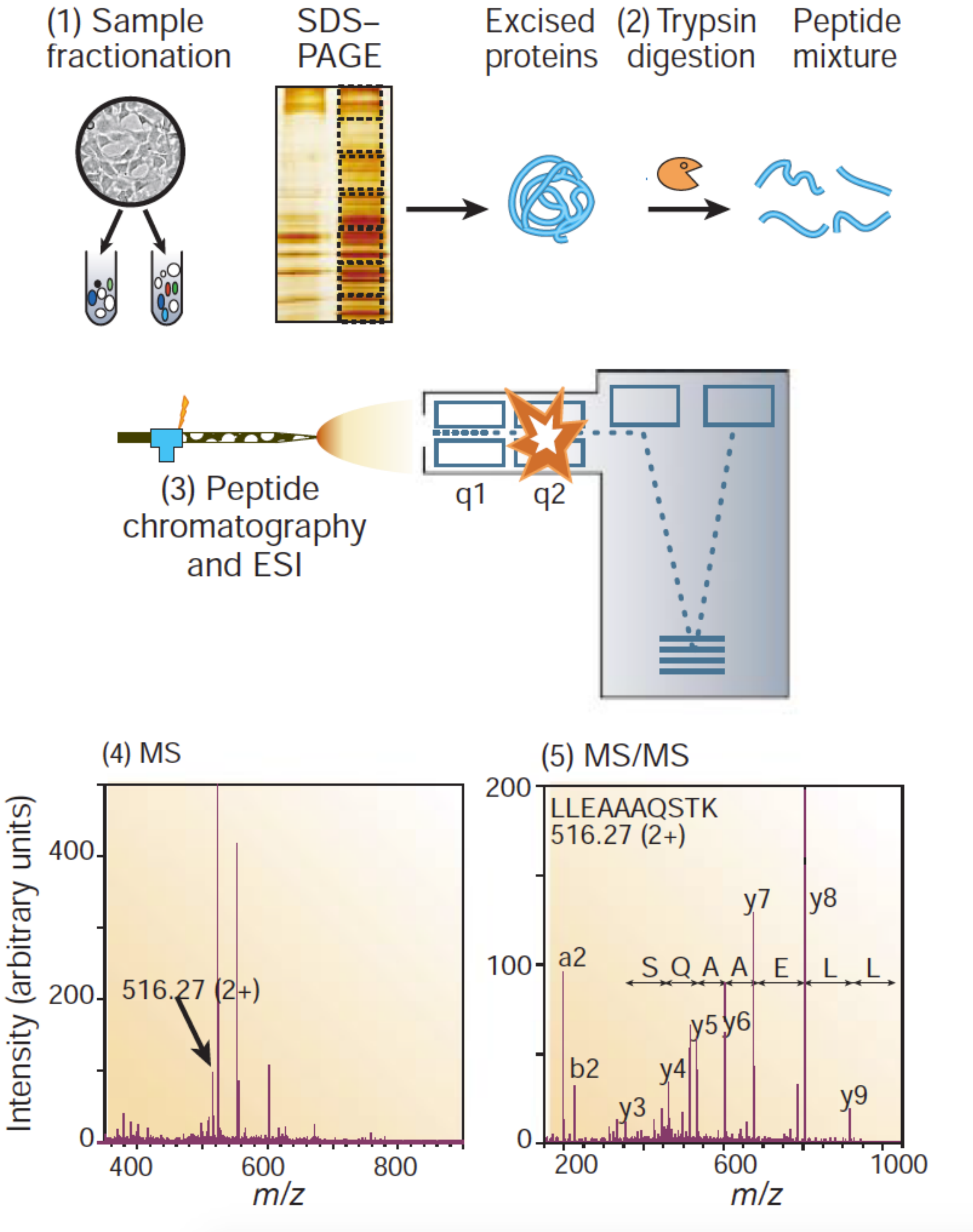


Figure 1.7: Schematic representation of mass spectrometry experiment. Reprinted by permission from Macmillan Publishers Ltd: Nature, (Aebersold and Mann, 2003), copyright 2003.

of the series of peaks may unambiguously identify which residue is modified and the peptide sequence. It is important to highlight that there are many different protocols and types of mass-spectrometers.

There are several obstacles to mapping *O*-GlcNAc sites reliably. The modification has low abundance (Roquemore et al., 1992) and is ten times less common than protein phosphorylation (Hahne et al., 2012). Accordingly, the unmodified version of the peptide can suppress the *O*-GlcNAcylated peptide mass/charge signal. The dynamic nature of the modification and the activity of unspecific hexosaminidases, lysosomal enzymes that remove sugars, enhance the problem. Also, methods that enrich *O*-GlcNAcylated peptides in samples have limited specificity (Hahne et al., 2012; Ma and Hart, 2014) and the β-glycosidic bond is labile under the peptide fragmentation step by collision-induced dissociation, which determines the modification's position within the peptide fragment (Khidekel et al., 2004).

Recent advances in mass spectrometry are numerous and outside of the scope of this thesis (Aebersold and Mann, 2003; Jensen, 2006; Miller and Blom, 2009; Ma and Hart, 2014). It is important, however, to highlight the advances that enabled the large scale identification of *O*-GlcNAc sites. Several methods swap the GlcNAc moiety for a chemical group with stable linkage, as reviewed recently by Banerjee et al. (2013). For example, Wells (2002) developed a the chemical tool that substitutes the *O*-GlcNAc with a sulphide adduct that is stable under peptide fragmentation conditions. The collision-induced dissociation is a standard mode of peptide fragmentation. As mentioned before, the glycoside bond between GlcNAc and the S/T is labile and breaks during peptide fragmentation by collision-induced dissociation (Chalkley and Burlingame, 2001). Later, Chalkley et al. (2009) showed that the glycosidic bond is not labile under electron transfer dissociation fragmentation, leading to a series of reports of higher-throughput *O*-GlcNAc sites identification from samples of different organisms and tissues (Alfaro et al., 2012; Trinidad et al., 2012; Kim et al., 2011).

One current problem is to control the specificity of the detection. The development of specific OGA (Dorfmueller et al., 2010) and OGT (Gross et al., 2005) inhibitors enhanced the ability to detect the modification in physiological conditions. A reduced OGA activity increases the intracellular level of the modification, and the inhibition of OGT reduces the probability of the digested peptides to be unspecifically modified.

Data quality is one central problem of proteomics (Wilkins, 2009). Mass spectrometry experimental protocols vary. Methods for mapping *O*-GlcNAc sites display the same challenges, considering that different studies use different experimental designs. Moreover, studies may have defined differently what is a genuine site, from what may be an experimental artefact (false positive) and the proteome heterogeneity should be considered. Mixing data from multiple proteomics studies is, therefore, challenging. Nonetheless, computational analysis of proteomics data is compelling and can reveal proprieties of the modification that were not observed in individual experiments.

1.6 Machine learning

In computer science, algorithms are defined as a set of instructions or operations to accomplish a task. Algorithms can be programmed to solve specific problems. Machine learning algorithms, on the other hand, learn from patterns in the input data, without being explicitly programmed for a problem. The methods are especially useful when the amount of data is too large and too complicated for humans to interpret. Accordingly, machine learning models learn from the training data and apply it to novel data and can also extract knowledge from datasets.

There are two approaches to machine learning: supervised and unsupervised learning. Supervised learning algorithms depend on an output variable, a categorical or continuous variable, for each example in the dataset, while unsupervised learning algorithms do not use it. Recently, two other methods have grown in popularity: reinforcement learning and semi-supervised learning, similar to positive-unlabelled learning. Reinforcement learning models apply continuous learning after the first learning iteration. For example, the AlphaGo model learns from playing against itself, after learning from recorded data of past games (Silver et al., 2016). The second mixes supervised and unsupervised approaches for classifying partly labelled data without requiring a negative dataset (Hao et al., 2015). Although this strategy may potentially change the machine learning field, their algorithms are still under scrutiny and testing.

The learning step is very specific to each machine learning method. Supervised learning models are detailed in Chapter 3 and Section 1.6.1 briefly introduces methods for unsupervised learning.

1.6.1 Unsupervised learning

Unsupervised learning can be applied to reduce the number of dimensions (features) or to form the subset of examples in datasets. The latter is also called clustering. There are two types of clustering algorithms: hierarchical and partitional clustering.

Hierarchical clustering provides a hierarchy among the examples in a dataset. The hierarchy is given by two factors: the distance metric and the linkage criteria. The distance metric is a mathematical function that measures the pairwise distance of examples. The Euclidean distance is the standard distance metric in most hierarchical clustering implementations, but the metric should vary depending on the problem. The linkage determines the distance among groups as a function of their distances.

Instead, to return the hierarchical relationship, partitional clustering methods divide a dataset with N elements in K clusters (or groups, partitions), where $K \leq N$ (Zeugmann et al., 2011). Every cluster has one or more elements, and each element belongs to a unique group. Single element groups are called singletons. Examples of partitional clustering algorithms that were applied to biology are:

- K-means clustering
- Markov Cluster Algorithm clustering (Dongen, 2000)
- Density-Based Spatial Clustering of Applications with Noise (DBSCAN) (Ester et al., 1996)

The methods mentioned above can be applied to similar problems; however, their implementations and results vary. One may consider hierarchical clustering slow since the methods calculate all pairwise relationships among examples in the training set. In the other hand, K-means is known for its speed and simplicity but requires the number of clusters beforehand. An alternative method is the pvclust method, which permutates with replacement the examples in the dataset hundreds or thousands of times, simulating new hierarchies, and detects groups that are statically significant (Suzuki and Shimodaira, 2006). Pvclust is a case that mixes partitional and hierarchical clustering. Furthermore, given a threshold hierarchy produced by clustering can generate groups.

The organisation of the machine learning algorithms in types and subtypes is necessary to understand the field. However, the real application of these methods is complementary, and many strategies use them in pipelines.

1.6.2 Supervised machine learning algorithms

The supervised approach to machine learning requires a labelled dataset. A model learns from the data, and there are two varieties of models, depending on the type of the labels. Regression methods have a continuous target variable, while classification methods have discrete target or classes. In general, trained models can predict from unseen data, although there are other applications for supervised models, such as prioritise explanatory features. The classification can be divided into binary or multi-class problems depending on the number of classes. PTM predictors aim to discover novel modification sites and, ideally, not predict an unmodified site as modified. Consequently, PTM prediction is a binary classification problem. Section 3.5 shows an example of how regression approach may be applied to PTM prediction. Unless otherwise stated, the subsequent references to the machine learning model will allude to a binary classifier.

Machine learning applications are prevalent in our everyday life and supervised learning has been applied to computational biology and the problem of PTM prediction. There is a standard workflow to the several machine learning methods (Domingos, 2014). In the first step, the data are encoded in a format compatible with the learning algorithm. Machine learning models depend heavily on the input data (Baldi and Brunak, 2001), which is also named the training dataset. Dataset encoding can also impact heavily on the performance of the classifier. Each example in the training set is represented and described by an array of attributes called features. The training set is a matrix with $m \times n$ elements, where m is the number of examples and n the number of features.

After encoding, the training set can undergo further processing, named pre-(learning)processing. This step aim operate the features, by reducing the number of features for example, to improve the prediction performance and training efficiency. The learning step is an iterative process, where the model optimises the prediction performance by minimising an error function, in case of the Artificial Neural Network (ANN) methods. Lastly, the model can undergo optimisation that involves tuning the parameters for the used algorithms.

Under- or overfitting are two main issues in training machine learning models. When a model is unable to learn (fit) from the data, as when it cannot detect any useful pattern in the training set, it predicts as randomly as a coin's flip. In the other extreme, overfitting happens when the model learns the specifics of the training data and is unable to generalise to new data. Overfitting can be decomposed into several subproblems. For example, a model can overfit by learning the noise of the dataset, which occurs when the training set has too many features or too few examples (cases of increased variance). Some machine learning algorithms are sensitive to parameters, and the lack of parameter tuning may lead to overfitting. Modern machine learning algorithms train in iterations and reserve part of the dataset for testing to minimise overfitting.

1.7 Genetic Variants

Errors in the replication machinery and other external factors lead to genome variability. A germinal genetic variant is a DNA mutation detected in the germinal tissue involved in sexual reproduction and, consequently, can be inherited by progeny. In contrast, somatic genetic variants are mutations usually detected in cancerous tissues and are not inherited.

The arrival of DNA Next-Generation Sequencing (NGS) technology has driven many large scale genomic studies since its popularisation. The higher-throughput NGS technologies enable the resolution and quantification of intrinsic or natural variations of genomes. Instances of DNA variation are formally called genetic variants and classified as either structural variants or Single Nucleotide Variant (SNV). Structural variants comprise complex changes or rearrangements in the genome sequence, such as insertions, deletions and duplications of DNA segments. SNVs are point mutations on the genome, where any other nucleotide substitutes a single nucleotide from the reference genome.

SNVs in DNA regions that encode proteins are further classified according to their effects on the encoded protein. 3 consecutive nucleotides forms a codon that encodes an amino acid. Since the genetic code is degenerate, meaning many codons can encode one amino acid, in a relationship many-to-one, a SNV may or may not yield an Amino Acid Substitution (AAS). A synonymous Single Nucleotide Variant (sSNV) does not alter the protein sequence, whilst a Non-Synonymous Single Nucleotide Variant (nsSNV) yields an AAS and changes the protein sequence. If the nsSNV leads to a new stop codon, then it is sub-classified as a nonsense variant, otherwise the variant is sub-classified as missense variant. Nonsense variants result in a truncation, a shorter protein product, which is normally considered to cause protein loss of function, although some exceptions are expected but yet unreported. In contrast, the effect of missense variants on protein function is harder to predict. An individual genome can have the order of 10 000 nsSNVs (Cargill et al., 1999), but less than 1 % may have effect on the protein function and lead to disease.

Most of the mutations will have little or no impact on the organism's ability to survive or reproduce, and are termed neutral variants (Nachman and Crowell, 2000). In contrast, mutations that have a positive or negative effect on maintaining its genetic information throughout the next generation are called beneficial or deleterious mutations, respectively. Deleterious mutations will undergo negative selection (or purifying selection), which reduces the frequency with which a variation is observed within a population until it is removed. That explains why genetic diseases are rare within the human population. On the other hand, mutations that confer an advantageous trait will be positively selected, and the observed frequency would tend to increase over generations.

A cancerous cell loses the cell growth checkpoints and starts uncontrolled cellular divisions. In this case, cancer driver mutations are the subtype of cancer mutations that increase the cell's division rate and will increase in frequency within the genomes of the affected cells. Passenger mutations may impact on the protein function but do not change cell's growth rate and accumulate over cancer growth (Stratton et al., 2009). Cancer biology is very complex and beyond the scope of this thesis. However, it may be useful to compare how germinal and somatic variants are distributed over proteins.

Two computational methods have been widely applied to predict the outcomes of genetic variants. Polymorphism Phenotyping (version 2) (PolyPhen-2) predicts the functional impact of nsSNVs based on the machine learning model trained on protein feature annotations, phylogenetic information and structural attributes, but only if the suitable protein structure has been determined. The method classifies the genetic variant as probably damaging, possibly damaging or benign. The Sorting Intolerant from Tolerant (SIFT) method calculates the residues' conservation on homologous sequences and the fact that highly conserved regions are less tolerant to AAS. This method also updates the confidence of the prediction based on the quality of the multiple sequence alignment. Unlike PolyPhen-2, SIFT does not use a source other than phylogenetic information. PolyPhen-2 and SIFT have been widely used to prioritise potentially disease-causing missense variants. However, the methods have their limitations. They yield poor predictions when the AAS results in a gain of protein function (Flanagan et al., 2010), which is particularly important in cancer. They also do not consider the different susceptibilities of genes to genetic variants (Petrovski et al., 2013) and that the interpretation of the scores should change depending on the gene in question (Itan et al., 2016). There are also other features not employed by these tools, such as protein PTM information and functional protein splicing variants.

In this thesis, PolyPhen-2 and SIFT were used because the methods are conveniently available at the Ensembl Variant Effect Predictor (VEP). However, these two tools might not be the most accurate and, in fact, have been outperformed by several other methods. An incomplete list of alternatives includes: SDM (Worth et al., 2007), Mutation Assessor (Reva et al., 2011), CONDEL (Gonzalez-Perez et al., 2012), FATHMM (Shihab et al., 2013) SuSPect (Yates et al., 2014), SAAPpred (Al-Numair and Martin, 2013), Ensembl VEP (McLaren et al., 2016).

More importantly, PolyPhen-2, SIFT and other tools that predict SNV effect on protein function do not replace manual interpretation of AAS by inspection of the protein's features, its three-dimensional structure and interactions. The background genetic variability of a protein sequence can also carry relevant information for protein function, by revealing regions of unknown function that do not tolerate variations. In conclusion, since the experimental characterisation of every AASs is unfeasible, the prioritisation of AAS in disease studies is critical, and new computational tools are needed to extend the functionality of the current ones.

1.8 Scope of the thesis

The molecular basis of OGT substrate recognition is poorly understood. The recent reports of OGT crystal structure in complex with the target peptide answered some questions on the issue; however a more comprehensive survey is needed. Thus Chapter 2 characterises the three-dimensional structure of *O*-GlcNAc sites and structural proprieties predicted from the protein sequences to extract some new information that might help to sort modified sites from unmodified ones. The current alternatives for classification of *O*-GlcNAc sites require a data-update, evaluation of new machine learning models and encoding strategies. So Chapter 3 investigates several supervised machine learning methods and data encoding strategies to predict *O*-GlcNAc sites. Chapter 4 analyses the results of the application of Predict *O*-GlcNAc Sites from Protein Sequences and Features (POGSPSF), including the proteome-wide analysis of predicted *O*-GlcNAc sites. Data analysis tools for protein structure integration with other features are currently lacking. Chapter 5 describes the development of Protein Feature Aggregator and Variants (ProteoFAV); ProteoFAV is used for the analysis of genetic variants over the OGT three-dimensional structure. Chapter 6 integrates the conclusions obtained in the thesis and suggests future work.

Chapter 2

Structural Characterisation of *O*-GlcNAc sites

Preface

This chapter briefly introduces the topic of kinase recognition and then characterises the three-dimensional structure of *O*-GlcNAc sites. Next, it extends the investigation by examining the structural features predicted from the proteins' sequences. The electron density maps of 32 structures were also examined for clues to the modification.

2.1 Introduction

Prior to protein *O*-GlcNacylation, OGT recognises the target protein. This interaction is temporary, specific and depends on other factors, such as the enzyme's cosubstrate UDP-GlcNAc. Reviewing the topic of protein-protein interaction is beyond the scope of this thesis; however, the kinase specificity problem may help understand the OGT recognition process.

The molecular basis of the molecular interaction of two molecules follows the rules described by Chothia and Janin (1975), over 40 years ago. Van der Waal's contacts, electrostatic forces, hydrogen bonds and hydrophobic effect are the 4 components that explain protein interactions. The authors demonstrated that the hydrophobic effect makes a significant contribution to protein-protein interaction. However, they also concluded that the hydrophobic effect was unspecific, and thus Van der Waal's contacts and hydrogen bonds ought to provide the specificity of protein-protein interactions.

Since protein phosphorylation is the most well-studied PTM, the interaction between protein substrates and kinases has been examined. The primary sequence of the kinase's substrates plays a major role in the kinase substrate recognition (Neuberger et al., 2007; Blom et al., 1999). Kemp et al. (1975) first reported the substrate specificity of the cyclic AMP-dependent protein kinase (PKA). The study demonstrated that a simple substitution of an Arg to a Ser in the sequence of β casein could decrease the modification rate 100 fold. At the time, the known PKA sites lacked a known sequon. The authors and others (Cohen et al., 1975) suspected that the tertiary structure of the substrate protein could participate in the kinase recognition like the substrate structure carries a three-dimensional signature. Pinna and Ruzzene (1996) commented that the tertiary structure might have overridden any primary structure propensity due to structural limitations in the active site of the kinase. More recently, Duarte et al. (2014) showed that residues far away from the site in the three-dimensional structure also mediate the substrate recognition by protein kinase C (PKC), building up evidence that the role of three-dimensional structure in PTM should be further studied.

Figure 2.1 compares the relative sequence entropy for sites modified by OGT and three protein kinases with highest number of known sites in the PhosphoSitePlus database (Hornbeck et al., 2015). The observed relative entropy for OGT sites is lower than the site relative entropy for PKA, PKC and casein kinase 2 (CK2) sites. Indicating the sequence in the sites recognised by OGT carries less primary sequence information than those recognised by PKA, PKS or CK2 and so are harder to distinguish from unmodified sites by sequence alone.

Most kinase substrate sites are intrinsically disordered (Iakoucheva et al., 2004). Protein intrinsic disorder can occur locally or globally and regions without an ordered three-dimensional structure often have specific functions (Wright and Dyson, 1999). Different types of intrinsic disorder vary from flexible to unfolded segments within the protein, so there is no single definition of the term. The flexibility can be estimated as, for example, B-factors (or temperature factors), which are atom's attributes present in the PDB file. The B-factors measure the uncertainty of the three-dimensional position of an atom based on the three-dimensional structure model. B-factors can be decomposed into the thermal vibrations component and the static disorder component. Another indication of protein disorder is missing electron densities that are annotated as missing residues (REMARK465) or missing atoms (REMARK470) in the PDB files. Although the intrinsic disorder is not the only explanation for missing regions in protein structures, the annotation is often used to classify disordered segments.

The crystal structure of OGT in a ternary complex with UDP-GlcNAc and a peptide substrate revealed that the OGT and the peptides' residues predominantly make contact via the peptide backbones (Lazarus et al., 2011; Schimpl et al., 2012). This fact reduces the importance of the peptide side chain in the enzyme active site, the cleft where the reaction occurs. A short structural motif, instead of sequence motif, could work as a point of molecular recognition, even with a degenerate sequence. Accordingly, this Chapter investigates the three-dimensional structure and features of OGT substrates to determine if they share tertiary or secondary structure similarities.

2.2 Methods

2.2.1 Data sources

A total of 1 533 modified sites from 676 proteins were selected by combining proteins curated from the literature up until 2011 (Wang et al., 2011) and from 2011-2013 (Jochmann et al., 2014). The sites were filtered to keep 7-residue long motifs with unique sequences. The resulting dataset contained 1 385 sites in 620 proteins. This dataset is referred to hereafter as the “modified sequence sites” (MSS). For comparison, 100 329 S/T from the same proteins, but not thought to be modified by OGT, were selected as a background and are referred to here as the “unmodified sequence sites” (USS).

2.2.2 Mapping *O*-GlcNAc sites to protein structures

Protein chains >30 residues long from structures determined by X-ray crystallography to ≤2.50 Å resolution were selected from the PDB on the 2nd August of 2015. Mapping the 1 385 OGT sites from 620 proteins to PDB structures by the Structure Integration with Function, Taxonomy and Sequence (SIFTS) software (Velankar et al., 2013)

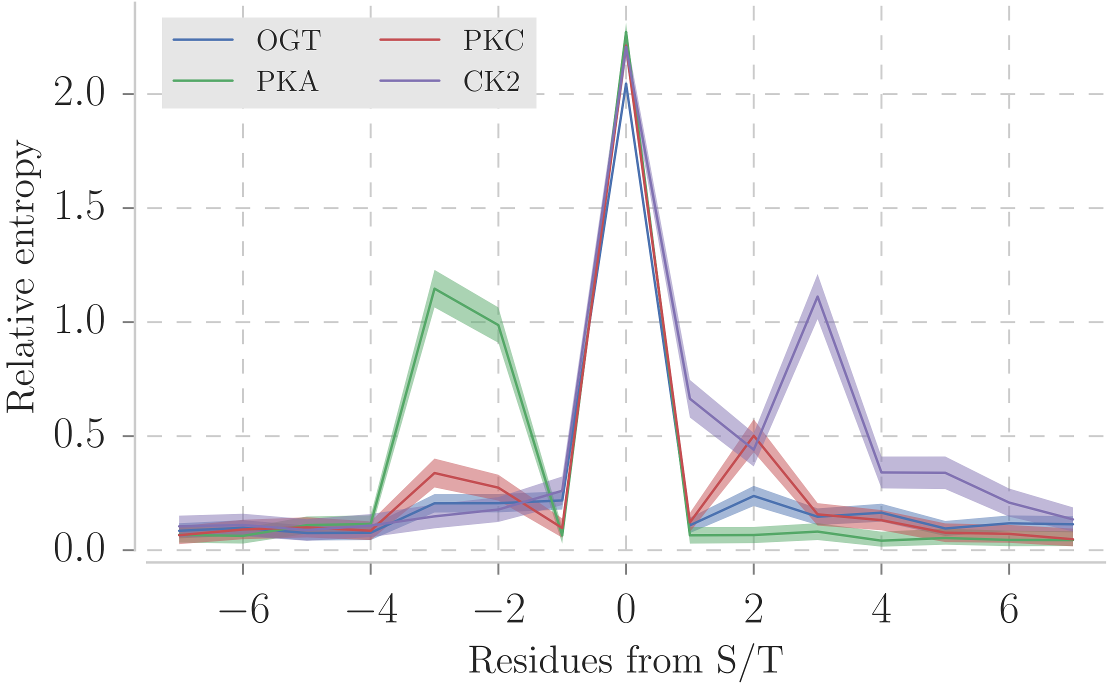


Figure 2.1: Sequence entropy of sites (±7 residues) modified by the three kinases with most sites in PhosphoSitePlus database (Hornbeck et al., 2015): protein kinase A (with 1 285 sites), protein kinase C (with 930 sites) and casein kinase 2 (CK2 with 742 sites). 1 530 OGT sites were compiled from the same database. The sequence entropy was calculated by using the Python library WebLogo (Crooks et al., 2004). Lines show mean relative entropy and the semi-transparent area represents 95% Confidence Intervals (CI).

located 45 sites in 24 proteins of known structure. The structures of a further 107 sites were identified by searching the sequences of *O*-GlcNAcylated proteins against the PDB chains with BLAST (Altschul et al., 1990) (release 2.2.18) and filtering by a conservative E-value of 1×10^{-25} to minimise erroneous matches. The cutoff of E-value $\leq 1 \times 10^{-25}$, found empirically, ensured the reliability of the match in the region of each site by inspecting all alignments between query and PDB sequence at different thresholds. Table 2.1 shows the number of matches obtained for each threshold and a less conservative threshold would minimally increase the number of sites including dubious matches. 336 proteins and 107 sites were matched by the Blast search. Selecting the protein chain with the highest coverage (SIFTS) or E-value (BLAST) left 143 sites in 107 proteins for further analysis, referred to hereafter as the "143 Structural Sites" (SS143). An alternative approach that represented a site as the mean or median attribute value from multiple protein structures were also tested with equivalent similar results to the ones shown here.

Table 2.1: Number of matches per E-value level.

E-value	Number of proteins	Number of sites
10^{-1}	403	159
10^{-5}	403	139
10^{-10}	403	126
10^{-15}	378	115
10^{-20}	353	109
10^{-25}	336	107
10^{-30}	312	100

2.2.3 Site definition and clustering

The three-dimensional structure of OGT with its substrates suggested that the region of contact between OGT and a modifiable S/T includes the residues and ±3 amino acids either side (Schimpl et al., 2012). From the structural sites returned in Section 2.2.2, 132 “Structural Sites” (hereafter SS132) had at least one match with all backbone atoms for the 7-residue long site and were retained for further analysis. Cα atoms of each residue and the Cα and the Cβ for the central S/T were superimposed for all pairs of sites. The resulting matrix of Root-Mean-Square Deviation (RMSD) values were clustered altogether with Euclidean distance and complete linkage and groups were produced by setting a 3 Å cutoff.

2.2.4 Structural properties of sites

Protein secondary structure assignments were obtained from Define Secondary Structure of Proteins (DSSP) (Kabsch and Sander, 1983). DSSP annotates 7 different secondary structure states: 3_{10} helix (G), α helix (H), π helix (I), bends (S), turns (T), isolated (B) and extended (E) β-bridge. These assignments were reduced to three states:

1. 3_{10} helix, α helix and π helix to H
2. isolated and extended β-bridge to E
3. all other, including residues with no assignment, to C

The solvent accessible area from DSSP was normalised by the residue’s maximum accessible area as described in Cuff and Barton (2000). A S/T was considered exposed if its Relative Solvent Accessibility (RSA) was $>$25%; partially buried if $>$5% and $\leq$ 25%; and buried if $\leq$ 5%.

Cα temperature factors or B-factors were standardised (Z-score normalised) over the B-factors for all Cα in the same chain. This operation is indicated because the

B-factors obtained in different X-ray crystallography experiments are not directly comparable.

2.2.5 Prediction of protein disorder and secondary structure

JPred4 (Drozdetskiy et al., 2015) produced the protein secondary structure predictions for the proteins in the MSS dataset. Since Jpred4 is limited to sequences shorter than 800 residues, 300 sequences were trimmed while ensuring the modified S/T was at least 100 residues away from the N- and C-terminus to avoid edge effects.

The intrinsic disorder was predicted by JRonn (Java implementation of Ronn (Yang et al., 2005)), IUpred (Dosztányi et al., 2005) and DisEMBL (Linding et al., 2003) through the JABAWS (Troshin et al., 2011) (release 2.1) command line application. These methods provide 6 disorder prediction scores, which are followed by the score cutoff (in parenthesis) as determined by the methods' authors: DisEMBL-REM465 (0.6), DisEMBL-COILS (0.516), DisEMBL-HOTLOOPS (0.1204), IUpred-Long (0.5), IUpred-Short (0.5) and JRonn (0.5). Disorder predictions were also performed on a background set of 1 164 S/T selected at random from globular proteins in the Astral dataset (Fox et al., 2014)] (release 2.04), referred to hereafter as the 'Globular Set' (GS). Figure 2.2 shows the relationship among the different datasets with the numbers of sites and proteins, which was also summarised in Table 2.2.

Table 2.2: Datasets summary. See Section 2.2 for details. Sites, number of sites; Proteins, number of proteins.

Dataset name	**Sites**	**Proteins**	**Short name**
Modified Sequence Sites	1 385	620	MSS
Unmodified Sequence Sites	100 329	620	USS
Structural Sites	143	106	SS143
Structural Sites with backbone	132	93	SS132
Globular Set	1 164	1 164	GS

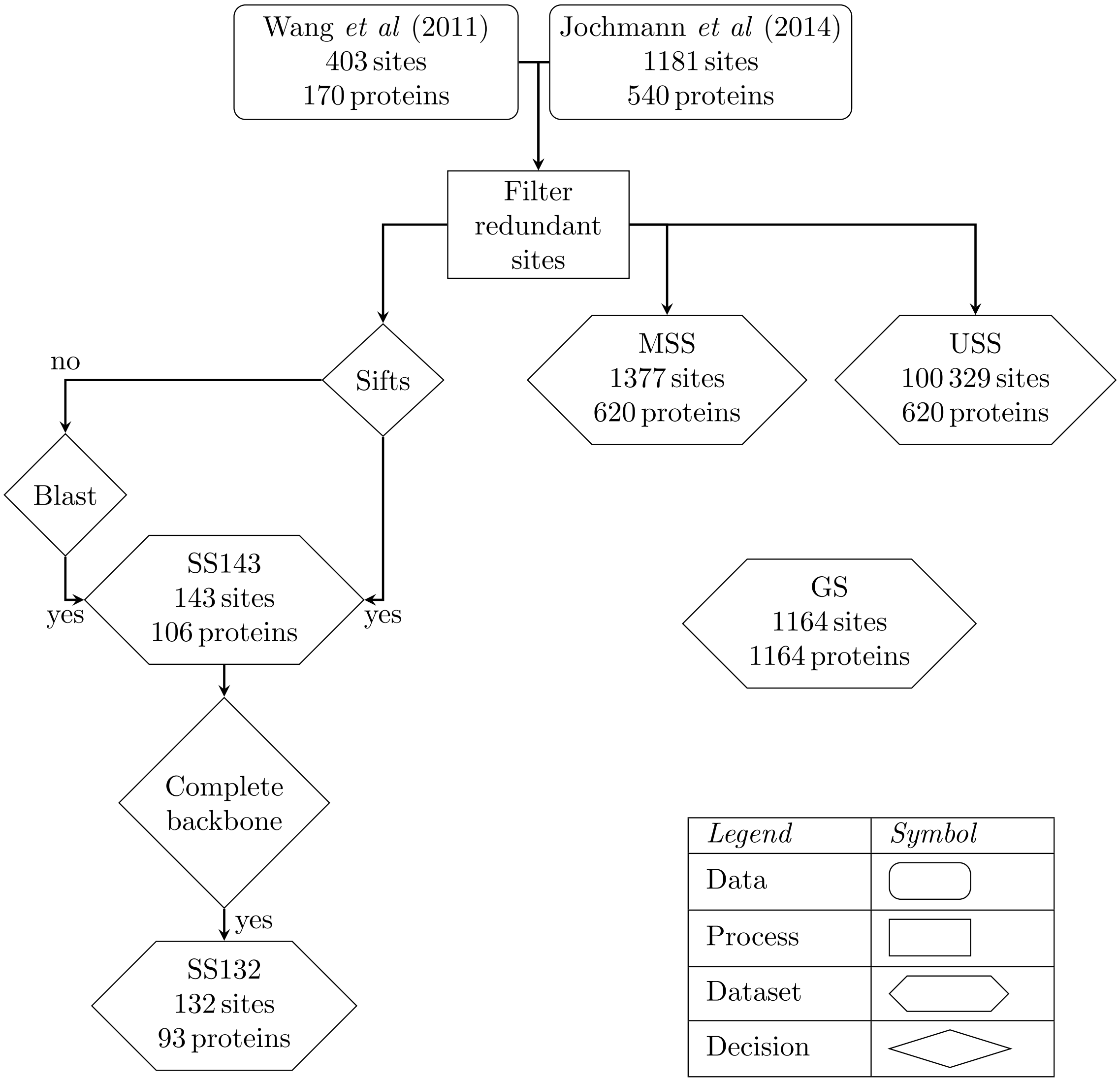


Figure 2.2: Datasets' relationships.

2.2.6 *O*-GlcNAc sites clues in electron density maps

In general, protein crystallisation demands large quantities of protein that is typically obtained from super expression in *Saccharomyces cerevisiae* and *Escherichia coli*, two organisms without *O*-GlcNAc cycling enzymes. However, when the protein is naturally abundant, there is no need for the superexpression in a heterologous system. Also, some protein structures are obtained from systems that might express OGT and OGA. The electron density models of 32 protein structures was then analysed for clues to the modification. The protein structures were obtained from organisms that contain *O*-GlcNAc cycling and Appendix A.1 lists these organisms and the site positions. Electron density was obtained from the Electron Density Server (`https://eds.bmc.uu.se/eds/`) and visually examined with Coot (Emsley and Cowtan, 2004).

For the examination, the Coot graphical interface was centred on the potentially modified residue with the command 'Draw/Go to atom'. The σ parameter was varied from 1.5 to 3.0 with the command 'HID/ScrollWheel/Attach scroll wheel'. The electron density map of OGT modified peptide TAB1 (PDB accession number 4ay5, chain I) centred at Ser11 was used as comparison.

2.2.7 Statistical analysis and code

The data collection, processing, analysis and the Cα clustering steps, were written in the Python programming language (Python Software Foundation, version 2.7 `http://www.python.org`) with the Pandas (version 0.17) (McKinney and Team, 2015), and Biopython (version 1.65) (Cock et al., 2009) libraries. Statistical tests were performed with the Statsmodels (version 0.6) and Scipy (version 0.16) libraries. A p-value threshold was set to 0.05.

2.3 Results

2.3.1 Analysis of *O*-GlcNAc sites in proteins structures

Previous reports have suggested that *O*-GlcNAc sites, like phosphorylation sites, are predominantly present in disordered regions of proteins (Trinidad et al., 2012). Increased B-factors is an indicative of structural flexibility. The standardised B-factor distribution is equivalent between modified and unmodified S/T (Kruskal-Wallis two-sample test p=0.12). Figure 2.3 shows the standardised B-factors distributions for each secondary structure element. However, the distribution of standardised B-factors is different (Kruskal-Wallis two-sample test p=0.02), while the distributions for residues in E and H are similar (Kruskal-Wallis two-sample test p=0.72 and 0.37, respectively). The shift on the B-factors distributions of modified residues in C indicates an increased uncertainty of the residues' spatial position in the analysed X-ray structures, and in consequence suggests that modified residues do not participate as much as unmodified ones in crystallographic contacts.

Of the 143 modified S/T mapped to protein structures in the present study, 26 are in parts of the protein structure annotated in the REMARK465 file of PDB file. In comparison, 553 of 4 811 unmodified S/T from the same protein structures are also found in missing regions. Accordingly, *O*-GlcNAcylated S/T in these proteins are 1.7 times more likely to be in REM465 regions (Fisher's exact test p=0.02). Since one cause of missing atoms in the three-dimensional structure of a protein is the high flexibility of the region, this finding is consistent with *O*-GlcNAcylated S/T occurring more frequently in disordered or highly flexible regions.

Table 2.3 summarises proportions of DSSP assigned secondary structure for the SS143 dataset comparing modified and unmodified S/T in the same proteins. The proportions of H, E and C are not different between the two groups, suggesting that there is no preference in the secondary structure for modified S/T in this dataset.

Table 2.3: The proportion of secondary structure types for modified and unmodified S/T in the SS143 dataset. Within parenthesis the number of S/T. Within square brackets the lower and upper 95% CI. SS, secondary structure type.

	Modified		**Unmodified**		**p-value**
SS	Proportion	95% CI	Proportion	95% CI	
C	0.55 (78)	[0.46, 0.63]	0.51 (2 475)	[0.50, 0.53]	0.36
H	0.25 (36)	[0.18, 0.32]	0.32 (1 525)	[0.31, 0.33]	0.06
E	0.2 (29)	[0.13, 0.27]	0.17 (811)	[0.16, 0.18]	0.27
Total	143		4 811		

Residues that are buried in the protein structure are not thought to be modified by phosphorylation, due to the structural constraints. Figure 2.4 shows the RSA for modified and unmodified S/T in the three levels of solvent accessibility. Although the overall distributions of unmodified and modified are in the limit of statistical significance (Kruskal-Wallis two-sample test p=0.06), there is no clear distinction for the three levels of RSA. The Kruskal-Wallis two-sample test for buried was 0.53, partially buried 0.25 and exposed p=0.80. So the distributions of RSA for modified and unmodified S/T are similar. Table 2.4 shows the contingency table for each level of RSA.

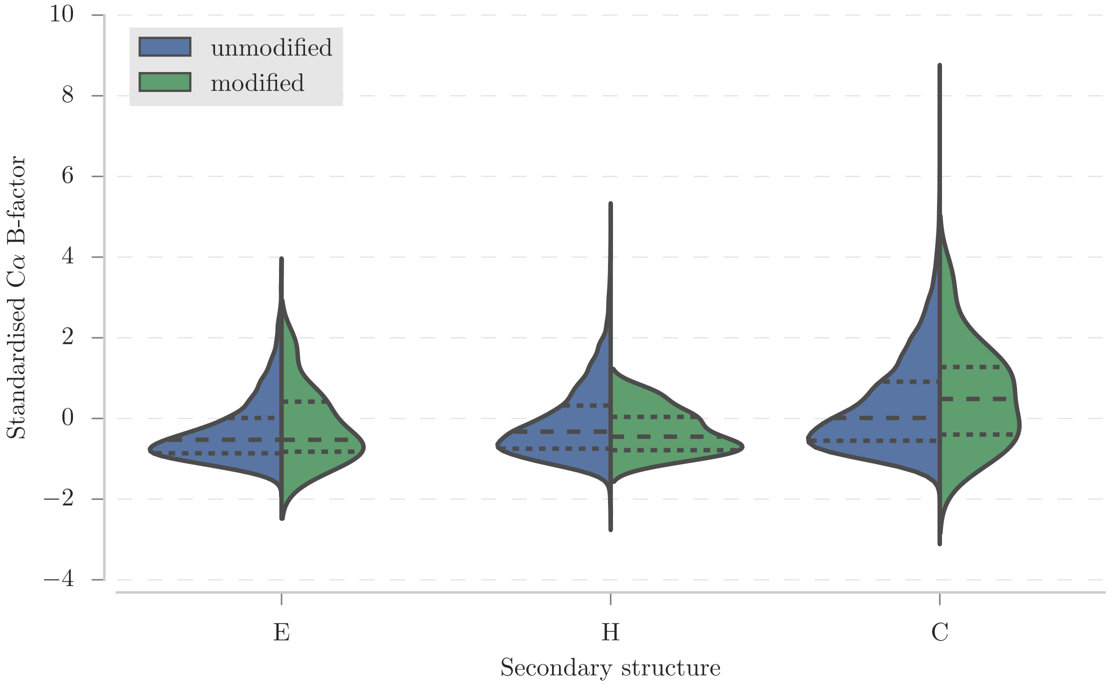


Figure 2.3: Modified and unmodified S/Ts ST have similar distributions of B-factors. The values were grouped by secondary structure, but no difference was observed without grouping. X-axis, DSSP secondary structure; y-axis, standardised Cα B-factor. All parameters refer to S/T, not to the site. The violin plots should be interpreted as the distributions for the B-factors values. Dashed lines represent 25%, 50% and 75% quantiles respectively.

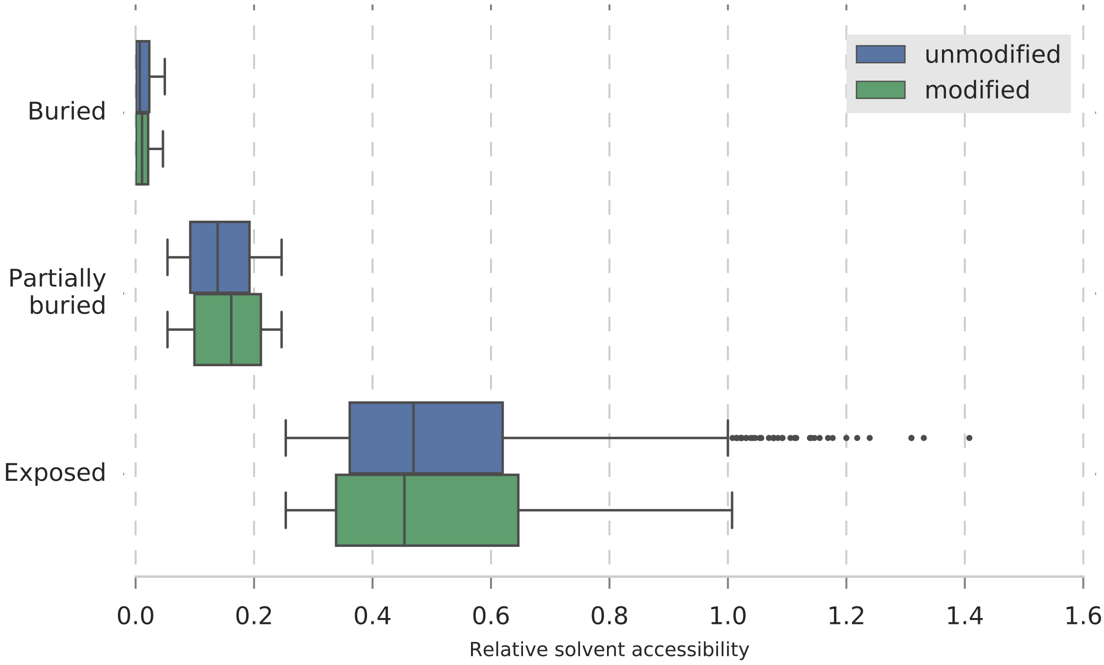


Figure 2.4: Relative solvent accessibility of modified and unmodified S/T of *O*-GlcNAcylated protein structures. DSSP calculated solvent accessibility was normalised by the residue theoretical maximum accessibility and the derived scores were reduced to three levels: buried (<0.05), partially buried (<0.25 and >.05) and exposed (>0.25) levels. The mean relative solvent accessibility is equivalent between modified and unmodified residues. The y-axis shows the solvent accessibility levels, and the x-axis the values of relative solvent accessibility.

Table 2.4: Frequencies for residues in each RSA level and the modified and unmodified groups. χ^2 test p=0.42.

	Buried	Partially Buried	Exposed
modified	26	27	65
unmodified	1066	1105	2084

2.3.2 Comparison of local structure around structural sites

Since no differences in secondary structure propensity were observed between modified and unmodified S/T, the local three-dimensional structure of the 7 residue peptides centred on S/T was investigated by pairwise superposition and clustering (see Methods). 36 sites produce singlet, while the remaining 96 sites fall into 10 clusters with the 3 Å RMSD cutoff. Figure 2.5 illustrates the superimposed structures for sites in clusters, where green, yellow and grey represent residues in H, E and C secondary structures elements, respectively. Table 2.5 shows the mean properties for each structural group. Sites are found in a wide range of secondary structure types. The sites in Clusters E, G and J, have consistent consensus secondary structures. Clusters A-D, F, H and I are all variants on coil-helix or coil-strand transitions. All buried sites, listed in Table 2.6, group in clusters D and G. Table A.2 in the Appendix section lists all sites and sites' properties in the SS132 dataset.

A close examination of the three-dimensional structure containing the buried sites confirms that these sites are indeed buried. The 3 sites in cluster D were improbable to be targeted by OGT due to their close proximity to the protein core. But the 4 sites in cluster G may be modified, since 2 of them (PDB accession number 3abm and 4y7y) are located at the interface between two protein chains, indicating the monomer could be modified. The other two (PDB accession number 2zxe and 4l3j) are in regions that might be accessible upon conformational change.

To test if the clusters found for the SS132 are features of *O*-GlcNAc modification or just reflect the natural composition of site centred in S/T, 132 unmodified S/T were randomly selected from the same proteins and clustered. Random selection and clustering was repeated 1 000 times and the resulting clusters compared with those clusters in the SS132 dataset. The number of clusters identified in each sample ranged from 10-14 (95% CI), which is consistent with the SS132 dataset. Furthermore, the structural clusters identified from the random sampling included structural clusters similar to those for the modified sites, suggesting that there are no dominant secondary structural or conformational patterns indicative of *O*-GlcNAc modified sites in the structural data currently available for these sites.

Table 2.5: Summary of structural groups properties. Members, number of members per group; B-factors, site Cα mean B-factors; SA, residue average relative solvent accessibility; SS, site DSSP secondary structure.

	Members	**B-factors**	**SA**	**SS**
Cluster				
A	8	0.06	0.35	[EH]CCHHHH
B	7	−0.35	0.28	HHHCCCC
C	2	1.26	0.43	-------
D	16	−0.37	0.23	[CE][CE]EEEEC
E	5	0.54	0.35	CCCCCCC
F	8	0.56	0.41	CCCCC[CH][CH]
G	25	−0.20	0.23	HHHHHHH
H	8	0.01	0.22	[CE]EE[CE]C[CE]E
I	6	0.22	0.32	CCCCHHH
J	11	0.22	0.36	CCCCCCC

Table 2.6: List of sites with buried sites in the SS132 dataset. PDB, PDB accession number; Chain, chain within the protein structure; Position, position within the protein structure chain; SA, site average relative solvent accessibility; SS, site DSSP secondary structure.

PDB	**Chain**	**Position**	**Cluster**	**SA**
1f4j	B	114	D	0.05
3cb2	B	170	D	0.02
4qvp	T	131	D	0.01
2zxe	A	366	G	0.02
3abm	R	63	G	0.01
4l3j	A	180	G	0.01
4y7y	Z	190	G	0.04

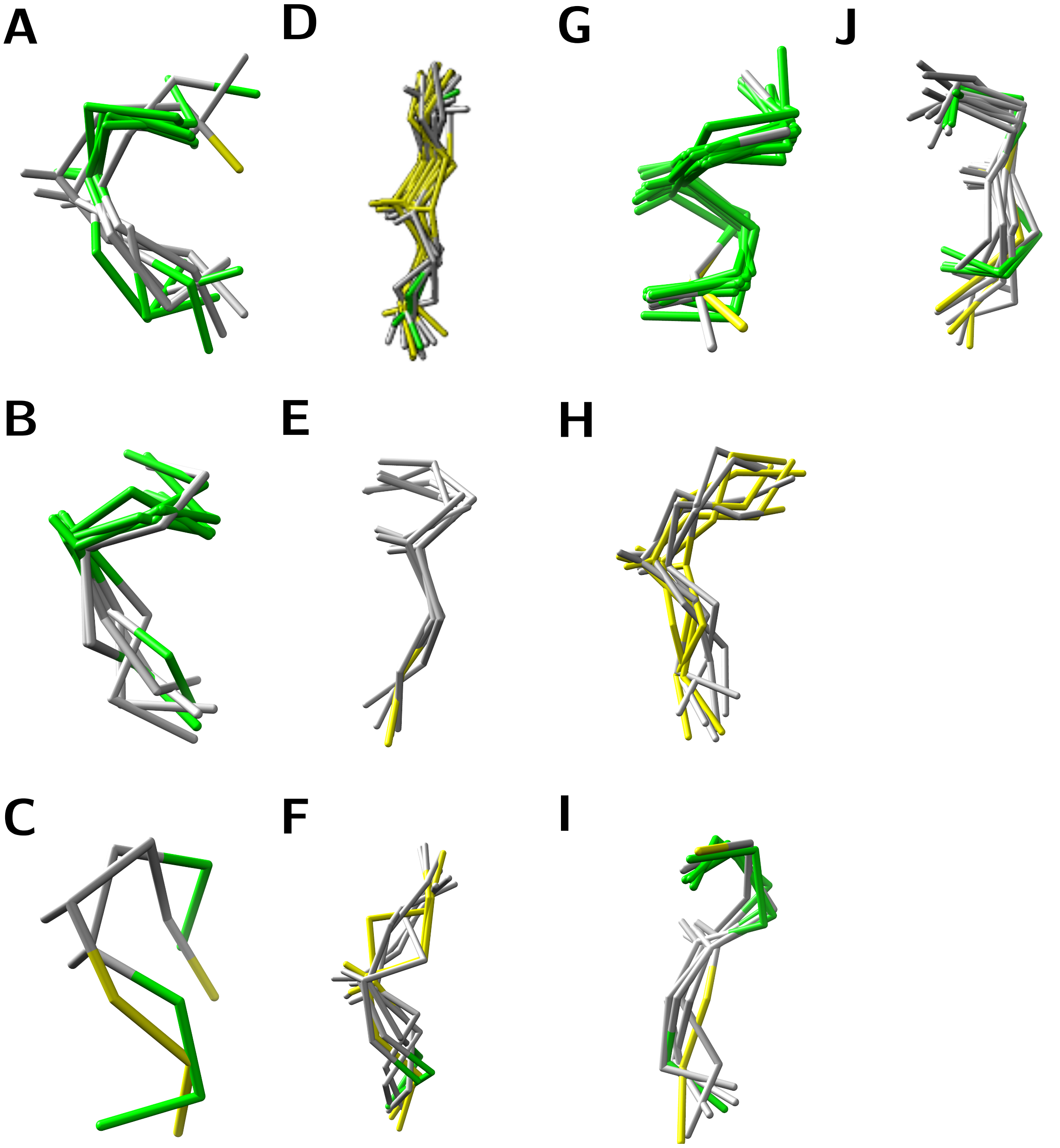


Figure 2.5: Structural groups of *O*-GlcNAc sites. Green residues participate in H, yellow in E and grey in C, following DSSP assignments. The sites' Cα atoms and the S/T Cβ were superimposed. Their RMSD were clustered with complete linkage and Euclidean distance, and groups were defined by the 3 Å threshold. The threshold, defined to minimise the diversity of secondary structure state per group, yielded 10 groups, and even within groups the structural conservation was minimal. Therefore, OGT target site has no single structural motif. Cβ for the target S/T is shown. Ribbon colour represents secondary structure elements: grey, C; green, H; and yellow, E.

2.3.3 Search for modification in electron density maps

Protein crystallography can be applied to the study of PTMs, focusing the investigation of a site's function and the modification's impact on the protein structure. Protein molecules in crystals are packed in a repetitive, symmetrical structure, whereas each protein and its solvation layer remains trapped. The electrons in the crystal scatter the energetic X-ray photons, resulting in a diffraction pattern. After processing, the diffraction pattern generates probabilistic maps of atom positions in the crystal lattice. Furthermore, if fractions of the proteins consistently carry a modified S/T, the electron density map may hold clues to its position and structural context.

Electron density maps are subject to interpretation. The crystallographer builds the protein structure based on the electron density maps, the protein sequence and other pieces of evidence. The electron density maps may contain an immense amount of information, and therefore the crystallographer often needs to ignore electron densities which are likely to be annotated as one of the molecules used in protein crystallisation. Thus, some protein structures could carry *O*-GlcNAc. However, the various chains in the 32 analysed protein structure did not exhibit any misannotated density or any clues regarding the modification.

2.3.4 Analysis of features predicted for the MSS dataset

The number of *O*-GlcNAc sites on proteins with known three-dimensional structure is limited to around 10% of the dataset. Therefore, to extend the analysis, prediction algorithms were applied to the sequences in the MSS and USS datasets, as detailed

Table 2.7: Jpred4 predicted solvent accessibility for S/T in the MSS and USS datasets. The proportions of buried S/T as predicted by the Jnetsol method in Jpred4. The proportions of buried S/T the same between the modified and unmodified groups in the three levels predicted by the method. Within parenthesis, the number of S/T; within square brackets, the lower and upper 95% CI. The p-value refers to the two-tailed z-score test between the modified and unmodified groups.

	Modified		**Unmodified**		**p-value**
Buried at	Proportion	95% CI	Proportion	95% CI	
0%	0.01 (7)	[0.00, 0.01]	0.01 (836)	[0.008, 0.009]	0.18
5%	0.04 (55)	[0.03, 0.05]	0.04 (3 917)	[0.038, 0.040]	0.86
25%	0.29 (403)	[0.27, 0.31]	0.35 (28 044)	[0.27, 0.28]	0.31

in Section 2.2.

One unanticipated finding was that modified and unmodified S/T were equally distributed for the three levels of solvent accessibility predicted by Jpred4. Table 2.7 shows the proportion of buried S/T in the MSS dataset, which is equivalent to the proportions of buried residues USS at the 0%, 5% and 25% levels. Modified residues are thought to be exposed to the solvent, so the lack of predicted accessibility was not expected.

According to Jpred4 secondary structure predictions, the composition secondary structure is different between modified and unmodified residues. Table 2.8 summarises the difference. An increase of the proportion of modified S/T occur in C ($p<0.01$) with a decrease in the H, but not in E. This indicates the Jpred4 predicts modified residues to be within C rather than structured secondary structure regions, which agrees with the argument that modified sites are more likely to occur in flexible regions.

The application of multiple disorder predictions, instead of a single one, is ideal

Table 2.8: Jpred4 predicted secondary structure proportions for S/T in the MSS and USS datasets. Modified S/T are significantly more likely to occur in C, compared to H and E. Within parenthesis the number of S/T. Within square brackets the lower and upper 95% CI. The p-value refers to the two-tailed z-score test between the modified and unmodified groups.

	Modified		**Unmodified**		**p-value**
SS	Proportion	95% CI	Proportion	95% CI	
C	0.88 (1212)	[0.86, 0.90]	0.829 (457848)	[0.826, 0.831]	<0.01
H	0.08 (107)	[0.07, 0.09]	0.126 (136489)	[0.124, 0.128]	<0.01
E	0.05 (66)	[0.04, 0.06]	0.045 (40243)	[0.043, 0.046]	0.6

Table 2.9: Predicted disorder between MSS and USS datasets. The mean scores ± standard error for each predictor is shown. All scores, excepting DisEMBL-HOTLOOPS, reveal a small but significant increase of mean disorder score for modified S/T over unmodified ones. The p-value refers to the two-tailed t-test between the modified and unmodified groups.

Method	**Modified**	**Unmodified**	**p-value**	**Effect size**
DisEMBL-REM465	0.48 ± 0.004	0.47 ± 0.001	0.01	0.07
DisEMBL-COILS	0.60 ± 0.004	0.58 ± 0.001	<0.01	0.09
DisEMBL-HOTLOOPS	0.10 ± 0.001	0.10 ± 0.001	0.45	0.02
IUpred-Long	0.59 ± 0.006	0.55 ± 0.001	<0.01	0.16
IUpred-Short	0.48 ± 0.005	0.45 ± 0.001	<0.01	0.11
JRonn	0.61 ± 0.004	0.62 ± 0.001	0.02	0.07

since the different tools base their algorithms on various definitions of protein intrinsic disorder. Three disorder prediction tools, resulting in 6 different scores, were applied to the sequence of *O*-GlcNAcyalted proteins. Table 2.9 shows a small, but significant, increase of mean predicted disorder for all scores predictions, except for DisEMBL-HOTLOOPS, a predictor method which is trained using structural B-factors. This result is consistent with the result obtained with the SS143 dataset.

To confirm that *O*-GlcNAc sites tend to be in disordered regions, the MSS dataset was compared with the GS, which contains proteins known to be predominantly globular, and hence mostly ordered structure. In Figure 2.6, the y-axis illustrates the $\log_{10}$ odds ratio of disordered residues in the MSS and GS datasets, for relative positions, in residues, to the central S/T (x-axis). The odds ratio measures the proportion of residues predicted as disordered, given each score threshold, for MSS/GS groups on a log scale. Consequently, a number close to 0 indicates no difference. DisEMBL-HOTLOOPS reports a relatively small increase ($> 0.5 \log_{10}$ odds of the proportion of disordered residues in each dataset) of the ratio of disordered residues around the modification sites; while DisEMBL-COILS and JRonn also indicate a small increase, but not in a particular region, but rather over the segment analysed. IUpred-Long and IUpred-Short and DisEMBL-REM465 show a bigger increase of the ratio of disordered residues in the MSS dataset and IUpred-Short and REM465 have their peak within -15 to +15 residues from the modification positions, detecting a clear increase of the ratio disordered residues in MSS / disordered residues in GS close to the modification.

2.4 Discussion

Protein phosphorylation and *O*-GlcNAcylation have many similarities. Both are reversible PTM that participate in cell signalling networks. Phosphosites are present in both ordered and disordered regions and the role of structural disorder of these sites has been studied and applied to computational predictors (Durek et al., 2009). However, despite its extensive involvement with human disease, little attention has been paid to protein *O*-GlcNAcylation as a molecular process. This work aimed to understand the three-dimensional structure of the *O*-GlcNAc sites.

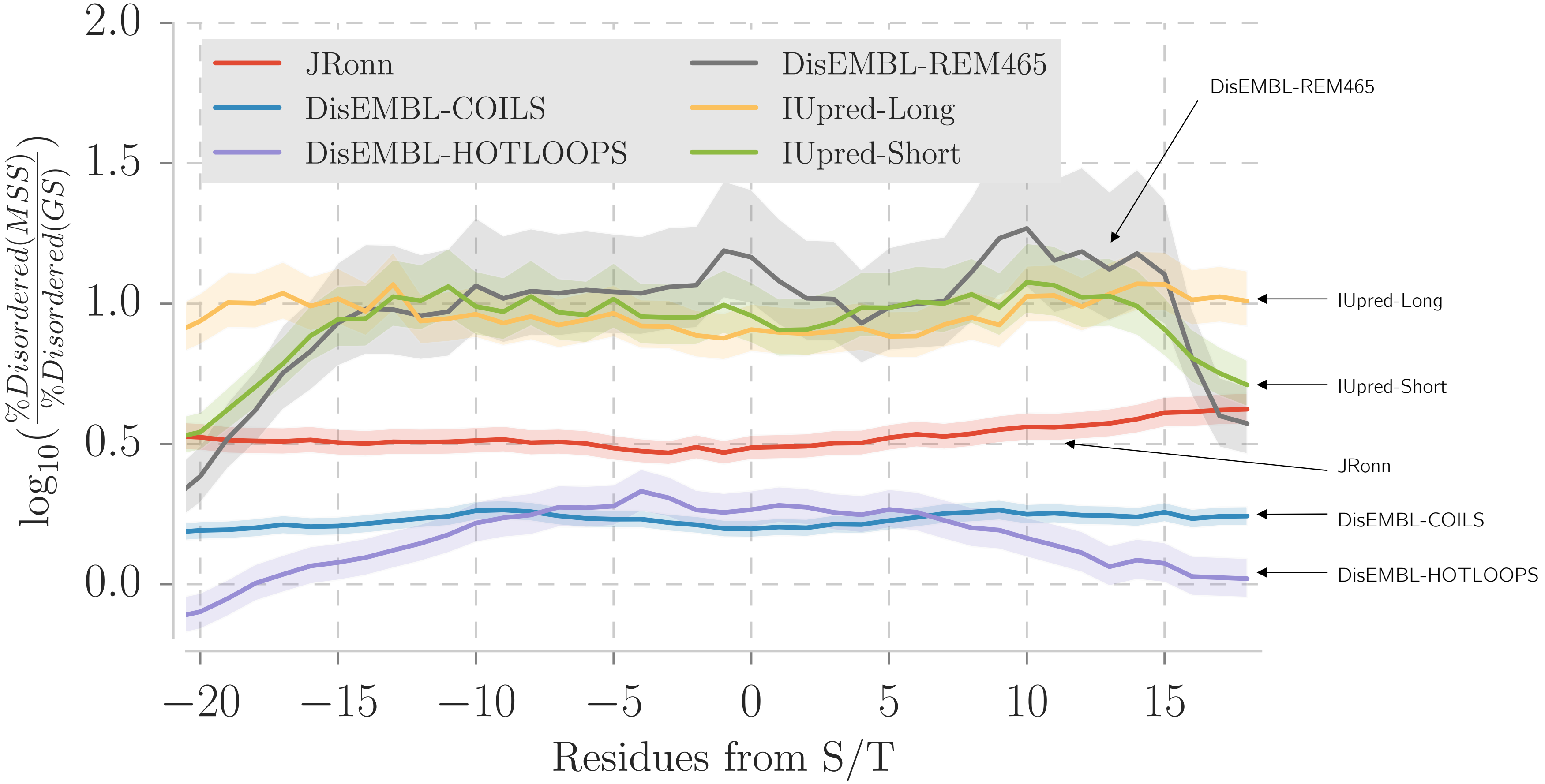


Figure 2.6: Protein predicted disorder around *O*-GlcNAc sites. Fold change (95% confidence interval) of predicted disorder in the *O*-GlcNAc sites compared to random S/T in globular proteins. The fold change shows the ratio of the probability of finding a disordered residue, given each predictor threshold. The x-axis represents the distance of the central residue, always a S/T. DisEMBL-REM465, IUpred-short predict protein structural disorder specifically around the modification site, while the other methods predict intrinsic disorder over *O*-GlcNAcylated proteins.

2.4.1 Modified and unmodified residues are equally exposed to the solvent

Solvent accessibility can be used to complement the prediction of phosphorylation sites (Zhou et al., 2004). But the characterisation of the structure of *O*-GlcNAc sites demonstrated that these sites are not more exposed than other S/T within *O*-GlcNAcylated proteins. *O*-GlcNAc modification of buried S/T is unlikely since the process requires ternary complex that includes the targeted protein before the reaction. However, Jiménez et al. (2007) found that around 15% of the analysed phosphosites are buried. Furthermore, a recent structural review of the *O*-GlcNAcylation of histones concluded that several sites are not on the nucleosome surface (Gambetta and Müller, 2015). So far, this issues has been unreported, and three potential explanation are possible. For example, Zhu et al. (2015) report Sp1 and Nup62 co-translational *O*-GlcNAcylation, which may explain the modification of residues not exposed to the solvent if the modification occurs before or during protein folding. Also, protein structures are not static entities as they are represented in protein crystal structures, and buried residues may become exposed in a different alternative conformation or by other forces, like molecular recognition by OGT. More likely, incorrect mapping from the mass spectrometry experiment cannot be disregarded. Anyhow, S/T are polar residues and are buried in only 5% of the sites in structures.

2.4.2 Sites targeted by OGT might be associated with secondary structure transitions

While the structural sites in the SS143 dataset have equal proportions of the secondary structure states, the result from secondary structure predictions on the MSS set showed that *O*-GlcNAc sites are likely to reside in coils. So the proportions of secondary structure assigned by DSSP and predicted by Jpred4 differ. While secondary structure prediction has limited accuracy, the number of samples in the SS143 dataset is limited and potentially biased. Also, clustering sites in the SS132 dataset highlight groups that are more likely to occur near to the transition between a secondary structure element and C, as observed in several members of clusters A-D, F and H.

2.4.3 *O*-GlcNAc sites also occur in disordered regions

The increased proportion of modified residues in disordered regions suggests that structural flexibility around the modification sites plays a role in the molecular recognition process. The idea that regions with increased structural mobility could mediate binding is not new (Janin and Chothia, 1990). But sites in structured regions of protein clearly indicate that not all mapped *O*-GlcNAc sites are in disordered regions. Furthermore, InterproScan (Zdobnov and Apweiler, 2001) (version 5.4 in September of 2014) analysis of *O*-GlcNAc sites assigned 19% of the sites to protein domains, the number that goes in agreement to what is known for phosphoserines and phosphothreonines in PFAM domains, which is around 25% (Hornbeck et al., 2012; Beltrao et al., 2012). Hence, like protein phosphorylation, protein *O*-GlcNAcylation targets sites in ordered and disordered regions.

2.4.4 Potential OGT specificity drivers

OGT targets specific peptides and the active site may change the three-dimensional structure of the substrate site (Pathak et al., 2015), what may explain why *O*-GlcNAc sites' tertiary structure is indistinguishable from unmodified sites. But other factors also may explain OGT specificity. OGT participates in macromolecular assemblies (Wells et al., 2004), and the role of adaptor proteins should not be ignored. Wells et al. (2004) report OGT as part of a functional complex that includes a S/T protein phosphatase recruited by the enzyme N-terminus. However, it is as yet unknown whether the complex interferes with the OGT interaction with its substrate. Besides, long-range residues, or residues that are far away in the protein sequence, but close in its structure are critical in protein phosphorylation, more specifically in PKC substrate recognition (Duarte et al., 2014). Other components, such as UDP-GlcNAc concentration and subcellular location-dependent interactions, modulate OGT activity and also may have a role in its substrate molecular recognition (Nagel and Ball, 2014). Altogether, site sequence and structure should influence, but are not able completely to distinguish, modified and unmodified sites. More strikingly features predicted from the sequence are better tools to sort *O*-GlcNAc sites from unmodified S/T than linear structural motifs.

2.5 Conclusions

- This work is the first comprehensive structural study of *O*-GlcNAc sites

- Modified sites mapped to proteins crystal structures do not seem to have a three-dimensional signature
- Surprisingly, some of the sites mapped to structures were found in buried regions
- Modified and unmodified S/T were compared, an excess of modified S/T was observed in regions annotated as REM465, suggesting that in fact *O*-GlcNAc sites are associated with protein intrinsic disorder
- The proportions of predicted disorder and secondary structure were different for modified and unmodified S/T

Chapter 3

Computational prediction of *O*-GlcNAc sites

Preface

Firstly, this chapter introduces machine learning algorithms and methods used to predict PTM sites, focusing on methods that predict *O*-GlcNAc sites. Next, it describes the stages for training a model and investigates the most appropriate training set and machine learning methods for the predictor. It also discusses optimal motif encoding strategies. Finally, a machine learning classifier of *O*-GlcNAc sites was trained on the motif sequences and predicted features and evaluated.

3.1 Introduction

The experimental validation of PTM sites demands time and resources that are often unavailable. Apart from studying the features that characterise the sites, as done in Chapter 2, computational methods can learn from the data by identifying patterns in a set of sites. Machine learning methods are widely used to prioritise potential PTM sites prior to experimental validation (see Section 1 for a review of experimental validation of *O*-GlcNAc sites). A trained model can systematically rank sites in a target protein, based on what is known about the sites. Another important application of PTM classifiers is the proteome-wide study of predicted sites, which can suggest new properties of the studied PTM that are not observed in the initial dataset due to its limited size.

Machine learning methods have been applied to predict several types of PTM sites. Examples are cleavage sites (Nielsen et al., 1997); cysteines that form disulfide bridges (Fariselli et al., 1999; Ceroni et al., 2006); phosphorylation sites (Jensen et al., 2002); and kinase-specific phosphorylation sites (Neuberger et al., 2007). Trost and Kusalik (2011) reviewed around 40 classifiers of phosphorylation sites. The differences among the tools include not only their predictive performance but also what machine learning algorithm was applied, how the training set was prepared, how the learning procedure was carried out, how to define the motif length, and whether to use the sequence only or to include other features. A new method that classifies PTM sites needs to address all these issues.

The importance of *O*-GlcNAcylated proteins was reviewed in Chapter 1. These proteins are involved in a wide range of functions, from gene transcription regulation (Brimble et al., 2010) and protein translation (Zeidan et al., 2010) to modulation of protein kinases (Dias et al., 2009; Wang et al., 2012). OGT is known for targeting proteins with the highest demands for chemical energy in specific tissues, such as P-type ATPases (Clark et al., 2003) and components of the cytoskeleton (Hédou et al., 2009). Overall, this PTM is thought to sense changes in nutrient availability and control adaptation to a new condition. More recently, the specific function of a few sites has been described.

For example, pairs of the Keratins 8/18 protein form the intermediate filament, critical for the epithelial organisation Kakade et al. (2016). The protein is phosphorylated (residues Ser33 and Ser52) and *O*-GlcNAcylated (residues Ser29, Ser30, and Ser48). Kakade et al. (2016) describe that the cells carrying the Ser30Ala mutant, which abolish *O*-GlcNAcylation of the Ser30, lose the ability to migrate under a condition that simulates wound closure. So, specific sites can have a higher level of importance to the cell and organism. The human proteome has around 2.50×10^6 S/T and machine learning methods can help sort potentially modified sites from unmodified ones. Thus, the sites can be further studied to establish whether their modification does or doesn't have a function or involvement in human diseases.

The next sections introduced the Position-Specific Scoring Matrix (PSSM), ANN, Support Vector Machines (SVM) and random forests supervised machine learning methods.

3.1.1 Position-Specific Scoring Matrix

A PSSM is not a machine learning method, although some authors treat them as so (Baldi and Brunak, 2001). PSSM represents the knowledge-based alignment of a series of motifs. The methods summarise a multiple sequence alignment and can be applied to describe motifs in DNA, RNA and protein sequences. PSSM can match new motifs that were not included in the training set (Stormo et al., 1982). The matrix captures the preferred amino acid in each position of the multiple sequence alignment, so the score for amino acid a in the position i is given by the Equation 3.1

$$S_{a,i} = log_{10}\frac{f_{a,i}}{B_a} \tag{3.1}$$

Where $f_{a,i}$ is the frequency of the amino acid a in the position i corrected by pseudo counts (arbitrarily set to = 1) and B_a is the background amino acid frequency[1]. The sum of each amino acid score gives the motif score (Durbin et al., 1998). The cutoff value that classifies the motif as modified or not was obtained under 10-fold cross-validation (see Section 3.2.6).

In this work, a PSSM was implemented with the Pandas DataFrame data structure (McKinney and Team, 2015), and adapted to work with SciKit-Learn, to take advantage of the library `RandomizedSearchCV` and metric evaluation pipeline. It is important to note that PSSM models did not include any extra features, detailed in Section 3.2.4, only the motif sequence.

The disadvantage of a PSSM over more advanced models is that the models cannot identify a non-linear combination of amino acids. A hypothetical training set contains examples dominated by two patterns, one always contains an Ala amino acid in the position -2 and the other always contains a Val in the position -3. An example with Ala in -2, and Val in -3 would have a high score, although this example never appears in the training set and the combination might be mutually exclusive. Other Machine learning methods deal with this issue better. Hence, PSSMs are indicated for motifs with a clear sequence pattern. In that case, PSSMs are useful because the result interpretation is direct and transparent, different from most machine learning

[1] The natural frequency of the amino acid in the human proteome

methods.

3.1.2 Artificial Neural Network

An ANN emulates a simplified biological network of neurons (Wu, 1997). Each node of the network is a neuron, and the neurons, which are organised in layers, are interconnected. A common ANN architecture has three layers: input, hidden and output. The systems learn by an iterative process where the weights of the neurons' interconnections update to adapt to the signals in the input layer and minimise learning error. Overall, the network aims to map the information from the input nodes to the output nodes.

The number of neurons per layer and the layers' dimensions varies in different implementations. Also, ANNs have several parameters that need to be optimised. So, a critical problem with this machine learning method is designing the network and tuning its parameters. The network design is problem specific, so different problems demand different architectures. Moreover, parameters tuning of ANN models lead to overfitting. Because of the limitations, other machine learning methods are preferred over ANN.

3.1.3 Support Vector Machines

SVMs work differently from ANNs. SVM were introduced in 1992 (Boser et al., 1992), later successfully applied to prioritisation of cancer genes (Guyon et al., 2002). Since then, the SVM has become one of the most popular machine learning methods for classifications in computational biology and has been applied to PTM site prediction (Kim et al., 2004).

There are two key concepts for understanding SVMs: kernel functions and the margin separation (Ben-Hur and Weston, 2010). For a binary classification problem that is linearly separable and with two features, or a feature space of two dimensions, the SVM solution is the line that maximises the separation between the two classes. The distance between the line and each class is the margin. For problems in higher dimensions, or more than two features, a hyperplane substitutes the line as the decision boundary. The members of each class that are closest to the decision boundary are the support vectors. The hyperplane can be interpreted as a decision surface, and the surface to member distance associated with the probability of the class membership.

As most problems are not linearly separable, the kernel functions work by transforming the feature space to separate members from the different classes, given their features similarities. The kernel function transforms the feature space into an equivalent but higher dimensional feature space, where the classes are separable. The Gaussian kernel Radial Basis Function (RBF) is the standard SVM kernel in SciKit-learn because of its popularity. The method also supports training on large datasets, that limits older machine learning methods. Kernel functions are critical for SVM, new kernels heavily impact on prediction performance (Leslie et al., 2002).

Overall, SVM models generalise better than ANN models, specially for training sets with high dimensional feature spaces. Also, SVM have fewer parameters and thus are simpler to tune. For example, the linear kernel has a single parameter, C, which tries to balance the number of misclassification and simplicity of the decision surface. So the higher the C parameter, the lower the misclassification rate, but more complicated the decision surface, which leads to overfitting. Apart from the C

parameters, the RBF kernel also have the γ parameter, which defines the shape of the Gaussian model. Smaller γ represents a more constrained model, where a single example has too much influence.

3.1.4 Random forests

Random forests are an ensemble machine learning method, while SVM and ANN are single decision classifiers. In general, ensemble methods train weaker classifiers that collectively outperform single decision ones. Decision trees are relative lightweight classifiers compared to other methods, like SVM.

The unit of the random forest algorithm is a decision tree. Decision trees use simple decision rules inferred from the features. There are multiple decision trees algorithms (Scikit-Learn website, 2016b), but most of them use a binary tree, which implies that each node splits in two.

The individual tree is trained with a bootstrap aggregating the training set. Bootstrapping is a statistical technique that subsamples N samples from a dataset with N samples with replacement, meaning that elements will be repeated in sampled data. The remaining samples are used to determine the out-of-bag error, which is used for internal estimation of the prediction error.

The learning occurs in iterations. In each step, the feature space is partitioned. One critical rule is how to determine the 'node impurity', or the features that best explain the mix of classes in one node. When the criterion is set and reached, the node is split into two child nodes. The process repeats until a secondary criterion is achieved and the terminal nodes assign the class.

Breiman (2001) demonstrated that the generalisation error converges to a minimum value. The author also showed that the new random forest implementations were more robust to noise and outliers than previous implementations, which makes the algorithm particularly useful for biological problems. Another significant advantage of random forests over SVMs and ANN is that the algorithm is not a 'black box'; thus some information regarding the feature importance can be extracted from the model. However, there is a trade-off between model transparency and generalisation error, because the higher the number of decisions tree the lower the generalisation error, but the model becomes harder to interpret.

The SciKit-learn implementations of random forests and SVMs were used in this work. The models were trained with the `class_weight` parameter set to 'balanced'. With this setup, a model trained with unbalanced datasets will penalise more the misclassification of the minority class (the positives) over the majority one. In the case of prediction of PTM sites, methods aim to predict novel sites, so the positive class is more important than the negative.

The random forest method has more parameters than SVMs and some other methods. However, it is less sensitive to parameter tuning than other methods. Two key parameters heavily impact the prediction performance, the `n_estimators` and the `max_features`. The other parameters define split criterion and other tree specific characteristics.

In conclusion, machine learning is an iterative process; adding features, engineering better encoder and optimising parameter can improve the prediction performance. Regardless, one golden rule of machine learning is "more data beats a cleverer algorithm" (Domingos, 2014). More data can always improve predictive performance.

3.1.5 *O*-GlcNAc site classifiers up to 2013

The aim of PTM sites classifiers is to sort modified sites from unmodified ones, ideally by some rank.

The YinOYang method was the first machine learning predictor of *O*-GlcNAc sites (Gupta and Brunak, 2002). YinOYang used an ANN trained on a dataset of 40 sites (Wang et al., 2011). YinOYang included phosphorylation prediction from NetPhosK (Hjerrild et al., 2004) since *O*-GlcNAc and phosphorylation can occur at the same site (Wang et al., 2008). Gupta and Brunak (2002) concluded that OGT does not recognise a clear sequence pattern, but they still highlight the position-specific over-representation of the following amino acids:

- Prolines at positions -4, -3, -2
- Valines at positions -1, +2, +4, +5
- Serines at positions +1, +4, +7

where position 0 is the modified S/T.

Wang et al. (2011) built a Database of *O*-GlcNAcylated Proteins and Sites (dbOGAP) comprising 172 proteins, 798 sites curated from the literature and 365 sites inferred from orthologous proteins. The sites in the database revealed the consensus motif [PV][PA][VT][S/T][TS][AS], which was not sufficient to define all the *O*-GlcNAc-site motifs in the database. So a SVM model called OGlcNAcScan was trained with the sites in the database. The model achieved an Area Under the Curve (AUC) of 74.30 % under a 5-fold cross-validation.

The prediction performance of OGlcNAcScan and YinOYang were assessed by

Jochmann et al. (2014), who collected a dataset of 1 181 new *O*-GlcNAc sites determined by mass-spectrometry from 520 proteins. On the new dataset, OGlcNAcScan and YinOYang yielded low sensitivity of 44 % and 30 %, respectively. The authors explain the low sensitivity as a consequence of both a small training set and the inclusion of a mixture of substrates of the 3 OGT isoforms. Recent evidence suggests that the nucleocytoplasmic isoform of OGT can target mitochondrial proteins (Trapannone et al., 2016) and so far, there is no experimental evidence indicating that the different isoforms have different substrates. For this reason, it is most likely that the small training dataset explains the lack of sensitivity of the OGlcNAcScan and YinOYang methods.

Gupta and Brunak (2002), Wang et al. (2011) and Jochmann et al. (2014) all agree that sequence alone does not contain enough information for training a predictor of *O*-GlcNAc sites. Section 3.2 describes the machine learning models and an approach to training a new classifier of *O*-GlcNAc sites that combines predicted secondary structure, solvent accessibility and disorder. Figure 3.1 summarises the stages of its implementation and testing.

3.2 Methods

Collecting and cleaning the training set is the first and maybe the most important step of training any machine learning model (Domingos, 2014). For models trained on data derived from proteins, this stage needs extra care to control for sequence redundancy (Cuff and Barton, 2000; Overton et al., 2011). Protein domains can be conserved even in proteins with different sequences, which may lead to a strong bias

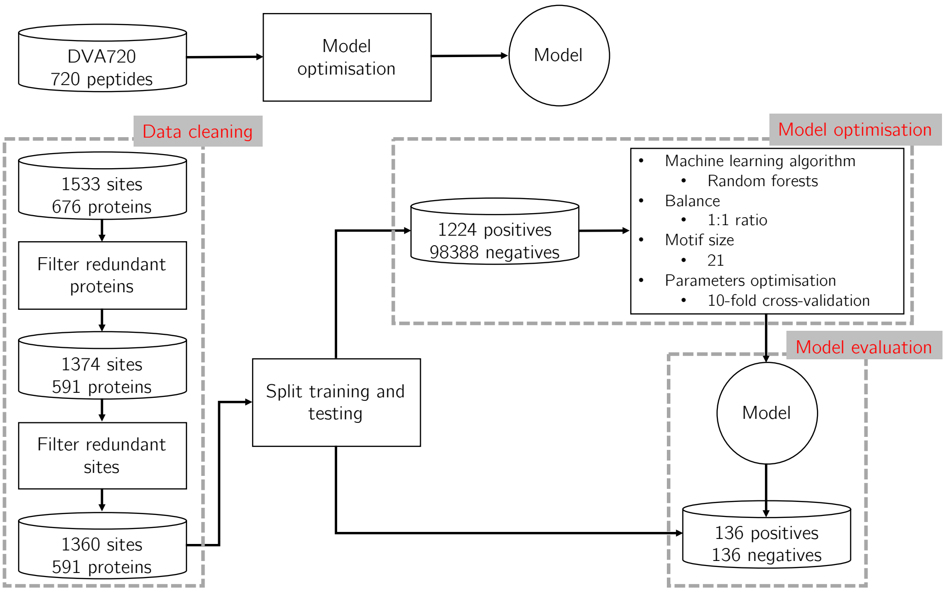


Figure 3.1: A systematic approach to building a predictor of *O*-GlcNAc sites. 2 datasets were studied. The DVA720 was kindly provided by Pathak et al. (2015) and analysed in Section 3.5, and the second dataset is based on the MSS and USS sets from Chapter 2. For the last dataset, the data cleaning step reduces redundant protein and site information. After splitting the non-redundant dataset in one training and one blind-testing set, the trained machine learning models included PSSM, SVM with linear and RBF kernels and random forests. The models were trained on datasets with 3 ratios of positive:negative examples (1:1, 1:10 and 1:70) and motifs of size 6, 7, 11, 15, 21, 31, 41 residues. Model parameters were optimised. The selected model was used to predict the blind test set. This test evaluates how well the predictor generalises by predicting unseen data.

in the data, resulting in over-training of the model (Baldi and Brunak, 2001; Miller and Blom, 2009).

3.2.1 True positive set

Binary classifiers of PTM sites should contain a set of positive (modified) examples and a set of negative (unmodified) examples. The predictor aims to discover new potential sites, so, in this case, comprises S/T with substantial evidence of modification by OGT. Thus, the dataset cleaning stages aimed to maximise the number of true positives, while minimising the number of redundant examples that do not add new information to the learning stage. Too few examples, or too many redundant examples, will lead to under- and overfitting, respectively, so a balance between both is needed.

Filtering redundant proteins

The dataset obtained in Chapter 2 contained 1 533 13-amino acid long sites with non-identical sequence from 676 proteins. The protein sequences were clustered with Blastclust (0.70 identity and 90 % coverage) (Altschul et al., 1990). Parameters were fine-tuned to control the proteins within clusters, which are shown in Appendix B.1. The protein with the most sites was selected, resulting in a dataset with 591 proteins containing a total of 1 374 sites. This stage of the cleaning is also necessary for dealing with redundant sites in homologous proteins (see Section 3.2.2).

Removing redundant sites

Mapping *O*-GlcNAc sites to protein structures in Chapter 2 revealed that some sites occur within similar domains. Since only a small percentage of *O*-GlcNAcylated proteins have a 3-dimensional structure in the PDB, InterproScan domain prediction on each protein sequence was executed (Zdobnov and Apweiler, 2001). Domains defined by InterPro (Hunter et al., 2009), CATH (Sillitoe et al., 2015), SCOP (Fox et al., 2014) and Pfam (Finn et al., 2014) were assigned to the sequence of *O*-GlcNAcylated proteins, and modified sites that occur in the same relative position or ±2 residues within a domain were analysed. Sites were only grouped if the domain assignment looked correct and the sequence alignment was unambiguous, i.e. without a large extension of gaps around the site. Appendix B.2 lists 12 groups and the 14 sites that were discarded. The discarded sites represent a small proportion of the positive dataset. However, their redundancy may have a significant impact on training machine learning models with predicted secondary structure and disorder; therefore the redundant sites were removed.

3.2.2 Negative dataset

The choice of a true negative set is often a problem in machine learning, in particular for models trained with PTM sites. There are no experimentally validated data on unmodified sites. Since there is no large, refined set of S/T that cannot be *O*-GlcNAcylated, here the negative dataset was sampled from the 89 771 unmodified S/T in *O*-GlcNAcylated proteins. This selection strategy of true negative examples has been previously applied to the phosphorylation classifiers (Neuberger et al., 2007). Although it might contain unreported true positives, it is expected that their proportion will be insignificant in comparison to the set. Negative selection from proteins known to be *O*-GlcNAcylated should provide the hardest examples for the classifiers to learn since positive and negative examples come from the same context and bias. If the model can distinguish the classes within this context, it should be able to generalise to unseen proteins.

3.2.3 Blind test dataset

A blind test set, also termed hold-out set elsewhere, was built by random sampling and setting aside 136 positive and 136 negative examples (10 %) from the training dataset. This dataset was not used during the training or optimisation stages and serves as an independent measure of generalisation to new data.

3.2.4 Dataset encoding

The majority of machine learning methods train on numerical features. Moreover, some learning tasks that are facilitated if features are pre-processed by normalisation, for example. Categorical features such as the amino acid sequence, and ordinal features need to be encoded as numbers.

The *O*-GlcNAc-site motif sequences were sparsely encoded. This encoding strategy represents each amino acid in the motif as a vector of 20 binary elements, where each element defines an amino acid. For example, the amino acid Alanine was represented by the $[1, 0, 0, \ldots, 0]$ vector. The absence of a residue, for motifs in the N- or C-terminus of proteins, were represented by the $[0, 0, 0, \ldots, 0]$ vector, with 20 zeros.

Extra features

The following additional features, analysed in Chapter 2, were added to the encoded motif sequence vector:

- Jpred4 secondary structure (Drozdetskiy et al., 2015)
- Jpred4 solvent accessibility (Cuff and Barton, 2000; Drozdetskiy et al., 2015)
- JRonn disorder (Yang et al., 2005) (missing JRonn predictions were substituted with zero)
- DiSEMBL disorder (Linding et al., 2003)
- IUpred disorder (Dosztányi et al., 2005)

The secondary structure elements H, E and C were also sparsely encoded. Each one of the 3 solvent accessibility levels (0%, 5% and 25%) was represented by a binary element 0, for exposed, and 1, for buried. The disorder scores were included as they are. Figure 3.2 illustrates the complete encoding scheme for one amino acid. For sites in proteins, a vector with 32 elements represents each amino acid. The number of amino acids per motif, or the motif length, varies from 6 to 41.

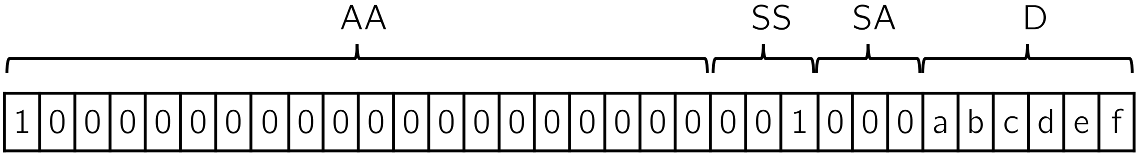

AA
SS
SA
D
1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 a b c d e f

Figure 3.2: Amino acid encoding for SVM and random forest models trained on *O*-GlcNAc sites data. The figure illustrates the encoding of a single example. The amino acid, AA in the image, and the secondary structure element (SS) were sparsely encoded. The 3 solvent accessibility levels (SA), 0%, 5% and 25%, were represented as 3 elements, 0 denoting exposed and 1 buried. Disorder scores (D) were input as they are: a, JRonn; b, DiSEMBL-HOTLOOPS; c, DiSEMBL-REM465; d, DiSEMBL-COILS; e, IUpred-short; and f, IUpred-long.

3.2.5 Dataset balance

For a 2-class problem, a balanced dataset has the same number of positive and negative examples in the training set. Equation 3.2 defines a balanced dataset.

$$\text{Ratio}_{\text{negative:positive}} = \frac{\text{Number of negatives}}{\text{Number of positives}} \approx 1 \tag{3.2}$$

In this work, $Ratio_{\text{negative:positive}} \approx 70$, which means there were approximately 70 unmodified S/T for each modified one. Data imbalance is a common problem for PTM classifiers that select the negative dataset from a set of unmodified sites.

One issue associated with dataset imbalance is the classifier performance evaluation. The accuracy metric, detailed in Section 3.2.6, is not indicated for evaluation of a model that learned from an imbalanced dataset because a naive predictor that always predicts the major class will obtain better accuracy than random choice.

A second issue is the elevated number of examples. After a critical point, additional negative examples will not improve the model's performance. Moreover, an increased number of negative examples will drop the model efficiency, defined as the time the method takes to learn from the data. As a result of an inefficient method, further developments, such as the optimisation of the method parameters' and potential improvements, and testing new encoding strategies, will be delayed. In conclusion, model efficiency also needs to be taken into account while deciding the most appropriate dataset balance.

A simple strategy to deal with class imbalance is to under-sample the majority class. So, the negative examples were randomly selected to match ratios equal to either 10:1 or 1:1. This strategy has the additional advantage of reducing the probability of a mislabelled true positive being included in the negative dataset.

The natural ratio of modified and unmodified S/T is unknown. From the data in the training set, the ratio in the human proteome was estimated to be approximately 36 000 modified S/T to the 2.50×10^6 S/T. However, this number should be smaller, since OGT will not target a subset of the proteins. If the under-sampling of the majority class is employed, the procedure will affect the classifier scoring and may also impact on the proportion of predictions, depending on the method used. So the artificial $Ratio_{\text{negative:positive}}$ may lead to over-prediction. The proportion of the classes is critical when the prediction is applied to the human proteome, in Chapter 4.

3.2.6 Training and testing

Since several models were trained with different methods in different setups, the performance of the models needs to be compared. The Section 3.2.6 discusses mainly the classification metrics for a binary classification task that needs to support unbalanced data. A metric for regression model evaluation is also briefly introduced, as it is used in Section 3.5.

Measuring the classifier performance

The accuracy metric (Equation 3.3) measures the proportion of correctly predicted examples (true positive and true negative) over all predictions.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.3}$$

As discussed before, accuracy is suboptimal for evaluation methods trained on unbalanced datasets. The True Positive Rate (TPR) (Equation 3.4), or sensitivity, and the False Positive Rate (FPR) (Equation 3.5), or specificity, measure the proportion of correctly assigned true positive and true negative respectively. Precision measures the fraction misclassification of the positive class.

$$\text{TPR} = \frac{TP}{TP + FN} \tag{3.4}$$

$$\text{FPR} = \frac{FP}{TN + FP} \tag{3.5}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3.6}$$

Where true positive (TP) and true negative (TN) are the correctly predicted positives and negatives, respectively. Negative examples misclassified as positives are termed false positives (FP), while positive examples misclassified as negative are called false negatives (FN).

The Receiver Operating Characteristic (ROC) curve plots the relationship between the TPR and FPR. The AUC is a summary statistic for the ROC curve. An AUC of 0.50 represents a random prediction, and the ideal ROC curve (all positive and negatives correctly classified) has AUC of 1.

Equation 3.7 shows the Matthews Correlation Coefficient (MCC), which also expresses the predictor performance. The score measures the correlation between expected and observed predictions and varies from -1 and 1, where 0 denotes as good as random predictions and 1 ideal predictions. Both MCC and AUC of the ROC curve are metrics indicated for binary imbalanced classification problem (Baldi and Brunak, 2001).

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{3.7}$$

Regression models in Section 3.5 were evaluated with the R^2 (Equation 3.8) metric:

$$\text{R}^2 = 1 - \frac{\sum\limits_i (y_i - f_i)^2}{\sum\limits_i (y_i - \bar{y})^2} \tag{3.8}$$

Where $\sum\limits_i (y_i - \bar{y})^2$ is the total sum of the squares and $\sum\limits_i (y_i - f_i)^2$ is the residual sum of the squares, for n samples $f = [f_1, \ldots, f_n]$, predicted values $y = [y_1, \ldots, y_n]$ and $\bar{y}$ is the mean predicted value. The metric measures the deviation of a predicted value from its target value. The closer to 1, the more accurate the regression model. This score is not symmetrical and may be negative in case the model fails to learn from the training set (Scikit-Learn website, 2016d).

Training stage

Models were trained with the 10-fold cross-validation. This approach divides the dataset into 10 parts and, for each iteration, trains with 9 parts and tests on 1. The mean MCC score was calculated from the 10 iterations. Since an algorithm's performance may be different for different motif lengths, the sizes 6, 7, 11, 15, 21, 31 and 41 amino acids were tested.

Parameters were optimised with the `RandomizedSearchCV` function implemented in SciKit-Learn. It works by fine-tuning the parameters, aiming to achieve the maximum MCC score during the cross-validation. An alternative method to `RandomizedSearchCV` is the `GridSearchCV` function, which exhaustively searches a pre-established list of parameters. The advantage of using `RandomizedSearchCV` is detecting parameters that impact on predictor performance and extends the search for optimal parameter values (Bergstra and Bengio, 2012) while limiting the search for parameters with less impact.

Each machine learning algorithm contains specific parameters that need to be empirically adjusted. Table 3.1 lists these parameters and their possible values. The random forest parameters ‘minimum samples per leaf’ (`min_samples_per_leaf`) and ‘minimum samples per split’ (`min_samples_per_split`) were sampled from a discrete uniform statistical distribution, while the other numeric values were sampled from a continuous uniform function, implemented in SciPy. In the ‘Possible values’ column, the number within parenthesis determines the range of the distribution. Random forest has an important parameter, `n_estimators`, that represents the number of decision trees in the forest. This parameter does not need to be optimised, because the higher the number, the better the predictive performance, until a critical point, where the performance stops increasing. The `n_estimators` was left at its default (10) until the final prediction, since this parameter also has a large impact on training efficiency.

3.2.7 Mining novel sites from abstracts

To compare POGSPSF with alternative predictors, a new dataset of *O*-GlcNAc sites was collected from abstracts of scientific articles in the Pubmed Central repository (Maloney et al., 2013). The repository was searched with the BioPython (Cock et al.,

Table 3.1: Machine learning algorithm parameters optimised with `RandomizedSearchCV`. The 'Possible values' column defines search the space.

Machine learning	**Parameter names**	**Possible values**
PSSM	cutoff	uniform distribution(-2, 12)
Random forests	bootstrap	True or False
	minimum samples per leaf	uniform distribution(1, 15)
	minimum samples per split	uniform distribution(1, 7)
	criterion	Gini or Entropy
	maximum features	uniform distribution(0.1, 1.0)
SVM with linear kernel	C	uniform distribution(0.1, 10 000)
SVM with RBF kernel	C	uniform distribution(0.10 , 10 000)
	γ	uniform distribution(0.01 , 100)

2009) Entrez Application Programming Interface (API) for articles containing the OGT term. The selected abstracts were processed with the following case insensitive regular expression:

```
\W(s|ser|serine|t|thr|threonine)([\d+]{1,3})([a-z]{1,3})?\W
```

The regular expression matches text such as Ser473 or T41A. 25 articles that match the regular expression were manually curated, yielding 19 novel *O*-GlcNAc sites in 8 proteins. The S/Ts in the proteins were classified with POGSPSF (this work), YinOYang (Gupta and Brunak, 2002) and OGlcNAcScan (Wang et al., 2011) to evaluated each classifier's performance.

3.3 Results

3.3.1 Machine learning algorithm selection

The efficiency and performance of methods were compared by running parameter optimisation with 10-fold cross-validation using the balanced training set and a time limit of 24 hours. Also, motif lengths from 6 to 41 amino acids were evaluated. The applied machine learning algorithms were PSSM, SVM with linear and RBF kernels, and random forest.

Figure 3.3 shows the parameter optimisation for the trained models. The SVM based models did not complete parameter optimisation within the time limit, except the model with the linear kernel and for motif lengths six and seven. All other models were optimised within the time limit.

The PSSM models obtain an MCC score ≤ 0.12 in the experiment. In fact, PSSM models were trained as a baseline since there is enough evidence to support that sequence alone poorly distinguishes modified and unmodified S/T. Hence PSSM models were expected to have the worst prediction performance of the methods.

Random forest models perform poorly for motif lengths of 6 or 7 amino acids but obtain a MCC score of 0.23 for motif lengths of 11, which is the top score in the experiment. Longer motif lengths did not increase the random forest performance. The result with the motif length of 7 residues was considered an outlier, and further investigation is needed to determine why the predictive performance of this point was so low.

Table 3.2 lists the performance metric for a combination of motif length and ratio ($Ratio_{\text{negative:positive}}$). Motif length < 11 amino acid have a negative impact on the

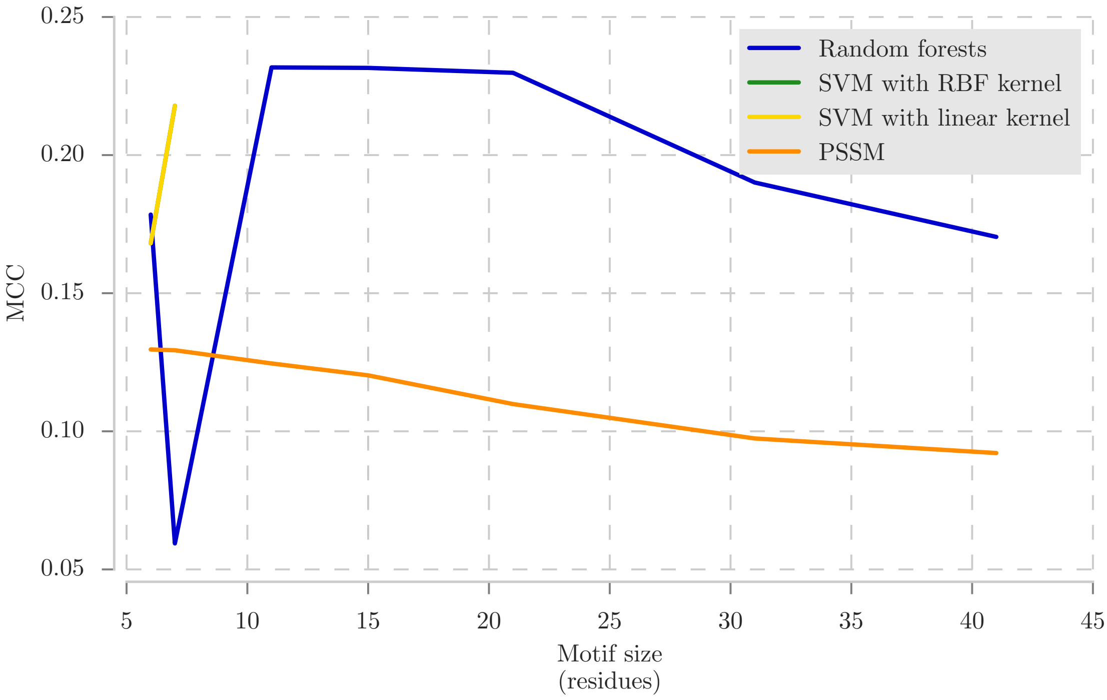


Figure 3.3: 4 machine learning algorithms were optimised for the same number of iterations for a maximum of 24 hours. In the y-axis, the mean MCC calculated from 10-fold cross validation; in the x-axis the motif length. The SVM based model did not train within the time limit, except the linear kernel with 6 and 7 amino acid motifs, and therefore, are missing from the plot.

predictor performance and are omitted. The balanced training set outperforms the unbalanced options for all tested motif lengths. 2 training sets obtained the top performance, 0.73 AUC. The simplest model, with the motif length of 21 residues, was selected for a longer round of parameter optimisation.

Table 3.2: Prediction performance for random forest models with a combination of training sets. The combination comprises example ratio and motif length. The models with 21 and 31 amino acids obtained the top scoring prediction performance of 0.73 AUC. Ratio, $Ratio_{\text{negative:positive}}$. Motif length in residues. AUC, MCC, sensitivity and specificity calculated with 10-fold cross-validation.

Motif length	**Ratio**	**AUC**	**MCC**	**Sensitivity**	**Specificity**
11	1	0.71	0.34	0.70	0.42
11	10	0.70	0.16	0.49	0.09
11	70	0.61	0.04	0.26	0.01
21	**1**	**0.73**	0.33	0.69	0.40
21	10	0.70	0.17	0.54	0.09
21	70	0.62	0.04	0.29	0.01
31	**1**	**0.73**	0.36	0.71	0.40
31	10	0.70	0.14	0.49	0.09
31	70	0.62	0.03	0.22	0.01
41	1	0.71	0.31	0.68	0.42
41	10	0.70	0.15	0.54	0.09
41	70	0.61	0.06	0.04	0.01

Figure 3.4 shows the ROC curve for the model trained with a motif length of 21, example ratio 1:1, optimised parameters and `n_estimators` to 10 000. The small increase in prediction performance indicates the model optimisation has a limited impact for the algorithm in this dataset. The AUC under the ROC curve for the blind test set is equal to 0.71, demonstrating that the model can generalise to new data. Table 3.3 shows model TPR and FPR for each class.

Table 3.3: POGSPSF TPR and FPR for each class.

Class	FPR	TPR
Unmodified	0.73	0.63
Modified	0.57	0.68
Mean	0.66	0.65

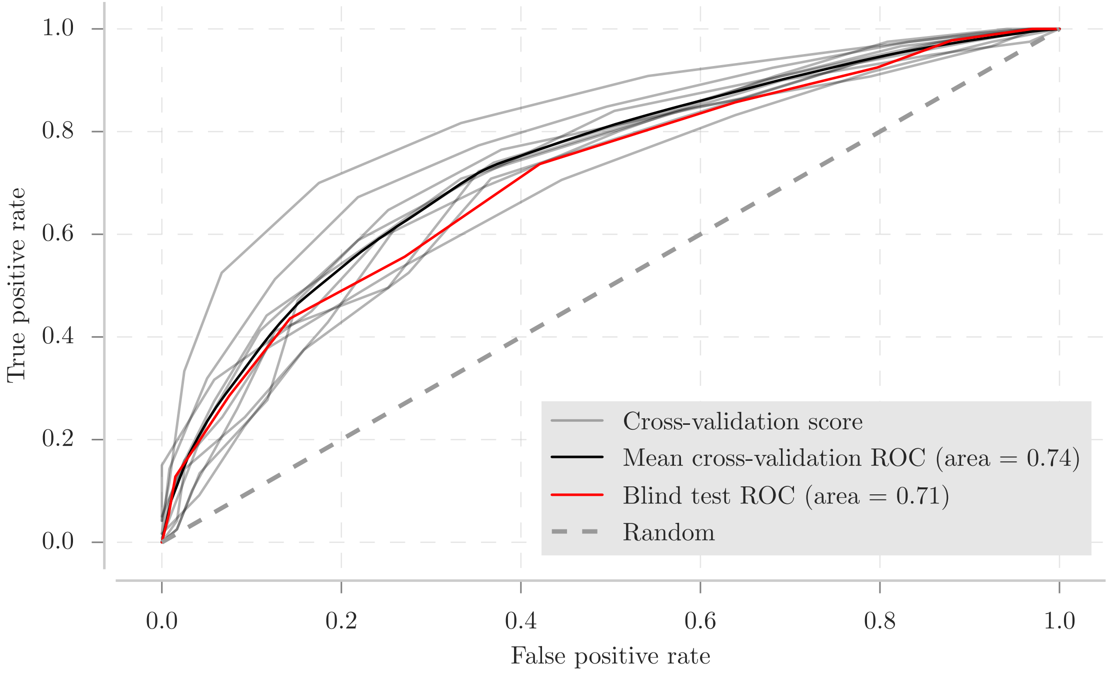


Figure 3.4: ROC curves for the final model. Each iteration of the cross-validation is shown by a grey curve. The combined results of the cross-validation iteration is shown by the black curve. The blind test result is shown by the red curve, which is within the cross-validation iterations. The dashed line represents the random predictor.

Table 3.4: POGSPSF performance comparison. YinOYang has highest MCC, however POGSPSF has the highest sensitivity among the methods, thus better for detecting truly modified *O*-GlcNAc sites, while challenged with 12 sites obtained from scientific literature. Precision and specificity were calculated for the positive class.

	Precision	**Specificity**	**MCC**
YinOYang	0.71	0.42	0.53
OGlcNAcScan	0.17	0.17	0.12
POGSPSF	0.23	0.58	0.31

note that this comparison is limited due to the small size of the dataset.

3.4 Discussion

This Chapter listed the common problems with computational classifiers of PTM sites. Then, the systematic study and development of a new machine learning model called POGSPSF was presented and compared to two broadly used tools for identification of potential of *O*-GlcNAc sites.

3.4.1 Problems with classification of PTM sites

Data redundancy is an overlooked issue for PTM classifiers based on machine learning algorithms, including OGlcNAcScan. To balance the maximum number of examples and to minimise the number of redundant examples is a challenge. Manual selection of sites is possible, but it can introduce undesirable biases. Since this work uses the motif sequence and its features for the prediction, a new pipeline was added to deal with motif redundancy. It first discards the sites with duplicated sequence, for a site defined as a sequence of 13 amino acids. Subsequently, the protein sequence was clustered with Blastclust. The clustering of protein sequences with Blastclust or CD-HIT (Li and Godzik, 2006) is a common step among the methods that aim to predict PTM sites. However protein sequence clustering with a very stringent parameter (identity $\leq 50\%$) may discard more sites than necessary, and the number of known *O*-GlcNAc sites is already limited. Since this work adds structural features calculated from the protein sequences, proteins were clustered with an identity threshold of parameter 70% and the sites occurring in the same relative position of a domain, i.e. with the same structural context, were manually analysed and removed. With this pipeline, only two pairs of sites, Q96L91-3027 and Q8CHI8-2940; P11798-306 and Q3TY93-178, had < 5 amino acids differences, for a motif length of 21 amino acids. Due to its size, it is much harder to apply the manual component of this pipeline to the negative dataset. The redundancy in the negative dataset may explain why models trained with the balanced dataset had superior prediction performance compared to the ones trained with the balanced dataset.

A second issue is the lack of high quality site information. In 2013, YinOYang and OGlcNAcScan were the only two computational predictors of *O*-GlcNAc sites available, contrasting to a plethora of tools that predict protein phosphorylation from the sequence. The YinOYang method was trained on a very limited dataset, which is not available despite the fact the method is still active on-line. Also, OGlcNAcScan learned from data in homologous proteins. However whether *O*-GlcNAc sites are conserved among close related species is an open question. Jochmann et al. (2014) reviewed theses methods and identified an issue: the small number of examples in the training set. The weak pattern observed in the sequence alignment of known *O*-GlcNAc sites might be due to lack of data, but also because the motif sequence itself only carries part of the information recognised by OGT. Chapter 2 indicates that predicted local structure, specifically secondary structure and disorder ought to carry part of this information.

One important aspect when selecting a machine learning implementation is the training efficiency. The training efficiency depends not only on the machine learning algorithm but also on the size of the training set, how the data are encoded and if the positive and negative classes are easy to separate. The training efficiency is important because several alternative models should be evaluated during the systematic development of machine learning model. In this work, random forest models had more efficient training times than other models.

3.4.2 Feature importance

Machine learning methods have a broad spectrum of transparency. In one extreme, methods such as PSSM, and decision trees provide details that the algorithm learned from the data. Other methods, such as ANNs, are 'black boxes' as they do not offer information on which features influence learning and how. Some machine learning methods enable the study of which features are determinant for a correct classification. Transparent machine learning models can be extended with new features to bring new light to the problem they describe. Moreover, random forest, and tree-based methods, enable study of feature importance.

3.4.3 POGSPSF development

POGSPSF was trained on a balanced dataset comprising 21 amino acids surrounding the modified S/T, plus features predicted from their protein sequences. The negative dataset contained randomly selected S/T from the same proteins. Careful evaluation of the predictor performance was carried out, and it achieved an AUC of the ROC curve of 0.71 on the blind test. This result is suboptimal, but the impact of structural features predicted from the protein sequence is positive in the classifier performance. Moreover, POGSPSF has higher specificity than OGlcNAcScan and YinOYang, and hence being useful for site prioritisation prior experimental validation.

The most significant contributions of POGSPSF model are the data update, the addition of features predicted from the sequence and training using random forests. Regardless, the model can be further optimised. Recent advances in machine learning are promising for the prediction of PTM sites. Specifically, semi-supervised learning approaches can handle the lack of an experimentally determined negative dataset (Hao et al., 2015; Yang et al., 2016). The first studies applying deep learning to the computational biology field appeared recently (LeCun et al., 2015). This modern machine learning method has been successfully applied to the prediction of DNA- and RNA-binding proteins (Alipanahi et al., 2015) and local protein properties (Qi et al., 2012).

Other machine learning methods were also tested, such as ANN with Stuttgart Neural Network Simulator (SNNS) (Zell et al.), Theano (The Theano Development Team et al., 2016), and Stochastic Gradient Descent implemented in SciKit-learn, but the ANN models could not learn from the data. In the case of the *O*-GlcNAc sites, an efficient training is needed to maximise the number of trained models for further selection.

Two methods that reduce the dimensionality of the feature space, principal component analysis and linear discriminant analysis implemented in SciKit-Learn, were tested. The feature dimensionality reduction methods aim to reduce the number of features to minimise noise in training and deal with with sparse datasets. Patterns are harder to spot in such datasets. However, neither method was able to improve the prediction performance.

Besides the experiments with alternative machine learning methods and preprocessing steps, different amino acid encoding strategies were also tested. The amino acid sparse encoding was extended with 3 encoding strategies. Firstly, the amino acids were represented by entries in the AAindex database (Kawashima et al., 1999). Neuberger et al. (2007) have successfully applied this approach to the prediction of protein kinase A sites. Secondly, the motif sequence was represented by the amino acid and dipeptide composition (1- and 2-mer kmer composition), an encoding strategy that has been extensively applied to machine learning tasks trained with protein information, similarly to the SVM Spectrum Kernel (Leslie et al., 2002) and the Pseudo Amino Acid Composition (Chou, 2001). Thirdly, a physico-chemical property-based encoding was based on work of Taylor (1986), which describes the amino acids set of 10 physico-chemical properties. None of the encoding strategies improved the random forest prediction performance. Nevertheless, amino acid sequence encoding to feature vectors is critical for prediction, and the *O*-GlcNAc-sites classification can benefit from new encoding strategies that are able to better represent the problem. Examples of encoding strategies that may be applied to this problem are amino acid substitution matrix such as BLOSUM 62 and reduced amino acid alphabets that incorporate structural information (Li et al., 2005; Wong et al., 2007). Chapter 4 also discusses additional features that may be included in the model.

3.5 The DVA720 dataset

3.5.1 Background

An alternative dataset was also studied. The dataset, called DVA720, comprises 720 13-residues-long peptides from a protein kinase library. The peptides were subjected to a large scale OGT activity assay (Pathak et al., 2015) to measure whether they are substrates to the enzyme. The assay consisted of incubating each of the substrate peptides in the presence of radioactive UDP-GlcNAc in the presence and absence of OGT. A positive control, CRYA1, and a negative control, CRYA1 with the modified Ser mutated to Ala, were present on each assay plate. The OGT activity was measured in duplicate for each peptide, and the radioactivity signal was converted to the activity of the positive control for each replicate, so the presented activity is relative to the CRYA1 peptides. The DVA720 dataset has 2 differences from the dataset used to train POGSPSF. The addition of features predicted from protein sequences would not be appropriate, because the OGT activity was measured from peptides and not from their full-length proteins. And, this dataset enabled the training of regression models targeting the OGT activity values. The dataset in its initial form does not have redundant sites, so cleaning was not required.

Due to the size of this dataset, the models were trained with 5-fold cross validation and no data was set apart for the blind-test.

3.5.2 Models trained with the DVA720 dataset

Regression models were trained on the sparsely encoded peptide sequences derived from DVA720, targeting the normalised OGT activity value. If the models could

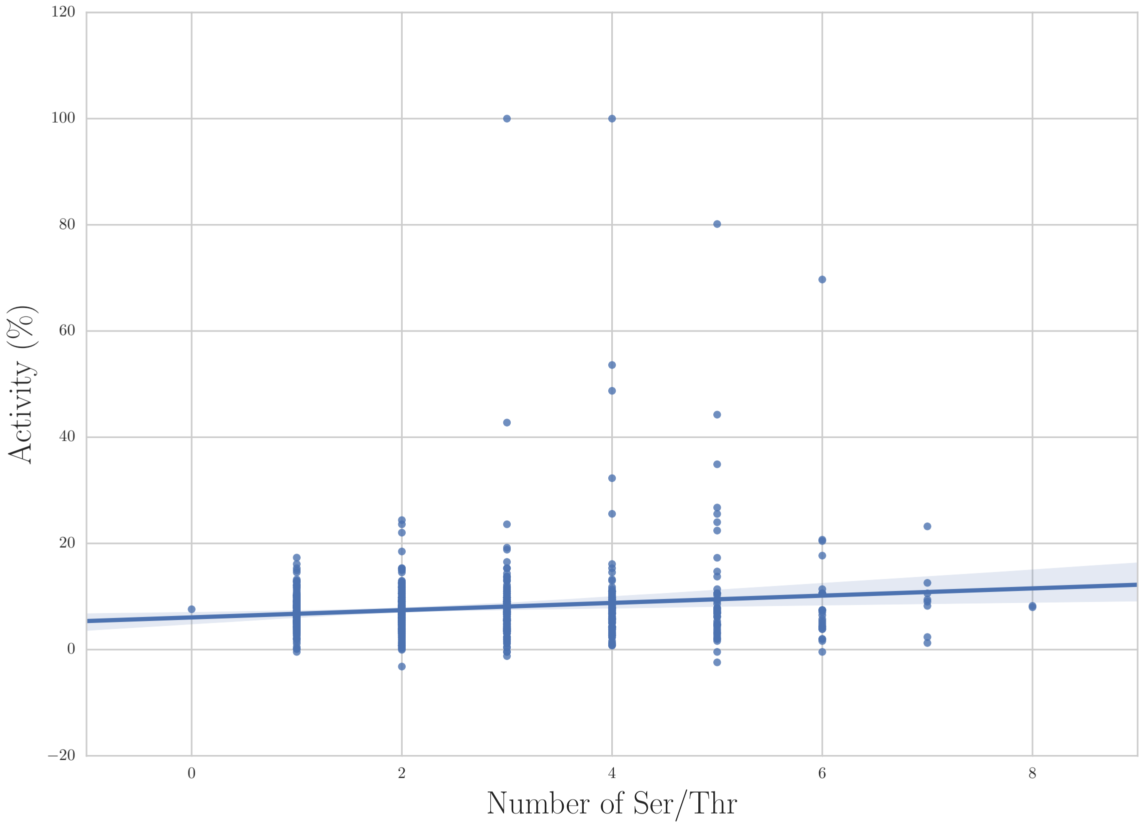


Figure 3.5: The number of S/T in peptides does not correlate with the OGT activity in the DVA720 dataset (Pearson's correlation=0.11). The activity values of 2 peptides, which were higher than 100 %, were omitted for visualisation purposes only.

learn from the data, they would be able to estimate the OGT activity based on a peptide sequence.

Classification models were also developed from the same dataset. 69 peptides yield activity greater than the activity observed for the reference peptide CRYA1, which is known as a poor OGT substrate (Roquemore et al., 1992). The modification of 36 of these peptides was confirmed by mass-spectrometry. The exploratory analysis of potential features that could explain the differences in activity reveals one peptide with no S/T and activity of 7.60 %. There is no relationship between the number of S/T in the peptides and OGT activity, as shown in Figure 3.5. So, it is safe to establish the 7.60 % value as a cutoff between unmodified peptides and the other peptides. Figure 3.6 shows the distribution of peptide activity. 3 classes were derived from the activity values: 'Positive' for the 70 peptides with activity $\geq$ 12.20 %; 'Negative' for the 410 peptides with activity $\leq$ 7.60 % (409 examples); and 'Other' for the 240 peptides with activity values between 7.60 % and 12.20 %. The 'Other' class comprises both modified and unmodified peptides. 2 binary classification models were trained: one considering Other and Negative peptides as the negative class and the other excluding the Other class from the training set.

3.5.3 Model performance

Table 3.5 summarises the results from the models trained with the DVA720 dataset. The low performance ($R^2 \leq 0$) of the regression models indicates that these models did not learn from the dataset. The classification model trained on the complete dataset, Negative and Other from Figure 3.6 as negative examples, had a lower AUC than the classification model trained on the dataset without examples classified as

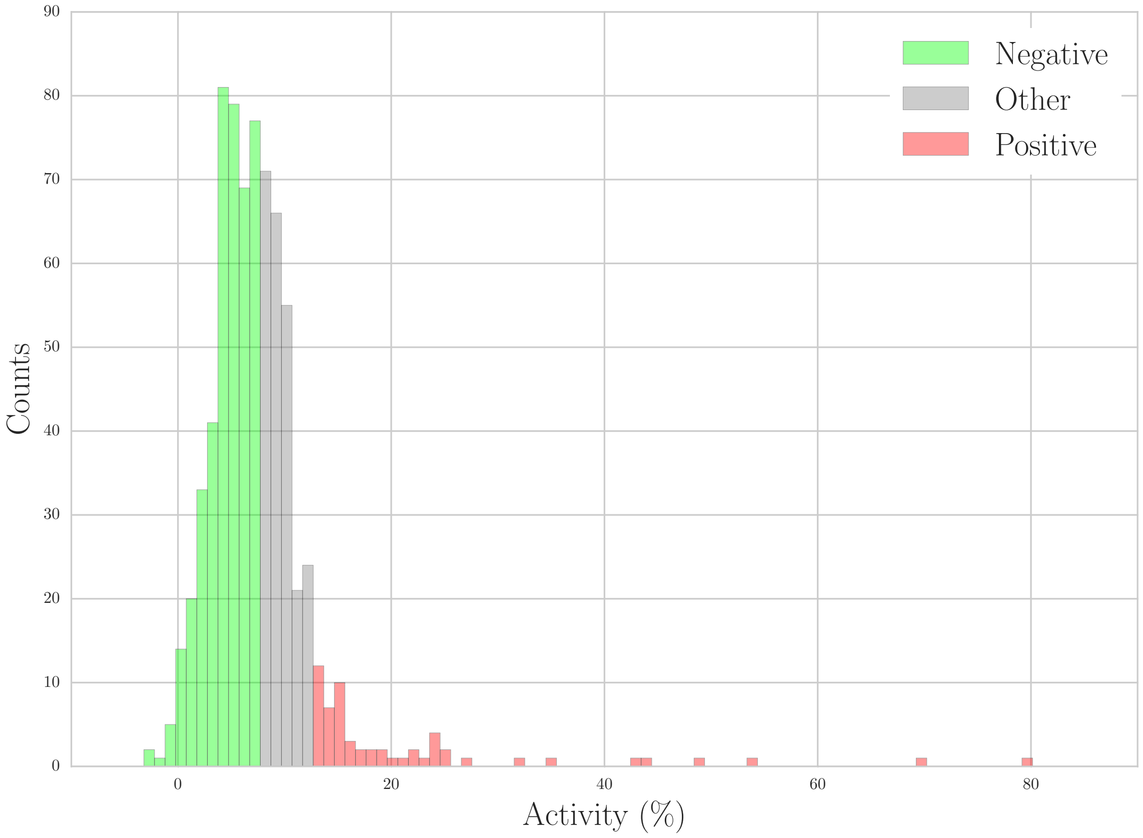


Figure 3.6: OGT activity distribution for peptides in the DVA720 dataset. The distribution is log-distributed and there is no clear separation of modified and unmodified peptides. The activity values were classified as Negative (unmodified), Other (maybe unmodified) and Positive (modified). 2 activity values $> 100\,\%$ were omitted from the plot.

Table 3.5: Prediction performance for models trained on the DVA720 dataset. Regression models were evaluated with the R^2 metric and classification models were evaluated with the AUC. The best prediction performance (bold) was obtained by the SVM model trained with the RBF kernel on the training without the Other class (see Figure 3.6).

	Method	**AUC**	R^2
Regression	Random forest		-8.96
	SVM with linear kernel		0.00
	SVM with RBF kernel		-0.22
	SVM with polynomial kernel		-0.01
Classification, all data	Random forest	0.57	
	SVM with linear kernel	0.53	
	SVM with RBF kernel	0.56	
	SVM with polynomial kernel	0.53	
Classification, without Other	Random forest	0.53	
	SVM with linear kernel	0.53	
	SVM with RBF kernel	**0.59**	
	SVM with polynomial kernel	0.55	

Other. For the model trained on the complete dataset, the random forest method obtains the highest performance of 0.57 AUC. The second classification model was trained excluding peptides with activity from $> 7.60\,\%$ and $< 12.20\,\%$, a range that might contain both modified and unmodified peptides. The SVM model with RBF kernel produces the highest performance of 0.59 AUC, which is a small gain on the first classification model. Overall, the classification model learns from the DVA720 dataset, but the models lack performance for real applications.

3.5.4 Discussion

The large scale OGT activity experiment demonstrated that OGT activity is specific to peptides (Pathak et al., 2015), and more than 400 peptides yield activity comparable to the peptide without a S/T. It is, however, unclear how this result extrapolates to full-length proteins. Despite that, the DVA720 dataset seemed to be a promising resource for training a predictor of *O*-GlcNAc sites.

Regression models targeting the OGT activity did not learn from the DVA720 dataset. The highest performance classification model, SVM with RBF kernel, only achieved 0.59 AUC. This dataset underwent extensive exploratory analysis, and several models were developed. With the dataset extended with the AAindex and di-peptide composition encoding and pre-processed with linear discriminant analysis, the prediction performance increased only to 0.61 AUC. The small increase in performance does not justify the over-complicated model. In conclusion, the models trained with the DVA720 dataset were not further studied, because of their low prediction performance.

3.6 Conclusions

- The chapter describes the development of a new classifier of *O*-GlcNAc sites
- Several motif encoding strategies and supervised machine learning methods were evaluated
- POGSPSF random forest model was trained with sparsely encoded motif sequence plus predicted secondary structure, solvent accessibility and disorder score
- the model achieved 0.71 AUC of the ROC and higher specificity for modified sites than YinOYang and OGlcNAcScan
- The DVA720 dataset was studied, but models trained with the dataset had

low performance.

Chapter 4

POGSPSF applications

Preface

The previous chapter described POGSPSF development; this chapter outlines the applications of the tool. Sites were predicted on 71 791 proteins in the human proteome. The chapter also analyses the proteome-wide prediction of sites, focusing on Gene Ontology (GO) term analysis of proteins with and without predicted *O*-GlcNAc sites. Next, the high and low-scoring S/Ts were also examined for phosphosites and genetic variants. Finally, the proteins with high-ranking predicted sites were manually curated to investigate potential novel OGT targets and the predictions were made available to users on a web application.

4.1 Introduction

POGSPSF has further applications in addition to ranking potential *O*-GlcNAc sites. The tool also allows the study of the predictions at a large scale. For example, known

O-GlcNAc sites occur in nuclear, cytoplasmic and mitochondrial proteins. So it is important to test whether the predictions preserve this profile. Furthermore, the analysis of pathways and processes represented by predicted sites could point to yet unreported targets and pathways.

Several factors limit the prediction of *O*-GlcNAc sites. The lack of high-quality *O*-GlcNAc site data is one of these factors. The limitations of training machine learning model on sites identified by mass-spectrometry were discussed in Chapter 1. Also, another factor that may limit the prediction performance is the suboptimal motif encoding method, since the few alternatives to sparse encoding did not improve the classifier performance. The application of the classifier on a large scale can identify problems with the tool.

Despite the limitations, it is still possible to extract new information from large scale predictions. The information may hold clues to potential associations that are difficult to observe from the mass-spectrometry data. For example, the predictions may identify pathways that include low abundance proteins. If the information is relevant and valid, it can be applied to the next version of the classifier and increase its predictive performance.

4.1.1 Proteome-wide predictions

In 1995, the term ‘proteome’ was first used to refer to the proteins complementary to a genome (Wilkins et al., 1996; Wilkins, 2009). At the time, the two-dimensional electrophoresis technique could observe around 200 proteins per experiment. Proteomics studies could detect the change from cell-type to cell-type and also different cell conditions such as stress and infection, revealing that the proteome is a dynamic entity.

Current proteomics techniques are widespread in life sciences research. These methods can determine the protein abundance, subcellular locations, interactors and post-translational state. The protein PTM state defines the multi-site modification at a given cellular moment. Due to the large scale nature of the methods, proteomics studies apply several techniques and tools from computational biology. The studies also depend on databases and information curated in the scientific literature. So, proteomics and computational biology complement each other.

Overall methods that predict PTM sites tend to overpredict (Blom et al., 2004). Overprediction can be defined as a high number of false positives meaning that unmodified sites will be classified as modified. Gupta and Brunak (2002) applied the YinOYang classifier to ≈5 500 proteins sequences and 4 600 had at least one predicted site, indicating overprediction. The analysis of this issue can be broken into several subcomponents. Some proteins, such as secreted and integral membrane proteins, will never encounter OGT and cannot be modified. Since POGSPSF predicts at the motif level, rather than the protein level, the tool might predict sites in proteins that are not in the cytoplasm, nucleus and mitochondria if the site's sequence and structural context are similar to examples in the training set. Users need to have this in mind during the selection of potential protein targets. For model organisms, this information is readily available and could be added to POGSPSF. Another component of the overprediction problem is the small number of examples in the training set, and the possible presence of unmodified sites reported as modified. One can argue that the number of known sites modified by OGT is greater than the number of known sites modified by kinases like PKA, which has been extensively studied (Pinna and Ruzzene, 1996; Ptacek et al., 2005; Neuberger et al., 2007; Hornbeck et al., 2015). Although this is true, kinases have a well-defined modification site. Furthermore, dataset size and quality is a common problem for classifiers trained with data obtained predominantly from mass spectrometry. A structure based filter was applied to phosphosites (Vandermarliere and Martens, 2013), to discard probable false positives and improve the data quality. Hence, for example, buried sites were discarded and the dataset cleaned. However, the use of protein structure as a filter is limited by the number of structures for the proteins in the dataset, as discussed in Chapter 2.

Lastly, POGSPSF was trained on the same number of positive and negative examples. The importance of dataset balance has been established for many computational biology problems including PTM site prediction and it is commonly observed that balanced datasets outperform unbalanced datasets (Wei et al., 2013). As discussed in Chapter 3, although undersampling the majority class increased the predictive performance, it also leads to overprediction by changing the prior ratio between classes. This problem is not restricted to random forest models and might occur with other machine learning methods (Scikit-Learn website, 2016c).

Instead of directly predicting class, the SciKit-Learn random forest implementation can provide a class probability score with the `predict_prob` method. For a binary classifier, the score indicates the likelihood that a target belongs to the positive class (modified). The random forest classifier uses a probabilistic model to weight each feature's importance and the classes are given by the score cutoff of 0.5. Different implementations use majority voting when aggregating the results from the forest (also called the ensemble) of decision trees (Breiman, 2006).

To overcome the overprediction issue, a new dataset of *O*-GlcNAc sites is needed, which could be used to calibrate the class probability score. However, this is not yet available. An alternative approach is to observe the prediction over the proteome and heuristically define a more conservative cutoff aiming to minimise the number of false positives. This strategy was applied in this chapter.

4.1.2 Gene ontology

The GO project provides a hierarchical annotation for genes (Gene Ontology Consortium). It facilitates the large scale comparison of sets of genes (Gaudet and Dessimoz, 2016). The project provides a controlled and consistent vocabulary of terms subdivided into cellular component terms, which identify the subcellular location of the gene products; the biological process terms, which list the gene participation on pathways, cell physiology and disease; and the molecular function terms, which describe the gene product's main actions. Apart from these three types of terms, GO also describes the term's relationship, as groups of terms and terms that regulate other terms.

Different sources of evidence tags support the GO term assignments. Evidence tags can be linked to experimental data, inferred by curators or automatically assigned. Experiment-supported evidence tags connect the gene to references in the scientific literature. Evidence tags inferred by curators are often associated with sequence alignments, protein models, orthology or genomic context. Automatically assigned evidence tags are made by algorithms without curator interference. Although predictions contain some level of inaccuracy, they are critical for the annotation of genes in non-model organisms with poorly annotated genomes.

A plethora of tools and analysis pipelines use GO terms for the high level functional characterization of proteins (Binns et al., 2009; Mi et al., 2013; Alexa et al., 2006). GO term analysis has been extensively applied to the investigation of large scale experiments, such as differential expression and mass-spectrometry in different conditions, to identify changes of the GO term profile between conditions. In this chapter GO enrichment analysis was applied to profile proteins with and without predicted sites.

4.2 Methods

4.2.1 Data collection and processing

Proteome-wide disorder and secondary structure prediction

The human proteome was retrieved from the UniProt KnowledgeBase (UniProtKB) in November 2015 (release 2015_11). Sequences longer than 800 residues were excluded because of the maximum sequence length accepted by Jpred4. The resulting dataset comprises 79 180 sequences. For comparison, the current UniProt release (2016_08) contains 92 910 sequences from which 20 980 sequences represent the canonical sequences for human.

Phosphorylation data

Phosphorylation sites were collected from PhosphoSitePlus (Hornbeck et al., 2015), dbPTM (Lu et al., 2013) and Phospho.ELM (Dinkel et al., 2011), in February 2016, resulting in 377 373 sites, after the duplicates were excluded. The phosphorylation dataset contains substrates for several kinases and different organisms. The intersection of the dataset of phosphosites and the human proteome included 13 254 proteins and 83 288 sites.

Human genetic variants

Human germinal variants were collected from Ensembl with ProteoFAV, which is described in Chapter 5. For this analysis, the protein sequences in UniProtKB and Ensembl databases were compared and proteins with different sequence length or mismatches were discarded. Three types of SNVs were collected, synonymous, missense, and nonsense, resulting in 442 780 S/Ts in 31 169 proteins in the human proteome.

POGSPSF class probability groups

Proteins or sites were grouped by the POGSPSF class probability score. For both cases, the score was divided into ten groups. Other numbers of groups were also tested but did not alter the interpretation of the results. Proteins were represented by their maximum score and ten score ranges were defined to keep the ten groups with the same number of proteins. Sites, also referred to as motifs, the 10 groups have equal width, so the number of motifs per group was used for normalisation to allow comparison. The difference between the two approaches is due to the difference of the two distributions. However, the interpretation is similar to either approach: the higher the quantile group to which a protein or motif belongs, the more likely it is to be modified by OGT according to the POGSPSF model. The 10 groups are referred as bins or quantiles.

4.2.2 GO term analysis

GO terms were investigated to verify differences regarding molecular function, biological process participation and cellular localisation among protein with top and low ranking POGSPSF scores. The analysis proceeded with GOATOOLS (Tang et al., 2015), a Python package with tools for GO analysis. The ontology file used was released on 14 December 2016 and the protein to gene associations for the human organism was validated on 20 of May 2016. Terms based on Inferred from Electronic Annotation (IEA) evidences were dropped to keep only human curated evidences. A total of 47 871 GO terms and 13 151 annotated proteins were compiled for the proteins in the human proteome.

The difference in GO term's composition was assessed over the 10 quantiles, each with 7 179 proteins. The GO term's composition was compared to the complete set of GO terms in the study. The reported p-value refers to the uncorrected Fisher's exact test p-value, implemented in Scipy. The advantage of Fisher's exact test over other statistical test, such as hypergeometric test, is the direct detection of over- and under-represented items. P-values for enriched and depleted terms were confirmed with Benjamini-Hochberg false discovery rate correction. The corrections is needed to increase the statistical power and avoid false discoveries from multiple statistical tests. A term was considered statistically significant if $p< 0.05$. The analysis significantly change terms in the ontology structure also considered clusters of groups (within 1 node of distance from each other) with uncorrected p-values. The Fisher's exact test odd ratio of selected terms was compared among quantiles to determine enrichment of terms for protein with and without high-scoring sites.

4.3 Results

4.3.1 Proteome prediction

The proteome-wide analysis provided predictions for 2 429 758 S/T residues, which belongs to 71 791 proteins. The default cutoff value for the random forest classifier, implemented in SciKit-Learn, is 0.5. Figure 4.1 shows the distribution of scores for potential sites in the human proteome. The rug plot (red) displays 111 known *O*-GlcNAc sites in human proteins. 99 of those sites have a POGSPSF class probability score > 0.6. Table 4.1 lists the known *O*-GlcNAc sites not predicted as modified by POGSPSF. Ser733 from the inhibitor of nuclear factor kappa-B kinase protein (UniProt O14920) has two Gln residues close to the modification site, which is uncommon for sites targeted by OGT, so the modification of this residue needs to be re-validated. Based on this distribution a class probability of 0.6 was used to define modified sites rather than the default, in order to limit overprediction.

Figure 4.1 shows the distribution of scores for the whole dataset as represented by their top scoring sites. The rug plot (red) illustrates 88 known *O*-GlcNAcylataded proteins. These modified proteins all have top scoring sites > 0.6 with one exception. The F-actin-capping protein subunit alpha-3 is a phosphoprotein expressed in the testes and sperm. The Thr2 residue is modified by OGT as identified by mass spectrometry (Wang et al., 2010). Mouse and rat homologous proteins have Ser instead of Thr in position 2. One report indicates the phosphorylation of Ser2 in the rat protein. This residue is not in the POGSPSF database and the issue is under investigation but the problem seems linked to the residue's close proximity to the protein N-terminus.

Table 4.1: Known *O*-GlcNAc sites predicted as unmodified by POGSPSF. Note that all but 5 motifs would pass with the default threshold of 0.5. Protein names: O00429, Dynamin-1-like protein; O14920, Inhibitor of nuclear factor kappa-B kinase subunit beta; P10636-8, Microtubule-associated protein Tau-4; P15586, N-acetylglucosamine-6-sulfatase; P68431, Histone H3.1; Q02818, Nucleoquantiledin-1; Q13492, Phosphatidylinositol-quantileding clathrin assembly protein; Q16566, Calcium/calmodulin-dependent protein kinase type IV; Q96JB8, MAGUK p55 subfamily member 4.

UniProt accession number	Position	Class probability
O00429	585	0.59
O14920	733	0.28
P10636-8	356	0.41
P15586	404	0.40
P68431	11	0.58
Q02818	47	0.42
Q13492	387	0.36
Q16566	57	0.54
	189	0.58
	344	0.58
	356	0.53
Q96JB8	436	0.59

Table 4.2: Summary of the number of significant GO terms per quantile. Proteins, number of protein per quantile; terms, number of terms per quantile; significant terms, number of significant corrected GO terms per quantile.

Quantile range	Proteins	Terms	Significant terms
(0.834, 0.951]	1 620	4 378	62
(0.794, 0.834]	1 560	4 507	21
(0.758, 0.794]	1 488	4 233	19
(0.722, 0.758]	1 496	4 399	5
(0.686, 0.722]	1 394	4 396	3
(0.648, 0.686]	1 332	4 266	1
(0.607, 0.648]	1 212	3 812	1
(0.56, 0.607]	1 084	3 447	14
(0.496, 0.56]	922	2 866	31
[0.0641, 0.496]	984	1 743	89

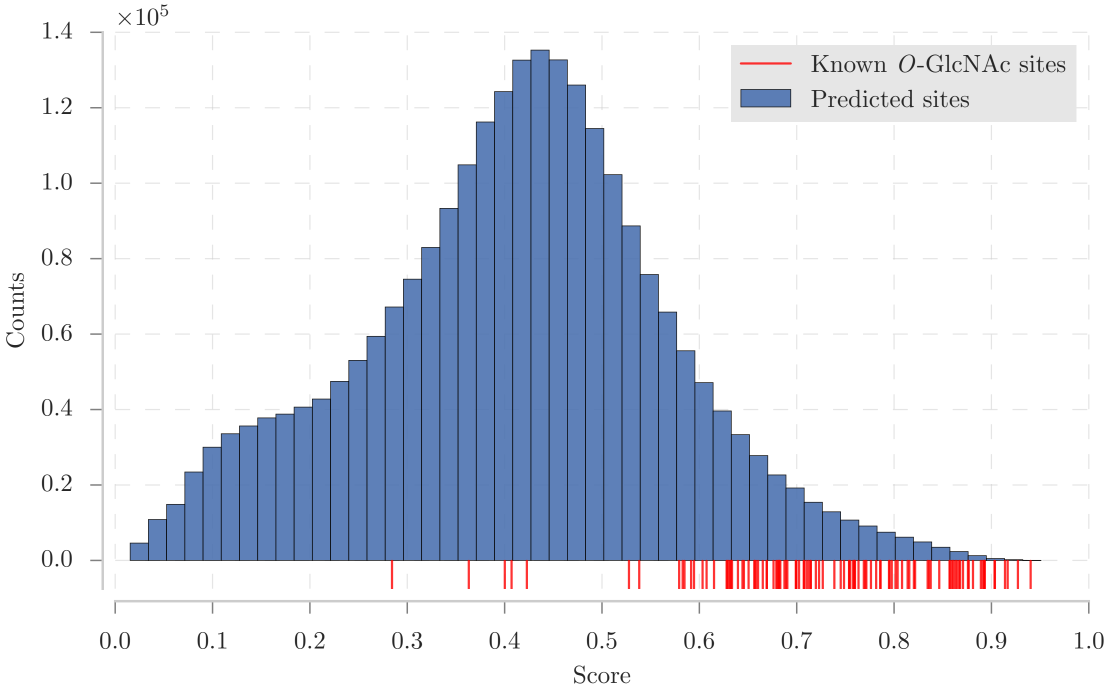


Figure 4.1: POGSPSF class probability distribution for 2 429 758 motifs in the human proteome. Know sites are represented by the rug plot in red. Most of the known sites have class probability score> 0.60 and exceptions are listed in Table 4.1.

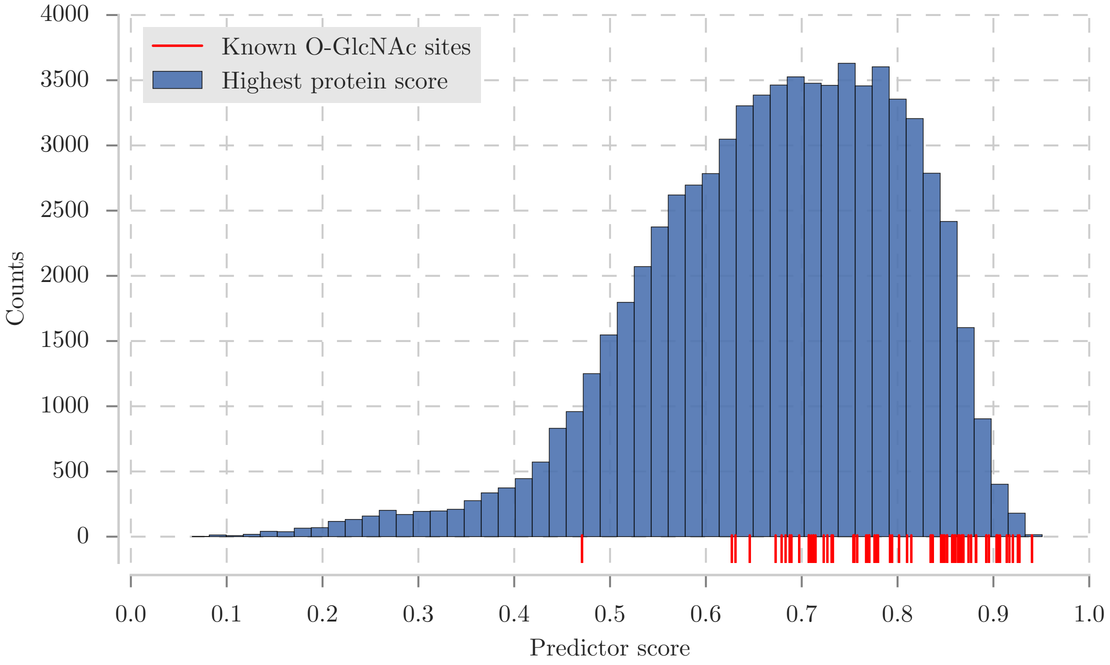


Figure 4.2: POGSPSF class probability score for 71 791 human proteins. The rug plot (red) shows 88 human proteins modified by OGT. Each protein's maximum motif score was selected to represent each protein.

4.3.2 GO term analysis

GO terms describe the functional attributes of genes. Analysis of the overrepresentation of terms on the GO ontology structures may reveal biological characteristics of a list of genes or proteins in large scale experiments.

Table 4.2 shows the number of significant terms per quantile. Interestingly, the number of statistically significant terms from the GO analysis was drastically higher in the two extremes of the distribution in Figure 4.2, suggesting that certain types of proteins were separated by the POGSPSF score. So, the investigation focused the top quantile, which ranges from 0.834 to 0.951 and contains the protein with high-scoring sites. In addition, the changes in the odd ration among quantiles was also analysed to determined the shift on representation among the POGSPSF score range.

Cellular component

Protein *O*-GlcNAcylation occurs in nuclear, cytoplasmic and mitochondrial proteins. However, the modification has also been found in proteins common to other subcellular locations, including the secretory pathway (Jochmann et al., 2014). Table 4.3 lists 19 cellular component GO term significantly changed on the quantile (0.834, 0.951]. All the terms were enriched. The list includes terms related to nucleus and cytoplasm, but also specific protein complexes such as RNA polymerase II transcription factor complex. However, terms not directly associated with protein *O*-GlcNAcylation have also detected: the external side of plasma membrane, immunoglobulin complex, lamellipodium and axon. These terms could relate to process yet to be associated with the modification or artefacts of the analysis. There is no specific mention of mitochondria-associated terms. The term external side of plasma membrane (GO:0009897) is also significantly enriched in the quantile, an unanticipated result since secreted proteins are not known to be modified by *O*-GlcNAcylation. Figure 4.4 illustrates the complete ontology structure for significantly enriched cellular component terms.

Figure 4.3 shows the change of the odds ratio among the quantiles for selected GO terms. Proteins annotated with the nucleus (GO:0005634) and cytoplasm (GO:00058860) terms are enriched in quantiles with high-scoring sites and depleted in quantiles with low-scoring sites. By contrast, proteins annotated with plasma membrane (GO:0005886) and extracellular region (GO:0005576) are not enriched nor depleted in the (0.834, 0.951] quantile, but significantly enriched in quantile containing low-scoring sites: $p< 0.01$ for integral component of membrane at [0.0641, 0.496]; $p< 0.01$ for extracellular region term at (0.607, 0.648]. The graph shows a clear trend for the decrease of the Fisher's exact odd ratio from top to bottom quantiles. However, there is no depletion of membrane proteins within the (0.834, 0.951] range, what is expected since some membrane proteins are modified by OGT. Thus, the analysis confirms the POGSPSF scoring conserves the nucleus/cytoplasm profile for proteins with potential sites.

Biological process

The biological process GO terms represent series of the events, such as cellular pathways. For example, the biological process GO terms can identify proteins that positively or negatively regulate the transcription of genes. Table 4.4 lists the significantly under- and overrepresented items in the namespace. Enriched terms

Table 4.3: Cellular component GO terms identified as significant on quantile with the top scoring proteins (POGSPSF score range from (0.834, 0.951]). All cellular component terms were enriched. p-value, corrected Fisher's exact test p-value; odds ratio, Fisher's exact test odds ratio, depth, the length of the longest path from the top term (Cellular component).

GO	Name	Depth	p-value	Odds ratio
GO:0005654	nucleoplasm	5	1.26×10^{-25}	1.76
GO:0005634	nucleus	5	1.54×10^{-23}	1.59
GO:0005737	cytoplasm	3	4.40×10^{-21}	1.57
GO:0015629	actin cytoskeleton	6	2.13×10^{-9}	2.93
GO:0005794	Golgi apparatus	5	6.93×10^{-8}	1.86
GO:0005813	centrosome	6	1.20×10^{-7}	2.11
GO:0005829	cytosol	4	1.42×10^{-7}	1.35
GO:0045111	intermediate filament cytoskeleton	6	3.26×10^{-6}	4.23
GO:0005882	intermediate filament	6	1.15×10^{-5}	4.80
GO:0005730	nucleolus	5	1.56×10^{-5}	1.62
GO:0009897	external side of plasma membrane	3	1.95×10^{-5}	2.76
GO:0030027	lamellipodium	3	2.12×10^{-5}	2.93
GO:0000922	spindle pole	5	1.15×10^{-4}	3.36
GO:0005925	focal adhesion	5	1.28×10^{-4}	1.79
GO:0042571	immunoglobulin complex, circulating	4	1.63×10^{-4}	4.48
GO:0043234	protein complex	2	1.67×10^{-4}	1.92
GO:0030424	axon	4	1.75×10^{-3}	2.80
GO:0005814	centriole	6	2.05×10^{-3}	3.18
GO:0090575	RNA polymerase II transcription factor complex	6	2.23×10^{-3}	3.60

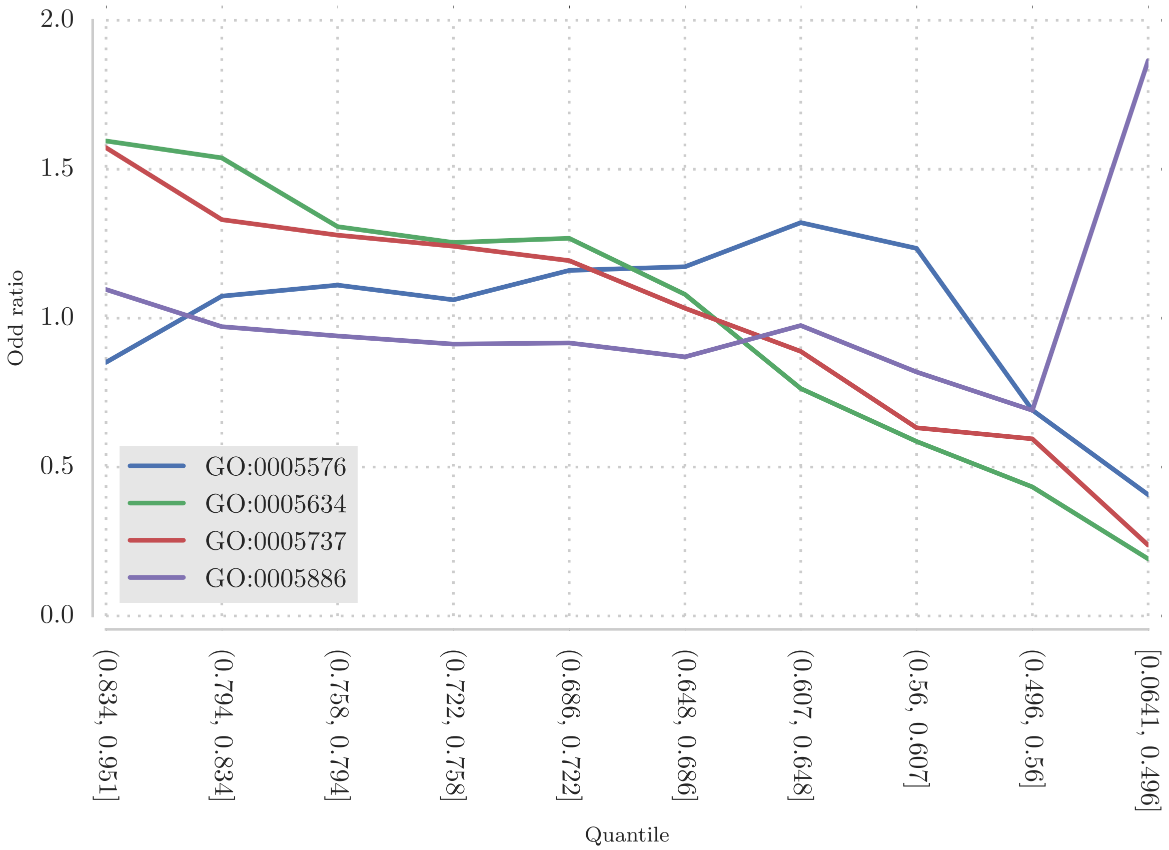


Figure 4.3: Trends of the Fisher's exact test odd ratio from cellular component terms among quantiles. The odd ratio for nucleus and cytoplasm terms decrease with the decrease of the score. Odds ratio enrichment peaks for the extracellular region and plasma membrane terms occurs in quantiles containing protein with low-scoring sites. GO:0005576, extracellular region; GO:0005634, nucleus; GO:0005737, cytoplasm; GO:0005886, plasma membrane.

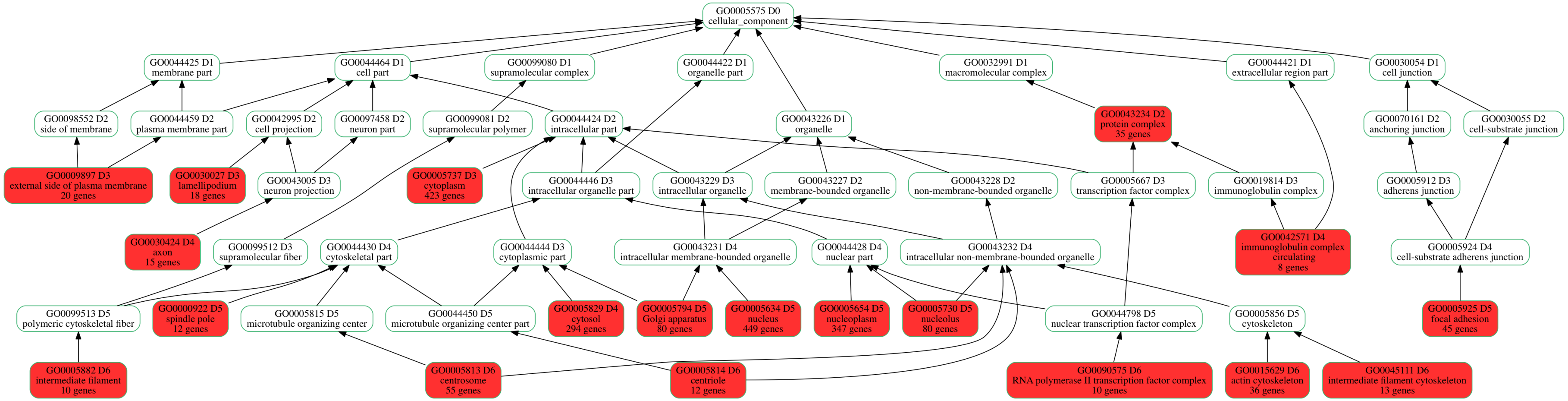


Figure 4.4: Complete ontology network for enriched and significant cellular component terms in the quantile with top scoring proteins. Red coloured nodes represented the statistically significant ones. D, the depth or the longest path until the cellular component term.

include positively and negatively regulation of the RNA polymerase II mediated transcription (GO:0045944 and GO:0000122), protein sumoylation (GO:0016925), negative regulation of phosphatase activity (GO:0010923) and negative regulation of apoptotic process (GO:0043066). Two terms related to membrane protein processes were depleted from the (0.834, 0.951] range: G-protein coupled receptor signaling pathway (GO:0007186) and sensory perception of smell (GO:0007608). The list also comprises terms that represent physiological process yet to be experimentally associated with protein *O*-GlcNAcylation modification.

Due to the great numbers of biological process GO terms, Figure 4.5 only illustrate the ontology structure for two terms: negative regulation of apoptotic process (GO:0043066) and protein sumoylation (GO:0016925). To the author's knowledge, protein *O*-GlcNAcylation has not previously been found controlling cell death processes; thus, because of the importance of the process, this term deserves further investigation. The two terms in the figure are distant from the biological_process node with the depth of 7 and 9 nodes, respectively. The majority of significantly changed terms were also distant from the biological_process node (depth $\geq$ 6 nodes), with exception of 7 other terms that are in intermediate range (depth between 4 and 6 nodes). From the intermediate depth, enriched biological process terms include glomerular filtration (GO:0003094), cell motility (GO:0048870) and synapse assembly (GO:0007416). Figure 4.6 illustrate trends on three biological process known to be associated with protein *O*-GlcNAcylation, intracellular signal transduction (GO:0035556), protein phosphorylation (GO:0006468), regulation of transcription, DNA-templated (GO:0006355).

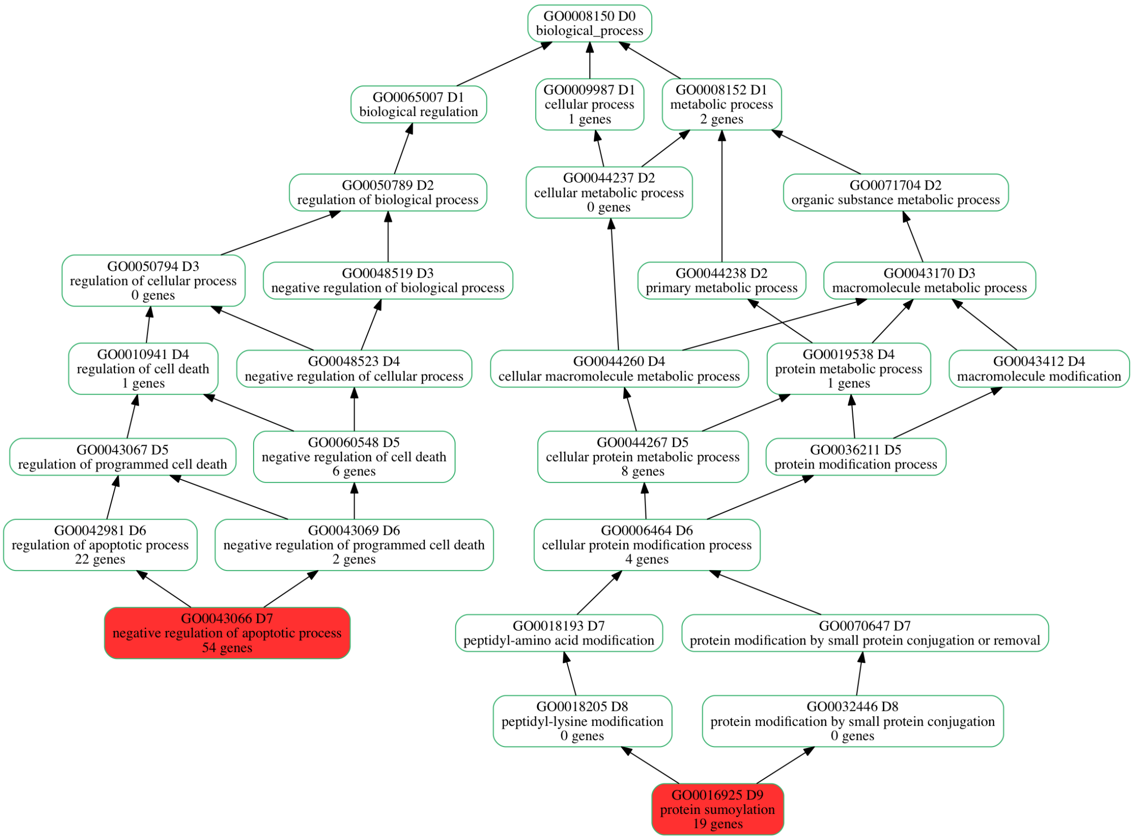

GO0008150 D0 biological_process
GO0065007 D1 biological regulation
GO0009987 D1 cellular process 1 genes
GO0008152 D1 metabolic process 2 genes
GO0050789 D2 regulation of biological process
GO0044237 D2 cellular metabolic process 0 genes
GO0071704 D2 organic substance metabolic process
GO0050794 D3 regulation of cellular process 0 genes
GO0048519 D3 negative regulation of biological process
GO0044238 D2 primary metabolic process
GO0043170 D3 macromolecule metabolic process
GO0010941 D4 regulation of cell death 1 genes
GO0048523 D4 negative regulation of cellular process
GO0044260 D4 cellular macromolecule metabolic process
GO0019538 D4 protein metabolic process 1 genes
GO0043412 D4 macromolecule modification
GO0043067 D5 regulation of programmed cell death
GO0060548 D5 negative regulation of cell death 6 genes
GO0044267 D5 cellular protein metabolic process 8 genes
GO0036211 D5 protein modification process
GO0042981 D6 regulation of apoptotic process 22 genes
GO0043069 D6 negative regulation of programmed cell death 2 genes
GO0006464 D6 cellular protein modification process 4 genes
GO0043066 D7 negative regulation of apoptotic process 54 genes
GO0018193 D7 peptidyl-amino acid modification
GO0070647 D7 protein modification by small protein conjugation or removal
GO0018205 D8 peptidyl-lysine modification 0 genes
GO0032446 D8 protein modification by small protein conjugation 0 genes
GO0016925 D9 protein sumoylation 19 genes

Figure 4.5: Selected statistically significant biological process terms of proteins in the (0.834, 0.951] quantile. Proteins with high-scoring sites belongs to this quantile. Red coloured nodes represented the statistically significant ones. D, the depth or the longest path until the top term (biological_process).

Table 4.4: Biological process GO terms identified as significant on quantile with the top scoring proteins (POGSPSF score range from (0.834, 0.951]). p-value, corrected Fisher's exact test p-value; Odds ratio, Fisher's exact test odds ratio, depth, the length of the longest path to the top term (Biological process).

GO	Name	Depth	p-value	Odds ratio
GO:0045944	positive regulation of transcription from RNA polymerase II promoter	11	5.20×10^{-17}	2.28
GO:0007186	G-protein coupled receptor signaling pathway	5	1.35×10^{-13}	0.18
GO:0045893	positive regulation of transcription, DNA-templated	10	1.20×10^{-7}	1.97
GO:0043066	negative regulation of apoptotic process	7	2.62×10^{-7}	2.06
GO:0045892	negative regulation of transcription, DNA-templated	10	3.40×10^{-7}	1.96
GO:0007416	synapse assembly	4	4.52×10^{-7}	4.55
GO:0000122	negative regulation of transcription from RNA polymerase II promoter	11	6.72×10^{-7}	1.87
GO:0006910	phagocytosis, recognition	4	3.93×10^{-6}	4.82
GO:0006366	transcription from RNA polymerase II promoter	10	4.04×10^{-6}	2.17
GO:0050871	positive regulation of B cell activation	8	6.60×10^{-6}	4.62
GO:0016339	calcium-dependent cell-cell adhesion via plasma membrane cell adhesion molecules	5	2.16×10^{-5}	6.42
GO:0007608	sensory perception of smell	6	2.97×10^{-5}	0.08
GO:0010923	negative regulation of phosphatase activity	9	3.90×10^{-5}	3.96
GO:0006911	phagocytosis, engulfment	6	3.90×10^{-5}	3.96
GO:0016925	protein sumoylation	9	4.45×10^{-5}	2.70
GO:0048870	cell motility	4	4.74×10^{-5}	5.88
GO:0030490	maturation of SSU-rRNA	10	4.74×10^{-5}	5.88
GO:0034332	adherens junction organization	5	1.37×10^{-4}	4.13
GO:0003094	glomerular filtration	5	1.69×10^{-4}	7.20
GO:0006355	regulation of transcription, DNA-templated	9	1.97×10^{-4}	1.61

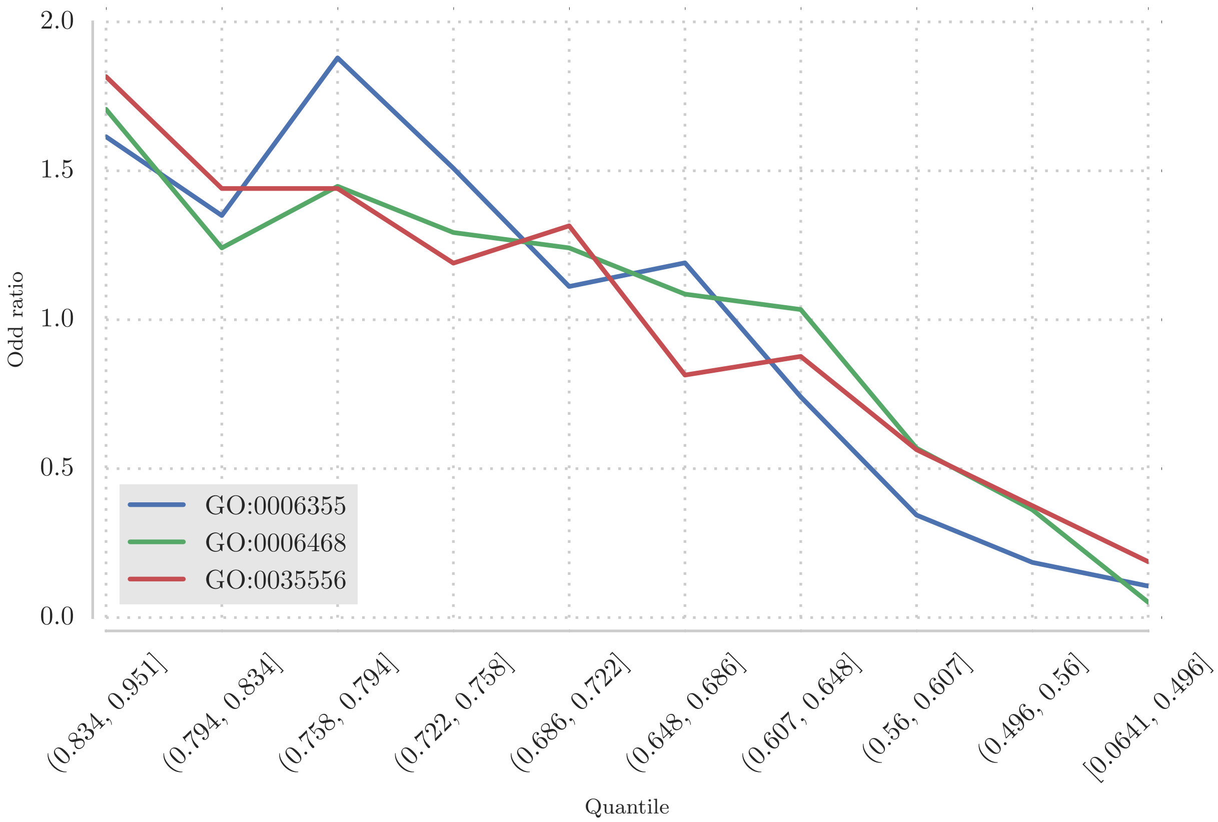


Figure 4.6: Trends of the Fisher's exact test odd ratio from biological process terms among quantiles. The three terms - intracellular signal, protein phosphorylation and regulation of transcription, DNA-templated - are enriched on proteins predicted to be modified. GO:0035556, intracellular signal transduction; GO:0006468, protein phosphorylation; GO:0006355, regulation of transcription, DNA-templated.

Molecular function

The molecular function terms describe the fundamental actions of gene products at the molecular level. These terms identify actions, such as chemical reactions catalysed by enzymes or protein-protein binding. Table 4.5 lists the statistically significant molecular function GO terms of proteins in the (0.834, 0.951] range. Only two terms are depleted, olfactory receptor activity (GO:0004984) and odorant binding (GO:0005549) in the analysed quantile, with 0 proteins annotated with such terms in the quantile. The vast majority of the enriched terms are related to regulation of the transcriptional activity mediated by RNA polymerase II and DNA binding. Other terms, which are relevant in the protein *O*-GlcNAcylation context are protein kinase binding (GO:0019901), structural constituent of cytoskeleton (GO:0005200) and chromatin DNA binding (GO:0031490). So far, the impact of protein *O*-GlcNAcylation on protein-protein and protein-ligand interaction has been neglected, but for a few exceptions (Roos et al., 1997). The enrichment of the immunoglobulin receptor binding (GO:0034987) term was unexpected, which was further confirmed by manual curation described in Section 4.3.5.

4.3.3 Analysis of phosphorylation in residues predicted as modified

Extensive cross-talk between protein phosphorylation and *O*-GlcNAcylation has been reported (Griffith and Schmitz, 1999; Wang et al., 2008; Dias et al., 2009). Recent mass-spectrometry studies show that the interplay between the two modifications might be more extensive than was thought before. The global levels of

Table 4.5: Molecular function GO terms identified as significant on quantile with the top scoring proteins (POGSPSF score range from (0.834, 0.951]). p-value, corrected Fisher's exact test p-value; Odds ratio, Fisher's exact test odds ratio, depth, the length of the longest path from the top term (molecular function).

GO	Name	Depth	p-value	Odds ratio
GO:0005515	protein binding	2	1.37×10^{-38}	1.48
GO:0003700	transcription factor activity, sequence-specific DNA binding	2	9.61×10^{-14}	2.05
GO:0004984	olfactory receptor activity	5	4.19×10^{-13}	0
GO:0043565	sequence-specific DNA binding	5	5.14×10^{-10}	2.69
GO:0000978	RNA polymerase II core promoter proximal region sequence-specific DNA binding	9	2.09×10^{-9}	2.46
GO:0000981	RNA polymerase II transcription factor activity, sequence-specific DNA binding	3	2.13×10^{-9}	2.93
GO:0044822	poly(A) RNA binding	5	1.16×10^{-8}	1.63
GO:0001077	transcriptional activator activity, RNA polymerase II core promoter proximal region sequence-specific binding	5	2.14×10^{-8}	2.84
GO:0019901	protein kinase binding	5	5.72×10^{-7}	2.22
GO:0097110	scaffold protein binding	3	3.62×10^{-6}	5.31
GO:0044212	transcription regulatory region DNA binding	6	1.04×10^{-5}	2.25
GO:0019899	enzyme binding	3	1.65×10^{-5}	1.98
GO:0042802	identical protein binding	3	3.06×10^{-5}	1.68
GO:0003677	DNA binding	4	3.07×10^{-5}	1.73
GO:0034987	immunoglobulin receptor binding	4	4.82×10^{-5}	4.20
GO:0008134	transcription factor binding	3	7.74×10^{-5}	2.03
GO:0005200	structural constituent of cytoskeleton	2	1.09×10^{-4}	2.91
GO:0005549	odorant binding	2	1.42×10^{-4}	0
GO:0031267	small GTPase binding	5	1.69×10^{-4}	7.20
GO:0003723	RNA binding	4	1.80×10^{-4}	1.79
GO:0046982	protein heterodimerization activity	4	1.90×10^{-4}	1.95
GO:0031490	chromatin DNA binding	6	2.23×10^{-4}	3.60

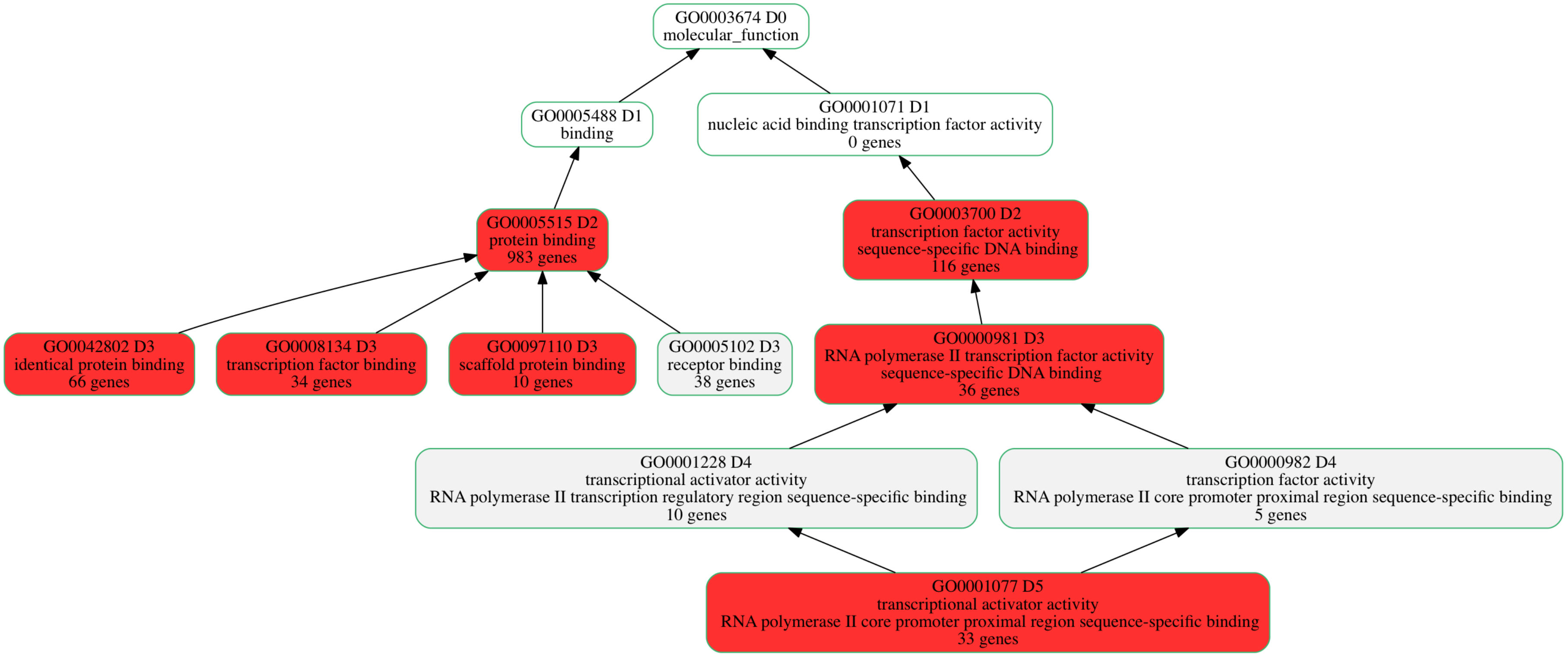

GO0003674 D0
molecular_function
GO0005488 D1
binding
GO0001071 D1
nucleic acid binding transcription factor activity
0 genes
GO0005515 D2
protein binding
983 genes
GO0003700 D2
transcription factor activity
sequence-specific DNA binding
116 genes
GO0042802 D3
identical protein binding
66 genes
GO0008134 D3
transcription factor binding
34 genes
GO0097110 D3
scaffold protein binding
10 genes
GO0005102 D3
receptor binding
38 genes
GO0000981 D3
RNA polymerase II transcription factor activity
sequence-specific DNA binding
36 genes
GO0001228 D4
transcriptional activator activity
RNA polymerase II transcription regulatory region sequence-specific binding
10 genes
GO0000982 D4
transcription factor activity
RNA polymerase II core promoter proximal region sequence-specific binding
5 genes
GO0001077 D5
transcriptional activator activity
RNA polymerase II core promoter proximal region sequence-specific binding
33 genes

Figure 4.7: Selected statistically significant molecular function terms in the quantile with top scoring proteins. Red coloured nodes represented the statistically significant ones. D, the depth of the longest path until the molecular_function term.

phosphorylation and *O*-GlcNAcylation appear to be associated (Hart et al., 2011; Wang et al., 2012; Bullen et al., 2014; Copeland et al., 2008). The interplay between phosphorylation and *O*-GlcNAcylation was analysed on the data used to train POGSPSF. Figure 4.8 shows the differences between the fraction of phosphosites around *O*-GlcNAc sites and unmodified S/Ts. If the modified and unmodified S/Ts are aligned, more phosphosites are observed at the modified group. The modified S/T is in the centre (0), where the mean fraction of phosphosites is 20% (95% CI from 18% to 23%) for modified and 14% (95% CI from 14% to 15%) for unmodified S/T. The mean values and CI were calculated from bootstrapping the dataset 1 000 times. Interestingly, other positions in proximity to the *O*-GlcNAc-site (±10 residues) also have increased fractions of phosphosites: sites -3 and -4 and the range from +1 to +15. So, in these datasets, the fraction of phosphosites is increased around the position target by OGT, including the target S/T.

Figure 4.9 shows the distribution of POGSPSF scores for phosphorylated and other S/Ts in the proteome. Although the number of sites in each group is different, no difference was observed in kernel densities, indicating the classifier does not have a bias for phosphosites. However, modified S/T (POGSPSF score$>$ 0.6) are 1.15 times more likely to be phosphorylated (two-sided Fisher's exact test p$<$ 0.01). To explore this result the motifs were grouped in 10 quantiles of equal width by their POGSPSF class probability score and the fraction of motifs with phosphosites was plotted. Figure 4.10 and Table 4.6 summarise the relationship between the fraction of phosphosites and the POGSPSF score. There is a discernible positive trend between the two variables. The trend comprises two components. A strong increase over the 3 first quantiles, from 0 to 0.3, and weaker component over [0.5, 1]. Due to the

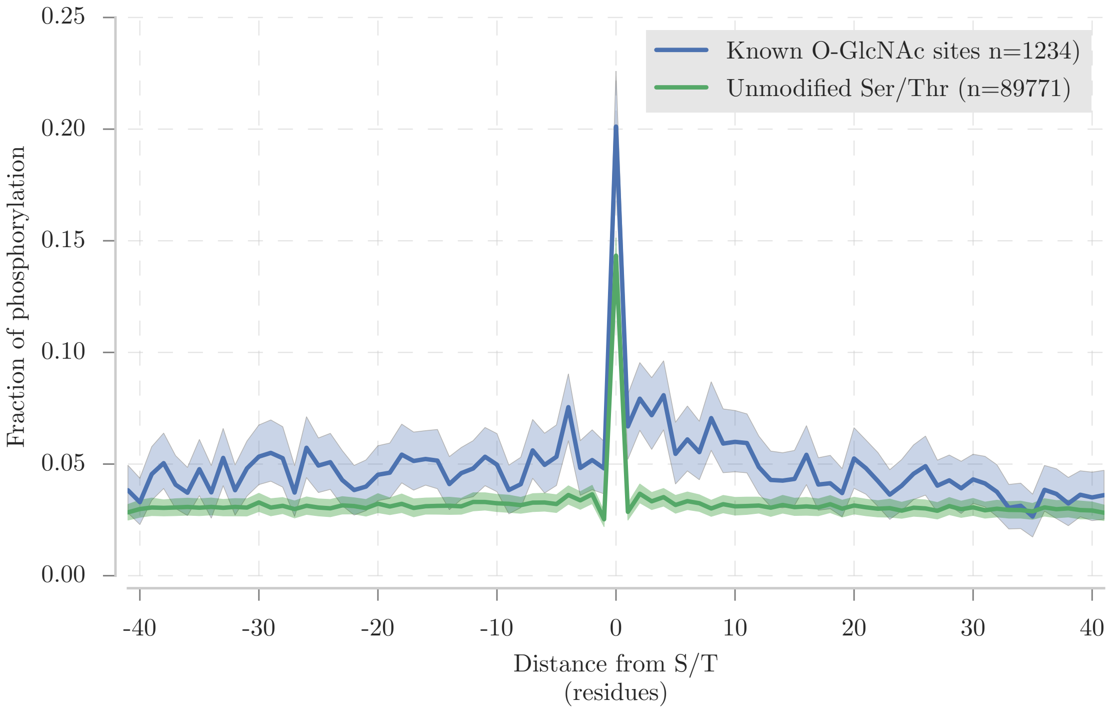


Figure 4.8: Comparison of the fraction phosphosites around *O*-GlcNAc sites and unmodified S/T of *O*-GlcNAcylated proteins. Semi-transparent colours represent the 95% CI obtained from bootstrapping the each dataset 1 000 times. 20% (95% CI from 18% to 23%) of the *O*-GlcNAc sites are phosphosites, while 14% (95% CI from 14% to 15%) of the unmodified S/T are phosphosites.

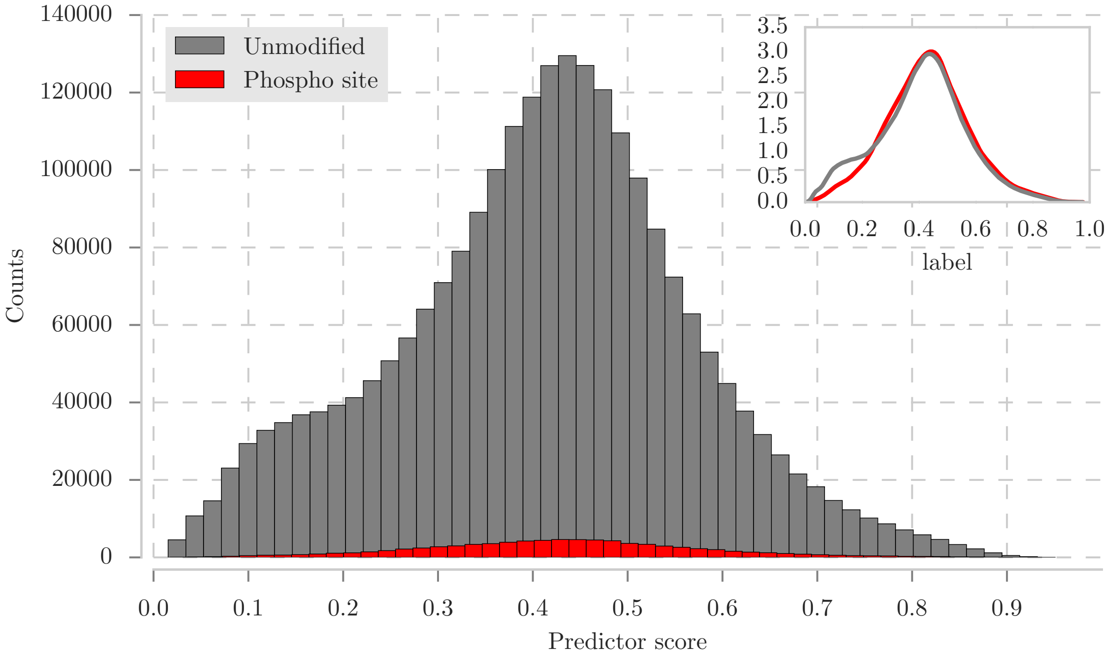


Figure 4.9: POGSPSF motif score distribution for phosphosites and S/T not targeted by kinases. The inset shows the kernel density distribution for the two quantities, which are practically equivalent.

reduced number of motifs in the two last quantiles ((0.8, 0.9]; (0.9, 1]), the CI is wider.

4.3.4 Analysis of genetic variants for residues predicted as modified

Germinal mutations were collected from Ensembl variation (Ensembl variation website, 2016). Three types of variants were selected and the number of SNVs mapped to S/Ts in the dataset are as follows: 266 964 missense; 171 712 synonymous; and nonsense 171 712. The odds-ratio for the co-occurrence of SNVs in S/Ts classified as modified in relation to unmodified is 0.93 for missense, 0.94 for nonsense and 0.92 for synonymous, two-sided Fisher's exact test $p< 0.01$, 0.22, < 0.01, respectively. So the test determines that residues predicted as modified by POGSPSF are significantly less likely to co-occur with known human missense and synonymous mutations, indicating these sites are less tolerant to variation.

The fraction of motifs with SNVs, in each POGSPSF score bin were investigated to better understand the relationship between the score and the co-occurrence of the mutations. Figure 4.11 shows the fraction of motifs with SNVs per POGSPSF score bin. Bins with high-scoring motifs have a lower fraction for the three types of variants. Note the discontinuity in the y-axis, since the number of nonsense variants is small. Also, as shown in Table 4.7, there are 0 nonsense variants in the [0.9, 1] bin, so the result should be interpreted with care.

As described in Section 1.7, only a small percentage of the SNVs may have an impact on protein function. Since Ensembl conveniently provides SIFT and

Table 4.6: Number of phosphosites in the proteome grouped by the POGSPSF class probability score.

Bins	**Number of phosphosites**	**Motifs per quantile**	**Proportion** (95% CI)
[0.0, 0.1]	851	68598	0.01 (0.01, 0.01)
(0.1, 0.2]	4044	194551	0.02 (0.02, 0.02)
(0.2, 0.3]	10169	288304	0.04 (0.03, 0.04)
(0.3, 0.4]	19170	520435	0.04 (0.04, 0.04)
(0.4, 0.5]	24141	681044	0.04 (0.04, 0.04)
(0.5, 0.6]	15123	408236	0.04 (0.04, 0.04)
(0.6, 0.7]	6468	168262	0.04 (0.04, 0.04)
(0.7, 0.8]	2445	61993	0.04 (0.04, 0.04)
(0.8, 0.9]	797	18623	0.04 (0.04, 0.05)
(0.9, 1.0]	25	525	0.05 (0.03, 0.07)

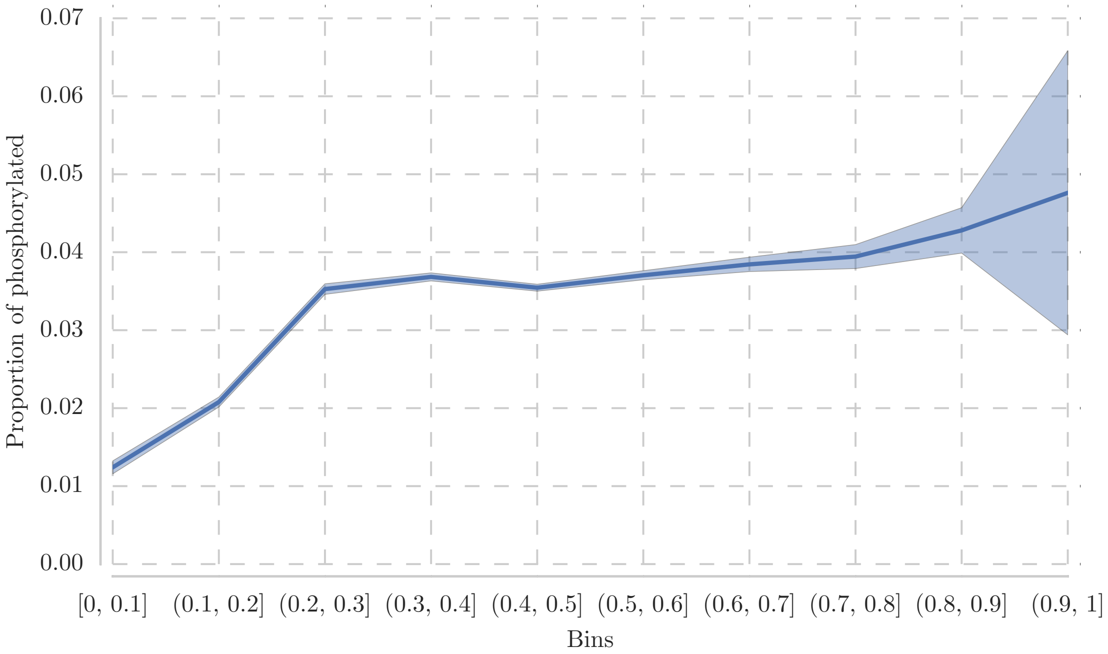


Figure 4.10: Fraction of S/T that are phosphorylated per bin of POGSPSF score quantile. The x-axis shows the 10 quantiles of equal-width. The y-axis shows the fraction of S/Ts that are known targets for kinases. Semi-transparent colour represents the 95% CI of the fraction for each quantile, calculated from the binomial proportion with StatsModel. There is a small increase of the fraction of phosphosites with the increasing of the score, which confirms the result that predicted modified sites are phosphorylated than the unmodified ones.

PolyPhen-2 scores, these two SNV impact classifiers were studied. SNVs classified as damaging by SIFT (score<0.05) and PolyPhen-2 (score>0.5) decrease for the bins with high-scoring motifs, with higher likelihood of being modified by OGT, while SNVs classified as benign by the two tools remain constant over the POGSPSF score bins. Surprisingly, this indicates that the *O*-GlcNAc sites are protected from genetic variation. Alternatively, the classifier may be able to incorporate structural and phylogenetic information not directly encoded in the training set.

4.3.5 Curated proteins with top ranking sites

The 3 000 proteins with top scoring sites were manually curated to highlight potential new OGT targets. The list included proteins identified in UniProt as 'fragments', which may represent an artefact of the proteome-wide prediction. If the protein sequence is too short, the secondary structure and disorder predictors might classify proteins as mostly disordered. Also, several of protein fragments have not been derived from experimental evidence of the protein, but from evidence of a transcript or inferred from homology. The list is also enriched in transcription factors and proteins containing zinc-finger domains, which are known OGT targets. The examination focuses on protein kinases and other enzymes involved in PTMs, interesting targets; secreted and membrane proteins, which should not be modified; and other potential targets involved in processes that were not previously associated with OGT. Protein fragments and proteins without information in UniProt were ignored.

Table 4.7: Number of germinal genetic variants from Ensembl per POGSPSF score bin. Each column shows the counts for a SNV type within a score bin. The Motifs columns has the total number of motifs for a given bin.

Bins	Missense	Nonsense	Synonymous	Motifs
[0.0, 0.1]	9573	136	6618	72349
(0.1, 0.2]	24485	363	16422	204570
(0.2, 0.3]	32971	510	21308	302171
(0.3, 0.4]	56929	842	36210	544638
(0.4, 0.5]	73631	1155	46716	711851
(0.5, 0.6]	43438	698	27907	426652
(0.6, 0.7]	17650	286	11270	176108
(0.7, 0.8]	6391	91	3978	64755
(0.8, 0.9]	1847	23	1252	19482
(0.9, 1.0]	49	0	31	545

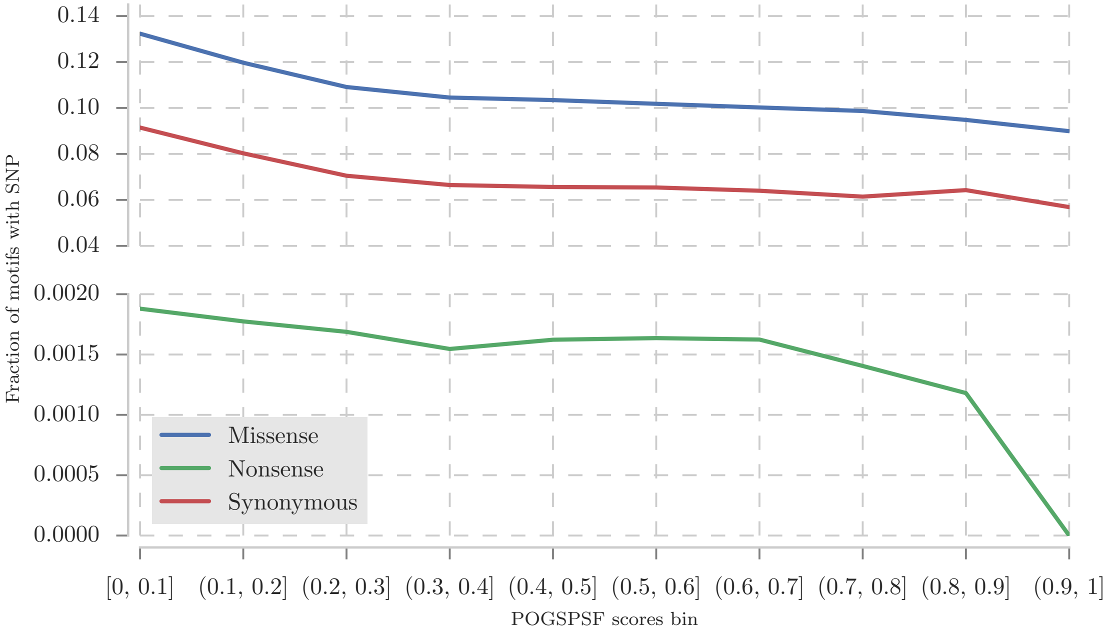


Figure 4.11: Fraction of motifs with missense, nonsense, and synonymous SNV in S/T in different POGSPSF score bins. The y-axis shows the fraction of S/T affected by the three types of genetic variants; and the x-axis shows the POGSPSF score bins. Overall, the fraction of SNVs tends to decrease with the increasing POGSPSF score bins, which denotes higher likelihood to a motif been targeted by OGT. Note the y-axis is discontinuous.

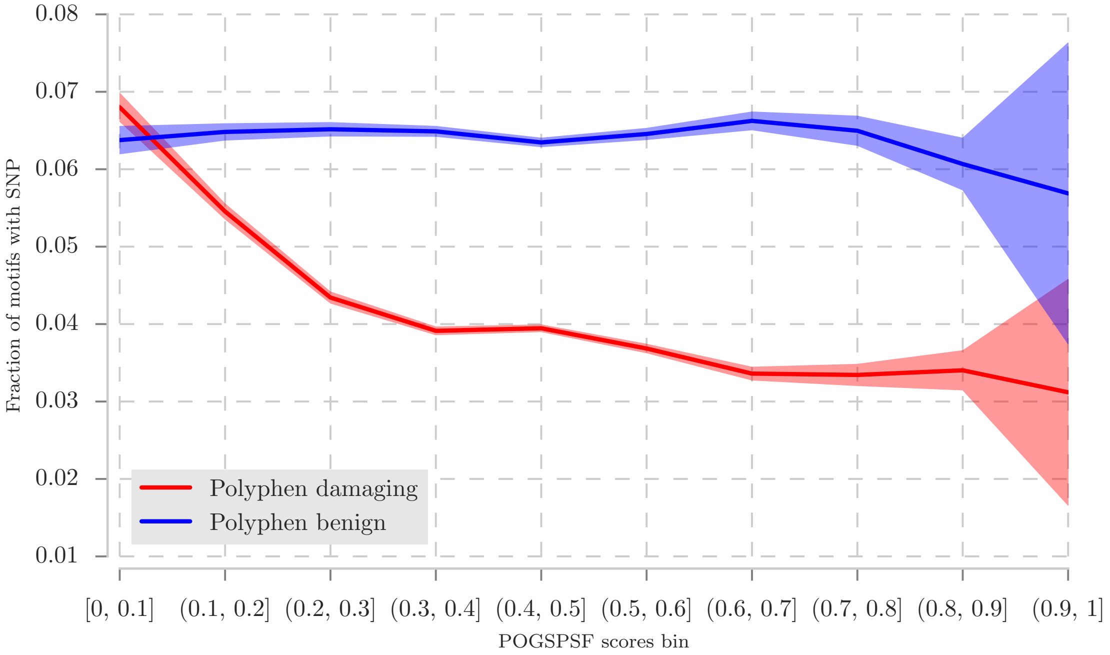


Figure 4.12: Fraction of SNVs with impact classified by PolyPhen-2. The fraction of SNVs classified by PolyPhen-2 as damaging (score ≥ 0.5) and as benign (score> 0.5) have different relationship with POGSPSF score. While the number of SNVs classified as benign remains constant over the bins, the proportion of damaging SNVs decrease. Semi-transparent colours represent the 95% binomial proportion CI of the values.

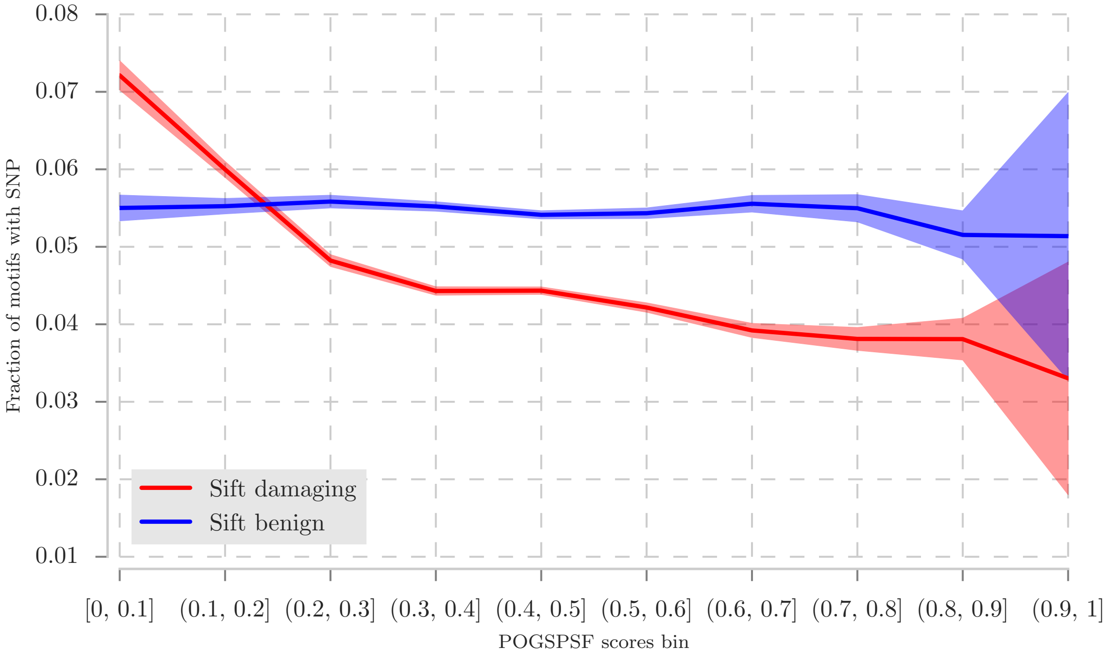


Figure 4.13: Fraction of SNV with impact classified by SIFT. Similar to the result above, the fraction of SNV classified as damaging by SIFT (score ≤ 0.05) reduce with the increasing score.

Histone deacetylase

Histone deacetylase complex subunit SAP130 (UniProt Q9H0E3) participates in the mSin3A complex, which acts as a corepressor of transcription. This protein is not only interesting as a biological target, but also an example of overprediction, since POGSPSF predicts 26 potential sites out of 99 S/Ts.

Kinase associated

The INCA1 protein (UniProtKB Q0VD86) has 7 S/Ts with POGSPSF class probability score> 0.6, none of which are known phosphosites. This protein negatively regulates CDK activity by binding. Moreover, the protein has an extensive list of protein-protein binding partners and so it may be a possible target for testing whether OGT modification mediates protein-protein interactions.

The Megakaryocyte-associated tyrosine-protein kinase (UniProt P42679; in POGSPSF with the obsolete accession number A0A087WUR1) caries high-scoring residues in the extremities of its SH3 domain. It is notable that the potential sites do not occur within the domain, but in its extremities.

Subunit alpha-1 of the 5′-AMP-activated protein kinase (UniProt Q13131) has 6 high-scoring residues, 4 of which are known phosphosites. The protein is extensively modified by phosphorylation, ubiquitination and acetylation. OGT and AMP-kinase regulate each other (Cheung and Hart, 2008; Bullen et al., 2014), but at this time, the alpha-1 subunit of the AMP kinase has no mapped *O*-GlcNAc sites.

Associated with nucleic acids

The Zinc finger protein 286A (UniProtKB J3KSW0) contains three potential sites in its KRAB domain. This domain is enriched in charged amino acids and folds in two α-helices. If the sites are truly modified, the modification may disrupt the fold and therefore transcription regulation.

Death-inducer obliterator 1 is a putative transcription factor with pro-apoptotic activity. The protein is extensively modified by phosphorylation and POGSPSF predicted 15 residues as potential new *O*-GlcNAc sites in the isoform DIDO1 (Q9BTC0-2). The DNA polymerase iota (UniProtKB J3KSW2) has 12 potential sites and is example of possible new biological association for OGT since the enzyme involvement in replication has not been previously reported.

Ubiquitination

Ubiquitination is an important post-translational process that can lead to protein degradation. The process involves the modifier, the protein ubiquitin, and three enzymes: E1 (ubiquitin-activating); E2 (ubiquitin-conjugating); and E3 (ubiquitin ligase). Like protein phosphorylation, ubiquitination cross-talks with *O*-GlcNAc. As reviewed by Ruan et al. (2013), the two PTMs have an antagonistic effect: while ubiquitination leads to degradation via the proteasome, protein *O*-GlcNAcylation increases protein stability. OGT is known for targeting ubiquitin itself, and various isoforms of the E3 enzyme.

POGSPSF predicts 14 potential modification sites in E3 ubiquitin-protein ligase Mdm2 (UniProt Q00987), 5 of which are known phosphosites. Additional E3 ubiquitin ligases were observed in the list of proteins with high-scoring sites. RNF43 (UniProt

Q68DV7) contains 12 high-scoring sites, none of which are phosphosites. Interestingly, this protein is a single-pass membrane protein, and the 12 high-scoring sites reside in the cytoplasmic portion of the protein. The E3 ubiquitin-protein ligase makorin-1 protein contains 10 sites with POGSPSF class probability score> 0.6, including a PVSAA site from residue 142 to residue 146. Ubiquitin-conjugating enzyme E2 Q2 (UniProt Q8WVN8) contains 4 potential sites. The sites are close to the first annotated secondary structure element in UniProtKB.

Secreted or membrane-bound proteins

Proteins in the secretory pathway and integral components of the membrane should not be modified by OGT. However, Jochmann et al. (2014) identified proteins with known sites in the secretory lumen. Also, some membrane bound proteins contain cytoplasmic portions or loops and are often regulated by kinases (Valverde et al., 2011) and OGT, for example the Sarcoplasmic/endoplasmic reticulum calcium ATPase 1 (UniProt Q8R429). Thus, this class of proteins was also analysed. Also, if a clear pattern is detected on the sequence of such proteins, the pattern could be applied to the next version of POGSPSF to avoid those sites yielding high scores.

The Signal-regulatory protein beta-1 (UniProtKB H3BQ21) is an integral component of the membrane with 10 high-scoring sites, among 36 possible S/Ts. The 10 sites are located in the enzyme's C-terminus, a region containing an immunoglobulin-like fold, which in general mediates interaction with other proteins with the same fold.

Another protein involved in antigen binding has one of the top scoring sites, the Ig kappa chain V-ID region 16. Interestingly, PhosphoSitePlus reports phospho- and acetylation sites for the region facing the extracellular side. Within the proteins with top scoring sites 35 other proteins contained the ‘immunoglobulin’ term and other proteins, such as Interleukin 18 binding protein (UniProt G3V1C5) that contain an immunoglobulin-like domain. To this author’s knowledge, OGT does not target secreted proteins or proteins with an immunoglobulin-like fold. Despite protein *O*-GlcNAcylation being discovered 30 years ago in immune cells, the modification’s role in immunity was only recently revealed. OGT is essential for response to infection mediated by T cells (Swamy et al., 2016) and also regulates the innate response to pathogens (Ryu and Do, 2011).

The probable palmitoyltransferase ZDHHC20 (UniProt Q5W0Z9) is a multi-pass membrane protein that catalyses the protein palmitoylation process, the addition of the palmitoyl group to cysteines. POGSPSF predicts two residues, 307 and 328 as potential *O*-GlcNAc sites. The sites are close to known phosphosites, Ser305 and Ser330. Other palmitoyltransferases were identified in the proteins with top ranking sites; for example, Palmitoyltransferase ZDHHC5 (UniProt Q9C0B5) and the probable palmitoyltransferase ZDHHC1 (Q8WTX9).

Mitochondrial

The Acyl-CoA synthetase family member 2 protein (UniProt Q96CM8) catalyses the metabolism of fatty acids in the mitochondria. The protein contains 11 phosphosites and 2 residues which have high POGSPSF class probability, neither of which are phosphosites. The top scoring site is a `PVTXX` site that is refered to as the ‘canonical motif’ for OGT in the literature.

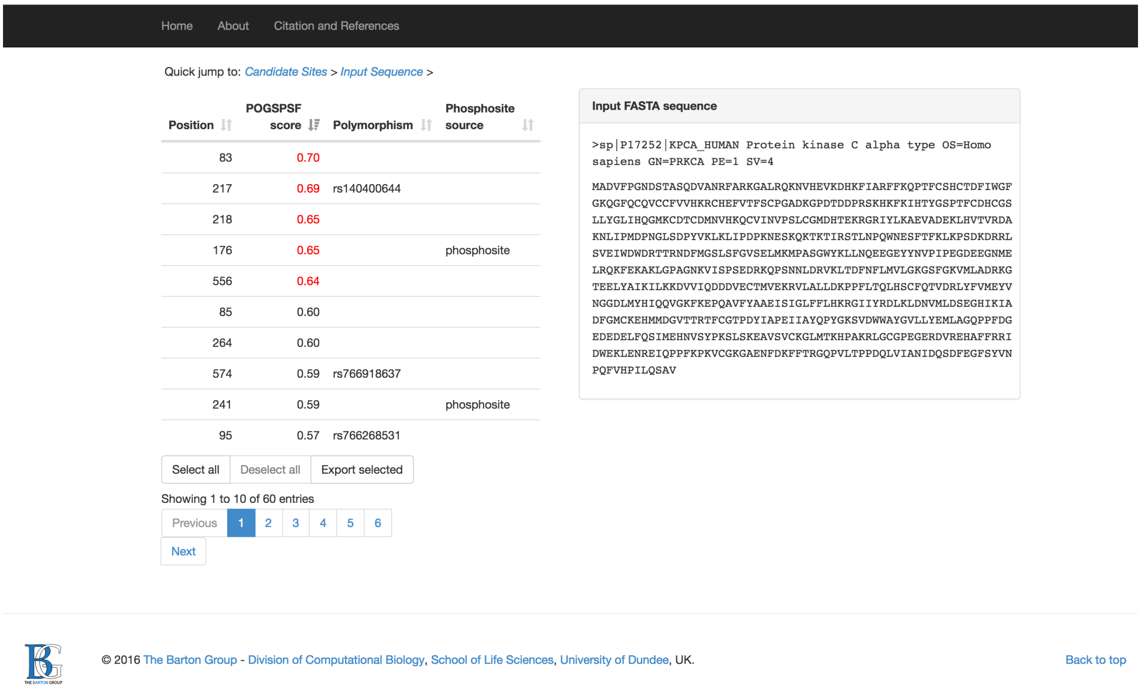
Bcl-2-binding component 3 protein (UniProt Q9BXH1) is critical to the apoptotic signalling pathway. Residues Thr69 and Ser98 are predicted as potential *O*-GlcNAc sites by POGSPSF. Although protein *O*-GlcNAcylation has been associated with the apoptotic process (Liu et al., 2000), the molecular mechanisms for the association have yet to be identified.

4.3.6 Web application

Nowadays, computational classifiers of PTMs are a helpful set of tools for lab researchers and computational biologists. The tools can be used to rank potential sites or to provide further evidence when validation methods are not available (Valverde et al., 2011; Cardoso et al., 2014). Nevertheless, the tools' application is only possible if the software is accessible to users.

Software developers can provide access to software via a standalone application, a web server/application or a web service with programmatic access. The standalone application requires installation and, generally, targets users trying to replicate the study or make large scale predictions. Users studying only a few specific proteins demand access in an easy way; this is best achieved within a web application. Some users may prefer to access results via a web service or API, so they can programmatically retrieve results for method comparison or to develop a meta-classifier. Meta-classifiers are important in machine learning, applying a jury-based approach and its variations to achieve more reliable prediction based on multiple predictors. Often, meta-classifier perform better than single decision classifiers (Madeira et al., 2015).

The POGSPSF classifications can be accessed from `www.compbio.dundee.ac.uk/pogspsf`. The website was implemented with Flask, SQLite, Bootstrap and



Quick jump to: *Candidate Sites* > *Input Sequence* >

Position	POGSPSF score	Polymorphism	Phosphosite source
83	0.70		
217	0.69	rs140400644	
218	0.65		
176	0.65		phosphosite
556	0.64		
85	0.60		
264	0.60		
574	0.59	rs766918637	
241	0.59		phosphosite
95	0.57	rs766268531	

Showing 1 to 10 of 60 entries

Input FASTA sequence

```
>sp|P17252|KPCA_HUMAN Protein kinase C alpha type OS=Homo
sapiens GN=PRKCA PE=1 SV=4

MADVFPGNDSTASQDVANRFARKGALRQKNVHEVKDHKFIARFFKQPTFCSHCTDFIWGF
GKQGFQCQVCCFVVHKRCHEFVTFSCPGADKGPDTDDPRSKHKFKIHTYGSPTFCDHCGS
LLYGLIHQGMKCDTCDMNVHKQCVINVPSLCGMDHTEKRGRIYLKAEVADEKLHVTVRDA
KNLIPMDPNGLSDPYVKLKLIPDPKNESKQKTKTIRSTLNPQWNESFTFKLKPSDKDRRL
SVEIWDWDRTTRNDFMGSLSFGVSELMKMPASGWYKLLNQEEGEYYNVPIPEGDEEGNME
LRQKFEKAKLGPAGNKVISPSEDRKQPSNNLDRVKLTDFNFLMVLGKGSFGKVMLADRKG
TEELYAIKILKKDVVIQDDDVECTMVEKRVLALLDKPPFLTQLHSCFQTVDRLYFVMEYV
NGGDLMYHIQQVGKFKEPQAVFYAAEISIGLFFLHKRGIIYRDLKLDNVMLDSEGHIKIA
DFGMCKEHMMDGVTTRTFCGTPDYIAPEIIAYQPYGKSVDWWAYGVLLYEMLAGQPPFDG
EDEDELFQSIMEHNVSYPKSLSKEAVSVCKGLMTKHPAKRLGCGPEGERDVREHAFFRRI
DWEKLENREIQPPFKPKVCGKGAENFDKFFTRGQPVLTPPDQLVIANIDQSDFEGFSYVN
PQFVHPILQSAV
```

© 2016 The Barton Group - Division of Computational Biology, School of Life Sciences, University of Dundee, UK.

Figure 4.14: POGSPSF result web-page for the Protein kinase C (UniProt P17252). Left panel JQuery-DataTables, where the score is > 0.6 highlighted in red. Human SNV and phosphosite data are incorporated in the table if available. The information can be exported as a CSV file. Right panel shows the protein sequence, and the S/T can be highlighted when selected in the table. `http://www.compbio.dundee.ac.uk/pogspsf/uniprot/P17252`

JQuery DataTables. Custom JavaScript code was developed to communicate from the DataTables to the protein sequence. Figure 4.14 shows the result page for protein kinase C (UniProtKB P17252). The 5 scores > 0.6 are highlighted in the table (red). The user can select proteins from 71 791 human proteins in the database. Phosphosites and human germinal variant data obtained in Section 4.2.1, are provided in the results table, if available. On-line prediction service for protein sequences not in the database will be provided in the near future. An API for programmatic access of the results is under development.

4.4 Discussion

The analysis of the results from the application of POGSPSF to the proteome had three main aims. Firstly, to assess the applicability of the classification model. Secondly, to identify issues in the application and potential improvements for the future versions. Finally, the study of top ranking sites could reveal new potential targets and roles for protein *O*-GlcNAcylation.

4.4.1 Technical challenges

The application of the classifier over the human proteome is technically challenging, and problems were reported here. The absence of predictions for residues close (within 3 residues) to the protein termini was a critical issue. However, very few reported *O*-GlcNAc sites occur close to the protein C- and N-termini, so this problem should not interfere with the results described here.

Jpred4's 800 amino acid cap also limited the proteome-wide study, since POGSPSF depends on secondary structure for classification. After discussion with the Jpred4 maintainer, it was decided that omitting proteins with more than 800 residues was the most practical approach at this stage. Section 6.2 further discuss this issue.

Another interesting observation is the presence of protein fragments within the proteins containing the high-scoring sites. ‘Fragment protein’ is a UniProtKB nomenclature for a shorter version of the canonical protein. Often there is no empirical confirmation of these protein fragments *in vivo*, so the expression of such small polypeptide are unknown. Early POGSPSF prototypes that also used secondary structure and disorder predictions had several fragment proteins within the top-scoring motifs. Further study of the secondary structure and disorder propensities of the protein fragments could reveal why these proteins are overrepresented within the top ranking proteins.

Overprediction is a common problem for PTM classifiers, but developers do not often discuss this issue. POGSPSF classifies 251 904 S/Ts as modified, for class probability of 0.6, at least one order of magnitude more than the estimated number of sites, 36 000, calculated from the POGSPSF training-set. The number calculated from the training set might be underestimating the total numbers of modified S/T in the human proteome, due to the incompleteness of the dataset. However, a number between 10 000 and 100 000 is a reasonable estimate based on the number of *O*-GlcNAc sites being less abundant than phosphosites.

One last problem was the lack of extra data to recalibrate the class probability score and minimise overprediction. Figure 4.1 shows that the majority of known sites have a class probability score$>$ 0.6, so this threshold seemed reasonable, but this value can be recalibrated when more data become available.

4.4.2 Feature importance

Another typical application of machine learning methods is to determine specific features that separate two classes. SciKit-Learn implements a few tools to provide feature selection and extraction from models. The random forest classifier falls between decision forests and ANNs regarding model transparency, so it is easier to extract the feature importance for a random forest model if compared to ANNs and SVMs. In fact, model interpretation is an overlooked attribute, when one decides which learning algorithm is appropriate for a model. It is much harder, if even possible, to extract the model preference from 'Black box' methods, best represented by ANNs. In fact, there is a trade-off between model interpretability and prediction performance (Kuhn and Johnson, 2013). Modern implementations, such as SciKit-Learn random forests, train models that are not complete black boxes and, therefore, can be interpreted.

Feature interpretation was attempted but unfortunately did not reveal any novelty. The motif sequence was revealed as the most important feature, followed by predicted disorder and secondary structure. Also, the analysis of the positions shows that residues close to the modification site contain more relevant information than those distant to it. Alternative methods for feature selection and extraction may be tested in the future.

4.4.3 Novel findings

Machine learning methods can generalise from examples (Domingos, 2014). So the POGSPSF application can contribute to (a) prioritising S/Ts within proteins for experimental validation; (b) improvements of PTM site prediction; and (c) offer an overview of the protein *O*-GlcNAcylation process. The proteome-wide prediction, described here, was the first practical application of the tool and revealed its main caveats and potential.

The GO term analysis demonstrated that the POGSPSF model captures the high-level biological role of protein *O*-GlcNAcylation. The enrichment of certain cellular component terms, specifically nucleoplasm, cytoplasm and nucleus, and molecular function, such as protein kinase binding and transcription factor binding, within the proteins with high-scoring sites suggest that predictions incorporate the high-level attributes from OGT modified proteins. In addition, multiple GO terms not directly related to the proteins in the training set were detected as enriched in the proteins more likely to be modified, as defined by the POGSPSF model. These could be new biological features of *O*-GlcNAc, but some discoveries might be related to with artefacts from the large scale predict and over-prediction.

Also, S/Ts classified as modified have an increased likelihood (odds ratio 1.15) of being phosphosites. The results obtained here show a particular increase of phosphorylation in the central S/T over the background protein frequency in the human proteome. Such high-level, i.e. proteome-wide, confirmation of this interplay is as yet unreported. The result agrees with the value observed for known *O*-GlcNAc sites. However, since POGSPSF overpredicts, this value may be underestimated. It would, therefore, be interesting to re-analyse the association of POGSPSF score and the fraction of phosphosites in future versions of the model, to check whether the observed trend changes.

Analysis of human genetic variants in S/Ts classified as modified by POGSPSF show a small decrease in proportions of germinal SNVs. It is notable that such a small depletion of SNVs was detected. Despite the magnitude of the effect, this result has interesting biological implications. Li et al. (2009) found that only a small percentage of variant affect known PTM sites and indicated that different PTMs have different observed frequencies for known germinal, somatic and disease-causing polymorphisms. Uyar et al. (2014) recently established the role of short linear motifs, unstructured segments of protein known to be associated with protein phosphorylation (Iakoucheva et al., 2004) and *O*-GlcNAcylation (see Chapter 2), in cancer. Uyar et al. (2014) shows a significant increase of somatic mutations in short linear motifs within disordered segments, compared to germinal variants. Gray et al. (2014) confirmed that lysines modified by multiple PTMs are more likely to be associated with a disease phenotype. More recently, Reimand et al. (2015) demonstrated that PTM sites are enriched in disease causing mutations, by eliminating other possible genetic confounders. The authors also identified that PTM sites obtained from experiments (modified residue ±7 residues) have a lower ratio of non-synonymous to synonymous variants, indicating less deleterious variants in the sites. These studies mostly focused on phosphorylation data and indicate that PTM sites are important, and often associated with disease causing genetic variants. This is the first time sites targeted by OGT have been associated with fewer genetic variants. What remains to be answered is why the analyses of deleteriousness by SIFT and PolyPhen-2 clearly reveals that the proportions of deleterious variants, but not benign, SNVs decrease with the likelihood of OGT modification. Chapter 5 discusses OGT connection with cancer and disease-causing variants. The study of somatic and other disease-associated genetic variants is suggested for future work.

Interestingly, regarding the relationship of POGSPSF with phosphosites and genetic variants, the association observed for a fraction of phosphosites is directly proportional, while the one obtained for SNVs predicted as damaging is inversely proportional. If further confirmed, this result denotes that some important sites, like *O*-GlcNAc sites and phosphosites, are protected from human genetic variation. The 10 000 (UK10K) Human Genomes very recently reported a reduction of SNV distribution in segments that are biologically significant, like transmembrane domains (Telenti et al., 2016).

Phosphosite and damaging SNV information may be incorporated in the POGSPSF model. For now, the POGSPSF web application provides the information so that the user can decide whether to consider it.

In conclusion, the use of machine learning models is a critical tool to the study of PTM sites. The technical limitations need to be well defined and integrated with the results interpretation. Also, the investigation needs to be followed up with experimental validation.

4.5 Conclusions

- POGSPSF was applied to 71 791 proteins in the human proteome generating scores for 2 429 758 S/Ts
- Problems with the classifier were detected and potential solutions were proposed
- GO term analysis indicated that the predictions could capture the subcellular location, process and functions known to be associated with protein *O*-GlcNAcylation

- S/Ts classified as modified by POGSPSF are 15% more likely to be phosphosites and 7% less likely to co-occur with known synonymous and missense SNVs

- Further analysis indicates a decrease of the fraction of motifs predicted as modified and possessing SNVs classified by SIFT or PolyPhen-2 as damaging, but no such effect was observed for those classified as benign

- Several potential novel targets were identified from the high-scoring sites; the experimental validation of these sites could help define OGT substrate recognition processes

- In summary, the application of POGSPSF provided validation to the method in the biological context of the modification; additionally, novel biological associations regarding of protein *O*-GlcNAcylation were suggested

- A web application was developed to provide external access to the predictions; phosphosite and genetic variant data were integrated to the application's result page so that the user can decide to consider it.

Chapter 5

Analysis of genetic variants on OGT structure

Preface

This chapter introduces the data integration problem and presents ProteoFAV, a Python library that integrates protein structural data with features and genetic variants. As an example of ProteoFAV, the chapter also investigates the occurrence of missense genetic variants over OGT in the context of its 3-dimensional structure.

5.1 Introduction

5.1.1 Data Integration

Since the Human Genome Project released the first human genome draft(Lander et al., 2001), DNA sequencing technologies have become cheaper, faster and more accurate, as reviewed by van Dijk et al. (2014). Other techniques in the field of Life

Sciences have also followed these advances. For example, the use of high-content compound screening led to the discovery of a new antimalarial agent (Baragaña et al., 2015). Also, low-throughput methods have been scaled to work in the large scale. For example the method for measuring the OGT activity was scaled to hundreds of substrates of a single one (Pathak et al., 2015). All of these methods are very data intensive, as they produce large amounts of data in different file formats. So new computational methods and tools are required to maximise the information extracted from these data-intensive methods.

Heterogeneous data integration is one important and challenging task of the analysis of data from experiments, including the Structural Biology field (Samish et al., 2015). It involves integrating data from multiple sources, which might be critical to making new discoveries, unseen when the data are outside each other's context. Such tools are absent from the current toolset of the data analyst working in the structural biology field. Thus, this chapter describes the development of a Python library that allows the integration of protein structural data with features and genetic variants. The genetic variants were introduced in Section 1.7

5.1.2 ProteoFAV

ProteoFAV is an open-source Python library that integrates protein structural data, genetic variants and features. The library combines three-dimensional coordinates with genetic variants and protein features from the Universal Protein Resource (UniProt) database. It tackles three main challenges. First, the mapping between genetic variants, which map to a position within the Ensembl reference genome, and protein structures, which refer to the UniProt database. Second, the lack of a light-weight representation, specifically a Python data structure, for structural data and other features able to be processed on a large scale. Third, the minor sequence inconsistencies observed among different databases. The library is modular and flexible for various use cases. Fábio Madeira and Stuart MacGowan collaborated in the development of this tool.

5.1.3 ProteoFAV implementation

ProteoFAV was implemented in Python, one of the most popular programming languages in computational biology. Python readable and easy to maintain and extend source code is a critical feature of the programming language. Python can be limited concerning speed and is slower than some other languages (Fourment and Gillings, 2008). For ProteoFAV development, the problem was overcome by building the tool over Pandas (McKinney and Team, 2015), a high-performance DataFrame implementation. Pandas has a wide user-base and is becoming the central library for data analysis in the Python programming language. Data structures implemented in the Pandas library - Panel, DataFrame and Series - are compliant with Numpy multi-dimensional arrays and can be processed by methods from the Numpy and Scipy libraries, which introduced scientific algorithms to the Python programming language.

ProteoFAV works as a command line application and as Python module. As a command line application the user can select a protein from UniProt or a protein structure from the PDB. The integrated data can be stored in one of the several formats supported by Pandas, as a comma-separated file, or as a Jalview Annotation file for easy annotation of multiple sequence alignments in Jalview (Waterhouse et al.,

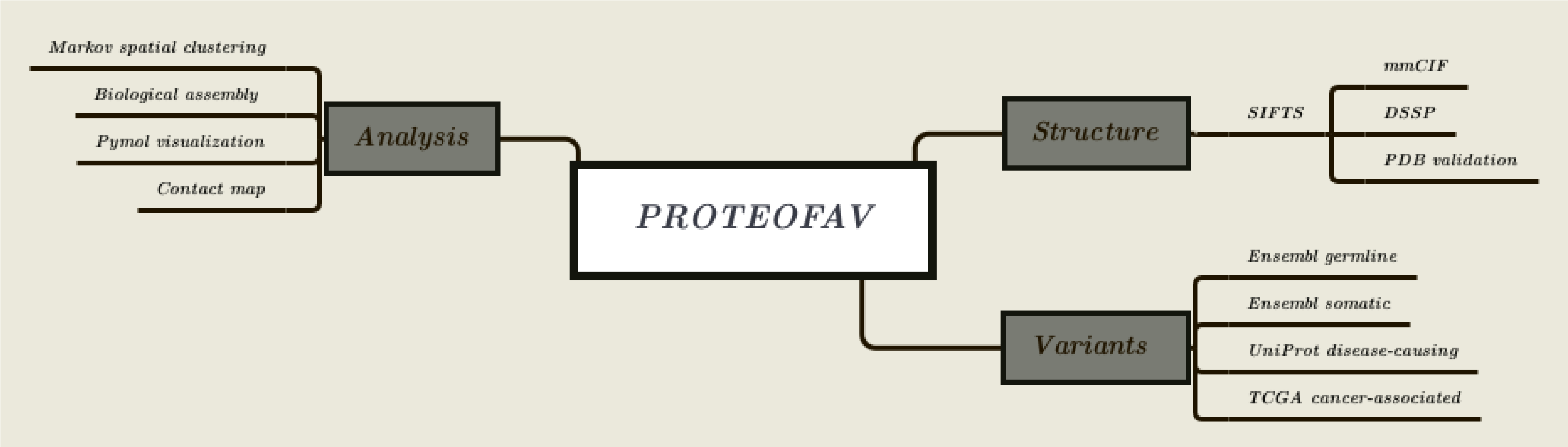


Figure 5.1: ProteoFAV: A Python open-source library for integration of protein structural with genetic variants and other features. The tools also comprise protocols to common tasks such as contact map calculation, custom visualisation with PyMol and Chimera and spatial clustering.

2009). Since Pandas is a data analysis library, importing ProteoFAV as a Python module allows lots of flexibility for data integration and processing. In this thesis, for example, the structural characterisation of *O*-GlcNAc sites in Chapter 2 and mapping genetic variants and phosphorylation sites to UniProt protein sequences in Chapter 4 were performed with ProteoFAV.

5.1.4 Data sources and integration

The ProteoFAV implementation relies on the Pandas join method, which emulates a Structured Query Language (SQL) join. A join procedure combines two datasets using set logic rules, linking the intersection of specific columns of the datasets. With standard parameters, ProteoFAV outputs a table where each row represents a protein residue, but use cases where each row represents an atom or specific collections of atoms - such as the backbone atoms - are also supported. The next section details how the data source outputs were simplified to achieve this tabular data format.

5.1.5 Structural module

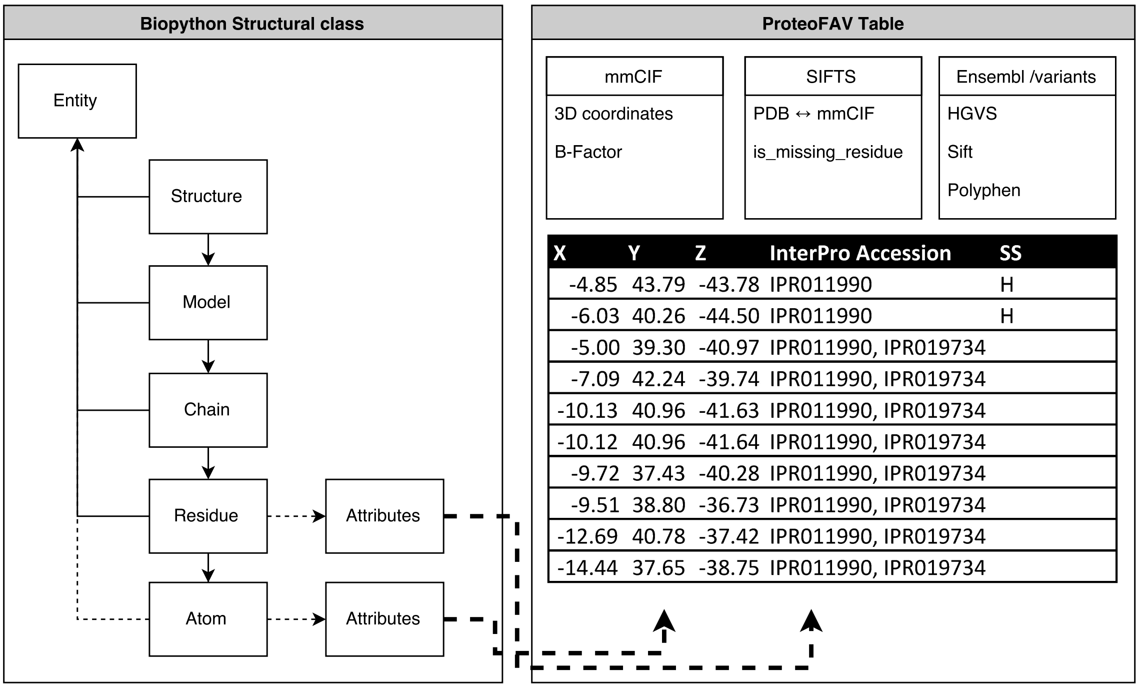
The mmCIF format

The macromolecular Crystallographic Information File (mmCIF) format describes protein macromolecular structural data, including its three-dimensional atom coordinates, and the experimental material and methods (Westbrook and Bourne, 2000). In 2014, the mmCIF format replaced the PDB format, which was not flexible and could not support the major advances in protein structure determination. For example, PDB files do not support structures with more than 99 999 atoms. The

PDBx Exchange Dictionary is an ontology adopted by mmCIF and it defined a convention of names, data types and data relationships (Bourne et al., 1997). An ontology, in this case, is a controlled vocabulary designed to represent the data items, data categories and their relationship. The PDBx Exchange Dictionary enables the annotation of macromolecular structures and experiments to keep up with advances in the field of macromolecular determination.

In mmCIF, data is represented as key-value pairs or in a tabular format. Attributes are named data items and grouped into data categories. The `ATOM_SITE` table describes the attributes of the atoms in the structure, including the x, y, and z coordinates. Other data categories, such as the structure references, are represented in key-value format.

Other Python libraries have PDB/mmCIF file parsers, notably the Biopython Structural package (Hamelryck and Manderick, 2003). The Biopython Structural Class represents macromolecular structures as a hierarchical data structure. This data structure can be inconvenient for data analysis and processing, due to the computational cost of retrieving attributes from the deeply nested data structure. ProteoFAV fixes this problem by storing data in a flat tabular format. Each column defines a residue attribute, and the table rows represent each residue. With default parameters, the residues contain its carbon α atom attributes. However, ProteoFAV also support other configurations, such as an arbitrary list of atom type or only backbone atoms. Figure 5.2 summarises the difference between Biopython Structural data structure and ProteoFAV table.



X	Y	Z	InterPro Accession	SS
-4.85	43.79	-43.78	IPR011990	H
-6.03	40.26	-44.50	IPR011990	H
-5.00	39.30	-40.97	IPR011990, IPR019734	
-7.09	42.24	-39.74	IPR011990, IPR019734	
-10.13	40.96	-41.63	IPR011990, IPR019734	
-10.12	40.96	-41.64	IPR011990, IPR019734	
-9.72	37.43	-40.28	IPR011990, IPR019734	
-9.51	38.80	-36.73	IPR011990, IPR019734	
-12.69	40.78	-37.42	IPR011990, IPR019734	
-14.44	37.65	-38.75	IPR011990, IPR019734	

Figure 5.2: Data accession in the Biopython Structural module compared to ProteoFAV Table. The Biopython Structural module (Hamelryck and Manderick, 2003) is a popular method to parse and access protein structural information among Python programmers. The diagram in the left panel shows a simplified version on how the Biopython class is organised. The top level object, Entity, contains a collection of Structures, which contain the Models and so on until the Atoms objects. Because of the deeply nested architecture, the data access for residue and atom attributes is computationally costly. An excerpt of a ProteoFAV table example is shown in the right panel. ProteoFAV parses the mmCIF directly into a table, where each row represents a residue and columns contain the attributes from residues and atoms. The table provides faster and more convenient access to the data, in addition to data integration and processing with Pandas. Left panel x, y and z, Atom three-dimensional coordinates; SS, secondary structure.

The SIFTS file format

The SIFTS Project (Velankar et al., 2013) is a joint effort of the PDB Europe and UniProt that provides residue-level or chain-level mappings among the following databases: Pfam (Finn et al., 2014), InterPro (Hunter et al., 2009), SCOP (Fox et al., 2014), CATH (Sillitoe et al., 2015), PubMed (Maloney et al., 2013), Gene Ontology (Gene Ontology Consortium), PDB (Velankar et al., 2010) and UniProt (Bateman et al., 2015). The SIFTS database serves Extensible Markup Language (XML) files with a complex hierarchical structure, where the higher-level element is a PDB entry and lowest-level element a residue or a domain. Apart from the database mapping, SIFTS also flags missing residues, engineered constructs, expression tags and other sequence heterogeneities in the protein structure. These problems and discontinuities of the amino acid sequence are very common, therefore SIFTS is a critical resource to resolve the differences between the sequence of the deposited protein structure and the sequence of the UniProt entry. ProteoFAV merges the mmCIF file and the SIFTS by joining the `auth_seq_id` and `auth_asym_id` to the `PDB_dbResNum` and `PDB_dbChainId` of each respective file. An additional step checks the integrity of the merged data by comparing the amino acid sequence obtained from each data source.

The DSSP file format

The DSSP algorithm assigns secondary structure elements to a protein structure based on the hydrogen bonding pattern among the backbone atoms of consecutive residues (Kabsch and Sander, 1983). The DSSP program was recently re-factored (Joosten et al., 2011) to support changes to the PDB file format, which is the input file to the DSSP program. DSSP also measures the peptide backbone torsion angles, the residue solvent accessible surface and other geometrical attributes of residues in a protein structure all of which are output in tabular format with fixed column widths. In ProteoFAV the data are merged via the `auth_seq_id` and `auth_asym_id` of the mmCIF and `icode` and `chain_id` from the DSSP file.

5.1.6 Variant module

The ProteoFAV variant module integrates genetic variant data retrieved from three databases via their APIs. The Ensembl (Flicek et al., 2014) and UniProt (Bateman et al., 2015) databases store both human germinal and somatic mutations from several projects. The Cancer Genome Atlas (TCGA) Pan-Cancer Data portal (Cline et al., 2013) stores genetic variants from cancerous tissues. Currently, the API responses are JavaScript Object Notation (JSON) files with no common standard, hence ProteoFAV has specific methods for processing the data from each database. By a process called JSON normalisation, a nested JSON file can be flattened into a table. The variant data are then represented by their Human Genome Variation Society (HGVS) notation (Den Dunnen and Antonarakis, 2000). For example a mutation of the Tryptophan in the position 26 to a Cysteine would be represented as p.Trp26Cys. The notation can be input to the Ensembl VEP (McLaren et al., 2016) to confirm the variant genomic location; enumerate known variants at the same position; retrieve the variant minor allele frequency and the SIFT and PolyPhen-2 scores for the variant. The default URL for the structural and variant modules are listed in Table 5.1.

Protein identifier mappings between the UniProt and Ensembl databases are well-established and easy to obtain. However, approximately 20 % of human proteins

Table 5.1: Current web addresses for ProteoFAV data resources. Note that for APIs different endpoints may be used. For example, in Ensembl the `variation` endpoint serves the genetic variant data.

Resource name	**Current web address**
Structure module	
mmCIF	`http://www.ebi.ac.uk/pdbe/entry-files/`
DSSP	`ftp://ftp.cmbi.ru.nl//pub/molbio/data/dssp/`
SIFTS	`ftp://ftp.ebi.ac.uk/pub/databases/msd/sifts/xml/`
PDB validation	`http://www.ebi.ac.uk/pdbe/entry-files/download/`
Variant module	
UniProt	`http://www.ebi.ac.uk/uniprot/api/`
TCGA Pan-cancer	`https://dcc.icgc.org/api/v1/`
Ensembl	`http://grch37.rest.ensembl.org/`

sequence had at least one amino acid mismatch, as the sequences of the two databases were compared for the genetic variant study in Chapter 4. ProteoFAV checks the protein sequence between the two databases before merging the structural and genetic variant data and alerts in case of errors.

5.1.7 Testing module

ProteoFAV includes a full suite of test code with using the Python UnitTest module. Currently, the testing module covers more than 80 % of the library methods. Moreover, if the process fails for a particular protein structure, the library is adapted to handle the exception and new test cases are added to it. For example, ProteoFAV can also join protein structures that use insertion codes for indexing, among other corner cases. The tests are also necessary to detect API changes that break the methods that process genetic variant data.

5.1.8 Data format limitations

The analysis of macromolecular structures in large scale demands a simplified data structure that can carry the maximum amount of information while keeping its structure. The Pandas DataFrame implementation seemed to be the potential solution but representing the protein structural data, and annotations in a tabular format does have its limitations. For example, data redundancy, caused by attributes that carry the same information is common and should be eliminated with a post-processing step. The second potential limitation is that the comprehensive integration of many data sources leads to a large number of attributes (columns). However, the Pandas DataFrame implementation is optimised to handle such tables. The third limitation is the attribute type ambiguity. As the solution for this problem depends on the use-case, the user should define the data type for the attributes with ambiguous types. ProteoFAV documents the attributes with their expected types - whether the type of the attribute is a Boolean, String, Integer or Float number - to assist in handling attributes and attribute type ambiguities. Despite these limitations, structural data analysis and integration of heterogeneous data sources work seamlessly with ProteoFAV, which is a convenient solution for the integration of structural, features and variant data.

5.2 Methods

5.2.1 OGT three-dimensional structure

The analysis of the AAS found in the OGT structure requires a complete three-dimensional model of the enzyme. Currently, the PDB contains only constructs of the OGT's N- and C-terminus. Hence, an entire model was obtained by combining two protein structures determined by X-ray crystallography. The structure 1w3b (chain A) (Jínek et al., 2004) covers the enzyme's N-terminus region from residue 4 to 388 (coverage 39 %). The OGT catalytic domain structure, which resides in the enzyme's C-terminus, has been determined more than once and the protein structure with the highest resolution, 4gyw (chain A) (Lazarus et al., 2012), was selected. This structure covers from residue 323 to 1 041 (coverage 69 %). The full-length OGT structure was assembled with the MultiDomain Assembler (Hertig et al., 2015) via the Chimera visualization program (Pettersen et al., 2004). This algorithm merges protein or domain structures by aligning the three-dimensional positions of the common residues from both structures using least-squares fitting. It also prepares the combined structure for model refinement with Modeller (Eswar et al., 2008). Modeller reports a normalised Discrete Optimized Protein Energy (Z-DOPE) score of -0.81. The Z-DOPE score measures how likely the model represents the native structure with more negative scores indicating better representations of the proteins native fold.

5.2.2 Collecting genetic variants

OGT genetic variants were retrieved from the three databases in Table 5.1. These databases contain variants mapped to the Ensembl human genome assembly Genome Reference Consortium human (version 37 from March 2009) (GRCh37) (Lander et al., 2001). Since the OGT protein sequence is identical between the Ensembl and UniProt databases no additional steps were needed to map the genetic variants to the protein sequence. The UniProt OGT protein (UniProtKB accession number `O15294`) maps to the Ensembl stable protein `ENSP00000362824` and transcript `ENST00000373719`, which was queried against the databases. Table 5.2 summarises the 957 genetic variants mapped to OGT. The sSNV and nonsense variants were removed from the dataset, resulting in the dataset in Appendix C.1. Three entries that yield insertions instead of missense variants were discarded (in bold on the Appendix table).

Table 5.2: Summary of genetic variants per resource and type. Total: number of variants before filtering. Missense: number of missense variants. The number of unique AAS was 242, and the difference of this number to the number in the Total versus Missense cell is due to more than one resource serving the same variant information.

Data resource	Variant type	Total	Missense
UniProt	Somatic	60	57
	Germinal	68	67
Ensembl	Somatic	247	184
	Germinal	154	58
TCGA	Somatic	428	111
Total		**957**	**477**

5.2.3 AAS spatial clustering and visualization

The resulting AAS dataset was mapped to the OGT structure obtained in Section 5.2.1. Chapter 1 describes the protein structure and function in detail. The

AAS spatial distribution was analysed visually with Chimera.

To determine whether a region or domain was more susceptible to a type of genetic variant, the coordinates of the centroid, defined as the mean atom position for every atom in a given residue, of amino acids affected by AAS were clustered with DBSCAN algorithm, implemented in the SciKit-learn library. DBSCAN clusters data based on the density given a distance metric. The algorithm has two parameters `min_samples` and `epf`. The `min_samples` parameter establishes the minimum number of elements - i.e. AAS - to seed a cluster whilst `epf` is the cut-off Euclidean distance for cluster membership in Å. Different from hierarchical clustering, which outputs the hierarchical the cluster hierarchies and the cluster definition depends on a cut-off, DBSCAN returns cluster partitions, which identifies the membership of each cluster and elements clustered that are not clustered, also called singletons. Among the SciKit-learn clustering algorithms, DBSCAN is best able to resolve clusters from noisy data (Scikit-Learn website, 2016a).

5.3 Results

5.3.1 AAS over OGT primary structure

A total of 242 AAS affecting 144 amino acids were retrieved, where 81 amino acids were affected by both somatic and germinal variants. Figure 5.3 illustrates the distribution of AAS over the protein sequence. Green and yellow rectangles represent the TPR domain (from residue 21 to residue 496) and the glycosyltransferase 41 domains (from residue 556 to residue 1 024), respectively. The plot reveals segments with no AAS, which are enumerated in Table 5.3. These segments which may be

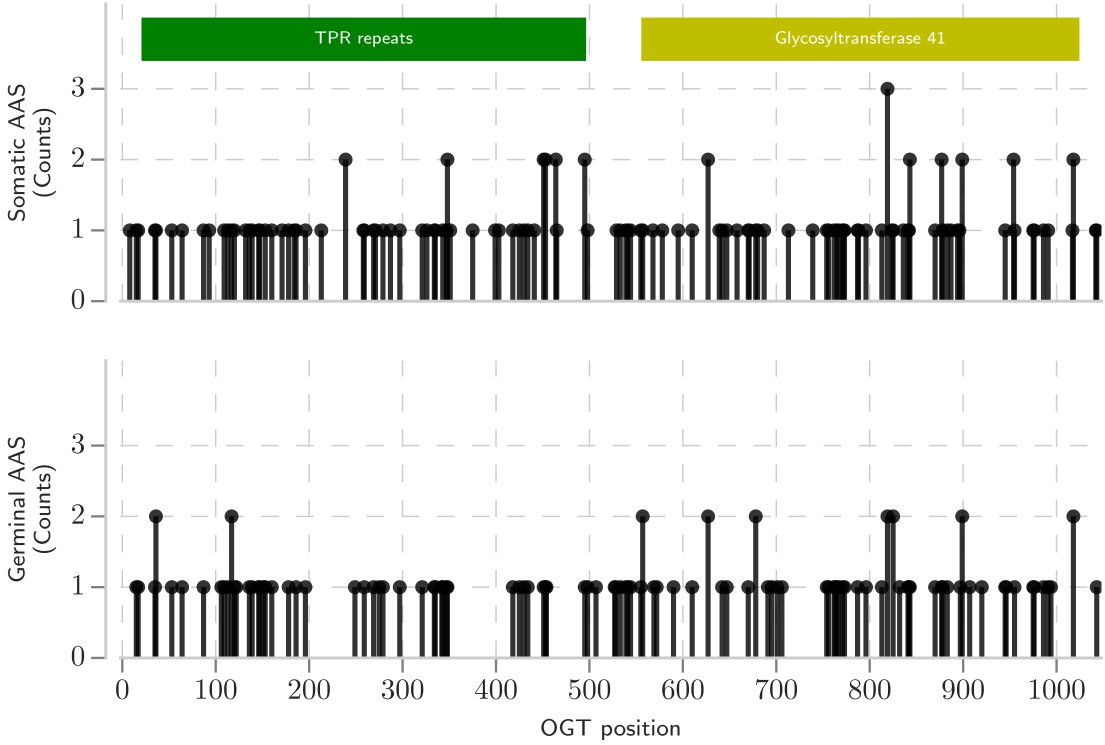


Figure 5.3: AAS over OGT primary structure. The x-axis shows the amino acid position within the OGT sequence; y-axis shows the number of AAS. Green and yellow rectangles represent the TPR and the glycosyltransferase 41 domains, respectively. Comparing top and bottom panels (somatic and germinal variant type) reveals that some regions are more susceptible to one genetic variant type. The three longest segments without any AAS are listed in Table 5.3.

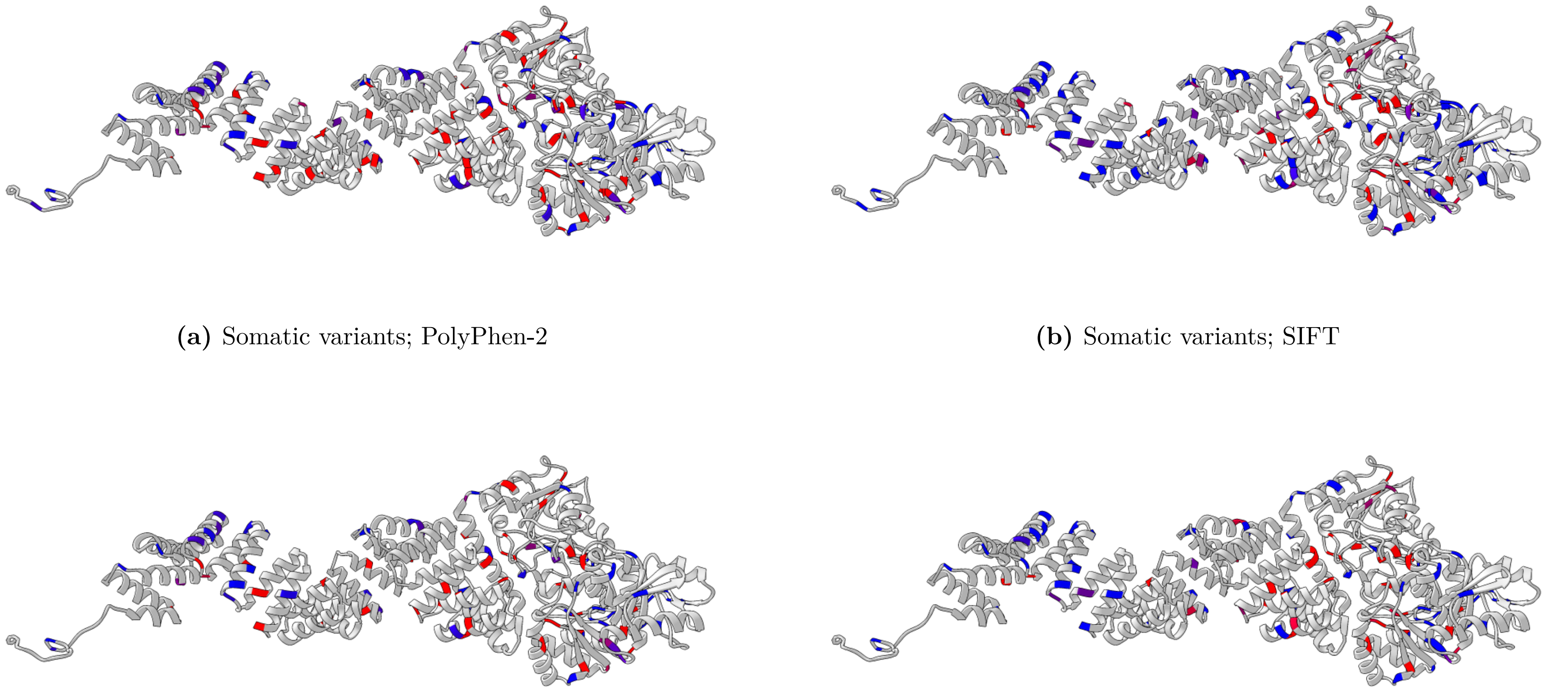
protected or intolerant to AAS could indicate important regions for the protein function.

Table 5.3: Top three longest segments not affected by observed AAS. Start and end refer to the amino acid position within the OGT sequence; the segments are ordered by size. 3 regions in the TPR domain are less likely to be affected by germinal variants, while 2 regions in the in the TPR domain and one in the Glycosyltransferase 41 are less likely to be affected by somatic variants.

Rank	**Mutation type**	**Region**	**Start**	**End**
First	Somatic	Glycosyltransferase 41	899	944
Second	Somatic	Interdomain	498	528
Third	Somatic	TPR	465	494
First	Germinal	TPR	348	417
Second	Germinal	TPR	196	248
Third	Germinal	TPR	454	494

5.3.2 AAS over OGT tertiary structure

The three-dimensional distribution of AAS was also analysed. Figure 5.4 compares the predictions from PolyPhen-2 (left) and SIFT (right) for somatic (top) and germinal (bottom) genetic variants. Amino acids affected by AAS predicted to be deleterious by SIFT (score <0.05) and PolyPhen-2 (score >0.50) were coloured red. Other amino acids affected by AAS predicted to be benign were coloured blue. Amino acids with both benign and deleterious variants received both colours. SIFT predicts more deleterious AAS than PolyPhen-2. Since SIFT is based upon the conservation of protein sequence, the higher number of deleterious variants might be due to the OGT primary sequence being highly conserved among homologous proteins. Overall, PolyPhen-2 predicts several variants in the TPR domain to be benign, but these results should be interpreted with care since this region modulates essential protein-protein interactions, which are essential for OGT function (Iyer



(a) Somatic variants; PolyPhen-2

(b) Somatic variants; SIFT

(c) Germinal variants; PolyPhen-2

(d) Germinal variants; SIFT

Figure 5.4: AAS derived from somatic (top panels) and germinal (bottom panels) variants over OGT three-dimensional structure. The difference between PolyPhen-2 (left panels) and SIFT (right panels) are small in specific residues. Red ribbon SIFT intolerant PolyPhen-2 possible or probably damaging; blue ribbon SIFT tolerant PolyPhen-2 benign;

and Hart, 2003). The methods have concordant predictions for only 36 % somatic and 34 % germinal variants leading to inconclusive prediction if the two methods are combined.

To identify segments in OGT, that are more often impacted by AAS the three-dimensional coordinates of residues affected by germinal and somatic variants were clustered with DBSCAN. The observed cluster distribution was different between the two variant types. The largest cluster for germinal variants (Figure 5.5, top panel, blue cluster) comprises 23 residues in the enzyme catalytic domain, in contrast to the largest cluster for somatic variants (Figure 5.5, bottom panel, purple cluster), which groups 14 residues in the TPR domain. Table 5.4 enumerates clusters in the two variant types.

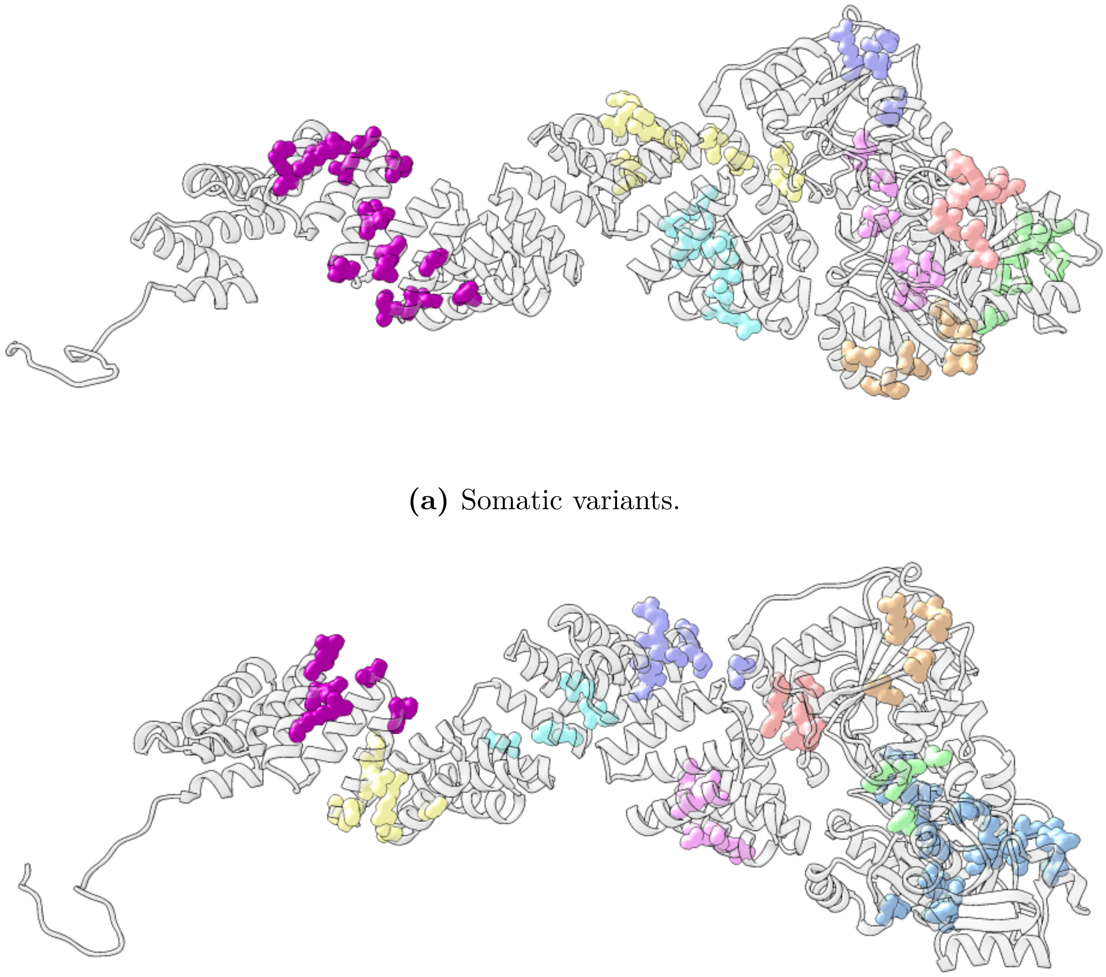
5.4 Discussion

The OGT gene is among the top 3 230 genes most sensitive to germinal genetic variants (Lek et al., 2016), calculated from the ratio between the expected and observed number of missense variants from the The Exome Aggregation Consortium (ExAC). It is also in the top 6 % of the genes ranked by the Residual Variation Intolerance Score (RVIS) (Petrovski et al., 2013) metric, which also sorts genes for their tolerance to genetic variation. RVIS scores the gene based on the ratio of the common functional variation, which is variants with minor allele frequency bigger than 1 % and effect as severe as missense variation, and the total number of variants. Thus, in comparison to other genes, the OGT gene is intolerant to genetic variation.

Moreover, the UniProtKB only lists 2 disease associated mutations in OGT. The

Table 5.4: Colour codes for clusters affecting domains in Figure 5.5. Clusters are ordered from the N-terminus to the C-terminus. The domain or feature column assigns clusters containing at least one variant affecting the domain or feature assigned by UniProt. The number of residues per cluster is in The Residues column. [A] Proximal to Histidine 508.

Variant type	Cluster colour	Domain or feature	Residues
Somatic	Purple	TPR 2, 3, 4 and 5	14
Somatic	Yellow	TPR 9 and 10	12
Somatic	Cyan	TPR 11, 12 and 13	9
Somatic	Median purple	TPR 9	4
Somatic	Pink	Phosphatidylinositol binding site	9
Somatic	Magenta	-	6
Somatic	Green	-	7
Somatic	Golden	UDP binding site	8
Germinal	Purple	TPR 2 and 3	5
Germinal	Yellow	TPR 3 and 4	7
Germinal	Cyan	TPR 6 and 7	4
Germinal	Median purple	TPR 9	7
Germinal	Pink	TPR 11 and 12	6
Germinal	Magenta	Active site[A]	5
Germinal	Golden	-	4
Germinal	Green	-	4
Germinal	Blue	Phosphatidylinositol binding site	23



(a) Somatic variants.

(b) Germinal variants.

Figure 5.5: Variant clusters over OGT. Each colour represents a cluster from DBSCAN with `min_samples` = 4 and `epf` = 10 and each residue is represented by its density map, which was generated with the Chimera `molmap` command. Note that the two structures were tilted to avoid cluster superimposition in the two-dimensional perspective.

germinal mutation p.Ala319Thr is annotated as probably linked to inherited X-linked intellectual disability (Bouazzi et al., 2015). The somatic mutation p.Leu538Pro was found in a patient with renal carcinoma. None of those mutations was in the dataset collected in this work.

So far, only one report experimentally studied the effects of AAS in OGT. The germinal variants p.Leu254Phe was detected in a family with X-linked intellectual disability (Wells, 2016). The mutation led to a higher rate of enzyme degradation and was not the dataset studied in this work, but it is close to the Cyan cluster for the germinal variants, which affects the TPR 6 and 7. From 5 variants near these regions, SIFT predicts 2 AAS to be deleterious, and PolyPhen-2 predicts 3.

A total of 111 germinal and 131 somatic variants were collected from 3 databases. The distribution over OGT's primary structure reveals several segments without mutations, including a segment from residue 843 to 870 that comprises the UDP binding site and does not contain missense, nonsense and synonymous SNV.

However, the statistical determination of whether the segment is protected or not from genetic variants is not a trivial task. An appropriate statistical test should incorporate the observed frequency of the mutations. There is no simple way to combine or compare the frequency of variants sourced from different projects, nor is it trivial to examine the functional impact of variants in cancer and normal tissue. Somatic variants are linked to the number of (tissue) donors, while for most, but not all, projects calculate the minor allele frequency for a germinal variant. For variants sourced from Ensembl, no germinal variant had a minor allele frequency > 0.1, meaning that no common functional variant, as defined by Petrovski et al. (2013) was observed for OGT.

The clusters in Figure 5.5 highlights regions more often affected by genetic variants. The groups were observed in three critical regions of the enzyme. One in TPR repeats 12-13, which provides a hinge mechanism in OGT (Lazarus et al., 2011). Another is near the catalytic pocket, which mediates interaction with the target substrate protein during catalysis. Also another is near the phosphatidylinositol binding site, which is involved in trafficking and insulin signalling (Yang et al., 2008). Nonetheless, it is difficult to draw conclusions from these results. Variant clustering does not directly imply that a region is more or less tolerant to a genetic variant. Thus, this method needs to be combined with a more robust statical test for comparing genetic variants obtained from random sampling and the scores obtained from SIFT and PolyPhen-2 to help define regions with increased density of AAS that may be associated with diseases.

A computational study (Kamburov et al., 2015) highlighted eight amino acids (Asn335, Gln372, Asp396, Tyr418, Asp430, Phe428, Pro569 and Thr643) mediating the interaction between OGT and the Host cell factor 1 (UniProt accession `P51610`) and affected by somatic missense variants. These residues are within or close to the Yellow or Pink clusters for the somatic variants. The authors conclude that these somatic variants may disturb the regulation of the Host cell factor 1 in cancer and other transcription factors in cancer. Protein *O*-GlcNAcylation has been extensively associated with breast cancer, prostate cancer and other cancers as reviewed by de Queiroz et al. (2014), but no molecular mechanism has been described so far. The Yellow and Cyan clusters for somatic variants comprise part of the TPR near to the hinge that links the N- and C-terminus domains. The full extension (including unclustered AAS) comprises 22 and 24 AAS, out of 39, which are predicted to be deleterious by PolyPhen-2 and SIFT, respectively.

ProteoFAV was developed as a simple and convenient data analysis solution for the study of features and genetic variants within the context of the protein structure. More details on the method future developments are provided in Chapter 6.

5.5 Conclusions

- ProteoFAV is early in its development. A few milestones need to be achieved before it is publicly released, such as full documentation and Python3 support. Regardless, the tool has proven effective for the study of sites in structures (see Chapter 2) and more convenient than the currently available alternatives in Python

- Despite its critical importance to the human organism, the OGT gene has 957 known genetic variants reported in databases

- The exploratory analysis and the use of SIFT and PolyPhen-2 did not reveal any clear patterns on the protein structure

- The investigation reveals no AAS and almost no genetic variants in the OGT UDP binding site. More elaborate statistical procedures may be developed from this observation

- The distribution of spatial cluster formed from germinal and somatic variants is distinct. The analysis of the two variants types can also be summed with genetic variants that cause diseases and analysed with protein features

- ProteoFAV will be available at `https://github.com/bartongroup/ProteoFAV`.

Chapter 6

Future directions

The overall aim of the thesis was (a) to study the structural features of *O*-GlcNAc sites and (b) to predict potential sites in the sequences of human proteins. The thesis described the structure and features of sites modified by OGT (Chapter 2). The features and the motif sequence were used to train a new random forest model, POGSPSF (Chapter 3). The model predicted potential sites in the human proteome, and the results were studied from the proteome to the protein scale (Chapter 4). A Python library was developed in collaboration with Fábio Madeira and Stuart MacGowan to tackle the integration of protein structure data with features like UniProt annotation and genetic variants (Chapter 5).

This chapter summarises the results obtained in the thesis. Additionally, suggestions are offered on the future direction of the computational study and prediction of *O*-GlcNAc sites. Future developments of ProteoFAV are also mentioned.

6.1 Further characterization of sites

The current model, in 2016, dictates that protein-protein interactions direct OGT specificity for the protein substrate. Databases of protein-protein interactions or tools that infer the interaction between two protein could thus help understand the OGT specificity at a protein level. Also, the molecular recognitions process could be studied by molecular docking experiments and other simulations. The study of OGT normal modes of motion might unravel structural constraints that might also help understand the OGT specificity.

One important aspect that was not addressed by this work is the site conservation. Many non-conserved phosphosites, which appear more recently in the evolution of an organism, do not have function (Nishi et al., 2011; Lienhard, 2008; Beltrao et al., 2012). The functional site prioritisation is not trivial even more for the limited amount of *O*-GlcNAc site data; however, the study of non-functional *O*-GlcNAc sites could reveal new properties of the modification.

The pipeline used during the site characterization could be applied to the characterisation of other PTM sites. PTM such as palmitoylation have a more conserved motif sequence than *O*-GlcNAc; therefore it would be interesting to compare the results for the two modification types.

6.2 Future development for POGSPSF

Development of a machine learning method for the classification of *O*-GlcNAc sites was the biggest challenge in this thesis. The pattern in the modified sites is weak. The secondary structure and disorder dependencies make the classifier slower. Besides all the alternative motif encoding strategies, preprocessing steps and machine learning methods did not improve the predictive performance. The classifier application over the human proteome was also challenging. As mentioned in Chapter 4, the large scale study had a few technical problems. Regarding the continuous development of the tool, the top priority is to fix the problems. Next, the web application will be able to predict from protein sequence not in the database, since now it just matches results to sequences in the database.

Ongoing work aims to include sequences with more than 800 residues to the web application database. To do so, the method described in Chapter 3 will be applied to the remaining sequences. The method consists of slicing the protein sequence in segments of 800 residues, leaving a 100 residue overlap for the segment ends. The segments are submitted to Jpred4 and, subsequently, predictions are joined, after removing 50 residues in the extremities, to avoid incorrect prediction of protein extremities. Disorder scores for DisEMBL, JRonn and IUpred methods were calculated with Jabaws, as described in Chapter 2. Failed predictions resulted from missing disorder data for some sequences. Ongoing work also targets to calculate the POGSPSF score for the proteins without disorder predictions.

Validation of the prediction performance is essential. Recently, new classifiers of *O*-GlcNAc sites have been published (Jia et al., 2013; Wu et al., 2014; Pejaver et al., 2014; Kao et al., 2015; Zhao et al., 2015). The publications focus on building a new machine learning methods and neglected the biological role of the modification. POGSPSF has increased in specificity when compared to YinOYang and OGlcNAcScan with 12 sites mined from papers abstracts. However, the size of the dataset limits the comparison, which will gain robustness with a larger dataset. In addition, further optimisations of the model, such as score calibration, may increase the prediction performance for the POGSPSF.

Work in this thesis shows that the regions around known sites and the predicted S/T are more likely to be phosphorylated. It also suggests that S/T classified as modified are less liable to have damaging genetic variants. What remains unanswered is whether the POGSPSF's overprediction impacts these two findings. Regardless, the proteome-wide experiment can be used to check the potential interplay of other modifications and with disease-causing variants, like the ones in the ClinVar dataset (Landrum et al., 2014) and somatic variants from TCGA and COSMIC (Forbes et al., 2015; Stratton et al., 2009).

Advances in the machine learning field should accelerate the improvement of predictors of PTM sites. The use of semi-supervised learning could deal with the lack of proper negative examples (unmodified sites) in training. Classifiers of PTM sites trained with random forest models are not very common. However, implementations of the random forest algorithm support categorical variables in the training set; therefore the use of amino acid motif as categorical data without sparse encoding will reduce the number of features, reducing the training time and potentially increasing the prediction performance.

The developments of machine learning methods occur in iterations. The tool's preliminary comparison and the proteome-wide analysis show promising results. Next, the web application should be finalised. Further study on the importance of the features and adding heterogeneous data, such as phosphorylation, may be observed in the next iteration of the method.

6.3 Future directions for ProteoFAV

With the enormous amount of data currently being produced, data integration solutions are becoming more common. Python is one popular language in computational biology; however, there are few Python libraries for working with protein structure. One notable exception is Biopython (Cock et al., 2009). However, the Python Structural module is not convenient. Therefore the ProteoFAV library was developed. The library aims to provide a set of tools that allow integration of protein structures, annotations and genetic variants. ProteoFAV itself was used to the characterisation of the sites in Chapter 2.

Although ProteoFAV has a test suite and works, it will demand some effort in documentation and more tests before it is publicly released. ProteoFAV is still under development. Future improvements will add a statistical permutation test to confirm or refute the lack of genetic variants within proteins regions.

Appendices

Appendix A

Appendices to Chapter 2

A.1 Electron densities manually examined

Table A.1: Uniprot, UniProt accession number; Up, position in the UniProt sequence; PDB, PDB accession number; C, protein structure chain; PP, position in the PDB; Expression organism, scientific name of the expression system.

UniProt	**Up**	**PDB**	**C**	**Pp**	***Expression organism***
P68871	73	4x0i	B	73	*Spodoptera frugiperda*
P68871	73	4x0l	B	73	*Homo sapiens*
P68871	73	3w4u	F	72	*Mus musculus*
P68871	73	1yvt	B	72	*Mus musculus*
P68871	73	1yvt	D	72	*Mus musculus*
P68871	85	4x0i	B	85	*Spodoptera frugiperda*
P68871	85	4x0l	B	85	*Homo sapiens*
P68871	85	3w4u	F	84	*Mus musculus*
P68871	85	1yvt	B	84	*Mus musculus*
P68871	85	1yvq	D	84	*Mus musculus*
P69905	134	4x0i	A	134	*Spodoptera frugiperda*
P69905	134	4x0l	A	134	*Homo sapiens*
P69905	134	1yvt	A	133	*Mus musculus*
P69905	134	1yvq	C	133	*Mus musculus*
P69905	4	4x0i	A	4	*Spodoptera frugiperda*
P69905	4	4x0l	A	4	*Homo sapiens*
P69905	4	1yvt	A	3	*Mus musculus*
P69905	4	1yvq	C	3	*Mus musculus*
P69905	36	4x0i	A	36	*Spodoptera frugiperda*
P69905	36	4x0l	A	36	*Homo sapiens*
P69905	36	1yvt	A	35	*Mus musculus*
P69905	36	1yvq	C	35	*Mus musculus*
P27601	59	3ab3	C	59	*Spodoptera frugiperda*
P27601	59	1zcb	A	59	*Spodoptera frugiperda*
P27601	59	3cx7	A	59	*Spodoptera frugiperda*
P27600	66	1zca	B	66	*Spodoptera frugiperda*

UniProt	Up	PDB	C	Pp	*Expression organism*
P68871	50	4x0i	B	50	*Spodoptera frugiperda*
P68871	50	4x0l	B	50	*Homo sapiens*
P68871	50	3w4u	F	49	*Mus musculus*
P68871	50	1yvt	B	49	*Mus musculus*
P68871	50	1yvq	D	49	*Mus musculus*
P31749	308	4ejn	A	308	*Spodoptera frugiperda*

A.2 Sites' properties for SS132 dataset

Table A.2: Sites' properties for SS132 dataset. List of all entries in the SS132 dataset. PDB, PDB accession number; Chain, chain in the PDB file; Position, residue position within the chain; Cluster, cluster id. RSA, relative solvent accessibility; SS, secondary structure.

PDB	Chain	Position	Cluster	B-factor	RSA	SS
1bab	C	3	-	2.31	0.51	CCCCCHH
1f2j	A	3	-	3.04	0.77	CCCCCCC
1kcx	B	17	-	1.20	0.33	CCCEEEE
1oy3	B	193	-	0.08	0.45	CCCCCEC
1pk8	B	115	-	-0.06	0.27	CCEEEEE
1pk8	E	114	-	0.70	0.39	CCCEEEE
1pk8	F	115	-	0.73	0.26	CCEEEEE
1px2	B	115	-	0.29	0.24	CCEEEEE
1sfc	D	133	-	0.80	0.35	CCCCCCC
1wua	A	7	-	0.02	0.29	CCCCEEE
2f2u	A	29	-	2.94	0.61	CCCCHHH
2ftw	A	9	-	1.30	0.21	CCCCEEE
2i1y	A	693	-	2.47	0.47	CCHHHHH
2j4o	A	18	-	2.42	0.56	CCCCCCC
2odv	A	305	-	-1.83	0.47	CCHHHHH
2q9p	A	12	-	0.35	0.57	CCCECCC
2qz4	A	306	-	0.69	0.49	CCCCCCC
2w4o	A	36	-	1.65	0.54	CCCCCEC
3cb2	A	4	-	-0.45	0.08	CCCCEEE
3dxe	C	538	-	2.65	0.67	CCCCCCC
3ig3	A	1248	-	0.81	0.65	CCHHHHC
3l1e	A	61	-	1.02	0.54	CCCCEEE
3lm5	A	20	-	0.23	0.57	CCCCCCE
3q05	D	96	-	0.35	0.31	CCCCCCC
3w6p	B	1	-	-0.23	0.28	CCCCCCC
4b1u	B	7	-	0.42	0.30	CCCCEEE
4b90	B	10	-	1.21	0.25	CCEEEEE
4ceg	A	286	-	0.80	0.33	ECCCCCC
4czt	D	21	-	1.37	0.58	CCCCCCC

PDB	Chain	Position	Cluster	B-factor	RSA	SS
4d9t	A	417	-	2.07	0.60	CCCCCCC
4ejn	A	6	-	2.00	0.54	CCCEEEE
4gv1	A	145	-	1.89	0.53	CCCCCHH
4mk0	A	32	-	-0.48	0.50	CCCCCCC
4n78	F	2	-	0.65	0.52	CHHHHHC
4pjl	A	640	-	0.76	0.39	CCCCCEH
5av9	A	39	-	2.66	0.84	CCCCCCC
1k8k	A	405	A	-0.36	0.32	HHCHHHH
2wbs	A	472	A	-0.71	0.32	ECCHHHH
3r7d	A	197	A	-0.70	0.28	ECCHHHH
3ud1	B	960	A	2.94	0.53	HHCCCCC
4l79	A	314	A	-0.80	0.31	HHHCCCH
4m9e	A	414	A	0.50	0.47	ECCHHHH
4o4h	B	298	A	-0.04	0.22	HCCHHHE
4y5q	A	86	A	-0.36	0.34	ECCCCCE
1bab	C	36	B	-0.63	0.21	HHHHCHH
1cun	A	30	B	1.16	0.49	HHCCCCC
1w7j	B	115	B	-1.04	0.20	HHHHCCC
1wua	A	232	B	0.09	0.54	HHHHCCC
2xtz	A	49	B	-0.95	0.11	CCCCCHH
3cx8	A	59	B	-1.29	0.10	CCCCCHH
4pa0	A	748	B	0.24	0.30	HHHCCCC
2w4o	A	57	C	1.52	0.45	ECCCEEE
4cbx	A	366	C	1.00	0.41	HHCCHHH
1btn	A	57	D	0.10	0.45	CCEECCC
1f4j	B	114	D	-0.89	0.05	EEEECCC
1k8k	A	113	D	-0.96	0.09	EEEEECC
2ci1	A	260	D	-0.21	0.19	EECCCHH
2ci3	A	260	D	-0.73	0.18	EECCCHH
2foy	A	217	D	-0.77	0.21	CCEEECH
2p9i	B	243	D	-1.20	0.30	CCEEEEC
2zv2	A	167	D	0.81	0.41	CCEEEEE
3cb2	B	170	D	-1.30	0.02	EEEEEEC
3ids	B	244	D	-0.80	0.17	CCEEEEE
4c69	X	101	D	0.17	0.28	EEEEEEC
4c69	X	234	D	-0.46	0.22	EEEEEEE
4cbx	A	199	D	1.42	0.50	CCCCCCC
4d8o	A	1337	D	-0.96	0.32	EEECCCE
4ky9	A	162	D	0.64	0.25	CCEECHH
4qvp	T	131	D	-0.77	0.01	CCEEEEE
1k8k	B	338	E	2.07	0.54	CCCCCCC
2gwf	A	218	E	-1.06	0.13	ECCCEEC
3bjf	B	37	E	0.55	0.25	CCCCCCC
3vln	A	13	E	-0.05	0.47	CCCCCCC
4ky9	A	224	E	1.17	0.37	CCCCCEE

PDB	Chain	Position	Cluster	B-factor	RSA	SS
1dxt	B	50	F	0.54	0.52	CCCCCHH
1j4n	A	238	F	1.17	0.59	CCCCCHH
1tki	B	324	F	0.55	0.49	CCCCCCC
1w0j	B	33	F	0.10	0.20	EEEEECC
2foy	A	129	F	0.64	0.38	CCCCCHH
3ud1	B	1162	F	0.71	0.45	CCCCCCC
4c69	X	165	F	0.13	0.37	CEEEEEE
4mk0	B	136	F	0.60	0.31	CEEEEEE
1bab	C	134	G	-0.64	0.23	HHHHHHH
1dxt	B	73	G	-0.10	0.25	HHHHHHH
1dxt	B	85	G	0.21	0.36	HHHHHHH
1f4j	B	254	G	0.86	0.28	HHHHHHC
1l0l	A	183	G	0.34	0.27	HHHHHHH
1okc	A	6	G	0.64	0.58	CHHHHHH
1qmv	A	112	G	-0.49	0.20	CHHHHHC
1ryp	R	195	G	0.09	0.08	HHHHHHH
1usu	A	422	G	-0.60	0.20	HHHHHHH
1w7j	A	182	G	0.89	0.16	HHHHHCC
2wbs	A	415	G	0.82	0.50	CCHHHHH
2zxe	A	366	G	-0.68	0.02	HHHHHHH
2zxe	A	668	G	-0.68	0.27	EEHHHHC
3abm	R	63	G	-0.11	0.01	CHHHHHH
3kn5	A	467	G	-0.96	0.21	HHHHHHH
3pry	A	452	G	-0.68	0.19	HHHHHCC
3pry	B	452	G	-0.84	0.24	HHHHCCC
3udu	B	196	G	-0.51	0.08	CHHHHHH
4aif	B	303	G	0.22	0.36	HHHHHHH
4b1u	B	89	G	-0.51	0.22	HHHHHHH
4eo9	A	188	G	-0.69	0.05	CHHHHHH
4htm	A	21	G	-0.59	0.51	HHHHHHH
4l3j	A	180	G	-0.61	0.01	HHHHHHH
4y5q	A	119	G	-0.08	0.40	CHHHHHH
4y7y	Z	190	G	-0.17	0.04	HHHHHHH
1gmi	A	132	H	0.60	0.35	EEEEEEE
1k8k	A	170	H	-0.12	0.20	EEEECCC
1pk8	F	262	H	0.64	0.51	CCCCCCC
1qmv	A	18	H	0.46	0.17	CEEEEEE
1wpg	D	210	H	-1.08	0.07	EECCCCE
3msu	A	417	H	-0.70	0.19	EEECCCC
3sde	B	147	H	1.29	0.19	CCCCCEE
3ufx	I	185	H	-1.00	0.11	CCEEEEE
1i7n	A	264	I	0.91	0.44	CCCCCCE
1yhw	A	427	I	-0.01	0.23	CCCCHHH
2zbd	A	8	I	-0.37	0.32	CCCCHHH
3q5i	A	369	I	-0.75	0.27	HHCCHHH

PDB	**Chain**	**Position**	**Cluster**	**B-factor**	**RSA**	**SS**
4is4	G	254	I	0.44	0.14	EEECHHH
4w8p	A	1487	I	1.13	0.50	CCCCHHH
1nug	B	76	J	1.49	0.41	EEEECCC
1okc	A	41	J	-0.24	0.22	HHHCCCC
1pk8	B	261	J	1.45	0.49	CCCCCCC
1u5p	A	1732	J	0.11	0.51	HHCCCCC
2w4o	A	137	J	-0.34	0.33	CCCCHHH
3ar4	A	625	J	-0.93	0.18	EEEECCC
3brv	A	733	J	-0.34	0.43	CCCCHHH
3cb2	A	289	J	0.34	0.41	CCCCHHH
3kn5	A	669	J	1.15	0.23	HHHCCCC
3ose	A	703	J	0.39	0.53	CCCCCCC
4i4t	B	174	J	-0.63	0.19	EECECCE

Appendix B

Appendices to Chapter 3

B.1 Redundant proteins sequence in the dataset

Table B.1: Protein sequence clusters from Blastclust in Chapter 3. Blastclust paramenters: 70% identity and 0.90 coverage. The protein with the most sites was selected.

UniProt accession	Protein name	*Organism name*	Cluster number
Q9ERD7	Tubulin beta-3 chain	*Mus musculus*	cluster 1
Q922F4	Tubulin beta-6 chain	*Mus musculus*	
P68372	Tubulin beta-4B chain	*Mus musculus*	
Q7TMM9	Tubulin beta-2A chain	*Mus musculus*	
Q9CWF2	Tubulin beta-2B chain	*Mus musculus*	
Q9D6F9	Tubulin beta-4A chain	*Mus musculus*	
P62737	Actin, aortic smooth muscle	*Mus musculus*	cluster 2
P68035	Actin, alpha cardiac muscle 1	*Rattus norvegicus*	
P68134	Actin, alpha skeletal muscle	*Mus musculus*	
Q8BFZ3	Beta-actin-like protein 2	*Mus musculus*	
P60710	Actin, cytoplasmic 1	*Mus musculus*	
P63260	Actin, cytoplasmic 2	*Mus musculus*	
P08752	Guanine nucleotide-binding protein G subunit alpha-2	*Mus musculus*	cluster 3
B2RSH2	Guanine nucleotide-binding protein G subunit alpha-1	*Mus musculus*	
P18872	Guanine nucleotide-binding protein G subunit alpha	*Mus musculus*	
Q9DC51	Guanine nucleotide-binding protein G subunit alpha-3	*Mus musculus*	
Q60974	Nuclear receptor corepressor 1	*Mus musculus*	cluster 4
O75376	Nuclear receptor corepressor 1	*Homo sapiens*	
E9Q2B2	Nuclear receptor corepressor 1	*Mus musculus*	
Q7Z3K3	Pogo transposable element with ZNF domain	*Homo sapiens*	cluster 5
Q8BZH4	Pogo transposable element with ZNF domain	*Mus musculus*	

UniProt accession	Protein name	*Organism name*	Cluster number
Q7Z3K3-2	Pogo transposable element with ZNF domain	*Homo sapiens*	
O88532	Zinc finger RNA-binding protein	*Mus musculus*	cluster 6
Q96KR1	Zinc finger RNA-binding protein	*Homo sapiens*	
Q562A2	Zinc finger RNA-binding protein	*Rattus norvegicus*	
Q8VDN2	Sodium/potassium-transporting ATPase subunit alpha-1 alpha-1	*Mus musculus*	cluster 7
Q6PIE5	Sodium/potassium-transporting ATPase subunit alpha-2 alpha-2	*Mus musculus*	
Q6PIC6	Sodium/potassium-transporting ATPase subunit alpha-3 alpha-3	*Mus musculus*	
P07197	Neurofilament medium polypeptide	*Homo sapiens*	cluster 8
P08553	Neurofilament medium polypeptide	*Mus musculus*	
P12839	Neurofilament medium polypeptide	*Rattus norvegicus*	
P63319	Protein kinase C gamma type	*Rattus norvegicus*	cluster 9
P05696	Protein kinase C alpha type	*Rattus norvegicus*	
P68403	Protein kinase C beta	*Rattus norvegicus*	
Q8VHR5	Transcriptional repressor p66-beta	*Mus musculus*	cluster 10
Q4V8E1	GATA zinc finger domain containing 2B	*Rattus norvegicus*	
Q8WXI9	Transcriptional repressor p66-beta	*Homo sapiens*	
P31749	RAC-alpha serine/threonine-protein kinase	*Homo sapiens*	cluster 11
P31750	RAC-alpha serine/threonine-protein kinase	*Mus musculus*	
Q8INB9	RAC serine/threonine-protein kinase	*Drosophila melanogaster*	

UniProt accession	Protein name	*Organism name*	Cluster number
P02470	Alpha-crystallin A chain	*Bos taurus*	cluster 12
P02505	Alpha-crystallin A chain	*Rhea americana*	
P02488	Alpha-crystallin A chain	*Macaca mulatta*	
A8DUV1	Alpha-globin	*Mus musculus*	cluster 13
P01942	Hemoglobin subunit alpha	*Mus musculus*	
P69905	Hemoglobin subunit alpha	*Homo sapiens*	
O55042	Alpha-synuclein	*Mus musculus*	cluster 14
P37377	Alpha-synuclein	*Rattus norvegicus*	
P37840	Alpha-synuclein	*Homo sapiens*	
Q9QYX7	Protein piccolo	*Mus musculus*	cluster 15
Q9QYX6	Protein piccolo	*Mus musculus*	
O88737	Protein bassoon	*Mus musculus*	cluster 16
O88778	Protein bassoon	*Rattus norvegicus*	
Q96T58	Msx2-interacting protein	*Homo sapiens*	cluster 17
Q62504	Msx2-interacting protein	*Mus musculus*	
Q12830	Nucleosome-remodeling factor subunit BPTF	*Homo sapiens*	cluster 18
A2A654	Protein Bptf	*Mus musculus*	
Q9Y520	Protein PRRC2C	*Homo sapiens*	cluster 19
Q3TLH4	Protein PRRC2C	*Mus musculus*	

UniProt accession	Protein name	*Organism name*	Cluster number
O75179	Ankyrin repeat domain-containing protein 17	*Homo sapiens*	cluster 20
Q99NH0	Ankyrin repeat domain-containing protein 17	*Mus musculus*	
Q9QX47	Protein SON	*Mus musculus*	cluster 21
P18583-3	Protein SON	*Homo sapiens*	
Q9H4A3	Serine/threonine-protein kinase WNK1	*Homo sapiens*	cluster 22
P83741	Serine/threonine-protein kinase WNK1	*Mus musculus*	
Q01082	Spectrin beta chain, non-erythrocytic 1	*Homo sapiens*	cluster 23
Q62261	Spectrin beta chain, non-erythrocytic 1	*Mus musculus*	
D3YZU1	SH3 and multiple ankyrin repeat domains protein 1	*Mus musculus*	cluster 24
Q9WV48	SH3 and multiple ankyrin repeat domains protein 1	*Rattus norvegicus*	
P35658	Nuclear pore complex protein Nup214	*Homo sapiens*	cluster 25
Q80U93	Nuclear pore complex protein Nup214	*Mus musculus*	
Q61191	Host cell factor 1	*Mus musculus*	cluster 26
P51610	Host cell factor 1	*Homo sapiens*	
A2AQ25	Sickle tail protein	*Mus musculus*	cluster 27
Q8BHY1	Sickle tail protein	*Mus musculus*	
P15146	Microtubule-associated protein 2	*Rattus norvegicus*	cluster 28
P20357	Microtubule-associated protein 2	*Mus musculus*	
Q9NYV4	Cyclin-dependent kinase 12	*Homo sapiens*	cluster 29

UniProt accession	Protein name	*Organism name*	Cluster number
Q14AX6	Cyclin-dependent kinase 12	*Mus musculus*	
P49790	Nuclear pore complex protein Nup153	*Homo sapiens*	cluster 30
E9Q3G8	Protein Nup153	*Mus musculus*	
Q6NXI6	Regulation of nuclear pre-mRNA domain-containing protein 2	*Mus musculus*	cluster 31
Q5VT52	Regulation of nuclear pre-mRNA domain-containing protein 2	*Homo sapiens*	
Q6P4R8	Ubiquitin carboxyl-terminal hydrolase isozyme L5	*Homo sapiens*	cluster 32
Q6PIJ4	Nuclear factor related to kappa-B-binding protein	*Mus musculus*	
O35927	Catenin delta-2	*Mus musculus*	cluster 33
B7ZNF6	Ctnnd2 protein	*Mus musculus*	
E9Q828	Calcium-transporting ATPase	*Mus musculus*	cluster 34
D1FNM8	Calcium-transporting ATPase	*Mus musculus*	
Q80X50	Ubiquitin-associated protein 2-like	*Mus musculus*	cluster 35
Q14157	Ubiquitin-associated protein 2-like	*Homo sapiens*	
P19246	Neurofilament heavy polypeptide	*Mus musculus*	cluster 36
P16884	Neurofilament heavy polypeptide	*Rattus norvegicus*	
Q8WWM7	Ataxin-2-like protein	*Homo sapiens*	cluster 37
Q7TQH0	Ataxin-2-like protein	*Mus musculus*	

UniProt accession	Protein name	*Organism name*	Cluster number
Q5SFM8	RNA-binding protein 27	*Mus musculus*	cluster 38
Q9P2N5	RNA-binding protein 27	*Homo sapiens*	
O55143	Sarcoplasmic/endoplasmic reticulum calcium ATPase 2	*Mus musculus*	cluster 39
Q8R429	Sarcoplasmic/endoplasmic reticulum calcium ATPase 1	*Mus musculus*	
P97836	Disks large-associated protein 1	*Rattus norvegicus*	cluster 40
Q9D415	Disks large-associated protein 1	*Mus musculus*	
Q8NDX5	Polyhomeotic-like protein 3	*Homo sapiens*	cluster 41
Q8CHP6	Polyhomeotic-like protein 3	*Mus musculus*	
Q6PFD5	Disks large-associated protein 3	*Mus musculus*	cluster 42
Q6PFD6	Kinesin-like protein KIF18B	*Mus musculus*	
Q8CC35	Synaptopodin	*Mus musculus*	cluster 43
Q8N3V7	Synaptopodin	*Homo sapiens*	
Q9QZQ0	Neuronal PAS domain-containing protein 3	*Mus musculus*	cluster 44
Q0IJ77	Npas3 protein fragment	*Mus musculus*	
Q69ZI1	E3 ubiquitin-protein ligase SH3RF1	*Mus musculus*	cluster 45
Q7Z6J0	E3 ubiquitin-protein ligase SIAH2	*Homo sapiens*	
Q05BC3	Echinoderm microtubule-associated protein-like 1	*Mus musculus*	cluster 46
B9EKL9	Eml1 protein	*Mus musculus*	
Q9H1B7	Interferon regulatory factor 2-binding protein-like	*Homo sapiens*	cluster 47

UniProt accession	Protein name	*Organism name*	Cluster number
Q8K3X4	Interferon regulatory factor 2-binding protein-like	*Mus musculus*	
A0JNY3	Gephyrin	*Mus musculus*	cluster 48
Q8BUV3	Gephyrin H	*Mus musculus*	
P07901	Heat shock protein HSP 90-alpha	*Mus musculus*	cluster 49
P11499	Heat shock protein HSP 90-beta	*Mus musculus*	
O88935	Synapsin-1	*Mus musculus*	cluster 50
P09951	Synapsin-1	*Rattus norvegicus*	
Q96BD5	Zyxin	*Homo sapiens*	cluster 51
Q6ZPK0	PHD finger protein 21A	*Mus musculus*	
Q8C2Q3	RNA-binding protein 14	*Mus musculus*	cluster 52
Q96PK6	RNA-binding protein 14	*Homo sapiens*	
Q7M6Y3	Phosphatidylinositol-binding clathrin assembly protein	*Mus musculus*	cluster 53
Q13492	Phosphatidylinositol-binding clathrin assembly protein	*Homo sapiens*	
Q86YP4	Transcriptional repressor p66-alpha	*Homo sapiens*	cluster 54
Q8CHY6	Transcriptional repressor p66 alpha	*Mus musculus*	
Q6UN15	Pre-mRNA 3'-end-processing factor FIP1	*Homo sapiens*	cluster 55
Q9D824	Pre-mRNA 3'-end-processing factor FIP1	*Mus musculus*	
Q15723	ETS-related transcription factor Elf-2	*Homo sapiens*	cluster 56
Q9JHC9	ETS-related transcription factor Elf-2	*Mus musculus*	

UniProt accession	Protein name	*Organism name*	Cluster number
Q7Z739	YTH domain-containing family protein 3	*Homo sapiens*	cluster 57
Q8BYK6	YTH domain-containing family protein 3	*Mus musculus*	
O08553	Dihydropyrimidinase-related protein 2	*Mus musculus*	cluster 58
P97427	Dihydropyrimidinase-related protein 1	*Mus musculus*	
Q4KLH5	Arf-GAP domain and FG repeat-containing protein 1	*Rattus norvegicus*	cluster 59
Q8K2K6	Arf-GAP domain and FG repeat-containing protein 1	*Mus musculus*	
P08551	Neurofilament light polypeptide	*Mus musculus*	cluster 60
P19527	Neurofilament light polypeptide	*Rattus norvegicus*	
P18146	Early growth response protein 1	*Homo sapiens*	cluster 61
P08046	Early growth response protein 1	*Mus musculus*	
B4DJQ5	cDNA FLJ59211, highly similar to Glucosidase 2 subunit beta	*Homo sapiens*	cluster 62
P14314	Glucosidase 2 subunit beta	*Homo sapiens*	
P37231	Peroxisome proliferator-activated receptor gamma	*Homo sapiens*	cluster 63
D2KUA6	Peroxisome proliferative activated receptor gamma	*Homo sapiens*	
Q15750	TGF-beta-activated kinase 1 and MAP3K7-binding protein 1	*Homo sapiens*	cluster 64
Q8CF89	TGF-beta-activated kinase 1 and MAP3K7-binding protein 1	*Mus musculus*	

UniProt accession	Protein name	*Organism name*	Cluster number
Q6PHZ2	Calcium/calmodulin-dependent protein kinase type II subunit delta	*Mus musculus*	cluster 56
P11798	Calcium/calmodulin-dependent protein kinase type II subunit alpha	*Mus musculus*	
Q00566	Methyl-CpG-binding protein 2	*Rattus norvegicus*	cluster 66
Q9Z2D6	Methyl-CpG-binding protein 2	*Mus musculus*	
P08670	Vimentin	*Homo sapiens*	cluster 67
P20152	Vimentin	*Mus musculus*	
Q02818	Nucleobindin-1	*Homo sapiens*	cluster 68
Q02819	Nucleobindin-1	*Mus musculus*	
Q9NR12	PDZ and LIM domain protein 7	*Homo sapiens*	cluster 69
Q9Z1Z9	PDZ and LIM domain protein 7	*Rattus norvegicus*	
Q16186	Proteasomal ubiquitin receptor ADRM1	*Homo sapiens*	cluster 70
Q9JKV1	Proteasomal ubiquitin receptor ADRM1	*Mus musculus*	
P63094	Guanine nucleotide-binding protein G subunit alpha isoforms short	*Mus musculus*	cluster 71
Q8CGK7	Guanine nucleotide-binding protein G subunit alpha	*Mus musculus*	
Q02614	SAP30-binding protein	*Mus musculus*	cluster 72
Q9UHR5	SAP30-binding protein	*Homo sapiens*	
P48962	ADP/ATP translocase 1	*Mus musculus*	cluster 73

UniProt accession	Protein name	*Organism name*	Cluster number
P51881	ADP/ATP translocase 2 [Cleaved into: ADP/ATP translocase 2, N-terminally processed]	*Mus musculus*	
Q9DBJ1	Phosphoglycerate mutase 1	*Mus musculus*	cluster 74
O70250	Phosphoglycerate mutase 2	*Mus musculus*	
P28066	Proteasome subunit alpha type-5	*Homo sapiens*	cluster 75
Q9Z2U1	Proteasome subunit alpha type-5	*Mus musculus*	
P02511	Alpha-crystallin B chain	*Homo sapiens*	cluster 76
P23928	Alpha-crystallin B chain	*Rattus norvegicus*	

B.2 Redundant sites in domains

Table B.2: Sites mapping to the same relative of domains inferred by InterproScan in Chapter 3. The first entry of each group was kept.

Domain name	Interpro accession	UniProt accession	Site position	Protein name
Protein kinase domain	IPR000719	P11798	253	CaM-kinase II
		P63319	591	Protein kinase C gamma
Protein kinase domain	IPR000719	Q16566	57	CaM-kinase IV
		P09216	418	Protein kinase C epsilon
Intermediate filament domain	IPR001664	Q3TTY5	118	Keratin II
		P08553	28	Neurofilament medium
Protein kinase-like domain	IPR011009	P63319	82	Protein kinase C gamma
		P09217	159	Protein kinase C
P-type ATPase, transmembrane domain	IPR023298	D1FNM8	456	Calcium-transporting ATPase
		Q8K314	159	Atp2b1
Protein kinase domain	IPR000719	P63319	253	Protein kinase C gamma
		P11798	591	CaM-kinase II
ATPase, F1 complex alpha/beta subunit	IPR004100	P56480	128	ATP synthase subunit beta
		Q03265	134	ATP synthase subunit alpha
G protein alpha subunit	IPR001019	P63094	51	Guanine nucleotide binding protein alpha short
		P08752	44	Guanine nucleotide binding protein alpha-2
		P27600	66	Guanine nucleotide binding protein alpha-12
		P27601	59	Guanine nucleotide binding protein alpha-13
Synuclein	IPR001058	Q91ZZ3	71	Beta-synuclein
		O55042	72	Alpha-synuclein
Cation-transporting P-type ATPase	IPR001757	Q64436	626	H+/K+ exchanging ATPase
		Q6PIE5	614	Na+/K+ transporting ATPase
Intermediate filament DNA binding region	IPR006821	P08551	27	Neurofilament light polypeptide
		P08553	28	Neurofilament medium polypeptide
Intermediate filament DNA binding region	IPR006821	P08670	34	Vimentin
		P08553	37	Neurofilament medium polypeptide

Appendix C

Appendices to Chapter 5

C.1 AAS mapped to OGT 3D

Table C.1: Missense variants over OGT structure. Bold rows represent dircarded entries, which were classified as missenses variants by Ensembl, but in fact are insertions.

Position	Variant type	AAS	Source	Polyphen score	Sift score
8	Somatic	V/M	TCGA	0.12	0.25
15	Germinal	T/M	Ensembl, UniProt	0.01	0.13
15	Somatic	T/M	Ensembl	0.01	0.13
17	Germinal	R/C	UniProt	0.08	0.08
17	Somatic	R/C	TCGA	0.08	0.08
35	Germinal	G/V	UniProt	0.90	0.02
35	Somatic	G/V	TCGA	0.90	0.02
36	Germinal	D/V	UniProt	0.68	0.01
36	Germinal	D/Y	UniProt	0.87	0.12
36	Somatic	D/Y	TCGA	0.87	0.12
53	Germinal	D/G	Ensembl, UniProt	0.04	0.30
53	Somatic	D/G	Ensembl	0.04	0.30
64	Germinal	I/M	UniProt	0.24	0
64	Somatic	I/M	TCGA	0.24	0
87	Germinal	L/V	Ensembl, UniProt	0.01	0.28
87	Somatic	L/V	Ensembl	0.01	0.28
93	Somatic	S/L	TCGA	0.99	0.01
102	**Germinal**	**R/KX**	**Ensembl**		
102	**Somatic**	**R/KX**	**Ensembl**		
106	Germinal	Q/R	UniProt	0.10	0.47
109	Germinal	I/T	Ensembl, UniProt	0.15	0.08
109	Somatic	I/T	Ensembl	0.15	0.08
113	Germinal	R/Q	Ensembl, UniProt	0.04	0.42
113	Somatic	R/Q	Ensembl	0.04	0.42
117	Germinal	R/C	UniProt	0.94	0
117	Germinal	R/H	UniProt	0.13	0.04

Position	Variant type	AAS	Source	Polyphen score	Sift score
117	Somatic	R/H	TCGA	0.13	0.04
120	Germinal	P/S	Ensembl, UniProt	0.98	0.05
120	Somatic	P/S	Ensembl	0.98	0.05
122	Germinal	F/I	UniProt	0.39	0
132	Somatic	A/T	TCGA	0.53	0.15
136	Germinal	A/V	Ensembl, UniProt	0.04	0.42
136	Somatic	A/V	Ensembl	0.04	0.42
139	Germinal	M/V	Ensembl, UniProt	0.01	0.08
139	Somatic	M/V	Ensembl	0.01	0.08
146	Germinal	Y/C	UniProt	0.92	0.29
146	Somatic	Y/C	TCGA	0.92	0.29
147	Germinal	V/I	Ensembl, UniProt	0.00	0.31
147	Somatic	V/I	Ensembl	0.00	0.31
151	Germinal	Q/H	UniProt	0.01	0.15
153	Germinal	N/S	Ensembl, UniProt	0.08	0.03
153	Somatic	N/S	Ensembl	0.08	0.03
160	Germinal	R/C	UniProt	1.00	0.03
160	Somatic	R/C	TCGA	1.00	0.03
171	Somatic	G/S	TCGA	0.32	0.01
178	Germinal	A/T	Ensembl, UniProt	0.06	0.41
178	Somatic	A/T	Ensembl	0.06	0.41
184	Somatic	I/V	TCGA	0.94	0.08
186	Germinal	T/M	Ensembl, UniProt	1.00	0.11
186	Somatic	T/M	Ensembl, TCGA	1.00	0.11
196	Germinal	N/K	Ensembl, UniProt	1	0
196	Somatic	N/K	Ensembl, TCGA	1	0
213	Somatic	H/Y	TCGA	1.00	0.53
239	Somatic	R/C	TCGA	1.00	0.05
239	Somatic	R/H	TCGA	1.00	0.25
249	Germinal	L/H	UniProt	1.00	0.13
258	Somatic	H/R	TCGA	0.87	0.07
259	Germinal	A/V	UniProt	0.99	0
259	Somatic	A/V	TCGA	0.99	0
269	Germinal	Y/C	UniProt	0.92	0.01
269	Somatic	Y/C	TCGA	0.92	0.01
271	Somatic	E/K	TCGA	0.98	0.02
275	Germinal	I/V	UniProt	0.12	0.09
279	Germinal	I/V	Ensembl, UniProt	0.06	0.12
279	Somatic	I/V	Ensembl	0.06	0.12
287	Somatic	E/K	TCGA	0.19	0.24
297	Germinal	C/W	Ensembl, UniProt	0.99	0.03
297	Somatic	C/W	Ensembl	0.99	0.03
298	**Germinal**	**N/MVCY**	**Ensembl**		
298	**Somatic**	**N/MVCY**	**Ensembl**		
321	Germinal	R/H	Ensembl, UniProt	0.00	0.12

Position	Variant type	AAS	Source	Polyphen score	Sift score
321	Somatic	R/H	Ensembl	0.00	0.12
326	Somatic	H/Y	TCGA	0.46	0.02
334	Germinal	A/G	UniProt	0.44	0.70
334	Somatic	A/G	TCGA	0.44	0.70
335	Germinal	N/T	UniProt	0.80	0.01
335	Somatic	N/T	TCGA	0.80	0.01
342	Germinal	N/D	UniProt	0.01	0.24
343	Germinal	I/T	Ensembl, UniProt	0.08	0.17
343	Somatic	I/T	Ensembl	0.08	0.17
347	Germinal	V/F	Ensembl, UniProt	0.16	0.13
347	Somatic	V/F	Ensembl	0.16	0.13
348	Germinal	R/C	Ensembl, UniProt	0.07	0.01
348	Somatic	R/C	Ensembl, TCGA	0.07	0.01
348	Somatic	R/S	TCGA	0.03	0.56
351	Somatic	R/C	TCGA	0.09	0.10
375	Somatic	G/R	TCGA	1.00	0
399	Somatic	S/F	TCGA	0.94	0.01
403	Somatic	N/S	TCGA	0.75	0.03
418	Germinal	Y/C	UniProt	1.00	0
418	Somatic	Y/C	TCGA	1.00	0
425	Germinal	N/I	UniProt	0.99	0.02
425	Somatic	N/I	TCGA	0.99	0.02
430	Germinal	D/N	UniProt	0.57	0
430	Somatic	D/N	TCGA	0.57	0
434	Germinal	N/I	Ensembl, UniProt	1.00	0
434	Somatic	N/I	Ensembl	1.00	0
441	Somatic	D/Y	TCGA	0.69	0
451	Germinal	S/F	Ensembl, UniProt	0.97	0.01
451	Somatic	S/F	Ensembl	0.97	0.01
451	Somatic	S/Y	TCGA	0.69	0.04
453	Germinal	R/C	Ensembl	0.97	0
453	Somatic	R/C	Ensembl	0.97	0
453	Somatic	R/G	TCGA	0.18	0.02
454	Germinal	T/K	UniProt	0.08	0.83
464	Somatic	D/H	TCGA	0.36	0
464	Somatic	D/V	TCGA	0.19	0.02
465	Somatic	A/V	TCGA	0.92	0.07
495	Germinal	D/E	Ensembl, UniProt	0	0.66
495	Somatic	D/E	Ensembl	0	0.66
495	Somatic	D/G	TCGA	0.16	0.04
498	Germinal	E/G	Ensembl, UniProt	0.00	0.16
498	Somatic	E/G	Ensembl	0.00	0.16
507	Germinal	P/L	UniProt	1.00	0
527	Germinal	H/Q	UniProt	0.93	0
528	Germinal	G/D	UniProt	0.05	0

Position	Variant type	AAS	Source	Polyphen score	Sift score
529	Somatic	N/K	TCGA	0.03	0.18
533	Germinal	D/H	UniProt	0.66	0.11
533	Somatic	D/H	TCGA	0.66	0.11
539	Germinal	H/Y	Ensembl, UniProt	0.26	0.38
539	Somatic	H/Y	Ensembl	0.26	0.38
541	Germinal	P/S	Ensembl, UniProt	0.00	0.08
541	Somatic	P/S	Ensembl	0.00	0.08
544	Germinal	E/K	Ensembl, UniProt	0.00	0.98
544	Somatic	E/K	Ensembl	0.00	0.98
555	Germinal	R/W	Ensembl, UniProt	0.96	0
555	Somatic	R/W	Ensembl	0.96	0
557	Germinal	R/C	Ensembl, UniProt	0.98	0
557	Germinal	R/H	UniProt	0.96	0
557	Somatic	R/C	Ensembl, TCGA	0.98	0
568	Somatic	H/Y	TCGA	0.75	0
569	Germinal	P/A	UniProt	0.19	0
572	Germinal	H/D	UniProt	1	0
578	Somatic	P/S	TCGA	0.93	0.52
590	Germinal	C/S	UniProt	0.99	0.02
595	Somatic	P/L	TCGA	0.03	0.02
610	Germinal	N/S	Ensembl, UniProt	0.01	0.02
610	Somatic	N/S	Ensembl	0.01	0.02
627	Germinal	R/C	Ensembl, UniProt	0.95	0.05
627	Germinal	R/H	Ensembl, UniProt	0.89	0.17
627	Somatic	R/C	Ensembl	0.95	0.05
627	Somatic	R/H	Ensembl	0.89	0.17
639	Somatic	M/I	TCGA	0.88	0.01
642	Germinal	Y/H	UniProt	0.94	0
642	Somatic	Y/H	TCGA	0.94	0
647	Germinal	R/Q	UniProt	0.99	0
647	Somatic	R/Q	TCGA	0.99	0
658	Somatic	I/M	TCGA	0.66	0
670	Germinal	G/C	Ensembl, UniProt	1	0
670	Somatic	G/C	Ensembl	1	0
671	Somatic	A/V	TCGA	0.32	0
678	Germinal	I/F	UniProt	0.56	0
678	Germinal	I/V	Ensembl, UniProt	0.00	0.69
678	Somatic	I/V	Ensembl	0.00	0.69
680	Somatic	D/N	TCGA	0.99	0
687	Somatic	E/Q	TCGA	0.02	0.13
691	Germinal	Q/K	UniProt	0.05	0.29
695	Germinal	K/E	UniProt	0.27	0
701	Germinal	H/L	UniProt	0.03	0
706	Germinal	G/C	UniProt	1.00	0
713	Somatic	P/S	TCGA	0.13	0.09

Position	Variant type	AAS	Source	Polyphen score	Sift score
739	Somatic	I/V	TCGA	0.00	1
744	**Germinal**	**-/X**	**Ensembl**		
744	**Somatic**	**-/X**	**Ensembl**		
754	Germinal	V/I	Ensembl, UniProt	0	0.37
754	Somatic	V/I	Ensembl	0	0.37
756	Germinal	M/T	Ensembl, UniProt	0.00	0.57
756	Somatic	M/T	Ensembl	0.00	0.57
762	Germinal	G/R	Ensembl, UniProt	0.07	0.13
762	Somatic	G/R	Ensembl	0.07	0.13
763	Germinal	D/G	UniProt	0.03	0.25
765	Germinal	A/T	Ensembl, UniProt	0.00	0.39
765	Somatic	A/T	Ensembl	0.00	0.39
767	Germinal	S/N	Ensembl, UniProt	0.00	0.48
767	Somatic	S/N	Ensembl	0.00	0.48
771	Germinal	A/T	Ensembl, UniProt	0.00	0.43
771	Somatic	A/T	Ensembl	0.00	0.43
772	Somatic	L/V	TCGA	0.00	0.88
773	Germinal	N/I	UniProt	0.01	0.18
773	Somatic	N/I	TCGA	0.01	0.18
787	Germinal	I/V	Ensembl, UniProt	0.00	0.55
787	Somatic	I/V	Ensembl	0.00	0.55
788	Somatic	E/G	TCGA	0.00	0.17
796	Germinal	Q/L	Ensembl, UniProt	0.01	0.06
796	Somatic	Q/L	Ensembl	0.01	0.06
813	Germinal	I/V	Ensembl, UniProt	0	1
813	Somatic	I/V	Ensembl	0	1
819	Germinal	T/A	UniProt	0.06	0.10
819	Germinal	T/I	UniProt	0.95	0
819	Somatic	T/A	TCGA	0.06	0.10
819	Somatic	T/I	TCGA	0.95	0
819	Somatic	T/P	TCGA	0.12	0.03
824	Somatic	P/L	TCGA	0.76	1
825	Germinal	R/C	UniProt	0.67	0.03
825	Germinal	R/H	Ensembl, UniProt	0.00	0.16
825	Somatic	R/H	Ensembl	0.00	0.16
832	Germinal	R/L	UniProt	1	0
836	Somatic	G/R	TCGA	0.96	0.07
841	Germinal	A/V	Ensembl, UniProt	0.16	0.03
841	Somatic	A/V	Ensembl, TCGA	0.16	0.03
842	Germinal	I/V	Ensembl, UniProt	0.16	0.59
842	Somatic	I/V	Ensembl	0.16	0.59
843	Germinal	V/I	UniProt	0.28	0.35
843	Somatic	V/I	TCGA	0.28	0.35
843	Somatic	V/L	TCGA	0.76	0.03
870	Germinal	N/K	Ensembl, UniProt	0.36	0.05

Position	Variant type	AAS	Source	Polyphen score	Sift score
870	Somatic	N/K	Ensembl	0.36	0.05
877	Germinal	R/C	UniProt	1	0
877	Somatic	R/C	TCGA	1	0
877	Somatic	R/H	TCGA	1	0
878	Germinal	F/L	UniProt	1.00	0.01
878	Somatic	F/L	TCGA	1.00	0.01
883	Germinal	E/Q	UniProt	1	0
883	Somatic	E/Q	TCGA	1	0
887	Somatic	Q/E	TCGA	0.09	0.07
894	Somatic	G/D	TCGA	0.88	0.02
896	Somatic	P/S	TCGA	0.08	0.69
897	Germinal	Q/R	UniProt	0.01	0.07
899	Germinal	R/C	Ensembl, UniProt	1.00	0
899	Germinal	R/H	Ensembl, UniProt	0.99	0
899	Somatic	R/C	Ensembl	1.00	0
899	Somatic	R/H	Ensembl	0.99	0
907	Germinal	P/S	UniProt	0.02	0.12
920	Germinal	V/I	UniProt	0.04	0.25
945	Germinal	T/I	Ensembl, UniProt	0.99	0
945	Somatic	T/I	Ensembl	0.99	0
946	Germinal	M/I	UniProt	0.04	0.01
954	Somatic	R/P	TCGA	0.99	0
954	Somatic	R/Q	TCGA	0.88	0
955	Germinal	V/I	Ensembl, UniProt	0.96	0
955	Somatic	V/I	Ensembl	0.96	0
975	Germinal	E/D	Ensembl	0	0.30
975	Somatic	E/D	Ensembl	0	0.30
976	Germinal	Y/F	Ensembl, UniProt	0.99	0
976	Somatic	Y/F	Ensembl	0.99	0
986	Germinal	D/G	Ensembl, UniProt	0.80	0.01
986	Somatic	D/G	Ensembl	0.80	0.01
991	Germinal	K/R	Ensembl, UniProt	0.01	0.35
991	Somatic	K/R	Ensembl	0.01	0.35
994	Germinal	R/C	UniProt	1.00	0
1017	Somatic	E/Q	TCGA	0.92	0
1018	Germinal	R/Q	Ensembl, UniProt	0.00	0.20
1018	Germinal	R/W	Ensembl, UniProt	0.36	0
1018	Somatic	R/Q	Ensembl	0.00	0.20
1018	Somatic	R/W	Ensembl, TCGA	0.36	0
1042	Somatic	V/I	TCGA	0.01	0.05
1043	Germinal	T/I	Ensembl, UniProt	0.03	0.10
1043	Somatic	T/I	Ensembl	0.03	0.10

Bibliography

Aebersold, R., and Mann, M. 2003. Mass spectrometry-based proteomics. *Nature* 422(6928):198–207.

Al-Numair, N. S., and Martin, A. C. 2013. The SAAP pipeline and database: tools to analyze the impact and predict the pathogenicity of mutations. *BMC Genomics* 14(3):1.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. 2010. *Molecular biology of the cell.* 5th ed. New York: Garland Sciences.

Alexa, A., Rahnenführer, J., and Lengauer, T. 2006. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* 22(13):1600–1607.

Alfaro, J. F., Gong, C.-X., Monroe, M. E., Aldrich, J. T., Clauss, T. R. W., Purvine, S. O., Wang, Z., Camp, D. G., Shabanowitz, J., Stanley, P., Hart, G. W., Hunt, D. F., Yang, F., and Smith, R. D. 2012. Tandem mass spectrometry identifies many mouse brain O-GlcNAcylated proteins including EGF domain-specific O-GlcNAc transferase targets. *Proceedings of the National Academy of Sciences* 109(19): 7280–7285.

Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J. 2015. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnoly* 33(8):831–838.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215(3):403–410.

Anfinsen, C. B. 1973. Principles that Govern the Folding of Protein Chains. *Science* 181(4096):223–230.

Baldi, P., and Brunak, S. 2001. *Bioinformatics - The machine learning approach.* MIT press.

Banerjee, P. S., Hart, G. W., and Cho, J. W. 2013. Chemical approaches to study O-GlcNAcylation. *Chemical Society reviews* 42(10):4345–57.

Baragaña, B., Hallyburton, I., Lee, M. C. S., Norcross, N. R., Grimaldi, R., Otto, T. D., Proto, W. R., Blagborough, A. M., Meister, S., Wirjanata, G., Ruecker, A., Upton, L. M., Abraham, T. S., Almeida, M. J., Pradhan, A., Porzelle, A., Martínez, M. S., Bolscher, J. M., Woodland, A., Norval, S., Zuccotto, F., Thomas, J., Simeons, F., Stojanovski, L., Osuna-Cabello, M., et al. 2015. A novel multiple-stage antimalarial agent that inhibits protein synthesis. *Nature* 522(7556):315–320.

Bateman, A., Martin, M. J., O'Donovan, C., Magrane, M., Apweiler, R., Alpi, E., Antunes, R., Arganiska, J., Bely, B., Bingley, M., Bonilla, C., Britto, R., Bursteinas, B., Chavali, G., Cibrian-Uhalte, E., Da Silva, A., De Giorgi, M., Dogan, T., Fazzini, F., Gane, P., Castro, L. G., Garmiri, P., Hatton-Ellis, E., Hieta, R., Huntley, R., et al. 2015. UniProt: A hub for protein information. *Nucleic Acids Research* 43(D1):D204–D212.

Beltrao, P., Albanèse, V., Kenner, L. R., Swaney, D. L., Burlingame, A., Villén, J., Lim, W. A., Fraser, J. S., Frydman, J., and Krogan, N. J. 2012. Systematic Functional Prioritization of Protein Posttranslational Modifications. *Cell* 150(2): 413–425.

Ben-Hur, A., and Weston, J. 2010. A User's Guide to Support Vector Machines. *Data Mining Techniques for the Life Sciences* 609:223–239.

Bergstra, J., and Bengio, Y. 2012. Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research* 13:281–305.

Binns, D., Dimmer, E., Huntley, R., Barrell, D., O'Donovan, C., and Apweiler, R. 2009. QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics* 25(22):3045–3046.

Blom, N., Gammeltoft, S., and Brunak, S. 1999. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *Journal of Molecular Biology* 294(5):1351–62.

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

Bond, M. R., and Hanover, J. a. 2013. O- GlcNAc Cycling: A Link Between Metabolism and Chronic Disease. *Annual Review of Nutrition* 33(1):205–229.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. 1992. A Training Algorithm for Optimal Margin Classifiers. In *Proceedings of the fifth annual acm workshop on computational learning theory*, 144–152. `arXiv:1011.1669v3`.

Bouazzi, H., Lesca, G., Trujillo, C., Alwasiyah, M. K., and Munnich, A. 2015. Nonsyndromic X-linked intellectual deficiency in three brothers with a novel MED12 missense mutation [c.5922G>T (p.Glu1974His)]. *Clinical Case Reports* 3(7):604–609.

Bourne, P. E., Berman, H. M., McMahon, B., Watenpaugh, K. D., Westbrook, J. D., and Fitzgerald, P. M. 1997. [30] Macromolecular crystallographic information file. *Methods in Enzymology* 277:571–590.

Breiman, L. 2001. Random Forest. *Machine Learning* 45(1):5–32.

Breiman, L. 2006. randomforest: Breiman and cutler's random forests for classification and regression. Tech. Rep. `https://cran.r-project.org/web/packages/randomForest/randomForest.pdf`.

Brimble, S., Wollaston-Hayden, E. E., Teo, C. F., Morris, A. C., and Wells, L. 2010. The Role of the O-GlcNAc Modification in Regulating Eukaryotic Gene Expression. *Current Signal Transduction Therapy* 5(1):12–24.

Bullen, J. W., Balsbaugh, J. L., Chanda, D., Shabanowitz, J., Hunt, D. F., Neumann, D., and Hart, G. W. 2014. Cross-talk between Two Essential Nutrient-sensitive Enzymes: O-GlcNAc TRANSFERASE (OGT) AND AMP-ACTIVATED PROTEIN KINASE (AMPK). *The Journal of Biological Chemistry* 289(15):10592–10606.

Cardoso, L. H., Britto-Borges, T., Vieyra, A., and Lowe, J. 2014. ATP7B activity is stimulated by PKCϵ in porcine liver. *The International Journal of Biochemistry & Cell Biology* 54:60–67.

Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Shaw, N., Lane, C. R., Lim, E. P., Kalyanaraman, N., Nemesh, J., Ziaugra, L., Friedland, L., Rolfe, a., Warrington, J., Lipshutz, R., Daley, G. Q., and Lander, E. S. 1999. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genetics* 22(3):231–238.

Cazy website. 2016. Glycoside hydrolase family 84. `http://www.cazy.org/GH84.html`. Accessed: 2016-08-20.

Ceroni, A., Passerini, A., Vullo, A., and Frasconi, P. 2006. Disulfind: A disulfide bonding state and cysteine connectivity prediction server. *Nucleic Acids Research* 34(suppl 2):W177–W181.

Chalkley, R. J., and Burlingame, A. L. 2001. Identification of GlcNAcylation sites of peptides and α-crystallin using Q-TOF mass spectrometry. *Journal of the American Society for Mass Spectrometry* 12(10):1106–1113.

Chalkley, R. J., Thalhammer, A., Schoepfer, R., and Burlingame, A. L. 2009. Identification of protein O-GlcNAcylation sites using electron transfer dissociation mass spectrometry on native peptides. *Proceedings of the National Academy of Sciences of the United States of America* 106(22):8894–8899.

Chen, D., Juárez, S., Hartweck, L., Alamillo, J. M., Simón-Mateo, C., Pérez, J. J., Fernández-Fernández, M. R., Olszewski, N. E., and García, J. A. 2005. Identification of secret agent as the O-GlcNAc transferase that participates in Plum pox virus infection. *Journal of Virology* 79(15):9381–9387.

Cheung, W. D., and Hart, G. W. 2008. AMP-activated protein kinase and p38 MAPK activate O-GlcNAcylation of neuronal proteins during glucose deprivation. *The Journal of Biological Chemistry* 283(19):13009–13020.

Cheung, W. D., Sakabe, K., Housley, M. P., Dias, W. B., and Hart, G. W. 2008. O-linked beta-N-acetylglucosaminyltransferase substrate specificity is regulated by myosin phosphatase targeting and other interacting proteins. *The Journal of Biological Chemistry* 283(49):33935–33941.

Chothia, C., and Janin, J. 1975. Principles of protein-protein recognition. *Nature* 256(5520):705–708.

Chou, C. F., Smith, A. J., and Omary, M. B. 1992. Characterization and dynamics of O-linked glycosylation of human cytokeratin 8 and 18. *The Journal of Biological Chemistry* 267(6):3901–6.

Chou, K. C. 2001. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins: Structure, Function and Genetics* 43(3):246–255.

Clark, R. J., McDonough, P. M., Swanson, E., Trost, S. U., Suzuki, M., Fukuda, M., and Dillmann, W. H. 2003. Diabetes and the Accompanying Hyperglycemia Impairs Cardiomyocyte Calcium Cycling through Increased Nuclear O-GlcNAcylation. *The Journal of Biological Chemistry* 278(45):44230–44237.

Cline, M. S., Craft, B., Swatloski, T., Goldman, M., Ma, S., Haussler, D., and Zhu, J. 2013. Exploring TCGA Pan-Cancer data at the UCSC Cancer Genomics Browser. *Scientific Reports* 3:2652.

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M. J. L. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25(11):1422–1423.

Cohen, P. 2000. The regulation of protein function by multisite phosphorylation - a 25 year update. *Trends in Biochemical Sciences* 25(12):596–601.

Cohen, P., Watson, D. C., and Dixon, G. H. 1975. The hormonal control of activity of skeletal muscle phosphorylase kinase. Amino-acid sequences at the two sites of action of adenosine-3':5'-monophosphate-dependent protein kinase. *European Journal of Biochemistry* 51(1):79–92.

Cohen, P. 2001. The role of protein phosphorylation in human health and disease: Delivered on June 30th 2001 at the FEBS meeting in Lisbon. *European Journal of Biochemistry* 268(19):5001–5010.

Comer, F. I., Vosseller, K., Wells, L., Accavitti, M. A., and Hart, G. W. 2001. Characterization of a Mouse Monoclonal Antibody Specific for O-Linked N-Acetylglucosamine. *Analytical Biochemistry* 293(2):169–177.

Comtesse, N., Maldener, E., and Meese, E. 2001. Identification of a Nuclear Variant of MGEA5, a Cytoplasmic Hyaluronidase and a β-N-Acetylglucosaminidase. *Biochemical and Biophysical Research Communications* 283(3):634–640.

Copeland, R. J., Bullen, J. W., and Hart, G. W. 2008. Cross-talk between GlcNAcylation and phosphorylation: roles in insulin resistance and glucose toxicity. *American Journal of Physiology-Endocrinology and Metabolism* 295(1):E17–28.

Crick, F. 1970. Central Dogma of Molecular Biology. *Nature* 227(5258):561–563.

Crooks, G. E., Hon, G., Chandonia, J.-M., and Brenner, S. E. 2004. WebLogo: a sequence logo generator. *Genome Research* 14(6):1188–90.

Cuff, J. A., and Barton, G. J. 2000. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins: Structure, Function, and Bioinformatics* 40(3):502–511.

D'Andrea, L. D., and Regan, L. 2003. TPR proteins: The versatile helix. *Trends in Biochemical Sciences* 28(12):655–662.

Das, A. K., Cohen, P. T. W., and Barford, D. 1998. The structure of the tetratricopeptide repeats of protein phosphatase 5: Implications for TPR-mediated protein-protein interactions. *EMBO Journal* 17(5):1192–1199.

Datta, B., Ray, M. K., Chakrabarti, D., Wylie, D. E., and Gupta, N. K. 1989. Glycosylation of eukaryotic peptide chain initiation factor 2 (eif-2)-associated 67-kda polypeptide (p67) and its possible role in the inhibition of eif-2 kinase-catalyzed phosphorylation of the eif-2 alpha-subunit. *The Journal of Biological Chemistry* 264(34):20620–20624.

Den Dunnen, J. T., and Antonarakis, S. E. 2000. Mutation nomenclature extensions and suggestions to describe complex mutations: A discussion. *Human Mutation* 15(1):7–12.

Dias, W. B., Cheung, W. D., Wang, Z., and Hart, G. W. 2009. Regulation of Calcium/Calmodulin-dependent Kinase IV by O-GlcNAc Modification. *The Journal of Biological Chemistry* 284(32):21327–21337.

van Dijk, E. L., Auger, H., Jaszczyszyn, Y., and Thermes, C. 2014. Ten years of next-generation sequencing technology. *Trends in Genetics* 30(9):418–426.

Dinkel, H., Chica, C., Via, A., Gould, C. M., Jensen, L. J., Gibson, T. J., and Diella, F. 2011. Phospho.ELM: A database of phosphorylation sites-update 2011. *Nucleic Acids Research* 39(SUPPL. 1):261–267.

Domingos, P. 2014. A Few Useful Things to Know about Machine Learning. *Communications of the ACM* 55(10):78–87.

Dong, D. L., and Hart, G. W. 1994. Purification and characterization of an O-GlcNAc selective N-acetyl-beta-D-glucosaminidase from rat spleen cytosol. *The Journal of Biological Chemistry* 269(30):19321–19330.

Dongen, S. M. 2000. Graph clustering by flow simulation. Ph.D. thesis, Utrecht University.

Dorfmueller, H. C., Borodkin, V. S., Schimpl, M., and van Aalten, D. M. F. 2009. GlcNAcstatins are nanomolar inhibitors of human O-GlcNAcase inducing cellular hyper-O-GlcNAcylation. *Biochemical Journal* 420(2):221–227.

Dorfmueller, H. C., Borodkin, V. S., Schimpl, M., Zheng, X., Kime, R., Read, K. D., and Van Aalten, D. M. F. 2010. Cell-penetrant, nanomolar O-GlcNAcase inhibitors selective against lysosomal hexosaminidases. *Chemistry and Biology* 17(11):1250–1255.

Dosztányi, Z., Csizmók, V., Tompa, P., and Simon, I. 2005. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *Journal of Molecular Biology* 347(4):827–839.

Drozdetskiy, A., Cole, C., Procter, J., and Barton, G. J. 2015. JPred4: A protein secondary structure prediction server. *Nucleic Acids Research* 43(W1):W389–94.

Duarte, M. L., Pena, D. A., Nunes Ferraz, F. A., Berti, D. A., Paschoal Sobreira, T. J., Costa-Junior, H. M., Abdel Baqui, M. M., Disatnik, M.-h. M.-H., Xavier-neto, J., Lopes de Oliveira, P. S., Schechtman, D., Augusto, F., Ferraz, N., Berti, D. A., José, T., Sobreira, P., Costa-Junior, H. M., Muhammad, M., Baqui, A., Disatnik, M.-h. M.-H., and Xavier-neto, J. 2014. Protein folding creates structure-based, noncontiguous consensus phosphorylation motifs recognized by kinases. *Science Signaling* 7(350):ra105–ra105.

Dumas, J. 2001. Protein kinase inhibitors: emerging pharmacophores 1997 - 2000. *Expert Opinion on Therapeutic Patents* 11(3):405–429.

Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. 1998. *Biological sequence analysis*. Cambridge: Cambridge University Press.

Durek, P., Schudoma, C., Weckwerth, W., Selbig, J., and Walther, D. 2009. Detection and characterization of 3D-signature phosphorylation site motifs and their contribution towards improved phosphorylation site prediction in proteins. *BMC Bioinformatics* 10(1):117.

EMBL-EBI website. 2016. Tetratricopeptide repeat (ipr019734). `https://www.ebi.ac.uk/interpro/entry/IPR019734/proteins-matched?species=9606`. Accessed: 2016-06-16.

Emsley, P., and Cowtan, K. 2004. Coot: model-building tools for molecular graphics. *Acta Crystallographica Section D Biological Crystallography* 60(12):2126–2132.

Ensembl variation website. 2016. About ensembl variation. `http://www.ensembl.org/info/genome/variation/index.html`. Accessed: 2016-01-20.

Ester, M., Kriegel, H. P., Sander, J., and Xu, X. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Second international conference on knowledge discovery and data mining*, vol. 96, 226–231. `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.71.1980`.

Eswar, N., Eramian, D., Webb, B., Shen, M.-Y., and Sali, A. 2008. Protein structure modeling with MODELLER. *Structural proteomics: high-throughput methods* Chapter 2:145–159.

Fariselli, P., Riccobelli, P., and Casadio, R. 1999. Role of evolutionary information in predicting the disulfide-bonding state of cysteine in proteins. *Proteins: Structure, Function and Genetics* 36(3):340–346.

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L. L., Tate, J., and Punta, M. 2014. Pfam: The protein families database. *Nucleic Acids Research* 42(D1).

Fischer, E. 2016. Reversible protein phosphorylation as a regulatory mechanism. `https://www.youtube.com/watch?v=8XKRAduG2Hk`. Accessed: July 2016 2016-07-01.

Flanagan, S. E., Patch, A.-M., and Ellard, S. 2010. Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations. *Genetic Testing and Molecular Biomarkers* 14(4):533–537.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Girón, C. G., Gordon, L., Hourlier, T., Hunt, S., Johnson, N., Juettemann, T., Kähäri, A. K., Keenan, S., Kulesha, E., Martin, F. J., Maurel, T., McLaren, W. M., Murphy, D. N., Nag, R., et al. 2014. Ensembl 2014. *Nucleic Acids Research* 42(D1):D749–D755.

Forbes, S. A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C., Ward, S., Kok, C. Y., Jia, M., De, T., Teague, J. W., Stratton, M. R., McDermott, U., and Campbell, P. J. 2015. COSMIC: Exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Research* 43(D1):D805–D811.

Fourment, M., and Gillings, M. R. 2008. A comparison of common programming languages used in bioinformatics. *BMC Bioinformatics* 9:82.

Fox, N. K., Brenner, S. E., and Chandonia, J.-M. 2014. SCOPe: Structural Classification of Proteins–extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Research* 42(D1):D304–D309.

Fu, Y., Dominissini, D., Rechavi, G., and He, C. 2014. Gene expression regulation mediated through reversible m6A RNA methylation. *Nature Reviews Genetics* 15(5):293–306.

Fushimi, K., Sasaki, S., and Marumo, F. 1997. Phosphorylation of serine 256 is required for cAMP-dependent regulatory exocytosis of the aquaporin-2 water channel. *The Journal of Biological Chemistry* 272(23):14800–14804.

Gambetta, M. C., and Müller, J. 2015. A critical perspective of the diverse roles of O-GlcNAc transferase in chromatin. *Chromosoma* 124(4):429–442.

Gao, Y., Wells, L., Comer, F. I., Parker, G. J., and Hart, G. W. 2001. Dynamic O-Glycosylation of Nuclear and Cytosolic Proteins. *The Journal of Biological Chemistry* 276(13):9838–9845.

Gaudet, P., and Dessimoz, C. 2016. Gene Ontology: Pitfalls, Biases, Remedies. *arXiv preprint arXiv:1602.01875.*

Gene Ontology Consortium. *Nucleic Acids Research* 43(D1):D1049–D1056.

Gonzalez-Perez, A., Deu-Pons, J., and Lopez-Bigas, N. 2012. Improving the prediction of the functional impact of cancer mutations by baseline tolerance transformation. *Genome Medicine* 4(11):89.

Gray, V. E., Liu, L., Nirankari, R., Hornbeck, P. V., and Kumar, S. 2014. Signatures of natural selection on mutations of residues with multiple posttranslational modifications. *Molecular Biology and Evolution* 31(7):1641–1645.

Griffith, L. S., and Schmitz, B. 1999. O-linked N-acetylglucosamine levels in cerebellar neurons respond reciprocally to pertubations of phosphorylation. *European Journal of Biochemistry* 262(3):824–831.

Gross, B. J., Kraybill, B. C., and Walker, S. 2005. Discovery of O-GlcNAc transferase inhibitors. *Journal of the American Chemical Society* 127(42):14588–14589.

Gupta, R., Birch, H., Rapacki, K., Brunak, S., and Hansen, J. E. 1999. O-GLYCBASE version 4.0: a revised database of O-glycosylated proteins. *Nucleic Acids Research* 27(1):370–372.

Gupta, R., and Brunak, S. 2002. Prediction of glycosylation across the human proteome and the correlation to protein function. In *Pacific symposium on biocomputing. pacific symposium on biocomputing*, vol. 3002, 310–322.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. 2002. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning* 46(1-3):389–422.

Hahne, H., Gholami, A. M., Kuster, B., Moghaddas Gholami, A., and Kuster, B. 2012. Discovery of O-GlcNAc-modified proteins in published large-scale proteome data. *Molecular & Cellular Proteomics* 11(10):843–50.

Hamelryck, T., and Manderick, B. 2003. PDB file parser and structure class implemented in Python. *Bioinformatics* 19(17):2308–2310.

Hanover, J. A., Krause, M. W., and Love, D. C. 2012. Bittersweet memories: linking metabolism to epigenetics through O-GlcNAcylation. *Nature Reviews. Molecular Cell Biology* 13(5):312–321.

Hao, Y., Colak, R., Teyra, J., Corbi-Verge, C., Ignatchenko, A., Hahne, H., Wilhelm, M., Kuster, B., Braun, P., Kaida, D., Kislinger, T., and Kim, P. M. 2015. Semi-supervised Learning Predicts Approximately One Third of the Alternative Splicing Isoforms as Functional Proteins. *Cell Reports* 12(2):183–189.

Hardivillé, S., and Hart, G. W. 2014. Nutrient regulation of signaling, transcription, and cell physiology by O-GlcNAcylation. *Cell Metabolism* 20(2):208–213.

Hart, G. W., Greis, K. D., Dong, L. Y., Blomberg, M. A., Chou, T. Y., Jiang, M. S., Roquemore, E. P., Snow, D. M., Kreppel, L. K., and Cole, R. N. 1995. O-linked n-acetylglucosamine: The 'yin-yang' of ser/thr phosphorylation? In *Glycoimmunology*, vol. 376, 115–123. Springer.

Hart, G. W., Slawson, C., Ramirez-Correa, G., and Lagerlof, O. 2011. Cross Talk Between O-GlcNAcylation and Phosphorylation: Roles in Signaling, Transcription, and Chronic Disease. *Annual Review of Biochemistry* 80(1):825–858.

Hédou, J., Bastide, B., Page, A., Michalski, J.-C., and Morelle, W. 2009. Mapping of O-linked β -N-acetylglucosamine modification sites in key contractile proteins of rat skeletal muscle. *Proteomics* 9(8):2139–2148.

Henderson, B., and Martin, A. C. R. 2014. Protein moonlighting: a new factor in biology and medicine. *Biochemical Society Transactions* 42(6):1671–1678.

Henzler-Wildman, K., and Kern, D. 2007. Dynamic personalities of proteins. *Nature* 450(7172):964–972.

Hertig, S., Goddard, T. D., Johnson, G. T., and Ferrin, T. E. 2015. Multidomain assembler (MDA) generates models of large multidomain proteins. *Biophysical Journal* 108(9):2097–2102.

Hilário-Souza, E., Valverde, R. H. F., Britto-Borges, T., Vieyra, A., and Lowe, J. 2011. Golgi membranes from liver express an ATPase with femtomolar copper affinity, inhibited by cAMP-dependent protein kinase. *International Journal of Biochemistry and Cell Biology* 43(3):358–362.

Hirano, T., Kinoshita, N., Morikawa, K., and Yanagida, M. 1990. Snap helix with knob and hole: essential repeats in s. pombe nuclear protein nuc2+. *Cell* 60(2): 319–328.

Hjerrild, M., Stensballe, A., Rasmussen, T. E., Kofoed, C. B., Blom, N., Sicheritz-Ponten, T., Larsen, M. R., Brunak, S., Jensen, O. N., and Ganuneltoft, S. 2004. Identification of phosphorylation sites in protein kinase A substrates using artificial neural networks and mass spectrometry. *Journal of Proteome Research* 3(3): 426–433.

Holt, G. D. 1987. Nuclear pore complex glycoproteins contain cytoplasmically disposed O- linked N-acetylglucosamine. *The Journal of Cell Biology* 104(5): 1157–1164.

Holts, G. D., and Hart, G. W. 1986. The Subcellular Distribution of Terminal N-Acetylglucosamine Moieties. *The Journal of Biological Chemistry* 261(17): 8049–4305.

Hornbeck, P. V., Kornhauser, J. M., Tkachev, S., Zhang, B., Skrzypek, E., Murray, B., Latham, V., and Sullivan, M. 2012. PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Research* 40(D1): D261–D270.

Hornbeck, P. V., Zhang, B., Murray, B., Kornhauser, J. M., Latham, V., and Skrzypek, E. 2015. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Research* 43(D1):D512–D520.

Hubbard, M. J., and Cohen, P. 1993. On target with a new mechanism for the regulation of protein phosphorylation. *Trends in Biochemical Sciences* 18(5): 172–177.

Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., Finn, R. D., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Laugraud, A., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistry, J., et al. 2009. InterPro: the integrative protein signature database. *Nucleic Acids Research* 37(D1):D211–D215.

Iakoucheva, L. M., Radivojac, P., Brown, C. J., O'Connor, T. R., Sikes, J. G., Obradovic, Z., Dunker, a. K., Connor, T. R. O., Sikes, J. G., Obradovic, Z., Dunker, a. K., and O'Connor, T. R. 2004. The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Research* 32(3):1037–49.

Itan, Y., Shang, L., Boisson, B., Ciancanelli, M. J., Markle, J. G., Martinez-barricarte, R., Scott, E., Shah, I., Stenson, P. D., Gleeson, J., Cooper, D. N., Quintana-murci, L., Zhang, S.-y., Abel, L., and Casanova, J.-l. 2016. The mutation significance cutoff: gene-level thresholds for variant predictions. *Nature Methods* 13(2):109–110.

Iyer, S. P. N., and Hart, G. W. 2003. Roles of the Tetratricopeptide Repeat Domain in O-GlcNAc Transferase Targeting and Protein Substrate Specificity. *The Journal of Biological Chemistry* 278(27):24608–24616.

Janin, J., and Chothia, C. 1990. The structure of protein-protein recognition sites. *The Journal of Biological Chemistry* 265(27):16027–16030.

Jeffery, C. J. 1999. Moonlighting proteins. *Trends in Biochemical Sciences* 24(1): 8–11.

Jensen, L. J., Gupta, R., Blom, N., Devos, D., Tamames, J., Kesmir, C., Nielsen, H., Stærfeldt, H. H., Rapacki, K., Workman, C., Andersen, C. A. F., Knudsen, S., Krogh, A., Valencia, A., and Brunak, S. 2002. Prediction of human protein function from post-translational modifications and localization features. *Journal of Molecular Biology* 319(5):1257–1265.

Jensen, O. N. 2006. Interpreting the protein language using proteomics. *Nature Reviews Molecular Cell Biology* 7(6):391–403.

Jia, C.-Z., Liu, T., and Wang, Z.-P. 2013. O-GlcNAcPRED: a sensitive predictor to capture protein O-GlcNAcylation sites. *Molecular bioSystems* 9(11):2909–13.

Jiménez, J. L., Hegemann, B., Hutchins, J. R. a., Peters, J.-M., and Durbin, R. 2007. A systematic comparative and structural analysis of protein phosphorylation sites based on the mtcPTM database. *Genome Biology* 8(5):R90.

Jínek, M., Rehwinkel, J., Lazarus, B. D., Izaurralde, E., Hanover, J. A., and Conti, E. 2004. The superhelical TPR-repeat domain of O-linked GlcNAc transferase exhibits structural similarities to importin α. *Nature Structural & Molecular Biology* 11(10):1001–1007.

Jochmann, R., Holz, P., Sticht, H., and Stürzl, M. 2014. Validation of the reliability of computational O-GlcNAc prediction. *Biochimica et Biophysica Acta - Proteins and Proteomics* 1844(2):416–421.

Johnson, L. N., and Barford, D. 1993. The Effects of Phosphorylation on the Structure and Function of Proteins. *Annual Review of Biophysics and Biomolecular Structure* 22(1):199–232.

Johnson, L. N., and Lewis, R. J. 2001. Structural Basis for Control by Phosphorylation. *Chemical Reviews* 101(8):2209–2242.

Joosten, R. P., Te Beek, T. A. H., Krieger, E., Hekkelman, M. L., Hooft, R. W. W., Schneider, R., Sander, C., and Vriend, G. 2011. A series of PDB related databases for everyday needs. *Nucleic Acids Research* 39(SUPPL. 1):D411–D419.

Kaasik, K., Kivimäe, S., Allen, J. J., Chalkley, R. J., Huang, Y., Baer, K., Kissel, H., Burlingame, A. L., Shokat, K. M., Ptáček, L. J., and Fu, Y. H. 2013. Glucose sensor O-GlcNAcylation coordinates with phosphorylation to regulate circadian clock. *Cell Metabolism* 17(2):291–302.

Kabsch, W., and Sander, C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22(12): 2577–2637.

Kakade, P. S., Budnar, S., Kalraiya, R. D., and Vaidya, M. M. 2016. Functional implications of O-GlcNAcylation dependent phosphorylation at proximal site on keratin 18. *The Journal of Biological Chemistry* 291(23):12003–12013.

Kamburov, A., Lawrence, M. S., Polak, P., Leshchiner, I., Lage, K., Golub, T. R., Lander, E. S., and Getz, G. 2015. Comprehensive assessment of cancer missense mutation clustering in protein structures. *Proceedings of the National Academy of Sciences of the United States of America* 112(40):E5486–E5495.

Kang, E.-S., Han, D., Park, J., Kwak, T. K., Oh, M.-A., Lee, S.-A., Choi, S., Park, Z. Y., Kim, Y., and Lee, J. W. 2008. O-GlcNAc modulation at Akt1 Ser473 correlates with apoptosis of murine pancreatic β cells. *Experimental Cell Research* 314(11-12):2238–2248.

Kao, H.-J., Huang, C.-H., Bretaña, N., Lu, C.-T., Huang, K.-Y., Weng, S.-L., and Lee, T.-Y. 2015. A two-layered machine learning method to identify protein O-GlcNAcylation sites with O-GlcNAc transferase substrate motifs. *BMC Bioinformatics* 16(18):1.

Karplus, M., and Kuriyan, J. 2005. Molecular dynamics and protein function. *Proceedings of The National Academy of Sciences of The United States of America* 102(19):6679–6685.

Kawashima, S., Ogata, H., and Kanehisa, M. 1999. AAindex: Amino acid index database. *Nucleic Acids Research* 27(1):368–369.

Kearse, K. P., and Hart, G. W. 1991. Lymphocyte activation induces rapid changes in nuclear and cytoplasmic glycoproteins. *Proceedings of the National Academy of Sciences of the United States of America* 88(5):1701–1705.

Kelly, W. G., Dahmus, M. E., and Hart, G. W. 1993. RNA polymerase II is a glycoprotein: Modification of the COOH-terminal domain by O-GlcNAc. *The Journal of Biological Chemistry* 268(14):10416–10424.

Kemp, B. E., Bylund, D. B., Huang, T. S., and Krebs, E. G. 1975. Substrate specificity of the cyclic AMP-dependent protein kinase. *Proceedings of the National Academy of Sciences of the United States of America* 72(9):3448–3452.

Khidekel, N., Ficarro, S. B., Peters, E. C., and Hsieh-Wilson, L. C. 2004. Exploring the O-GlcNAc proteome: direct identification of O-GlcNAc-modified proteins from the brain. *Proceedings of the National Academy of Sciences of the United States of America* 101(36):13132–13137.

Khoury, G. a., Baliban, R. C., and Floudas, C. a. 2011. Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. *Scientific Reports* 1:1–5.

Kim, J. H., Lee, J., Oh, B., Kimm, K., and Koh, I. 2004. Prediction of phosphorylation sites using SVMs. *Bioinformatics* 20(17):3179–3184.

Kim, Y.-C., Udeshi, N. D., Balsbaugh, J. L., Shabanowitz, J., Hunt, D. F., and Olszewski, N. E. 2011. O-GlcNAcylation of the Plum pox virus capsid protein catalyzed by SECRET AGENT: characterization of O-GlcNAc sites by electron transfer dissociation mass spectrometry. *Amino Acids* 40(3):869–876.

Kreppel, L. K., Blomberg, M. A., and Hart, G. W. 1997. Dynamic glycosylation of nuclear and cytosolic proteins. Cloning and characterization of a unique O-GlcNAc transferase with multiple tetratricopeptide repeats. *The Journal of Biological Chemistry* 272(14):9308–15.

Kreppel, L. K., and Hart, G. W. 1999. Regulation of a cytosolic and nuclear o-glcnac transferase role of the tetratricopeptide repeats. *The Journal of Biological Chemistry* 274(45):32015–32022.

Kuhn, M., and Johnson, K. 2013. *Applied predictive modeling.* New York: Springer.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* 409(6822):860–921.

Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., and Maglott, D. R. 2014. ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research* 42(D1):D980–D985.

Lazarus, M. M. B., Nam, Y., Jiang, J., Sliz, P., and Walker, S. 2011. Structure of human O-GlcNAc transferase and its complex with a peptide substrate. *Nature* 22(5):4109.

Lazarus, M. B., Jiang, J., Gloster, T. M., Zandberg, W. F., Whitworth, G. E., Vocadlo, D. J., and Walker, S. 2012. Structural snapshots of the reaction coordinate for O-GlcNAc transferase. *Nature Chemical Biology* 8(12):966–968.

LeCun, Y., Bengio, Y., Geoffrey, H., Rusk, N., LeCun, Y., Bengio, Y., and Hinton, G. 2015. Deep learning. *Nature Methods* 13(1):35–35.

Lefebvre, T., Baert, F., Bodart, J. F., Flament, S., Michalski, J. C., and Vilain, J. P. 2004. Modulation of O-GlcNAc glycosylation during xenopus oocyte maturation. *Journal of Cellular Biochemistry* 93(5):999–1010.

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., Tukiainen, T., Birnbaum, D. P., Kosmicki, J. A., Duncan, L. E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Berghout, J., Cooper, D. N., Deflaux, N., DePristo, M., Do, R., Flannick, J., Fromer, M., et al. 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536(7616):285–91.

Leslie, C., Eskin, E., and Noble, W. S. 2002. The spectrum kernel: a string kernel for SVM protein classification. In *Pacific symposium on biocomputing. pacific symposium on biocomputing*, vol. 7, 564–575.

Li, A., Wang, L., Shi, Y., Wang, M., Jiang, Z., and Feng, H. 2005. Phosphorylation site prediction with a modified k-nearest neighbor algorithm and blosum62 matrix. In *2005 ieee engineering in medicine and biology 27th annual conference*, 6075–6078.

Li, W., and Godzik, A. 2006. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22(13):1658–1659.

Li, X., Molina, H., Huang, H., Zhang, Y.-Y., Liu, M., Qian, S.-W., Slawson, C., Dias, W. B., Pandey, A., Hart, G. W., Lane, M. D., and Tang, Q.-Q. 2009. O-linked N-acetylglucosamine modification on CCAAT enhancer-binding protein beta: role during adipocyte differentiation. *The Journal of Biological Chemistry* 284(29): 19248–54.

Lienhard, G. E. 2008. Non-functional phosphorylations? *Trends in Biochemical Sciences* 33(8):351–352.

Linding, R., Jensen, L. J., Diella, F., Bork, P., Gibson, T. J., and Russell, R. B. 2003. Protein Disorder Prediction. *Structure* 11(11):1453–1459.

Liu, K., Paterson, a. J., Chin, E., and Kudlow, J. E. 2000. Glucose stimulates protein modification by O-linked GlcNAc in pancreatic beta cells: linkage of O-linked GlcNAc to beta cell death. *Proceedings of the National Academy of Sciences of the United States of America* 97(6):2820–2825.

Love, D. C. 2002. Mitochondrial and nucleocytoplasmic targeting of O-linked GlcNAc transferase. *Journal of Cell Science* 116(4):647–654.

Lu, C. T., Huang, K. Y., Su, M. G., Lee, T. Y., Bretaña, N. A., Chang, W. C., Chen, Y. J., Chen, Y. J., and Huang, H. D. 2013. DbPTM 3.0: An informative resource for investigating substrate site specificity and functional association of protein post-translational modifications. *Nucleic Acids Research* 41(D1):D295–D305.

Lubas, W. A., Frank, D. W., Krause, M., and Hanover, J. A. 1997. O-linked GlcNAc transferase is a conserved nucleocytoplasmic protein containing tetratricopeptide repeats. *The Journal of Biological Chemistry* 272(14):9316–9324.

Lubas, W. a., and Hanover, J. a. 2000. Functional Expression of O-linked GlcNAc Transferase. *The Journal of Biological Chemistry* 275(15):10983–10988.

Ma, J., and Hart, G. W. 2014. O-GlcNAc profiling: from proteins to proteomes. *Clinical Proteomics* 11(1):8.

Madeira, F., Tinti, M., Murugesan, G., Berrett, E., Stafford, M., Toth, R., Cole, C., MacKintosh, C., and Barton, G. J. 2015. 14-3-3-Pred: Improved methods to predict 14-3-3-binding phosphopeptides. *Bioinformatics* 31(14):2276–2283.

Maloney, C., Sequeira, E., Kelly, C., Orris, R., and Beck, J. 2013. PubMed Central. In *The ncbi handbook [internet]. 2nd edition.*, 1–31.

Manning, G. 2002. The Protein Kinase Complement of the Human Genome. *Science* 298(5600):1912–1934.

Marth, J. D., and Grewal, P. K. 2008. Mammalian glycosylation in immunity. *Nature Reviews Immunology* 8(11):874–887.

Martin, J. C., Fadda, E., Ito, K., and Woods, R. J. 2014. Defining the structural origin of the substrate sequence independence of O-GlcNAcase using a combination of molecular docking and dynamics simulation. *Glycobiology* 24(1):85–96.

McKinney, W., and Team, P. D. 2015. Pandas - Powerful Python Data Analysis Toolkit.

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., and Cunningham, F. 2016. The Ensembl Variant Effect Predictor. *Genome biology* 17(1):122.

Mi, H., Muruganujan, A., Casagrande, J. T., and Thomas, P. D. 2013. Large-scale gene function analysis with the PANTHER classification system. *Nature Protocols* 8(8):1551–1566.

Miller, M. L. M., and Blom, N. 2009. Phospho-Proteomics. *Phospho-Proteomics* 527(4):299–310.

Müller, R., Jenny, A., and Stanley, P. 2013. The EGF Repeat-Specific O-GlcNAc-Transferase Eogt Interacts with Notch Signaling and Pyrimidine Metabolism Pathways in Drosophila. *PLoS ONE* 8(5):e62835.

Nachman, M. W., and Crowell, S. L. 2000. Estimate of the mutation rate per nucleotide in humans. *Genetics* 156(1):297–304.

Nagel, A. K., and Ball, L. E. 2014. O-GlcNAc transferase and O-GlcNAcase: achieving target substrate specificity. *Amino Acids* 46(10):2305–2316.

Neuberger, G., Schneider, G., and Eisenhaber, F. 2007. pkaPS: prediction of protein kinase A phosphorylation sites with the simplified kinase-substrate binding model. *Biology Direct* 2(1):1.

Ngoh, G. A., Facundo, H. T., Zafir, A., and Jones, S. P. 2010. O-GlcNAc signaling in the cardiovascular system. *Circulation Research* 107(2):171–85.

Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijne, G. 1997. a Neural Network Method for Identification of Prokaryotic and Eukaryotic Signal Peptides and Prediction of Their Cleavage Sites. *Protein Engineering* 10(1):1–6.

Nishi, H., Hashimoto, K., and Panchenko, A. R. 2011. Phosphorylation in protein-protein binding: Effect on stability and function. *Structure* 19(12):1807–1815.

O'Donnell, N., Zachara, N. E., Hart, G. W., and Marth, J. D. 2004. Ogt-Dependent X-Chromosome-Linked Protein Glycosylation Is a Requisite Modification in Somatic Cell Function and Embryo Viability. *Molecular and Cellular Biology* 24(4):1680–1690.

Overton, I. M., van Niekerk, C. A. J., and Barton, G. J. 2011. XANNpred: Neural nets that predict the propensity of a protein to yield diffraction-quality crystals. *Proteins: Structure, Function and Bioinformatics* 79(4):1027–1033.

Pathak, S., Alonso, J., Schimpl, M., Rafie, K., Blair, D. E., Borodkin, V. S., Schüttelkopf, A. W., Albarbarawi, O., and van Aalten, D. M. F. 2015. The active site of O-GlcNAc transferase imposes constraints on substrate sequence. *Nature Structural & Molecular Biology* 22(9):744–750.

Patti, M. E., Virkamäki, A., Landaker, E. J., Kahn, C. R., and Yki-Järvinen, H. 1999. Activation of the hexosamine pathway by glucosamine in vivo induces insulin resistance of early postreceptor insulin signaling events in skeletal muscle. *Diabetes* 48(8):1562–1571.

Pauling, L., and Corey, R. B. 1951. Configurations of Polypeptide Chains With Favored Orientations Around Single Bonds: Two New Pleated Sheets. *Proceedings of the National Academy of Sciences of the United States of America* 37(11): 729–740.

Pauling, L., Corey, R. B., and Branson, H. R. 1951. The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences of the United States of America* 37(4):205–211.

Pejaver, V., Hsu, W.-L., Xin, F., Dunker, a. K., Uversky, V. N., and Radivojac, P. 2014. The structural and functional signatures of proteins that undergo multiple events of post-translational modification. *Protein Science* 23(8):1077–1093.

Perez-Cervera, Y., Dehennaut, V., Aquino Gil, M., Guedri, K., Solórzano Mata, C. J., Olivier-Van Stichelen, S., Michalski, J. C., Foulquier, F., and Lefebvre, T. 2013. Insulin signaling controls the expression of O-GlcNAc transferase and its interaction with lipid microdomains. *FASEB Journal* 27(9):3478–3486.

Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S., and Goldstein, D. B. 2013. Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes. *PLoS Genetics* 9(8):e1003709.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. 2004. UCSF Chimera–a visualization system for exploratory research and analysis. *Journal of Computational Chemistry* 25(13): 1605–1612.

PFAM website. 2016a. Summary: beta-n-acetylglucosaminidase. `http://pfam.xfam.org/family/PF07555.11`. Accessed: 2016-12-10.

———. 2016b. Summary: Tetratricopeptide. `http://pfam.xfam.org/family/PF00515`. Accessed: 2016-12-10.

Pinna, L. A., and Ruzzene, M. 1996. How do protein kinases recognize their substrates? *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* 1314(3):191–225.

Poizat, C., Puri, P. L., Bai, Y., and Kedes, L. 2005. Phosphorylation-dependent degradation of p300 by doxorubicin-activated p38 mitogen-activated protein kinase in cardiac cells. *Molecular and Cellular Biology* 25(7):2673–87.

Ptacek, J., Devgan, G., Michaud, G., Zhu, H., Zhu, X., Fasolo, J., Guo, H., Jona, G., Breitkreutz, A., Sopko, R., McCartney, R. R., Schmidt, M. C., Rachidi, N., Lee, S.-J., Mah, A. S., Meng, L., Stark, M. J. R., Stern, D. F., De Virgilio, C., Tyers, M., Andrews, B., Gerstein, M., Schweitzer, B., Predki, P. F., and Snyder, M. 2005. Global analysis of protein phosphorylation in yeast. *Nature* 438(7068):679–84.

Qi, Y., Oja, M., Weston, J., and Noble, W. S. 2012. A unified multitask architecture for predicting local protein properties. *PLoS ONE* 7(3).

de Queiroz, R. M., Carvalho, E., and Dias, W. B. 2014. O-GlcNAcylation: The Sweet Side of the Cancer. *Frontiers in Oncology* 4(June):132.

Radermacher, P. T., Myachina, F., Bosshardt, F., Pandey, R., Mariappa, D., Müller, H.-A. J., and Lehner, C. F. 2014. O-GlcNAc reports ambient temperature and confers heat resistance on ectotherm development. *Proceedings of the National Academy of Sciences of the United States of America* 111(15):5592–7.

Ramirez-Correa, G. A., Jin, W., Wang, Z., Zhong, X., Gao, W. D., Dias, W. B., Vecoli, C., Hart, G. W., and Murphy, A. M. 2008. O-linked GlcNAc modification of cardiac myofilament proteins: A novel regulator of myocardial contractile function. *Circulation Research* 103(12):1354–1358.

Rao, F. V., Schüttelkopf, A. W., Dorfmueller, H. C., Ferenbach, A. T., Navratilova, I., and van Aalten, D. M. F. 2013. Structure of a bacterial putative acetyltransferase defines the fold of the human O-GlcNAcase C-terminal domain. *Open Biology* 3(10):130021.

Reimand, J., Wagih, O., and Bader, G. D. 2015. Evolutionary Constraint and Disease Associations of Post-Translational Modification Sites in Human Genomes. *PLOS Genetics* 11(1):e1004919.

Reva, B., Antipin, Y., and Sander, C. 2011. Predicting the functional impact of protein mutations: Application to cancer genomics. *Nucleic Acids Research* 39(17).

Roos, M. D., Su, K., Baker, J. R., and Kudlow, J. E. 1997. O glycosylation of an Sp1-derived peptide blocks known Sp1 protein interactions. *Molecular and Cellular Biology* 17(11):6472–6480.

Roquemore, E. P., Dell, A., Morris, H. R., Panico, M., Reason, A. J., Savoy, L. A., Wistow, G. J., Zigler, J. S., Earles, B. J., and Hart, G. W. 1992. Vertebrate lens alpha-crystallins are modified by O-linked N-acetylglucosamine. *The Journal of Biological Chemistry* 267(1):555–63.

Ruan, H.-B., Nie, Y., and Yang, X. 2013. Regulation of protein degradation by O-GlcNAcylation: crosstalk with ubiquitination. *Molecular & Cellular Proteomics* 12(12):3489–3497.

Ryu, I. H., and Do, S. I. 2011. Denitrosylation of S-nitrosylated OGT is triggered in LPS-stimulated innate immune response. *Biochemical and Biophysical Research Communications* 408(1):52–57.

Samish, I., Bourne, P. E., and Najmanovich, R. J. 2015. Achievements and challenges in structural bioinformatics and computational biophysics. *Bioinformatics* 31(1): 146–50.

Schimpl, M., Zheng, X., Borodkin, V. S., Blair, D. E., Ferenbach, A. T., Schüttelkopf, A. W., Navratilova, I., Aristotelous, T., Albarbarawi, O., Robinson, D. a., Macnaughtan, M. a., and van Aalten, D. M. F. 2012. O-GlcNAc transferase invokes nucleotide sugar pyrophosphate participation in catalysis. *Nature Chemical Biology* 8(12):969–74.

Schweers, O., Schönbrunn-Hanebeck, E., Marx, a., and Mandelkow, E. 1994. Structural studies of tau protein and Alzheimer paired helical filaments show no evidence for beta-structure. *The Journal of Biological Chemistry* 269(39):24290–24297.

Scikit-Learn website. 2016a. Comparing different clustering algorithms on toy datasets. `http://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html`. Accessed: 2016-06-10.

———. 2016b. Decision trees. `http://scikit-learn.org/stable/modules/tree.html#tree-algorithms`. Accessed: 2016-06-10.

———. 2016c. Probability calibration. `http://scikit-learn.org/stable/modules/calibration.html`. Accessed: 2016-09-20.

———. 2016d. R2 (coefficient of determination). `http://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html`. Accessed: 2016-12-10.

Selvan, N., Mariappa, D., Van Den Toorn, H. W. P., Heck, A. J. R., Ferenbach, A. T., and Van Aalten, D. M. F. 2015. The early metazoan Trichoplax adhaerens possesses a functional O-GlcNAc system. *The Journal of Biological Chemistry* 290(19):11969–11982.

Shafi, R., Iyer, S. P., Ellies, L. G., O'Donnell, N., Marek, K. W., Chui, D., Hart, G. W., and Marth, J. D. 2000. The O-GlcNAc transferase gene resides on the X chromosome and is essential for embryonic stem cell viability and mouse ontogeny. *Proceedings of the National Academy of Sciences of the United States of America* 97(11):5735–9.

Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L. A., Edwards, K. J., Day, I. N. M., and Gaunt, T. R. 2013. Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models. *Human Mutation* 34(1):57–65.

Sibille, N., Huvent, I., Fauquant, C., Verdegem, D., Amniai, L., Leroy, A., Wieruszeski, J.-M. M., Lippens, G., and Landrieu, I. 2012. Structural characterization by nuclear magnetic resonance of the impact of phosphorylation in the proline-rich region of the disordered Tau protein. *Proteins: Structure, Function and Bioinformatics* 80(2):454–462.

Siddiqui, a. S., and Barton, G. J. 1995. Continuous and discontinuous domains: an algorithm for the automatic generation of reliable protein domain definitions. *Protein Science* 4(5):872–884.

Sikorski, R. S., Boguski, M. S., Goebl, M., and Hieter, P. 1990. A repeating amino acid motif in CDC23 defines a family of proteins and a new relationship among genes required for mitosis and RNA synthesis. *Cell* 60(2):307–317.

Sillitoe, I., Lewis, T. E., Cuff, A., Das, S., Ashford, P., Dawson, N. L., Furnham, N., Laskowski, R. A., Lee, D., Lees, J. G., Lehtinen, S., Studer, R. A., Thornton, J., and Orengo, C. A. 2015. CATH: Comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Research* 43(D1):D376–D381.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature* 529(7587):484–489.

Sinclair, D. A. R., Syrzycka, M., Macauley, M. S., Rastgardani, T., Komljenovic, I., Vocadlo, D. J., Brock, H. W., and Honda, B. M. 2009. Drosophila O-GlcNAc transferase (OGT) is encoded by the Polycomb group (PcG) gene, super sex combs (sxc). *Proceedings of the National Academy of Sciences of the United States of America* 106(32):13427–13432.

Stormo, G. D., Schneider, T. D., Gold, L., and Ehrenfeucht, A. 1982. Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Research* 10(9):2997–3011.

Stratton, M. R., Campbell, P. J., and Futreal, P. A. 2009. The cancer genome. *Nature* 458(7239):719–724.

Suzuki, R., and Shimodaira, H. 2006. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* 22(12):1540–2.

Swamy, M., Pathak, S., Grzes, K. M., Damerow, S., Sinclair, L. V., van Aalten, D. M. F., and Cantrell, D. A. 2016. Glucose and glutamine fuel protein O-GlcNAcylation to control T cell self-renewal and malignancy. *Nature Immunology* 17(6):712–20.

Tang, H., Klopfenstein, D., Pedersen, B., Flick, P., Sato, K., Ramirez, F., Yunes, J., and Mungall, C. 2015. Goatools: Tools for gene ontology. https://github.com/tanghaibao/goatools.

Taylor, R. P., Geisler, T. S., Chambers, J. H., and McClain, D. A. 2009. Up-regulation of O-GlcNAc Transferase with Glucose Deprivation in HepG2 Cells Is Mediated by Decreased Hexosamine Pathway Flux. *The Journal of Biological Chemistry* 284(6):3425–3432.

Taylor, W. R. 1986. The classification of amino acid conservation. *Journal of Theoretical Biology* 119(2):205–218.

Telenti, A., Pierce, L. T., Biggs, W. H., di Iulio, J., Wong, E. H., Fabani, M. M., Kirkness, E. F., Moustafa, A., Shah, N., Xie, C., Brewerton, S. C., Bulsara, N., Garner, C., Metzker, G., Sandoval, E., Perkins, B. A., Och, F. J., Turpaz, Y., and Venter, J. C. 2016. Deep Sequencing of 10,000 Human Genomes. *bioRxiv preprint http://biorxiv.org/lookup/doi/10.1101/061663*.

The Theano Development Team, Al-Rfou, R., Alain, G., Almahairi, A., Angermueller, C., Bahdanau, D., Ballas, N., Bastien, F., Bayer, J., Belikov, A., Belopolsky, A., Bengio, Y., Bergeron, A., Bergstra, J., Bisson, V., Snyder, J. B., Bouchard, N., Boulanger-Lewandowski, N., Bouthillier, X., de Brébisson, A., Breuleux, O., Carrier, P.-L., Cho, K., Chorowski, J., Christiano, P., et al. 2016. Theano: A Python framework for fast computation of mathematical expressions. *arXiv preprint http://arxiv.org/abs/1605.02688* 19.

Torres, C. R., and Hart, G. W. 1984. Topography and polypeptide distribution of terminal N-acetylglucosamine residues on the surfaces of intact lymphocytes. Evidence for O-linked GlcNAc. *The Journal of Biological Chemistry* 259(5): 3308–3317.

Trapannone, R., Mariappa, D., Ferenbach, A. T., and van Aalten, D. M. 2016. Nucleocytoplasmic human O-GlcNAc transferase is sufficient for O-GlcNAcylation of mitochondrial proteins. *The Biochemical Journal* 0:1693–1702.

Trinidad, J. C., Barkan, D. T., Gulledge, B. F., Thalhammer, A., Sali, A., Schoepfer, R., and Burlingame, A. L. 2012. Global Identification and Characterization of Both O-GlcNAcylation and Phosphorylation at the Murine Synapse. *Molecular & Cellular Proteomics* 11(8):215–229.

Troshin, P. V., Procter, J. B., and Barton, G. J. 2011. Java bioinformatics analysis web services for multiple sequence alignment–JABAWS:MSA. *Bioinformatics* 27(14):2001–2002.

Trost, B., and Kusalik, A. 2011. Computational prediction of eukaryotic phosphorylation sites. *Bioinformatics* 27(21):2927–35.

Uyar, B., Weatheritt, R. J., Dinkel, H., Davey, N. E., and Gibson, T. J. 2014. Proteome-wide analysis of human disease mutations in short linear motifs: neglected players in cancer? *Molecular BioSystems* 10(10):2626.

Valverde, R. H. F., Britto-Borges, T., Lowe, J., Einicker-Lamas, M., Mintz, E., Cuillel, M., and Vieyra, A. 2011. Two Serine Residues Control Sequential Steps during Catalysis of the Yeast Copper ATPase through Different Mechanisms That Involve Kinase-mediated Phosphorylations. *The Journal of Biological Chemistry* 286(9):6879–6889.

Vandermarliere, E., and Martens, L. 2013. Protein structure as a means to triage proposed PTM sites. *Proteomics* 13(6):1028–1035.

Varki, A., Cummings, R. D., Esko, J. D., Freeze, H. H., Stanley, P., Bertozzi, C. R., Hart, G. W., and Etzler, M. E. 2009. *Essentials of glycobiology.* Cold Spring Harbor Laboratory Press.

Velankar, S., Best, C., Beuth, B., Boutselakis, C. H., Cobley, N., Sousa Da Silva, A. W., Dimitropoulos, D., Golovin, A., Hirshberg, M., John, M., Krissinel, E. B., Newman, R., Oldfield, T., Pajon, A., Penkett, C. J., Pineda-Castillo, J., Sahni, G., Sen, S., Slowley, R., Suarez-Uruena, A., Swaminathan, J., van Ginkel, G., Vranken, W. F., Henrick, K., and Kleywegt, G. J. 2010. PDBe: Protein Data Bank in Europe. *Nucleic acids research* 38(Database issue):D308–17.

Velankar, S., Dana, J. M., Jacobsen, J., van Ginkel, G., Gane, P. J., Luo, J., Oldfield, T. J., O'Donovan, C., Martin, M.-J., and Kleywegt, G. J. 2013. SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Research* 41(D1):D483–D489.

Walsh, C. T., Garneau-Tsodikova, S., and Gatto, G. J. 2005. Protein posttranslational modifications: The chemistry of proteome diversifications. *Angewandte Chemie - International Edition* 44(45):7342–7372.

Wang, J., Torii, M., Liu, H., Hart, G. W., and Hu, Z.-z. 2011. dbOGAP - An Integrated Bioinformatics Resource for Protein O-GlcNAcylation. *BMC Bioinformatics* 12(1):91.

Wang, S., Huang, X., Sun, D., Xin, X., Pan, Q., Peng, S., Liang, Z., Luo, C., Yang, Y., Jiang, H., Huang, M., Chai, W., Ding, J., and Geng, M. 2012. Extensive crosstalk between O-GlcNAcylation and phosphorylation regulates Akt signaling. *PLoS ONE* 7(5).

Wang, Z., Gucek, M., and Hart, G. W. 2008. Cross-talk between GlcNAcylation and phosphorylation: site-specific phosphorylation dynamics in response to globally elevated O-GlcNAc. *Proceedings of the National Academy of Sciences of the United States of America* 105(37):13793–8.

Wang, Z., Udeshi, N. D. N., Slawson, C., Compton, P. D., Sakabe, K., Cheung, W. D., Shabanowitz, J., Hunt, D. F., and Hart, G. W. 2010. Extensive crosstalk between O-GlcNAcylation and phosphorylation regulates cytokinesis. *Science Signaling* 3(104):ra2.

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., and Barton, G. J. 2009. Jalview Version 2-A multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25(9):1189–1191.

Wei, Q., Dunbrack, R. L., Yue, P., Melamud, E., Moult, J., Ferrer-Costa, C., Gelpi, J., Zamakola, L., Parraga, I., de la Cruz, X., Bao, L., Zhou, M., Cui, Y., Bao, L., Cui, Y., Bromberg, Y., Rost, B., Bromberg, Y., Yachdav, G., Rost, B., Wainreb, G., Ashkenazy, H., Bromberg, Y., Starovolsky-Shitrit, A., Haliloglu, T., et al. 2013. The Role of Balanced Training and Testing Data Sets for Binary Classifiers in Bioinformatics. *PLoS ONE* 8(7):e67863.

Wells, L. 2002. Mapping Sites of O-GlcNAc Modification Using Affinity Tags for Serine and Threonine Post-translational Modifications. *Molecular & Cellular Proteomics* 1(10):791–804.

Wells, L. 2016. Mutations in O-GlcNAc Transferase Linked to X-linked Intellectual Disability. *The FASEB Journal* 30(1 Supplement):98.5–98.5.

Wells, L., Kreppel, L. K., Comer, F. I., Wadzinski, B. E., and Hart, G. W. 2004. O-GlcNAc Transferase Is in a Functional Complex with Protein Phosphatase 1 Catalytic Subunits. *The Journal of Biological Chemistry* 279(37):38466–38470.

Westbrook, J. D., and Bourne, P. E. 2000. STAR/mmCIF: an ontology for macromolecular structure. *Bioinformatics* 16(2):159–168.

Wilkins, M. 2009. Proteomics data mining. *Expert Review of Proteomics* 6(6): 599–603.

Wilkins, M. R., Sanchez, J.-C., Gooley, A. a., Appel, R. D., Humphery-Smith, I., Hochstrasser, D. F., and Williams, K. L. 1996. Progress with Proteome Projects: Why all Proteins Expressed by a Genome Should be Identified and How To Do It. *Biotechnology and Genetic Engineering Reviews* 13(1):19–50.

Wong, Y. H., Lee, T. Y., Liang, H. K., Huang, C. M., Wang, T. Y., Yang, Y. H., Chu, C. H., Huang, H. D., Ko, M. T., and Hwang, J. K. 2007. KinasePhos 2.0: A web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns. *Nucleic Acids Research* 35(SUPPL.2).

Worth, C. L., Bickerton, G. R. J., Schreyer, A., Forman, J. R., Cheng, T. M. K., Lee, S., Gong, S., Burke, D. F., and Blundell, T. L. 2007. A structural bioinformatics approach to the analysis of nonsynonymous single nucleotide polymorphisms (nsSNPs) and their relation to disease. *Journal of Bioinformatics and Computational Biology* 5(6):1297–318.

Wright, P. E., and Dyson, H. 1999. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *Journal of Molecular Biology* 293(2): 321–331.

Wu, C. H. 1997. Artificial neural networks for molecular sequence analysis. *Computers & Chemistry* 21(4):237–56.

Wu, H.-Y., Lu, C.-T., Kao, H.-J., Chen, Y.-J. Y.-J., Chen, Y.-J. Y.-J., and Lee, T.-Y. 2014. Characterization and identification of protein O-GlcNAcylation sites with substrate specificity. *BMC Bioinformatics* 15(16):1.

Xin, F., and Radivojac, P. 2012. Post-translational modifications induce significant yet not extreme changes to protein structure. *Bioinformatics* 28(22):2905–2913.

Yang, P., Humphrey, S. J., James, D. E., Yang, Y. H., and Jothi, R. 2016. Positive-unlabeled ensemble learning for kinase substrate prediction from dynamic phosphoproteomics data. *Bioinformatics* 32(September):252–259.

Yang, X., Ongusaha, P. P., Miles, P. D., Havstad, J. C., Zhang, F., So, W. V., Kudlow, J. E., Michell, R. H., Olefsky, J. M., Field, S. J., and Evans, R. M. 2008. Phosphoinositide signalling links O-GlcNAc transferase to insulin resistance. *Nature* 451(7181):964–969.

Yang, Z. R., Thomson, R., McNeil, P., and Esnouf, R. M. 2005. RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics* 21(16):3369–76.

Yates, C. M., Filippis, I., Kelley, L. A., and Sternberg, M. J. 2014. SuSPect: Enhanced Prediction of Single Amino Acid Variant (SAV) Phenotype Using Network Features. *Journal of Molecular Biology* 426(14):2692–2701.

Zachara, N. E., and Hart, G. W. 2004. O-GlcNAc a sensor of cellular state: The role of nucleocytoplasmic glycosylation in modulating cellular function in response to nutrition and stress. *Biochimica et Biophysica Acta - General Subjects* 1673(1-2): 13–28.

Zdobnov, E. M., and Apweiler, R. 2001. InterProScan - an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17(9):847–848.

Zeidan, Q., Wang, Z., De Maio, A., and Hart, G. W. 2010. O-GlcNAc Cycling Enzymes Associate with the Translational Machinery and Modify Core Ribosomal Proteins. *Molecular Biology of the Cell* 21(12):1922–1936.

Zell, A., Mache, N., Sommer, T., and Korb, T. In *Applications of artificial neural networks ii*, ed. Steven K. Rogers, 708–718. International Society for Optics and Photonics.

Zeugmann, T., Poupart, P., Kennedy, J., Jin, X., Han, J., Saitta, L., Sebag, M., Peters, J., Bagnell, J. A., Daelemans, W., Webb, G. I., Ting, K. M., Ting, K. M., Webb, G. I., Shirabad, J. S., Fürnkranz, J., Hüllermeier, E., Matwin, S., Sakakibara, Y., Flener, P., Schmid, U., Procopiuc, C. M., Lachiche, N., and Fürnkranz, J. 2011. Partitional Clustering. In *Encyclopedia of machine learning*, 766–766. Boston, MA: Springer US.

Zhao, X., Ning, Q., Chai, H., Ai, M., and Ma, Z. 2015. PGlcS: Prediction of protein O-GlcNAcylation sites with multiple features and analysis. *Journal of Theoretical Biology* 380:524–529.

Zhou, F.-F., Xue, Y., Chen, G.-L., and Yao, X. 2004. GPS: a novel group-based phosphorylation predicting and scoring method. *Biochemical and Biophysical Research Communications* 325(4):1443–1448.

Zhu, Y., Liu, T.-W., Cecioni, S., Eskandari, R., Zandberg, W. F., and Vocadlo, D. J. 2015. O-GlcNAc occurs cotranslationally to stabilize nascent polypeptide chains. *Nature Chemical Biology* 11(5):319–25.